INTEGRATING FINANCE DICTIONARY IN LEXICON-BASED APPROACH WITH MACHINE LEARNING ALGORITHM TO ANALYSE THE IMPACT OF OPEC NEWS SENTIMENT ON FINANCIAL MARKET

WU LING

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

Universiti Malaya

INTEGRATING FINANCE DICTIONARY IN LEXICON-BASED APPROACH WITH MACHINE LEARNING ALGORITHM TO ANALYSE THE IMPACT OF OPEC NEWS SENTIMENT ON FINANCIAL MARKET

WU LING

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SOFTWARE ENGINEERING (SOFTWARE TECHNOLOGY)

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

Universiti Malaya

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Matric No:

Name of Degree:

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Field of Study:

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

University of Malaya

INTEGRATING FINANCE DICTIONARY IN LEXICON-BASED APPROACH WITH MACHINE LEARNING ALGORITHM TO ANALYSE THE IMPACT OF OPEC NEWS SENTIMENT ON FINANCIAL MARKET

ABSTRACT

Since last few decades, machine learning algorithm which trains computers to learn from experience, is one of the most rapidly developing techniques which settles in the intersection research field of statistics and computer science. This research aims to build a properly trained machine learning classifier to study the impact of Organization of Petroleum Exporting Countries (OPEC) news sentiment on stock prices of six Malaysian public listed companies (energy sector) in the main board of Bursa Malaysia. The data used in this research are collected during the period 2012-2017. To carry out the research, firstly, lexicon-based approach is used to analyze the sentiment of sentences in the financial news articles. A sentiment dictionary from a finance domain is applied to improve the accuracy in labelling the financial news sentences. The labelled sentences are then used to train the supervised machine learning classifiers. The classifiers classify the OPEC news sentences into three different categories – negative (labeled with sentiment score -1), neutral (labeled with sentiment score 0), and positive (labeled with sentiment score 1). The performance of the supervised machine learning classifier is found to achieve 70% accuracy. The OPEC news article's sentiment score is calculated using relative proportional difference evaluating method: S = (P-N) / (P+N), whereby, P and N are the number of positive and negative sentences in the article, respectively. The sentiment score of each article ranges from -1 to 1. Using event study method, this sentiment score is used to compare with the historical stock prices of the six selected public listed energy sector companies. Results of the analysis show that OPEC news sentiment shows impact on the stock prices of these six companies. However, the impact did not occur on the news release date. During the event window period (i.e., five days

Universiti Malaya

before and after a news released), there is a negative correlation between OPEC news sentiment and the six companies' average cumulative abnormal return. Cumulative abnormal return is the average of daily abnormal return during the event window, which can be used to show the overall fluctuation of the stock prices. The findings of this research show that applying financial sentiment dictionary to train the supervised machine learning algorithm can enhance the performance of machine learning classifier. Results of statistical analysis in this research also provides a clear picture to the stock investors on the movement of the six Malaysian energy sector companies' stock prices during the event window period. This can help them to make better decisions in their trading in order to obtain profitable stock returns.

Keywords: Machine Learning Algorithm, Lexicon-based Labelling, News Sentiment Classification, Organization of Petroleum Exporting Countries, OPEC, Bursa Malaysia, Energy Sector

Universiti Malaya

MENGINTEGRASIKAN KAMUS KEWANGAN DALAM PENDEKATAN BERASASKAN-LEXIKON ALGORITMA PEMBELAJARAN MESIN UNTUK MENGANALISIS KESAN SENTIMEN BERITA OPEC DI PASARAN KEWANGAN

ABSTRAK

Sejak beberapa dekad yang lalu, algoritma pembelajaran mesin yang melatih komputer belajar dari pengalaman, adalah salah satu teknik yang paling pesat berkembang yang menetap di bidang penyelidikan persimpangan statistik dan sains komputer. Penyelidikan ini bertujuan untuk membina pengkelas pembelajaran mesin yang terlatih untuk mengkaji kesan sentimen berita Organization of Petroleum Exporting Countries (OPEC) terhadap harga saham enam buah syarikat awam Malaysia (sektor tenaga) di papan utama Bursa Malaysia. Data yang digunakan dalam penyelidikan ini dikumpul dalam tempoh 2012-2017. Untuk menjalankan penyelidikan ini, pendekatan berasaskan-lexikon digunakan untuk menganalisis sentimen ayat dalam artikel berita kewangan. Sentimen kamus daripada satu domain kewangan digunakan untuk meningkatkan ketepatan dalam pelabelan ayat berita kewangan. Kemudian, ayat yang dilabelkan digunakan untuk melatih pengklasifikasi pembelajaran mesin yang diselia. Pengklasifikasi mengklasifikasikan ayat berita OPEC kepada tiga kategori yang berlainan-negatif (dilabelkan dengan skor sentimen -1), neutral (dilabelkan dengan skor sentimen 0), dan positif (dilabelkan dengan skor sentimen 1). Prestasi ketepatan bagi pengklasifikasi pembelajaran mesin yang diselia didapati mencapai 70%. Skor sentimen artikel berita OPEC dikira dengan menggunakan kaedah penilaian perbezaan berkadar relatif: S = (P-N) / (P + N), di mana, P dan N adalah bilangan ayat positif dan negatif dalam artikel tersebut, masing-masing. Skor sentimen bagi setiap artikel berkisar dari -1 hingga 1. Dengan menggunakan kaedah pembelajaran peristiwa, skor sentimen ini digunakan untuk berbanding dengan harga saham sejarah bagi enam buah syarikat tersenarai awam yang

Universiti Malaya

dipilih dari sektor tenaga. Hasil analisa menunjukkan bahawa sentimen berita OPEC mempunyai kesan terhadap harga pasaran saham keenam-enam buah syarikat ini. Walau bagaimanapun, kesan tidak berlaku pada tarikh keluaran berita. Semasa tempoh tetingkap peristiwa (iaitu, lima hari sebelum dan selepas berita dikeluarkan), terdapat satu korelasi negatif di antara sentimen berita OPEC dengan pulangan purata kumulatif yang tidak normal bagi keenam-enam buah syarikat ini. Pulangan kumulatif yang tidak normal adalah purata pulangan harian yang tidak normal dalam tetingkap peristiwa, yang boleh digunakan untuk menunjukkan turun naik keseluruhan harga pasaran saham. Penemuan kajian ini menunjukkan bahawa menerapkan kamus sentimen kewangan untuk melatih algoritma pembelajaran mesin yang diselia dapat meningkatkan prestasi pengelasan pembelajaran mesin. Hasil analisis statistik dalam penyelidikan ini juga memberikan satu gambaran yang jelas kepada para pelabur saham mengenai pergerakan harga saham bagi enam buah syarikat sektor tenaga Malaysia dalam tempoh tetingkap peristiwa. Ini dapat membantu mereka membuat keputusan yang lebih baik dalam perdagangan mereka demi mendapatkan pulangan saham yang menguntungkan.

Kata kunci: Algoritma pembelajaran mesin, Pelabelan berasaskan-lexikon, Klasifikasi sentimen berita, Organization of Petroleum Exporting Countries, OPEC, Bursa Malaysia, Sektor tenaga

Universiti Malaya

Acknowledgements

First and foremost, praises and thanks to the God, the Almighty, for His showers of blessings throughout my research work to complete the research successfully.

I would like to express my deep and sincere gratitude to my research supervisor, Associate Prof. Dr. Ow Siew Hock for giving me the opportunity to do research and providing invaluable guidance throughout this research. Her dynamism, vision, sincerity and motivation have deeply inspired me. She has taught me the way to carry out the research and to present the research works as clearly as possible. It was a great privilege and honor to study under her guidance. I am extremely grateful for what she has offered me.

Also, I express my thanks to my parents for their love, prayers, caring and sacrifices for educating and preparing me for my future. My Special thanks also goes to my friend Ali Khan Ghumro for his constant encouragement and support.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

Wu Ling

Universiti Malaya

TABLE OF CONTENTS

Universiti Malaya

Universiti Malaya

Universiti Malaya

List of Figures

Universiti Malaya

List of Tables

Universiti Malaya

List of Symbols and Abbreviations

OPEC	:	Organization of Petroleum Exporting Countries
SPR	:	US Strategic Petroleum Reserve
TF	:	Term Frequency
IDF	:	Inverse Document Frequency
SVM	:	Support Vector Machine
H4N	:	Havard-IV-4 TagNeg
WSJ	:	Wall Street Journal
NLP	:	Natural Language Processing
ML	:	Machine Learning
GNB	:	Gaussian Naïve Bayes
MNB	:	Multinomial Naïve Bayes
CNB	:	Complement Naïve Bayes
BNB	:	Bernoulli Naïve Bayes
SGDC	:	Stochastic Gradient Descent Classifier
OVA	:	One Versus All
ANOVA	:	Analysis of Variances
SPSS	:	Statistical Package of for the Social Sciences
CAR	:	Cumulative Abnormal Return
NLTK	:	Natural Language Toolkit
R	:	Daily Return
AR	:	Abnormal Return
ER	:	Expected Return
Q-Q	:	Quantile-Quantile
IBM	:	International Business Machine Corporation

Universiti Malaya

List of Appendices

Universiti Malaya

CHAPTER 1: INTRODUCTION

The market participants can join the real-time market trading based on the high speed computing technology. Thus, there is more and more attention be paid on the analyzing how the news sentiment affects the stock prices (Li et al., 2014).

Organization of Petroleum Exporting Countries (OPEC) is an organization which has great influence on the market of world's most important commodity-petroleum (Colgan, 2014). Thus, OPEC news announcements have significant effect on the stock prices of the energy sector (oil & gas) companies. Since there are limited researches focusing on the impact of OPEC news sentiments on the Malaysian stock prices, this research is initiated to investigate the impact of OPEC news announcements on the stock prices of public listed energy sector (oil & gas) companies in Bursa Malaysia.

The following section highlights the background of this research.

1.1 Background of Research

In the mid-long term, the movement of the oil price has shown impact on the fluctuation of the stock prices globally (Phan, Sharma, and Narayan, 2015). Compared with other commodities, petroleum has significant influence on the world economy, especially when it comes to causing economy recessions (Elder and Serletis, 2010). Hence, the announcement of oil-related news can influence the stock market at large, which will affect the stock market participants' return (Narayan and Narayan, 2017).

With the fact that Organization of Petroleum Exporting Countries (OPEC) has great influence on the global oil prices, the OPEC news announcements catch more and more market participants as well as researchers' attention. To understand the pattern of the fluctuation caused by OPEC news sentiments can provide crucial information for share market investors to make better investment decisions. Thus, the number of studies on

Universiti Malaya

OPEC news sentiments analysis and its impact on stock prices of companies' is increasing. By analyzing the news announcements released by OPEC, those who are concerned about the crude oil markets can get pivotal information about the market because of the huge impacts those announcements have on the global oil price (Hanabusa, 2012).

However, there are yet limited research working on finding the movements of the stock prices of Malaysian public listed companies in the energy sector (oi & gas) in relation to the OPEC news announcements.

According to the researches on news sentiment analysis, there are two commonly used methods - lexicon-based techniques and supervised machine learning-based approaches (Saif et al., 2016). It is proven that by applying hybrid approaches which combines both lexicon-based and machine learning approaches, can achieve not only the stability from lexicon-based approach but also productivity from machine learning algorithms (Biltawi et al., 2016). Since training data plays an important role in machine learning, labelling training data properly is the key to ensure the performance of supervised machine learning classifiers (Tripathy, Agrawal, and Rath, 2016). There are mainly two types of labelling methods - manual annotation and automatic labelling. Manual labelling is laborious and requires a sufficient amount of domain knowledge (Pham et al., 2016). Lexicon-based approach is more productive compared to manual labelling. Since various lexicon resources can be used in labelling the data, the selection of lexicon resource also influences the results of data labelling (Soroka, Young, and Balmas, 2015).

1.2 Research Problems

The research problems of this study are as follow:

- Manually classify OPEC news data for sentiment analysis is time consuming.
- Unsuitable lexicon resources used in labelling data can cause low accuracy of the

Universiti Malaya

machine learning classifier.

- Limited research pertaining to the impact of OPEC news sentiment on the stock prices of public listed Malaysian energy sector (oil & gas) companies.

1.3 Research Objectives

The objectives of this research are defined as follow:

- To build an innovative classifier to classify the OPEC news sentiment.
- To improve the accuracy of the innovative classifier by using proper lexicon resource from finance domain to label training data.
- To find out how would the stock prices of public listed Malaysian energy sector (oil & gas) companies react to the OPEC news sentiments.

1.4 Research Scope

In this research, a sentiment dictionary from the finance domain is applied to train the supervised machine learning algorithms. The performance of seven commonly used supervised machine learning algorithm-based classifiers are tested. Among these classifiers, the classifier with the highest accuracy score will be used to analyze the sentiment of OPEC news.

Furthermore, altogether 28 energy sector (oil & gas) companies are listed on the Main Market Board of Bursa Malaysia. Six companies are randomly selected to study the impact of OPEC news sentiments on their stock prices fluctuation. These companies are as shown in Table 1.1: Bumi Armada Berhad (stock code: 5210), HengYuan Refining Company Berhad (stock code: 4324), Hibiscus Petroleum Berhad (stock code: 5199), Petron Malaysia Refining & Marketing Berhad (stock code: 3042), Sapura Energy Berhad (stock code: 5218) and Sumatec Resources Berhad (stock code: 1201).

Universiti Malaya

Table 1.1: Companies Selected for the Study (Energy Sector)

NO.	Stock Code	Company
1	5210	Bumi Armada Berhad
2	4324	HengYuan Refining Company Berhad
3	5199	Hibiscus Petroleum Berhad
4	3042	Petron Malaysia Refining & Marketing Berhad
5	5218	Sapura Energy Berhad
6	1201	Sumatec Resources Berhad

The historical stock prices data from 2012 to 2017 of these six companies are used in this research. Similarly, the OPEC official news releases in this same period of time are also collected and used in data analysis.

1.5 Techniques Used

Since this research aims to study the impact of OPEC news announcements on the stock prices of public listed energy sector (oil & gas) companies in Bursa Malaysia, it can be divided into two parts: 1) analyzing the OPEC news sentiment and classify it into positive, neutral and negative and 2) analyzing the fluctuation of stock prices of the six selected companies after the release of OPEC news, to determine whether there is any relationship between them.

The techniques applied in this research can be divided into 1) news sentiment classification techniques and 2) statistical analysis using event study method.

News sentiment classification aims to classify the OPEC news announcements based on its sentiment. This research uses the lexicon-based approaches together with machine learning algorithm-based techniques to build a machine learning classifier with good performance.

Universiti Malaya

This research also uses event study method to analyze the historical stock prices data of the six energy sector (oil & gas) companies.

1.6 Thesis Organization

Chapter 1 of this dissertation presents the background of this research, research questions, research objectives and research scope.

Chapter 2 covers literature review on the Organization of the Petrol Exporting Countries (OPEC), existing research methods used to analyze the news impact on stock markets prices, the commonly-used feature processing approaches, machine learning algorithm techniques and the event study method to analyze the stock prices.

Chapter 3 explains the research methodology used in this research. Firstly, the qualitative and quantitative research methodology are introduced and gives the reasons for using the combination of these two types of research. The research activities, financial terms, the research tools used in the study, the choosing of proper datasets, research design as well as the performance evaluation measures for machine learning classifiers are also included in this chapter.

Chapter 4 presents data collection and the results of data analysis. Chapter 5 discusses about the research findings, problems encountered, weaknesses of the study, recommendation of future works and concludes this research.

Universiti Malaya

CHAPTER 2: LITERATURE REVIEW

News impact on the financial market has been widely studied, but limited researches has been conducted on how OPEC news influence the energy sector (oil & gas) stock prices. Current studies on energy sector (oil & gas) stock market are focusing on the impact of OPEC events rather than OPEC news sentiment (Demirer & Kutan, 2010; Loutia, Mellios, & Andriosopoulos, 2016). Thus, research on this aspect is yet to be investigated.

In this chapter, an introduction of OPEC news and its relevant studies are presented. This chapter also discusses the commonly used techniques of news classification. The procedure of news classification can be divided into two parts: feature processing and classification based on machine learning algorithms. As the event study methodology is crucial in this study, it is also highlighted in this chapter. A summary of the literature review is presented in this chapter.

2.1 Organization of The Petrol Exporting Countries (OPEC)

In the mid-1960s, the Organization of Petroleum Exporting Countries (OPEC) was established. Initially, OPEC consisted of five oil-producing developing Middle-East Asia countries (Plante, 2015). Today, it has 14 members of the world's key oil-producing countries which account for 44 percent of the global oil production with a proven reservation of 81.5 percent of global oil. Thus, its influence on global oil prices is enormous since its establishment (Lin & Tamvakis, 2010).

Every year, OPEC hosts conferences to make decision on the policies about oil production among its members (Schmidbauer & Rösch, 2012). The announcements made in those conferences play a major role in the oil market, worldwide (Mensi, Hammoudeh, & Yoon, 2014).

Universiti Malaya

2.1.1 Impact of OPEC news

Compared with other commodities, petroleum plays the most important role and has great influence on the world economy (Elder & Serletis, 2010). Nevertheless, similar studies indicate that OPEC news announcements show significant impact on the global oil & gas markets. Mensi et al (2014) in their study about what causes the volatility of oil price has also expressed the important role of OPEC in the world crude oil market. OPEC usually provides announcements about their decisions on the overall goal of oil production for the cartel as well as the target of individual oil production of their members (OPEC Secretariat, 2003).

By analyzing the news announcements made by OPEC, those who are concerned about the crude oil markets can get crucial information about the market because of the huge impact that those announcements have on the global oil prices (Hanabusa, 2012).

Mensi et al (2014) conducted a research on the volatility of oil markets prices and the price of crude oil based on OPEC announcements released between May 1987 to December 2012. They found that the OPEC announcements about "cut" and "maintain" decision on oil production have great effect on the returns and volatility on crude oil markets. Demirer and Kutan (2010) studied both the US Strategic Petroleum Reserve (SPR) and OPEC's announcements released between 1983 to 2008 about the on spot and future oil prices. They found that after the OPEC announcements were released, an abnormal return of the related markets shown apparent fluctuations. Conversely, the announcements from SPR did not show any influence on the abnormal returns. Schmidbauer and Rösch (2012) conducted a research about the effect of OPEC announcements on the fluctuation of related stock prices by analyzing the daily data collected from OPEC announcements between 1986 to 2009. The result shows that the influence of OPEC news vary depending on whether it is before or after the

Universiti Malaya

announcements. These results illustrate that OPEC announcements have positive influence on the volatility before the announcements were made, and negative effect after the announcements. A recent research which analyzed the OPEC news data over the period from 2003 to 2014 indicates that negative news announced by OPEC have positive effect on the stock market returns of US energy companies. (Gupta & Banerjee, 2018).

2.2 Review on News Impact Study Methods

With the fact that the market participants can join the real-time market trading using high speed computing technology, news announcements can influence the stock market in very short time. Financial news articles as one of the major resource of market information, are analyzed widely by the researchers and investors (Li et al., 2014). Existing studies conducted by Engelberg et al (2011) and Wisniewski et al (2013) suggest that investors' sentiment is deeply influenced by news, which in turn, affects the price of stock market.

To study the news impact on the stock prices, it starts with news text classification. There are multiple algorithms provided by machine learning to classify the news text. To apply machine learning algorithms to do the news sentiment classification, the first step is about feature processing. Feature processing and classification are the two main stages in the classification of news text (Uysal & Gunal, 2014).

The approaches mentioned in the news impact on the stock market literature are different in three aspects: i) feature processing (a process to generate the information which can be analyzed based on the given data); ii) the machine learning algorithm which is used to classify the text based on the output of feature processing; and iii) data set from a certain field which consists of two parts: the news textual data and the corresponding data about the reaction of the stock market (Hagenau, Liebmann, & Neumann, 2013).

Universiti Malaya

2.3 Feature Processing

Feature processing procedure aims to adequately represent the text content to the information which can be further processed by machine learning algorithm. In a typical framework of text classification, feature processing is one of the crucial components which significantly influence the outcome of classification task (Uysal & Gunal, 2014). Generally, feature processing consists of two main parts, feature selection and feature extraction (Tan, Wang, & Wu, 2011). By performing feature processing, the words of the news text can be technically chosen to be used in training the machine learning algorithms. Feature processing has three main benefits in the news sentiment classification (Mejova, 2009).

1) Scalability: By selecting the fraction of the whole article data as input rather than every word of the text, can save the storage and computational time. Reducing the data dimension is one of the major goals to apply feature processing methods to the dataset. By proper feature processing, the irrelevant information of solving the problem is removed from the data set. Thus, it reduces the scalability problem (Khalid, Khalil, & Nasreen, 2014).

2) Accuracy: Without feature processing, the accuracy of machine learning algorithm can be distrustful. For example, in text classification, Naïve Bayes shows poor performance without feature processing (J. Chen, Huang, Tian, & Qu, 2009). By eliminating useless noise words and selecting the most related features, the accuracy of the machine learning algorithms can achieve a significant improvement. To establish a classifier which has higher accuracy, selecting those words with stronger signal-to-noise ratio is the key point (Ladha & Deepa, 2011).

3) Comprehension: A better understanding of data in machine learning or pattern recognition applications can be achieved by feature processing (Chandrasheka &

Universiti Malaya

Sahin, 2014). Feature processing produce a good feature which can efficiently describe the input news text data. At the same time, it can also reduce the computational time by eliminating irrelevant features (Hira & Gillies, 2015).

2.3.1 Methods of Feature Processing

Many methods can be used in feature processing. The following section describes three commonly used methods.

1) Terms Frequency (TF): The importance of term frequency has been widely noticed in the traditional information retrieval systems. The intuition in this method is that the more one term is repeatedly mentioned in the document, the more informative it is. Term Frequency – Inverse Document Frequency (TF-IDF) which is a famous method in modeling documents, have also been widely used in feature processing (Trstenjak, Mikac, & Donko, 2014). As one of the most recognized word weighing algorithms, TF-IDF has promising accuracy in classifying the text documents (Hakim, Erwin, Eng, Galinium, & Muliady, 2015). By applying this method, the document can be represented by those terms which most frequently appear in the document. However, it is insufficient to weigh the term only by calculating its frequency (Xia & Chai, 2011). For instance, in the research conducted by Yelena Mejova (2009), he found that in text sentiment classification, it is more beneficial to find the most unique terms of the documents rather than the most frequent ones.

2) N-Grams: In feature processing, the term's position is also crucial in document representation. The term's position determines, and sometimes reverses the polarity of the phrases (Mejova, 2009). Thus, the feature vector sometimes is encoded with the information of term's position. N-Grams are the sequences of those elements appear in the texts. Elements can be characters, words or any other

Universiti Malaya

elements which appear in the text one after another (Sidorov, Velasquez, Stamatatos, Gelbukh, & Chanona-Hernández, 2014). In N-Grams, "n" represents the number of elements in a sequence. Sidorov et al (2014) provided a method named SN-Grams -- a combination of syntactic relations in syntactic with N-Grams. Their research result shows that the SN-Grams outperformed the traditional N-Grams in machine learning tasks. N-Grams is commonly used with the combination of word-stem and part of speech techniques (Kalchbrenner, Grefenstette, & Blunsom, 2014).

3) Part-of-Speech: Part-of-speech is another state-of-the-art method in natural language processing. One of the classic topic of natural language processing is text classification (X. Zhang, Zhao, & LeCun, 2015). As the name suggests, by applying part-of-speech method, the text document can be represented by the words which are grouped by its syntactic functions such as verbs, nouns, noun phrase, adjectives, etc. The most commonly used approaches from part-of-speech are Bag-of-Words, Noun Phrases and Named Entities (Q. Li et al., 2014).

Bag-of-words is commonly used in financial text research (Gidófalvi, 2001). It is also one of the most famous approach from part-of-speech. However, bag-of-word approach has noise issue caused by seldom-used terms and scalability problem resulted from large number of terms (Schumaker, Zhang, Huang, & Chen, 2012).

Noun Phrases is an improved text representation system which extracts nouns and noun phrases from the text document and can sufficiently represent the important concept of the news text (Tolle & Chen, 2000). As Noun Phrases technique only uses the noun and noun phrase to represent the text, it reduces the dimension of the textual data which further results in a better article scaling (Schumaker & Chen, 2009a).

Universiti Malaya

Named Entities is also a technique from part-of-speech and it is an extension of the Noun Phrases. Named Entities selects those proper nouns which are located in well-defined categories only. To find out which categories those terms should be, it uses a semantic lexical hierarchy (Sekine & Nobata, 2004) and a syntactic tagging process (McDonald, Chen, & Schumaker, 2005). Named Entities also does not have scalability problem because it reduces the selected terms to the specific category of nouns.

2.4 Methods for Text Classification

News sentiment classification methods can be grouped into two categories: lexicon-based classification methods and machine learning algorithms classification methods. Lexicon-based approaches classify the news sentiment by using the external lexica such as dictionary or corpus. Machine learning algorithms in news sentiment classification are mainly supervised approaches, which relies on the labelled training documents (Biltawi et al., 2016). When it comes to classifying high dimension of textual data, machine learning classifiers are more effective (Lei et al., 2011)

2.4.1 Lexicon-based Classification Methods

After the text documents have been properly represented, the lexicon-based method can be used to further analyze whether the news text is negative, positive or neutral. This approach can measure the sentiment of text document by analyzing the sentiment of those words or sentences in the document (Chan & Chong, 2017). Lexicon-based sentiment classification approaches consist of two main categories: dictionary-based approach and corpus-based approach (Biltawi et al., 2016). Lexicon-based dictionaries can be built manually or automatically. In corpus-based approaches, a dataset of certain corpus can also be used for news sentiment classification. In the research conducted by Rao et al (2014), they proposed a word-level sentiment dictionary which is automatically generated by maximum likelihood estimation and Jensen's inequality. Esuli et al (2010) proposed a

Universiti Malaya

system named SentiWordNet which has better accuracy on analyzing sentiment of the words. It is a system based on the existing semantic analysis tool called WordNet. Simplex lexicon-based methods have shortcomings as it can easily ignore the linguistic conventions and external evidences of the natural language expression. In order to solve those problems, Xiaowen Ding et al (2008) proposed a holistic lexicon-based approach which is built on Opinion Observer and this approach can analyze words without ignoring the whole context.

2.4.2 Machine Learning Algorithms

Machine learning algorithms can also be used to classify the news text into different categories. Machine leaning techniques need two sets of data for classification - training set and test set. Machine learning classifiers can classify the test set data according to the classification model which is developed based on the training set data (Neethu & Rajasree, 2013). Like feature processing, there are also a variety of classifiers which are developed from machine learning algorithms. The following section explains three popular news text classification classifiers: Naïve Bayes classifier, Maximum Entropy classifier and Support Vector Machine classifier.

1) Naïve Bayes classifiers: Naïve Bayes classifier is popular in news text classification it is built upon an attribute independent assumption and Bayesian theorem (Dey, Chakraborty, Biswas, Bose, & Tiwari, 2016). The Naïve Bayes have been extensively studied in the text classification task and it has been proven to be a simple model and can classify the text very effectively (Farid, Zhang, Rahman, Hossian, & Strachan, 2014).

The existing researches on text classification with Naïve Bayes are mainly focusing on three aspects. Firstly, there are researches focusing on constructing and improving Naïve Bayes model. Secondly, some researchers discuss the ‘naïve hypothesis’ then present the corresponding improvement based on mathematics. Lastly, the feature selection for Naïve

University Malaya

Bayes was also studied since Naïve Bayes Algorithm is very sensitive to features (W. Zhang & Gao, 2011).The way how Naïve Bayes algorithm works in news text classification is explained below (H. Zhang, 2006).

Assume the news text document is represented by a vector of variables, D = < d_i >, i = 1, 2, …, n. d_i can be a letter, a word, or other features selected from the text. In addition, there is a set of C which is predefined classes. C = {c_1, c_2, …, c_k}. The task of classification in Naïve Bayes model is to assign a class label c_j, j = 1, 2, …, k from C to analyzed document. Given a document D, the probability of its class c_j can be calculated as:

$$P\,(c_j \mid \mathrm{D}) = \frac{P(c_j)P(D|c_j)}{P(D)} \tag{2.1}$$

$P(c_j)$ is the probability of class c_j appears in the document, $P(D)$ is the knowledge from the text document itself to be classified. $P(D|c_j)$ is the probability of document D is attributed to class c_j. Naïve Bayes classifier computes separately the posteriori of document D falling into each class c_j, and assign the document to the class with the highest probability, which is,

$$C^*(D) = arg_j \max P(\,C_j \mid D) \tag{2.2}$$

Assume the d_i of document D are independent with each other. The conditional probability of $P(D|c_j)$ cannot be computed directly in the practice. Thus,

$$P(D|c_j) = \prod_i \; P(d_i \,|c_j) \tag{2.3}$$

The model with the assumption above is called Naïve Bayes model, and formula (2.1) becomes

$$P\big(c_j \,\big|D\big) = \frac{P(c_j)\prod_i \; P(d_i|c_j)}{P(D)} \tag{2.4}$$

Because of the $P(D)$ is identical to each class c_j, j = 1, 2, …, k, formula (2) becomes

Universiti Malaya

$$C^*(D) = arg_j max P(c_j) \prod_i P(d_i | c_j) \quad (2.5)$$

In spite of its simplicity and the fact that its conditional independence assumption is clearly not existed in real-world situations, Naïve Bayes classifier is surprisingly performs well in text classification (Farid et al., 2014). Furthermore, based on the concept that Naïve Bayes classifiers are sensitive about features, Jang et al (2016) proposed a deep feature weighting approach for Naïve Bayes classifier, which significantly improves the performance of the classifier.

2) Maximum Entropy Classifiers: Unlike Naïve Bayes, Maximum Entropy does not make the independence assumptions for its features. This means that the features like bigrams and noun phrases can be added to Maximum Entropy's feature without causing feature overlapping (Go, Bhayani, & Huang, 2009). Maximum Entropy models are feature-based models. Other than estimating the probabilities based on imposed constraints, Maximum Entropy models prefer to make as few assumptions as possible to build the most uniform models (Perikos & Hatzilygeroudis, 2016).

In the text classification, Maximum Entropy assigns each word of the document d, a class c based on the training data D. It computes the conditional distributed $P(c|d)$ by taking the following formula:

$$P(c|d) = \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right) \quad (2.6)$$

In the equation (6), $Z(d)$ is a normalization function, which is computed as:

$$Z(d) = \sum_c exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right) \quad (2.7)$$

$\lambda_{i,c}$ is the feature parameter weights and it must be learned by estimation (El-halees, 2007). A large $\lambda_{i,c}$ means that feature f_i is considered a strong indicator for class c (Pang, Lee, Rd, & Jose, 2002).

Universiti Malaya

$F_{i,c}$ is a feature/class function for feature f_i and class c. It is a binary valued feature which can make the prediction of the outcome. It is defined as follows:

$$F_{i,c}(d, c') = \begin{cases} 1, n_i(d) > 0 \; and \; c' = c \\ 0, otherwise \end{cases} \quad (2.8)$$

When conditional independence assumptions are not met, Maximum Entropy classifiers may potentially outperform the other machine learning algorithms since Maximum Entropy makes no assumptions about the relationship between the features (Go et al., 2009; Turney, 2002). Shenghuo Zhu et al (2005), proposed a multi-labeled text classification system based on Maximum Entropy methods. Their research shows that their system significantly outperforms those systems which use combination of single label approach. More researches proved that in practice, even though Maximum Entropy performs better in handling the feature overlap, Naïve Bayes can still outperform the Maximum Entropy in various classification tasks (Perikos & Hatzilygeroudis, 2016).

3) Support Vector Machine Classifiers: Support Vector Machine (SVM), as a binary classifier, has been widely and successfully used in text classification tasks as well as many other supervised learning tasks (Li, Fong, Zhuang, & Khoury, 2015; Zhang, Dang, Chen, Thurmond, & Larson, 2009). By applying the training data, SVM classifier can find a hyperplane as its decision surface which separates the training sets into two parts, negative and positive (Vapnik, 2013). Unlike probabilistic classifiers such as Naïve Bayes and Maximum Entropy classifier, SVM classifiers are large-margin classifiers (Parikh & Shah, 2016).

In the task of classifying two categories of the documents, the training procedure of SVM classifier can find a hyperplane which is represented by vector $\vec{w}$. This hyperplane not only separates the document vectors into two parts, but also separates them with a as large as possible margin. The searching of the hyperplane with maximum margin

Universiti Malaya

corresponds to a constrained optimization problem (Pang et al., 2002).

Training documents are represented as pairs $(\overrightarrow{x_i}, y_i)$. $\overrightarrow{x_i}$ is the weighted feature vector of the training example and $y_i \in$ {-1,1} is the label of the training example. $||\vec{w}||$ denotes the L_2-norm of the $\vec{w}$, therefore, the maximizing margin is equivalent to minimizing $\vec{w} \cdot \vec{w}$, that is $\frac{1}{2}||\vec{w}||^2$ subject to

$$y_i(\vec{w} \cdot \overrightarrow{x_i} - b) \geq 1, \ \forall_i \tag{2.9}$$

Vector $\vec{w}$ defines the orientation of the hyperplane and b defines the location of hyperplane. The learned hyperplane is defined by positive and negative support vectors. After $\vec{w}$ and b are learned, then based on the feature vector $\vec{x}$ of an unlabeled document, the SVM uses function $f(\vec{x}) = \vec{w} \cdot \vec{x} - b$ to compute the score for this document. If $f(\vec{x}) \geq 0$, the document can be labeled positive, otherwise, the label of the analyzed document is negative. SVM takes $f(\vec{x}) = 0$ as a default thresholding in its classification function (Meyer & Wien, 2015).

Aixin Sun et al (2009) conducted a research about comparing the experimental results of classifying the text data by applying 10 commonly used methods. They found out that when it comes to imbalanced text classification, the best decision surface is often learned by SVM other than any other strategies.

In the research conducted by Wang et al (2012), it turns out for the long text sentiment classification task, Support Vector Machines significantly outperform the Naïve Bayes. Based on their experiment, the SVM variants perform better than most published results on analyzing the sentiment of datasets and even sometimes reach the new state-of-the art performance level.

Universiti Malaya

2.4.3 Hybrid Methods for Text Classification

The hybrid methods, by combining the lexicon-based and machine learning approaches together, has the potential to improve the performance of sentiment classification (D'Andrea, Ferri, Grifoni, & Guzzo, 2015).

Mukwazvure and Supreethi (2015) used hybrid approach in their study of analyzing sentiment of news comments. They applied AFFIN-111 word list to label the training data and used SVM machine learning algorithm to classify the sentiment of news comments from Technology, Politics and Business sections on the guardian website (www.theguardian.com). Their system's accuracy of analyzing the sentiment of news comments under the Technology section achieved 74%. Nasim (2018) also conducted a research on analyzing the sentiment of financial microblogs. He proposed a system which combined the machine learning algorithm XgBoost Regressor and a lexicon-based approach named Loughran and McDonald Financial Sentiment Dictionaries. This system is among the top scorers of those proposed solutions for SemEval[1] tasks.

2.4.4 Labelling approaches in Hybrid Methods

In hybrid methods, the lexicon-based approaches are applied in labelling the training data for machine learning algorithm-based classifiers. Compare manual labelling, using automatic labelling is less laborious. In a research which aims to classify the sentiment of financial microblogs (Cortis et al., 2018), it took four financial experts 120 hours (30 hours per expert) to annotate 5218 sample sentiment.

[1]SemEval (Semantic Evaluation) is an ongoing series of computational semantic evaluation systems. SemEval community holds the evaluation workshop annually in association with *SEM conference (SemEval Portal (n.d.). In ACLwiki. Retrived April 14, 2019 from https://aclweb.org/aclwiki/SemEval_Portal).

Universiti Malaya

Moreover, simply using the lexicon-based approaches in labelling the training data cannot ensure the performance of machine learning algorithm-based classifiers. Since there are various lexicon-based resources, different lexicon-based dictionary or corpus may result in different labels. Loughran & Mcdonald, (2011) proved that Harvard Psychosociological Dictionary, specifically, the Harvard-IV-4 TagNeg (H4N) which is a commonly used dictionary for sentiment analysis is not suitable for financial news sentiment analysis. In their research, they found that according to the Harvard list, almost three-fourths (73.8%) of the negative word counts are attributable to words that are typically not negative in a financial context. Due to the sentiment of article is highly influenced by the background of the text, to analyze sentiment of financial news, a sentiment dictionary in financial domain is required (Ito, Izumi, Sakaji, & Suda, 2017).

2.5 Event Study Methodology

The event study method was first introduced in 1969 (Fama, Fisher, Jensen, & Roll, 1969). Event study is a statistical analyze technique aims to estimate the stock market's reaction to certain events such as important personnel announcement of the company, mergers, dividend announcements and so on (Sorescu, Warren & Ertekin, 2017). There are two kinds of information that may cause the fluctuation in stock prices: Information that is released by company such as dividend announcement or personnel change announcement and the information that likely to affect the stock prices such as big flaw reported found in the product and influential news from third parties (Akita, Yoshihara, Matsubara, & Uehara, 2016). Thus, in order to analyze the OPEC news impact on selected companies stock prices in this research, event study method is also pivotal.

To study the effects of OPEC news on oil & gas companies stock prices, the event study methodology is needed (Loutia et al., 2016). Event studies examine the abnormal returns happen in the stock market around a relevant event time. It has been widely

Universiti Malaya

applied financial economics research but barely been used in pertaining to study OPEC news announcements, and energy sector (oil & gas) stock prices. The stock prices may react to the information immediately or over a certain period. Thus, choosing a proper event window is critical for the research. In order to prevent overlapping among OPEC news announcements, avoid the contamination from other events, and to capture the leakage of information before the OPEC events, there are existing studies which show that five days event window is more appropriate (Bina & Vo, 2007; Horan et al., 2004).

The event's impact on the stock prices can be measured by the abnormal return of the markets which happens in the event window period. The abnormal return can be calculated as follows:

$$AR_t = R_t - E(R_t) \tag{2.10}$$

R_t is the daily log return on energy sector (oil & gas) stock prices at date t. $E(R_t)$ is the normal return which is an expected return based on the assumption that the event does not occur (Ji & Guo, 2015).

In the study conducted by Lin and Tamvakis (2010), they applied event study methodology in studying the impact of OPEC announcements on crude oil prices. In their research, they found out that for the most of abnormal returns, the data series has zero mean. Thus, using the 'mean adjusted' return to calculate abnormal returns has no significant difference from zero mean.

2.6 Comparison of Existing News Studies

The existing studies about news analysis can be divided into two categories – content-oriented and sentiment-oriented. For content-oriented news analysis, researchers aimed to find the relationship between the content of financial news and the fluctuation of stock prices. On the other hand, in sentiment-oriented news studies, the fluctuation of stock

Universiti Malaya

prices was investigated based on news sentiments. Different techniques used in those studies achieved different level of accuracy in the classification. Table 2.1 and Table 2.2 present a brief comparison of research studies on financial news based on content-oriented and sentiment-oriented, respectively. Both tables list the dataset used, feature processing techniques, and the accuracy achieved in the classification method.

Table 2.1: Comparison of Existing News Studies (Content-Oriented)

Author	Dataset	Feature Processing		Classification		
		Feature Type	Selection Method	Method	Accuracy	
(Schumaker & Chen, 2009b)	US Financial News	Noun Phrases	Minimum occurrence per document	SVM	58.2%	Content Analysis-Oriented
(Groth & Muntermann, 2011)	German ad hoc announcement	Bag-of-Words	Only stop words removal	SVM	56.5%	
(Kaya, 2010)	US Financial News	Couple words	Chi-square	SVM	59%	
(Schumaker et al., 2012)	US Financial News	Noun Phrases	Minimum occurrence per document	SVR	59.0%	
(Hagenau et al., 2013)	DGAP (Deutschce Gesellschaft fur Adhoc-Publizitat) and EuroAdhoc	N-Grams	Bi-normal separation, Chi-square	SVM	65.4%	
(Atkins, Niranjan, & Gerding, 2018)	Reuters US News	Topic of the Article	Latent Dirichlet Allocation	Naïve Bayes	63%	

Universiti Malaya

Table 2.2: Comparison of Existing News Studies (Sentiment-Oriented)

Author	Dataset	Feature Processing		Classification		
		Feature Type	Selection Method	Method	Accuracy	
(Ranco, Aleksovski, Caldarelli, Grčar, & Mozetič, 2015)	Twitter	N-Grams	Human Annotation	SVM	76%	Sentiment Analysis-Oriented
(Jian Li, Xu, Yu, & Tang, 2016)	Thomson Reuters News	Bag-of-Words	—	Henry's-Specific dictionary	67%	
(Sinha, 2016)	Thomson Reuters News Scope	Noun-Phrases	Specialist Manually Label	Neural Network	75%	
(Seng & Yang, 2017)	Financial News from Knowledge Management Winner (KMW)	Bag-of-Words	Chi-square	Manually built dictionary	69.5%	
(Nasim, 2018)	Microblogs on financial domain	Bag-of-Words	TF-IDF	Loughran McDonald Financial Sentiment Dictionaries, XgBoost Regression	65.5%	

Universiti Malaya

2.7 Summary

Based on literature review, the commonly used methods of news sentiment classification can be classified into two categories as shown in Figure 2.1.

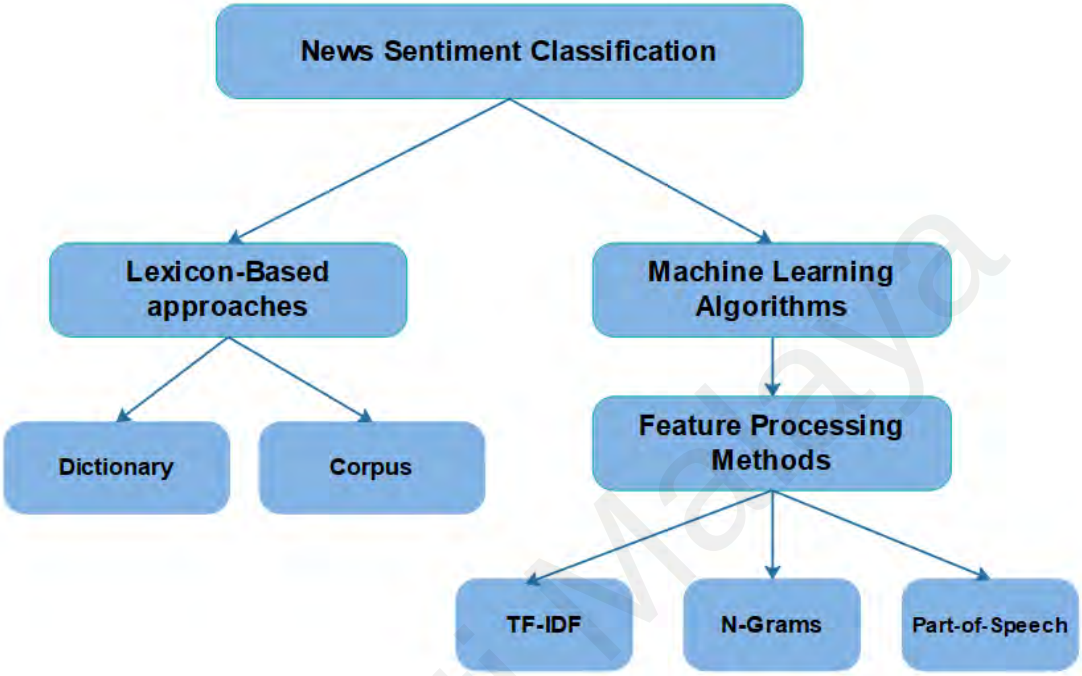

Figure 2.1: Summary of News Sentiment Classification

The first category: lexicon-based classification approaches are further divided into two different approaches – dictionary-based approach and corpus-based approach. In the second category: machine learning algorithms, despite of varieties of existing algorithms, the feature processing methods for machine learning also vary. The commonly used important feature processing methods are TF-IDF, N-Grams and Part-of-Speech approaches. It is also found that combining both lexicon-based and machine learning algorithm-based approaches, the classifier achieves better accuracy (Mukwazvure & Supreethi, 2015).

University Malaya

CHAPTER 3: RESEARCH METHODOLOGY

Research methodology is a scientific and systematic way to solve a research problem. It is very essential because it illustrates how the research is conducted and the researcher's logic in reaching the research goals. This chapter explains the methodology used and the research methods applied to answer the following research questions: (i) How to build an innovative classifier to analyze the OPEC news sentiment? (ii) How to improve the accuracy of the innovative classifier? (iii) How the OPEC news sentiment impact on the stock prices of public listed Malaysian energy sector (oil & gas) companies?

In this chapter, the rationale for using combination of qualitative and quantitative research design is described. The design of the research is then introduced and illustrated in a diagram. Also, the techniques used in this research are also explained. Finally, the sampling method and details of the datasets used in this research are explained.

3.1 Qualitative and Quantitative Research Method

Qualitative research is an inductive research which attempts to interpret certain phenomena or experience in a specific context, and at a particular period of time. Qualitative research data are collected directly from the research participants. The results of the research can be illustrated in the research participants' angle (McCusker & Gunaydin, 2015). Thus, qualitative research design is suitable for this study since (i) this research aims to analyze the OPEC news sentiment (text-related); (ii) it uses an inductive approach to analyze the data which were collected directly from the research targets.

Quantitative research aims to find the cause and effect relationship, and build a statistical model after analyzing the features. The data applied in quantitative research is normally numbers and statistics. Furthermore, the data is analyzed using mathematically-based methods (Almalki, 2016). This research also adopts quantitative methodology because (i) this study generates results by analyzing the historical data of the stock prices

Universiti Malaya

of selected public listed energy sector (oil & gas) companies (numerical data); (ii) an event study methodology (mathematically-based method) is applied in this research.

3.2 Research Activities

To answer the research questions, this research was carried out as follow:

Firstly, to get ideas about the solutions for the research, literature review was conducted that covered the background of the Organization of The Petrol Exporting Countries (OPEC), the influence of OPEC news, the commonly used techniques for classifying the news textual based on its sentiment and the methodology for analyzing the fluctuation of stock prices caused by certain events.

Secondly, to achieve the research goals, suitable data sets were collected and generated in a proper way for further analysis. This research uses two types of data - financial news textual data from both Wall Street Journal website and OPEC official press releases and stock prices (numerical data) of the six selected energy sector companies listed in Bursa Malaysia.

Thirdly, based on literature review, correct techniques were selected and applied to process the financial news textual datasets. To classify the OPEC news according to its sentiments, different machine learning algorithms are tested. The OPEC news sentiments are classified into three different categories: Negative, Positive and Neutral. Then, the six selected stock prices of the company datasets are analyzed using methods of the financial research domain.

Finally, the results from both the OPEC news classification and stock prices fluctuation are compared to answer the research questions and achieve research objectives defined in Chapter 1. Figure 3.1 shows the research activities of this research.

Universiti Malaya

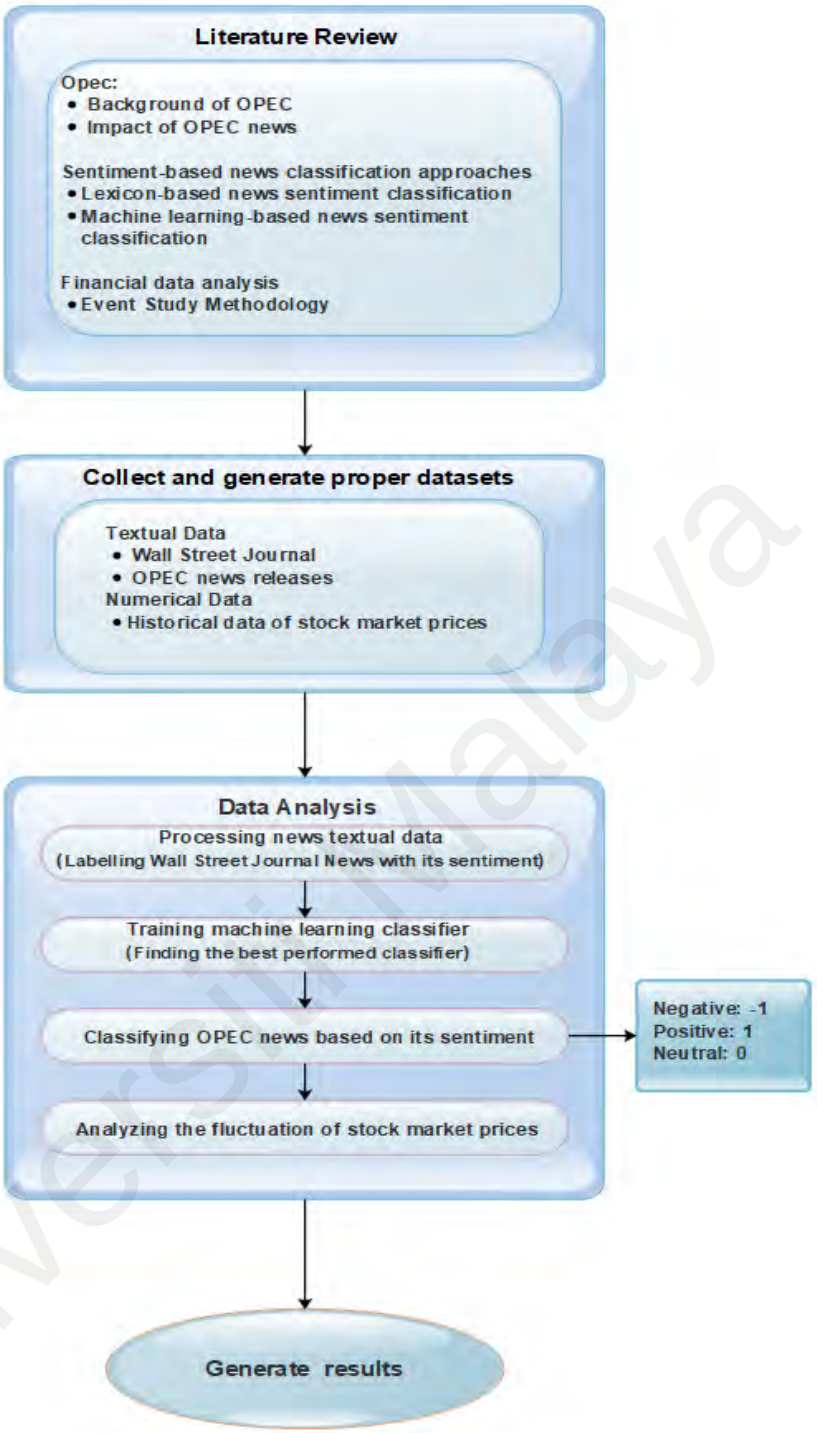

Figure 3.1: Research Activities

3.3 Selection of Textual Data

The Wall Street Journal financial news text (Chen, 2017) dataset is used in this research. The Wall Street Journal, printed since 1889 and has its online version since 1995, is one of the largest business-focused English-language newspaper in the United States by circulation (Salwen, Garrison, & Discoll, 2004).

Universiti Malaya

This dataset is used as training data for machine learning algorithms. This dataset aims to provide sufficient sentiment words in the financial news domain for ensuring accuracy in classifying the OPEC news. As this research aims to study the impact of OPEC news sentiment on stock prices of public listed Malaysian energy sector (oil & gas) companies, OPEC news textual data were also collected and used in this research.

3.4 Selection of Stock Market Data

Based on the information found on the Bursa Malaysia website, there are 28 energy sector (oil & gas) companies are listed on the Main Market Board of Bursa Malaysia (*Bursa Malaysia sectorial index series*, 2018). Proper sampling can provide suitable dataset for the research. At the same time, it can also reduce the time spent in data analysis without causing any undesired effect on the results. In this research, simple random sampling is used to generate the proper dataset. This sampling method not only has the highest generalizability but also gives the least bias (Bangi, 2007). Thus, the six companies chosen are randomly selected from the 28 energy sector companies. The historical stock prices of these six companies are collected from the Yahoo Finance website (https://finance.yahoo.com/).

Yahoo Finance website has been one of the top financial research site in the United States since 2008 which provides historical and current information about stock exchange rates, financial reports, stock quotes and corporate press releases (Bordino, Kourtellis, Laptev, & Billawala, 2014). There are many financial researches conducted based on the data collected from Yahoo Finance. For instance, Xu (2014) conducted a research about forecasting stock prices based on the information obtained from Yahoo Finance. Ko et al (2015) studied the relationship between economic policy uncertainty and stock prices based on the historical stock prices data collected from this website as well. There was also a research about social media sentiment's influence on stock prices which were based

Universiti Malaya

on the stock prices data published on Yahoo Finance (Nguyen, Shirai & Velcin, 2015). Therefore, the historical stock prices data of the six energy sector (oil & gas) companies are accurate and reliable.

3.5 Research Design

Figure 3.2 shows the design of this research. As shown in the figure, Wall Street Journal news textual data and OPEC news textual data were applied separately in machine learning classifiers. In this research, different machine learning algorithms-based classifiers were tested to find the outperforming algorithms to classify the OPEC news. After the historical stock prices have been analyzed, the results were compared with the OPEC news sentiment. Then, the outcome of the relationship between OPEC news sentiment and the fluctuation of stock prices of the six public listed energy sector (oil & gas) companies in Bursa Malaysia was generated. The following sections further explain the methods applied in this research.

Universiti Malaya

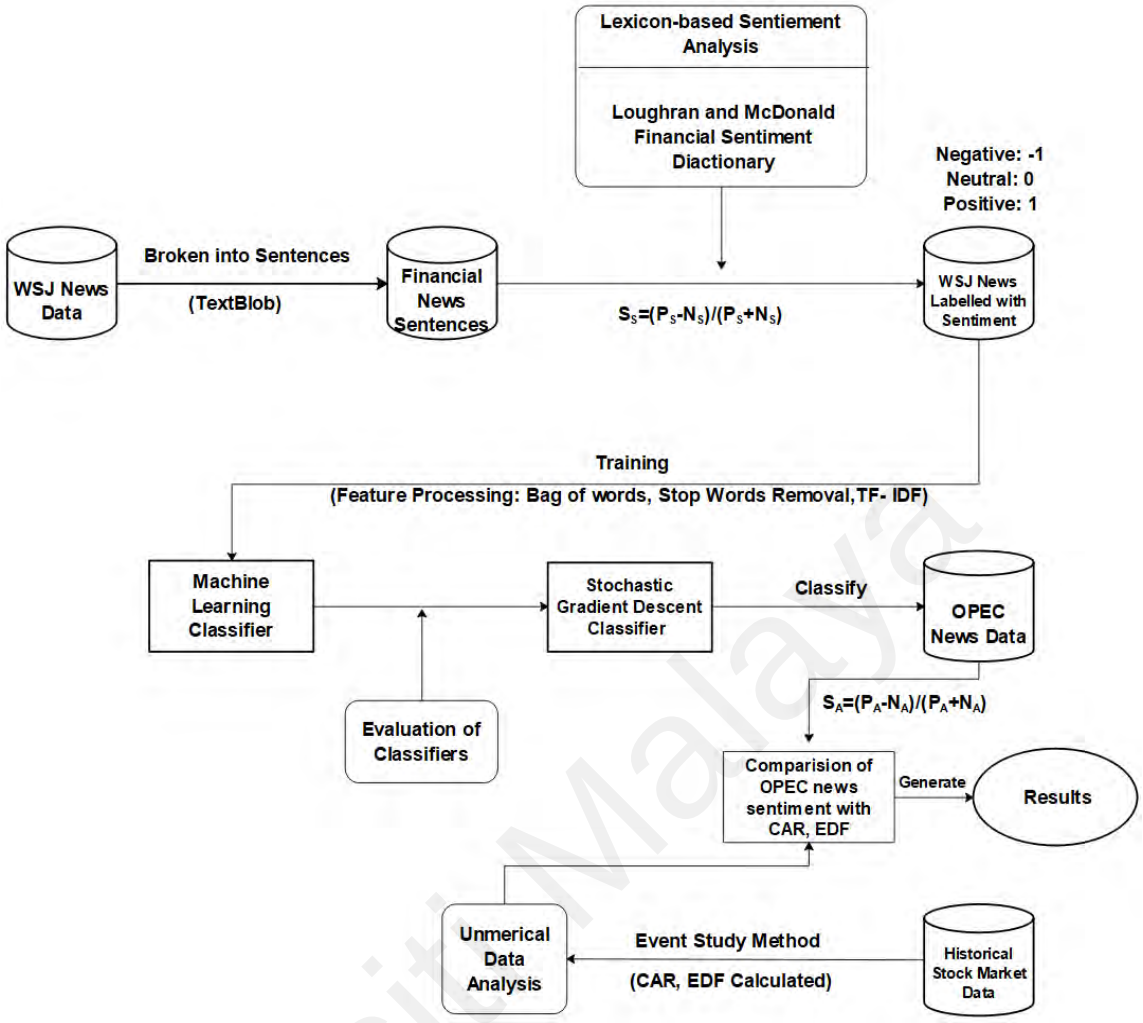

Figure 3.2: Research Design

3.6 Textual Data Processing Methods

Wall Street Journal (WSJ) news articles data is first broken into financial news sentences using the tool, Textblob (Loria et al., 2014). After being further processed by lexicon-based sentiment analysis, these sentences are used as training data for the machine learning classifier. In this research, the labelling method and techniques applied to process these financial news sentences are explained below.

3.6.1 Textual Data Labelling

The lexicon-based sentiment analysis approach determines the sentiment of the text by detecting those sentiment lexicon words in the text (Hutto & Gilbert, 2014). Lexicon-based labelling approach is used to preprocess the textual data for two reasons: 1) The

University Malaya

textual data needs to be labelled with its sentiments so that it can be further processed and applied to train the machine learning classifiers; 2) Comparing to manual labelling, using lexicon-based labelling is much more effective.

As the sentiment words in normal news articles and financial news articles may vary, Loughran and McDonald Financial Sentiment Dictionaries (Loughran & Mcdonald, 2011) is applied to label the textual training data. In this approach, each word in the sentence is analyzed by comparing it with the sentiment words stored in the dictionary to determine whether it is positive or negative. The sentiment of a sentence is determined by the difference in counts between the positive words and negative words. This research uses relative proportional difference evaluating method to calculate the sentiment of a sentence based on the positive and negative sentiment words exist in the sentence (Will, Benoit, Slava & Laver, 2011). The formula for calculating the sentiment of a sentence is as follow:

$$S_S = (P_S - N_S) / (P_S + N_S) \quad (3.1)$$

S_S: Sentiment score of a sentence.

P_S: The number of positive words in the sentence.

N_S: The number of negative words in the sentence.

The measure ranges from -1 to 1. If $S_S = 0$, the sentence's sentiment is neutral. If $S_S > 0$, it means that the sentence's sentiment is positive. Otherwise, it is a negative sentence.

As mentioned in section 3.6, the WSJ news textual data is analyzed in sentence level. News articles in Wall Street Journal dataset are firstly broken into sentences using Textblob. Textblob is an easy-to-use library in Python which can break news articles into sentences (Loria et al., 2014). The Loughran and McDonald Financial Sentiment

Universiti Malaya

Dictionaries based sentiment analysis function is implemented in a python library named pysentiment (Han, 2012). Hence, pysentiment was used to accomplish the lexicon-based labelling task. Based on the sentiment score obtained from formula 3.1, the Wall Street Journal dataset sentences are labelled with its sentiment – positive sentences are labelled with 1, neutral sentences are labelled with 0, and negative sentences are labelled with -1.

Similarly, the relative proportional difference evaluating method can also be used to calculate an article's sentiment. The sentiment of an article can be calculated by using the same formula:

$$S_A = (P_A - N_A) / (P_A + N_A) \quad (3.2)$$

S_A: Sentiment score of the article.

P_A: The number of positive sentences in the article.

N_A: The number of negative sentences in the article.

The measure is also ranging from -1 to 1. If S = 0, the article's sentiment is neutral. If S > 0, it means that the article's sentiment is positive. Otherwise, it is negative.

3.6.2 Natural Language Processing

Classifying the news text is part of the Natural Language Processing (NLP) task in Artificial Intelligence domain. NLP applies computational techniques to analyze, learn, understand and reproduce human languages (Hirschberg & Manning, 2015). Processing textual data needs Natural Language Processing (NLP) tools. Since Machine learning algorithms only process numerical data, feature processing methods are further applied to prepare textual data ready to be used in machine learning algorithms (Mukwazvure & Supreethi, 2015).

Universiti Malaya

It has been proven that both the feature processing and classification methods have significant influence on the performance of text classification (Uysal & Gunal, 2014). By applying the proper feature processing methods on the textual data, not only can reduce the scalability problem, increase the accuracy of classification but also prepare the textual data in a proper way which can further be applied to machine learning algorithms (Uysal, 2015).

The feature processing methods used in this research contains bag-of-words representation, stop words removal, TF-IDF weighting and vectorization. Using bag-of-words representation approach, the text is represented as a collection of its words, ignoring the grammar, order of the words, and the context (Le & Mikolov, 2014). As bag-of-words is an easy-to-use tool and can produce proper representation of the text, the bag-of-words representation is a popular method in natural language processing. Generally, bag-of-words representation contains tokenization, occurrence counting and normalization (Pedregosa et al., 2011).

After the textual data have been processed using bag-of-words representation, there is a large number of non-informative words such as prepositions, and conjunctions which need to be removed from the article. Therefore, the approach of stop words removal is applied to further processing the textual data. Stop words removal methods can reduce the dimension of data and improve the effectiveness of the machine learning classifiers (Vijayarani, Ilamathi, & Nithya, 2015). The stop words removal is accomplished by using NLTK natural language processing toolkit in Python (Perkins, 2010). The stop words list of NLTK (Perkins, 2010) is as follow:

'ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out', 'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'its', b 'from', 'him', 'each', 'the', 'themselves', 'until', 'below', 'are', 'we', 'these', 'your', 'his',

Universiti Malaya

'through', 'don', 'nor', 'me', 'were', 'her', 'more', 'himself', 'this', 'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all', 'no', 'when', 'at', 'any', 'before', 'them', 'same', 'and', 'been', 'have', 'in', 'will', 'on', 'does', 'yourselves', 'then', 'that', 'because', 'what', 'over', 'why', 'so', 'can', 'did', 'not', 'now', 'under', 'he', 'you', 'herself', 'has', 'just', 'where', 'too', 'only', 'myself', 'which', 'those', 'i', 'after', 'few', 'whom', 't', 'being', 'if', 'theirs', 'my', 'against', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', 'than'.

To identify the important and informative words of the text, Term Frequency-Inverse Document Frequency (TF-IDF) weighting approach is applied after removing the stop words. TF-IDF is a numerical value which represents the importance of a word for a document or a collection of corpora (Vijayarani et al., 2015). TF-IDF can be calculated using the following formula:

$$A_{ij} = Tf_{ij}/Df_i = Tf_{ij} * \log_2(\frac{N}{Df_i}) \quad (3.3)$$

A_{ij}: the weight of term i in document j

N: the number of documents

Tf_{ij}: the term frequency of term i in document j

Df_i: the document frequency of term i in the whole collection.

As it is shown in formula (3.3) above, the TF-IDF value increases proportionally to the frequency of a word appears in a document, but is offset by the number of times of that word appears in the corpus (Trstenjak, Mikac & Donko, 2014). In this way, TF-IDF can calculate whether a word is common or rare across all documents (Munot & Govilkar, 2014). TF-IDF not only can weigh the words based on its importance but also transform

Universiti Malaya

the textual data into numerical data. This essential step is also called “vectorization” (Tripathy et al., 2016). After the textual data have been transformed into numerical data, it can be applied to train the machine leaning classifiers. All these steps can be processed using the free python library named Scikit-Learn (Pedregosa et al., 2011). In this library, it contains a function named, TfidfVectorizer. Through this function the textual data can be processed with tokenization, stop words removing and TF-IDF weighting through TfidfVectorizer. Figure 3.3 shows the steps to processing the textual data.

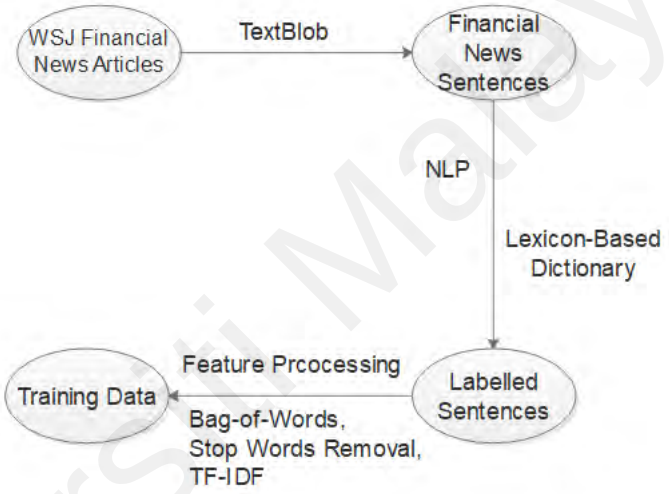

Figure 3.3: Training Data Preparing

3.7 Machine Learning Algorithms

Machine learning algorithms, with the purpose of making computers learn from experience, is one of the most rapidly developing techniques which settles in the intersection research field of statistics and computer science (Jordan & Mitchell, 2015). Machine Learning (ML) algorithms have shown a deep and diversify development due to its aim to solve various problems and to cover a wide variety of different kinds of data (Witten, Eibe, Hall & Pal, 2016).

Each instance in the dataset applied to machine learning algorithms is represented using the same set of features. In supervised machine learning algorithms, those instances

Universiti Malaya

are given with known labels. Without those labels, it is called unsupervised machine learning (Jordan & Mitchell, 2015). By applying unsupervised learning, researchers aim to find out the unknown but useful classes of items (Libbrecht & Noble, 2015). On the contrary, the supervised machine learning, aims to train the learner with known features and labels. There is also another kind of machine learning algorithm called reinforcement learning. Reinforced training data of learning algorithms is provided by the external trainer. Those data are in the form of a scalar reinforcement signal which constantly return the measure of system's performance. Therefore, the learner can discover which action has the best result by trying each action in turn (Sutton & Barto, 2018).

The supervised machine learning algorithms are applied in this research. The classifiers are trained by labelled textual data. By using supervised learning, the classifier can distribute the class labels to the unlabeled testing data in terms of learned features. To find the best performed supervised machine learning algorithm in classifying OPEC news, this research applied and tested multiple commonly used and state-of-the-art machine learning classifiers separately. The following sections explain the supervised machine learning algorithms used in this research in detail.

3.7.1 Naïve Bayes Classifiers

Naïve Bayes is one of the most effective and efficient inductive learning algorithms for machine learning. It has surprisingly outstanding performance in classification because of its conditional independence assumption which is rarely applicable in real-world situations (Raschka, 2014). Naïve Bayes classifiers have worked well in solving real-life problems such as document classification and spam filtering (Harisinghaney, Dixit, Gupta, & Arora, 2014).

Since there are different kinds of Naïve Bayes classifier, this research tested four of them to find the proper algorithm with best performance. Gaussian Naïve Bayes (GNB),

Universiti Malaya

Multinomial Naïve Bayes (MNB), Complement Naïve Bayes (CNB), and Bernoulli Naïve Bayes (BNB) are applied separately to train the classifiers.

Gaussian Naïve Bayes Classifier: It is a Naïve Bayes probabilistic-based model with assumption that the continuous values of each class is distributed by Gaussian distribution (Di Nunzio, 2014). Gaussian Naïve Bayes is one of the simplest classification algorithms. The formula of Gaussian Naïve Bayes (Ontivero, Castellanos, Valente, & Goebel, 2017) is as follow:

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \quad (3.4)$$

Whereby:

$p(x = v|C_k)$: The probability distribution of observation v given a class C_k

x: The continuous attribute in training data

μ_k: The mean of the values x associated with class C_k

σ_k^2: The variance of the values in x associated with class C_k

Multinomial Naïve Bayes Classifier: In this classifier, it models the distribution of words in a document as multinomial data. Therefore, the feature vectors of the training data which represent the frequencies of words appearance in the document are generated multinomially. And it assumes that word's position in the document is generated independently and it's irrelevant with other words' positions. The formula of Multinomial Naïve Bayes is as follow (Tang, He, Baggenstoss, & Kay, 2016):

$$p(x|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i Pki^{x_i} \quad (3.5)$$

Universiti Malaya

$p(x|C_k)$: The probabilistic statistic of observing feature vector X in class C_k

x_i: It is the count of event i appears in a particular instance

p_i: It is the probability that event i occurs

The parameters $p(x|C_k)$ is estimated by a smoothed version of maximum likehood:

$$\widehat{p(x|C_k)} = \frac{N_{yi} + a}{N_y + an} \tag{3.6}$$

$\widehat{p(x|C_k)}$: The probability of feature x appearing in a sample belonging to class C_k

$N_{yi} = \sum_{x \in T} x_i$ is the frequency of feature i appears in the training set T's class y

a : It is the smoothing priors, $a = 1$ is called Laplace smoothing, $a < 1$ is called Lidstone smoothing, $a \geq 0$ accounts for features not present in the learning samples and prevents zero probabilities in further computations.

Complement Naïve Bayes Classifier: It is built upon the standard multinomial Naïve Bayes algorithms but improved to particularly suit for imbalanced data sets. It uses the complement statistics for each class to compute the model's weights. The empirical results indicate that compared to the parameters estimate for Multinomial Naïve Bayes (MNB), parameters of Complement Naïve Bayes (CNB) algorithms are more stable. Thus, CNB normally outperforms the MNB in text classification task (Wang, Jiang, & Li, 2014). Since the CNB is based on the MNB explicated above, the difference between these two algorithms only lies in how they calculate the weights. The procedure for calculating the weights in CNB (Wang, Jiang, & Li, 2014) is as follow:

$$\widehat{p(x|C_k)} = \frac{a_i + \sum_{j:y \neq c} d_{ij}}{a + \sum_{j:y \neq c} \sum_k d_{kj}} \tag{3.7}$$

Universiti Malaya

In CNB, the summations are calculated over all documents j not just in class c.

$\widehat{p(x|C_k)}$: The probability of feature x appearing in a sample belonging to class C_k

d_{ij}: It is the count or TF-IDF value of term i in document j

a_i: it is the smoothing parameter found by the same way as in MNB algorithm

a: It is the summation of a_i, a=$\sum_1^i a_i$

Bernoulli Naïve Bayes Classifier: In this algorithm, the training data is distributed based on multivariate Bernoulli distributions. In the text classification which uses BNB algorithm, the training dataset is required to be represented as binary-valued feature vectors. As it is like Multinomial Naïve Bayes, BNB algorithm is also popular in solving the text classification tasks (Tang, Member, Kay, He & Member, 2016). With the benefit that BNB algorithm can explicitly model the absence of terms, BNB classifier is especially commonly-used to classify the short text (Manning, Raghavan & Schütze, 2010). In BNB classifier, the probability of a document belongs to class C_k is calculated as follow (Manning, Raghavan & Schütze, 2010) :

$$p(x|C_k) = \prod_{i=1}^{n} p_{ki}^{xi} (1 - p_{ki})^{(1-x_i)} \quad (3.8)$$

$p(x|C_k)$: The probability of a document x belongs to class C_k

x_i: It is a Boolean value whether the $i'th$ term from the vocabulary appeared in the document or not

p_{ki}: It is the probability of class C_k generating the term x_i

Universiti Malaya

3.7.2 Support Vector Machine Classifier

Support Vector Machines (SVM) learning method was firstly introduced by Vapnik and Cortes in 1995 (Cortes & Vapnik, 1995). SVM embodies the structure which aims to minimize the structural risk exists in the training error and reduce the modeling complication (Meyer & Wien, 2015). Geometrically, SVM algorithms classify the data by constructing a separating hyperplane with the maximal margin, and the larger the margin, the lower chance of generalizing error of the classifier. Based on the fact that only Support Vectors are used for classification and those samples which are far from the decision boundary can be removed without affecting the classification, the SVM classifier have better accuracy on moderately imbalanced data than other standard classifiers (Lilleberg, Zhu, & Zhang, 2015).

With the purpose of building a classifier with better accuracy, this research also tested Support Vector Machines learning algorithms-based classifier.

Support Vector Classifier: Assume training vectors $x_i \in R^p, i = 1,2,3 \dots, n$, in two classes, and a vector $y \in \{-1,1\}^n$, Support Vector Classifiers solves the following primal problem (Cortes & Vapnik, 1995):

$$\min_{\omega,b,\zeta} \frac{1}{2}\omega^T\omega + C\sum_i^n \zeta_i \tag{3.9}$$

Subject to $y_i(\omega^t \emptyset(x_i) + b) \geq 1 - \zeta_i$

$$\zeta_i \geq 0, i = 1,2,3, \dots, n$$

$\emptyset(x_i)$: It maps x_i into a higher dimension space, $C > 0$, is the regularization parameter. Since the dimension of vector variable ω may be high, usually the dual problem can be solved as (Cortes & Vapnik, 1995):

Universiti Malaya

$$\min_{a} \frac{1}{2} a^T Q a - e^T a \tag{3.10}$$

Subject to $y^T a = 0$,

$$0 \leq a_i \leq C, i = 1,2,3, \ldots, n$$

e: It is the vector of all ones, $C > 0$, is the upper bound of margin

Q: It is an n by n positive semi-definite matrix

$Q_{ij} \equiv y_i y_j K(x_i, x_j)$, where $K(x_i, x_j) = \emptyset(x_i)^T \emptyset(x_j)$ is the kernel.

By applying function $\emptyset$, training vectors are implicitly mapped into a higher dimensional space.

3.7.3 Stochastic Gradient Descent Classifier

Stochastic Gradient Descent Classifier can perform multi-class classification by combining multiple binary classifiers into a one versus all (OVA) scheme. It is a simple but also efficient approach to discriminatively learn linear classifier under convex loss functions such as Logistic Regression and linear Support Vector Machine. It has been successfully applied to solve large-scale and sparse machine learning problems in natural language processing and text classification (Scikit, 2011a).

When it comes to a large scale of training data, the Stochastic Gradient Classifier (Tong, 2004) classifies the text by learning random training example at each iteration of the training data (Kabir, Siddique, Kotwal, & Huda, 2015). Thus, learning algorithms built upon Stochastic Gradient approximations are known for good performance on machine learning tasks. At the same time, it is also known for its poor results in optimization tasks. This is due to the convergence of the Stochastic Gradient Descent Classifier which is significantly limited by the stochastic noise caused by choosing one

Universiti Malaya

random example at each iteration of the training data (Kabir et al., 2015). The mathematical formula of Stochastic Gradient Descent Classifier (Kabir et al., 2015) is as follow:

Assume that there is a set of training examples $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \in R^m$ and $y_i \in \{-1.1\}$. The goal is to learn a linear scoring function $f(x) = \omega^T x + b$ with model parameters $\omega \in R^m$ and $b \in R$. By simply looking at $f(x)$, the predictions can be made. A common formula to find the model parameters by minimizing the regularized training error is as follow:

$$E(\omega, b) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i f(x)\right) + aR(\omega) \quad (3.11)$$

L: It is a loss function that measures model fit

R: It is a regularization term which penalizes the model complexity

a>0, It is a non-negative hyper parameter

Different choices of L entail different classifiers, such as:

Hinge: (Soft-Margin) Support Vector Machines

Log: Logistic Regression

Least-Squares: Ridge Regression

Epsilon-Insensitive: (Soft-Margin) Support Vector Regression

Universiti Malaya

3.7.4 Radom Forest Classifier

Random Forest algorithm was developed by Breiman and Cutler (Breiman & Cutler, 2007). It is an algorithm ensembled with both classification and regression methods and it aims to solve classification problem (Akinyelu & Adewumi, 2014).

The Random Forest Classifier contains a collection of tree-structured classifiers. Each of these classifiers are independently distributed random vectors (Belgiu & Dragut, 2016). In Random Forest algorithm, predictions are made by decision tress. Every classifier in the Random Forest algorithm can vote for the most popular class of the input. Thus, by using Random Forest Classifier, the accuracy of prediction can be improved since it uses the average of the votes from multiple classifiers to decide the output. Moreover, in Random Forest algorithm, each classifier's sample size is the same as the input sample size, the over-fitting problem can be controlled (Scikit, 2011b). The mathematical formula of Random Forest Classifier (Belgiu & Dragut, 2016) is as follow:

Assume a collection of classifiers $h_1(X)$, $h_2(X)$,…, $h_K(X)$, and with the training set distributed randomly from the random distribution of vector Y, X, the margin function can be defined as:

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \qquad (3.12)$$

$mg(X,Y)$: The margin measures the extent to which the average number of votes at X

$I(\cdot)$: It is the indicator function.

Y: It is for the right class exceeds the average vote for any other class

The larger the margin, the more confidence in the classification.

The generation error is given by:

Universiti Malaya

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \quad (3.13)$$

The X, Y subscripts of P indicates that probability is over X, Y space

3.8 NLP and Machine Learning in Python

There are several programming languages suitable for preprocessing the textual data and building the news text classifier based on machine learning algorithms. After conducting an in-depth literature review, python is chosen in programming part of this research. Python is an interpreted, object-oriented and high-level programming languages. It becomes one of the most popular programming language in exploratory data analysis due to its developed ecosystem of scientific libraries (Raschka, 2015).

For the NLP part of this research, python has a specific module named Natural Language Toolkit (NLTK) for the NLP tasks. NLTK has an open-source license. It provides almost all the basic wrappers and functions for common NLP tasks. Moreover, NLTK contains not only modules but also both unprocessed and pre-processed standard corpora, which means, it also provides classic examples for learning how to use it before building a real project (Hardeniya, 2015).

Scikit-Learn (Pedregosa et al., 2011) is an easy-to-use python module which integrates those state-of-the-art machine learning algorithms for both supervised and unsupervised medium scale problems. Comparing to other machine learning algorithms modules in python such as MLPy (Albanese, Merler, & Jurman, 2008), Pymvpa (Hanke et al., 2009), MDP (Zito, Wilbert, Wiskott, & Berkes, 2009), Shogun (Sonnenburg et al., 2010) and PyBrain (Schaul, Bayer, Wierstra, & Sun, 2010), Scikit-Learn has better performance in the speed of computing which can significantly save the running time of the program, and hence, it is used in building the supervised machine learning classifier in this research.

Universiti Malaya

3.9 Performance Evaluation Measures for Classifiers

Evaluation of the machine learning algorithms classifiers mainly focuses on the ability of classifiers to classify the input data correctly. To evaluate the performance of text classifiers, proper evaluation measures are needed. Different evaluation measures indicate different characteristics of the machine learning algorithm-based classifier (Vafeiadis, Diamantaras, Sarigiannidis & Chatzisavvas, 2015). The following section introduces several commonly used performance evaluation measures for machine learning classifiers.

Confusion Matrix: It is normally used to summarize the performance of supervised machine learning algorithms-based classifiers (Tripathy, Agrawal & Rath, 2016). It contains the information about the actual classification and the predicted classification results from the examined classifier. A confusion matrix has two dimensions, one dimension indexed the actual class of classified objects, the other indexed prediction of objects' class from classifier (Deng, Liu, Deng & Mahadevan, 2016). Table 3.1 illustrates a basic form of confusion matrix for a multi-class classification task.

Table 3.1: Confusion Matrix for Multi-class Classification (Deng et al., 2016)

		Predicted		
		A_1	... A_j ...	A_n
Actual	A_1	N_{11}	N_{1j}	N_{1n}
	... A_i ...	N_{i1}	... N_{ij} ...	N_{in}
	A_n	N_{n1}	N_{nj}	N_{nn}

Universiti Malaya

Assuming there are classes $\boldsymbol{A_1}$, $\boldsymbol{A_2}$,…, $\boldsymbol{A_n}$. In Table 3.1 above, the confusion matrix $\boldsymbol{N_{ij}}$ represents the number of samples belonging to class $\boldsymbol{A_i}$ but being classified to class $\boldsymbol{A_j}$ by the classifier.

Other commonly used measures for the performance of classifier can be defined based on the confusion matrix. These include accuracy, precision, recall and F-score which are explained below.

Accuracy: the proportion of total correct predictions.

$$Accuracy = \sum_{i=1}^{n} N_{ii} / \sum_{i=1}^{n} \sum_{j=1}^{n} N_{ij} \tag{3.14}$$

Precision: the accuracy of certain specific class that has been predicted by classifier.

$$Precision_i = N_{ii} / \sum_{k=1}^{n} N_{ki} \tag{3.15}$$

Recall: shows the ability of the prediction model to choose instances of a certain class from the data set.

$$Recall_i = N_{ii} / \sum_{k=1}^{n} N_{ik} \tag{3.16}$$

F-score: a mean of recall and precision.

$$F - score_i = \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i} \tag{3.17}$$

In the open source machine learning Python library Scikit-Learn (Pedregosa et al., 2011), it has functions to analyze the confusion matrix, generate the classification report which includes the precision, recall and F-score, and calculate the accuracy score of the classifier.

Universiti Malaya

3.10 Event Study Methods

In financial event study, the basic goal is to detect the abnormal returns caused by certain event within a certain event period which is called event window (Dutta, 2014). By defining the event window within an appropriate period which includes the event day, the analysis of the stock market data can target on the impact of OPEC news announcement. A common issue in all event studies, there is no resolution for choosing the most appropriate event window, that is, deciding the pre- and post-event window length (Spencer & Bredin, 2019). Event window decides how many days share market prices data should be analyzed based on event date. Proper event window not only can include the impact of analyzed events but also avoid influence of irrelevant events. Therefore, a suitable event window ensures a reliable result of the analysis. When it comes to study certain event's impact on share market prices, the event window varies because the characteristics of those events are vary (Loutia et al., 2016).

Since the choice of event window may differ among different studies, and there are no formal rules to choose the event window, this research adopts an event window length of five trading days before and after the event day. This is based on a related research on titled "Do OPEC announcements influence oil prices?" (Loutia et al., 2016). The same event window was also applied in related researches conducted by Horan et al (2004) and Bina and Vo (2007). Hence, in this research, the event window includes five trading days before and after the event date. Thus, the length of event window is 11 days including the event day.

Furthermore, estimate period decides how many days share market prices data should be used to calculate the expected return based on the historical stock prices. Estimate period should exclude the event window so that the expected return is not affected by the event. Like event window, there is no guidelines for determining the best estimation

Universiti Malaya

period. Since the stock prices of energy sector (oil & gas) companies may fluctuate heavily, the estimate period cannot be too long, otherwise it will not reflect the influence of OPEC news announcements. On the other hand, if the estimation period is too short, there would not be enough observations for an adequate estimation of the relationship between OPEC news sentiment and fluctuation of stock prices from analyzed companies (Croese & Westerman, 2015).

Since estimation period of 30 days is a common choice in the literature, this research also uses 30 trading days before the starting day of the event window as the estimation period (Philipp & Andre, 2016). The expected return is the mean daily return over the estimation period (Guidi, Russell, & Tarbert, 2006). Figure 3.4 below shows how the event study methods are used in this research.

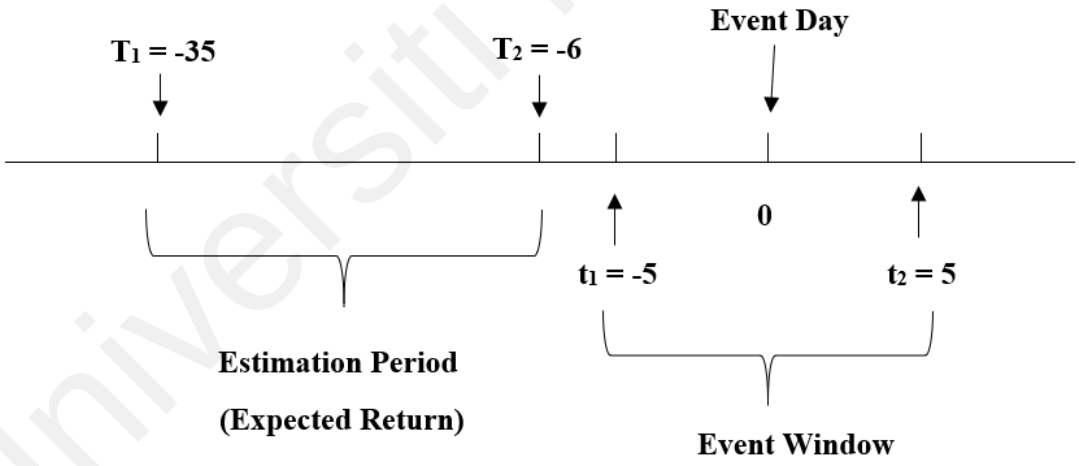

Figure 3.4: Event Study Methods Used in this Research

3.11 Analysis of Historical Stock prices

As this research analyzes the historical stock prices data of selected companies, the basic terms, mathematical formula and methods used are chosen from financial research domain. The financial terms and formulas used in this research include: Daily Return, Expected Return, Abnormal Return and Cumulative Abnormal Returns (CAR).

Universiti Malaya

Daily return: shows the magnitude of stock prices changes on daily basis. The increase of the stock prices results in a positive daily return. A negative daily return means the price of stock goes down (Bryan, 2018).

The formula of calculating Daily Return (R) (Bryan, 2018):

$$R = (R_2-R_1)/R_1 \quad (3.18)$$

R_2: Today's closing price

R_1: Previous day's closing price

Expected Return (ER): an important index for solving various investment problems (Bali & Zhou, 2016). The expected return can be calculated by the constant mean returns over the estimation period, as shown below (Guidi, Russell & Tarbert, 2006; Lin & Tamvakis, 2010; Philipp & Andre, 2016).

$$ER = \bar{R} \quad (3.19)$$

$\bar{R}$: Mean daily return over the estimation period

Abnormal Return (AR): gives the difference between actual daily return and expected return. It is an important index for analyzing the fluctuation of the stock prices. The evaluation of the OPEC news impact on the six selected companies stock prices can be measured by abnormal return (Loutia, Mellios & Andriosopoulos, 2016). Based on the study conducted by Kothari & Warner (1980) and the book entitled 'The econometrics of financial markets' (Campbell, Champbell, Wen-Chuan Lo, & MacKinlay, 1997), the formula of calculating Abnormal Return (AR) is as follows:

$$AR = R - ER \quad (3.20)$$

Universiti Malaya

R: Daily Return

ER: Expected Return

Cumulative Abnormal Return (CAR): A summary of daily abnormal return. In a study which needs to analyze the abnormal return that happens in a specific period, CAR can be used to represent the overall fluctuation of the stock prices of all the selected companies. If the CAR from the event window is statistically insignificant, it means that the OPEC news has nearly no impact on the stock prices of selected companies (Lin & Tamvakis, 2010; Spencer & Bredin, 2019). The formula of calculating CAR is:

$$\text{CAR}=\sum_t^T AR \tag{3.21}$$

AR: Abnormal Return

t: Starting day of analyze

T: Ending day of analyze

3.12 Statistical Analysis in IBM SPSS

SPSS (Statistical Package for the Social Sciences) which was released in 1968, and it was acquired by the International Business Machines Corporation (IBM) in 2009. SPSS changed its name to IBM SPSS since the 2015 version (Mocormick & Salcedo, 2017). SPSS is a versatile and responsive software package designed to manipulate, statistical analyze and present the data (Pallant, 2013). Until now, SPSS is still a widely used program for statistical analysis in social science. It is commonly used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners, and others (StatisticsSolutions, 2019).

Universiti Malaya

Statistics and extensions included in the base software (Leech, Barrett, & Morgan, 2014):

- Descriptive statistics: Cross Tabulation, Frequencies, Descriptive Ratio Statistics
- Bivariate statistics: Means, t-test, ANOVA, Correlation (bivariate, partial, distances), Nonparametric tests, Bayesian Analysis
- Prediction for Numerical Outcomes: Linear Regression Analysis
- Prediction for Identifying Groups: Factor Analysis, Cluster Analysis (two-step, K-means, hierarchical), Discriminant Analysis
- Geo Spatial Analysis, Simulation
- R Extension (GUI), Python Language Supported

SPSS is a preferred data analysis software and of great importance among students and professional researchers due to its capability of analyzing a wide scope as well as a large amount of data (Leech et al., 2014). Hence, this research uses IBM SPSS Statistics 22 software to perform the statistical analysis.

Universiti Malaya

CHAPTER 4: DATA COLLECTION AND ANALYSIS

Analyzing proper datasets using appropriate techniques and methods helps to achieve the research objectives of this research. As the machine learning algorithms are sensitive to training data's features, a slight difference in the datasets or the preprocessing methods may affect the results significantly (Khalid et al., 2014). Hence, the data sets and the analysis methods used in this research need to be considered carefully.

This chapter contains detailed explanation on data collection and data analysis to find out the relationship between OPEC news sentiment and the fluctuation of stock prices of public listed energy sector (oil & gas) companies in Bursa Malaysia. The results of data analysis are also presented clearly.

4.1 Data Collection

As mentioned in Chapter 3, this research uses two types of datasets: textual data from Wall Street Journal and OPEC, and historical stock prices data of six public listed energy sector (oil & gas) companies in Bursa Malaysia. The quality of data sets applied in the research defines how much reliable of research results can be. Hence, in order to ensure the authenticity and reliability of the research results, the source of data sets should be valid and reliable (Zhu & Cai, 2015).

4.1.1 Collection of Historical Stock prices Data

The historical stock prices data used in this research is collected from Yahoo Finance website which is a proven reliable financial research resource. This dataset includes six years historical stock prices data between 2012 to 2017 of the six selected energy sector (oil & gas) companies listed on the Main Market Board of Bursa Malaysia. These six companies are shown in the table 4.1 as follow.

Universiti Malaya

Table 4.1: The List of Stock prices Companies Dataset (Energy Sector)

No.	Name of Company	Stock Code	Year Established
1	Petron Malaysia Refining & Marketing Berhad	3042	1893
2	HengYuan Refining Company Berhad	4324	1960
3	Bumi Armada Berhad	5210	1995
4	Hibiscus Petroleum Berhad	5199	2007
5	Sumatec Resources Berhad	1201	1979
6	Sapura Energy Berhad	5218	2012

4.1.2 Collection of Textual Data

There are two textual data sets used in this research. The first textual dataset are collected from the Organization of the Petroleum Exporting Countries (OPEC) official press releases purposively for this research. This OPEC news textual dataset contains totally 116 news articles which were published on the OPEC's official website 'Press Release' column (https://www.opec.org/opec_web/en/press_room/28.htm), from the year 2012 to 2017. Table 4.2 shows the details of OPEC news dataset.

Table 4.2: OPEC News Released from 2012 to 2017

Year	Number of News	Year	Number of News
2012	9	2015	11
2013	9	2016	35
2014	9	2017	43
Total number of News:			116

The other textual dataset used in this research is a Wall Street Journal news articles dataset (Chen, 2017). This dataset captured the news articles released on the Wall Street

Universiti Malaya

Journal website's 'Markets' column (https://www.wsj.com/news/markets) published from 7 June until 26 July 2017 only. This dataset contains 2062 financial news articles. This dataset was processed and used to provide sufficient financial news sentiment words for training the machine learning classifiers.

4.2 Textual Data Analysis

Textual data are analyzed to build a machine learning classifier to classify the OPEC news announcements based on its sentiment. Besides, by analyzing the statistical data, the fluctuation of stock prices of selected companies happened during the OPEC news release data can be summarized. Thus, the relationship between OPEC news sentiment and the fluctuation of stock prices of public listed energy sector (oil & gas) companies in Bursa Malaysia can be studied by comparing the results from textual data analysis and statistical data analysis.

Universiti Malaya

4.2.1 Preparing Training Data

To build the machine learning classifiers, firstly, 2062 financial news articles from the Wall Street Journal dataset are broken into 37452 sentences. Then each sentence is tokenized as shown in Figure 4.1.

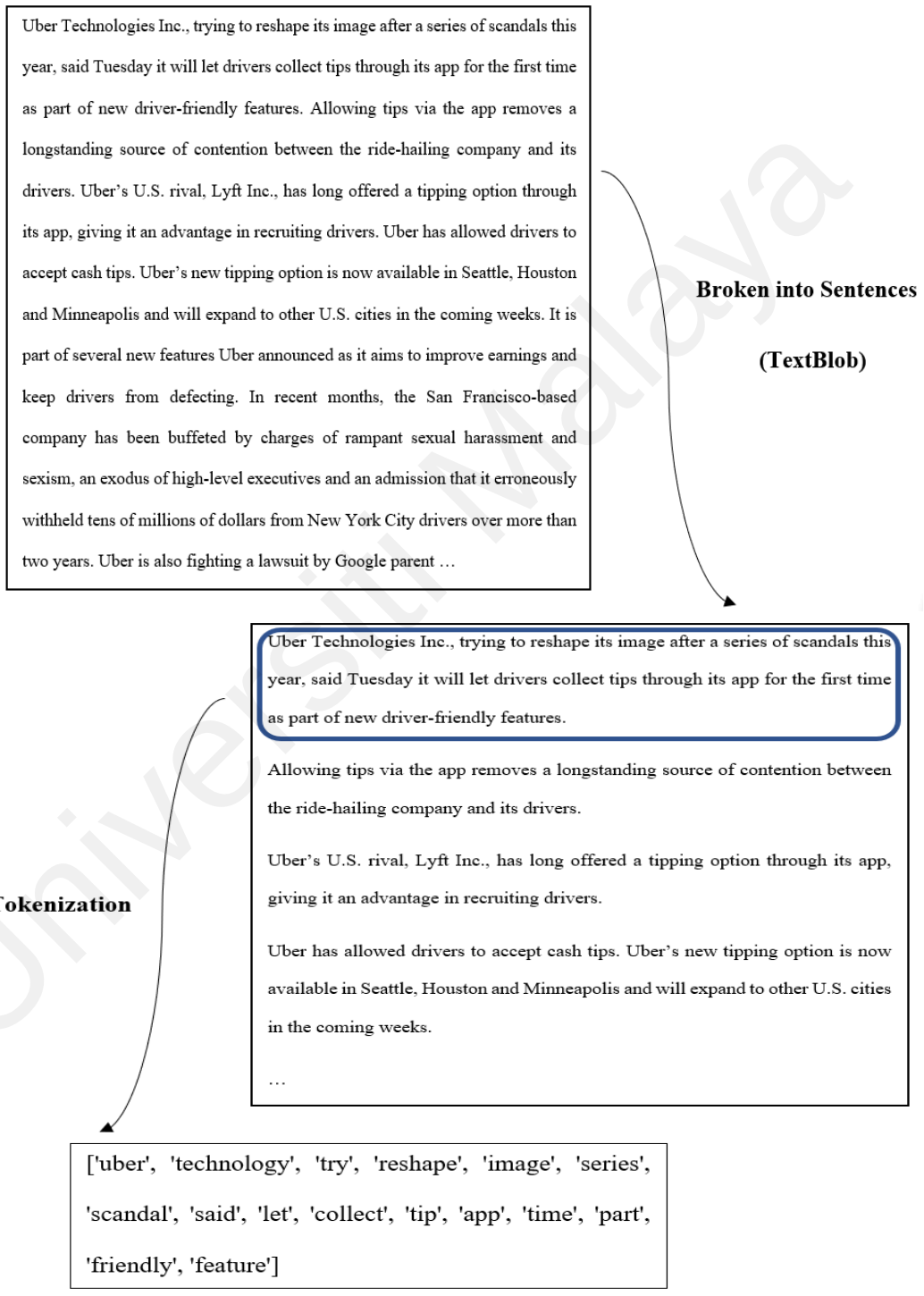

Figure 4.1: Preprocessing of WSJ dataset

Universiti Malaya

After being tokenized, the given sentence is divided into individual words which is also called token. In this way, the unnecessary characters such as punctuations of the sentence can be removed (Singh, 2019).

Each token of the sentence is compared to the words in Loughran and McDonald Financial Sentiment Dictionaries (Loughran & Mcdonald, 2011), then the whole sentence's sentiment is calculated by the number of positive and negative words. This is called relative proportional difference evaluating method (Will et al., 2011) as explained in Chapter 3. Formula 3.1 is used to calculate the sentiment of sentence. Figure 4.2 shows a continued analysis of example presented in Figure 4.1

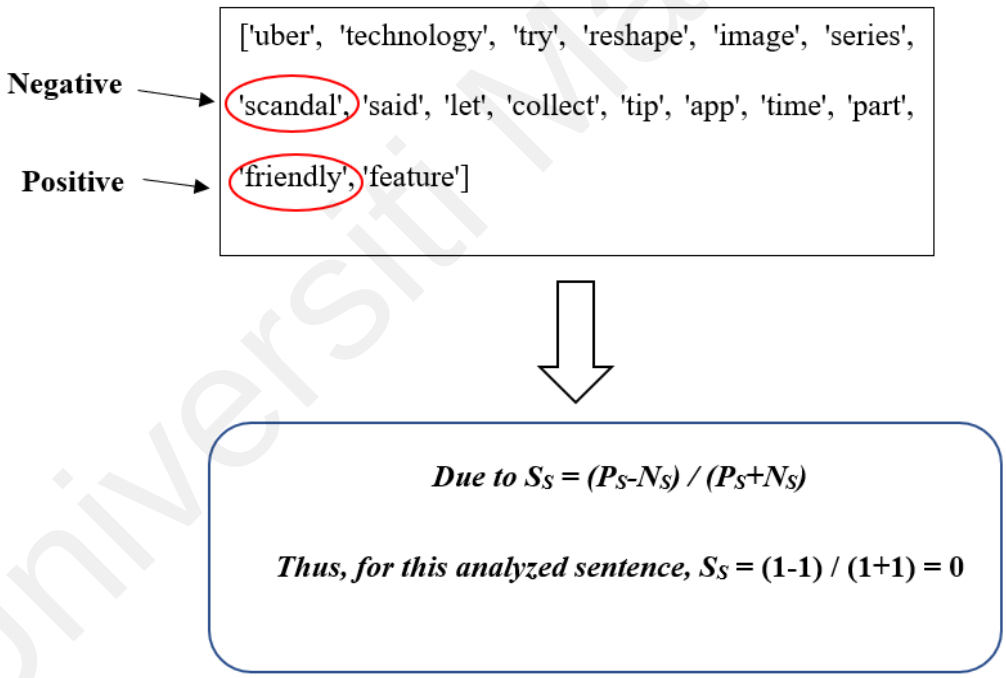

Figure 4.2: Example of Sentence's Sentiment Calculation

The sentiment of a sentence is classified either as neutral as shown in Figure 4.2 above, negative, or positive.

Based on the result of the above lexicon-based sentiment analysis, those sentences are further labelled with '-1', '1' or '0' to represent its sentiment is 'negative', 'positive' or 'neutral'. To balance the training data, from the 37,452 labelled sentences, this research

Universiti Malaya

uses same number of sentences in each category of sentiment. It is important to note that imbalanced data can cause different prediction confidence of the different classes in the target domain (Krawczyk, 2016). Hence, altogether 12,000 sentences which consists 4000 sentences for each kind of sentiment (positive, neutral and negative) are used to train the machine learning classifiers.

After every sentence from Wall Street Journal financial news articles are labelled with its sentiment, bag-of-words representation, stop words removing and TF-IDF feature processing method are applied to weigh the sentiment words' value of the whole data set and transfer the textual data into numerical data based on its weight. After applying feature processing procedures mentioned above, the Wall street Journal financial news textual data is used as proper training data set for the machine learning algorithms. The following section explains the program codes for feature processing steps and illustrates with an example of output.

Part 1: Bag-of-Words Representation, this creates a list of words with its frequency in the analyzed sentences. Figure 4.3 shows the programming codes for the Bag-of-Words Representation.

```
def create_freq_dict(sents):
        i = 0
        freqDict_list = []
        for sent in sents:
                i += 1
                freq_dict = {}
                words = word_tokenize(sent)
                for word in words:
                        word = word.lower()
                        if word in freq_dict:
                                freq_dict[word] += 1
                        else:
                                freq_dict[word] = 1
                        temp = { 'doc_id' : i, 'freq_dict' : freq_dict}
                freqDict_list.append(temp)
        return freqDict_list
```

Figure 4.3: Programming Codes for the Bag-of-Words Representation

Universiti Malaya

Part 2: Stop Words Removal, this removes the informational words of the sentences. The detail of stop words list applied in this research can be found in Chapter 3. Figure 4.4 shows the programming codes for stop words removal.

```
def remove_stop_words(sent):
        stop_words = set(stopwords.words('english'))
        word_tokens = word_tokenize(sent)
        filtered_sentence = [w for w in word_tokens if not w in stop_words]
        filtered_sentence = []
        for w in word_tokens:
                if w not in stop_words:
                        filtered_sentence.append(w)
                my_str = str(filtered_sentence)
        punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~"'
        no_punct = ""
        for char in my_str:
                if char not in punctuations:
                        no_punct = no_punct + char
        return no_punct
```

Figure 4.4 : Programming Codes for Stop Words Removal

Part 3: TF-IDF Score Calculation, TF-IDF score is a statistical measure which shows how important a certain word is to a document in a collection of documents. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Figure 4.4 shows the programming codes for TF-IDF score calculation.

```
def computeTF(doc_info, freqDict_list):
        TF_scores = []
        for tempDict in freqDict_list:
                id = tempDict['doc_id']
                for k in tempDict['freq_dict']:
                        temp = {'doc_id' : id,
                                'TF_score':tempDict['freq_dict'][k]/doc_info[id-1]['doc_length'],
                                'key': k}
                        TF_scores.append(temp)
        return TF_scores

def computeIDF(doc_info, freqDict_list):
        IDF_scores = []
        counter = 0
        for dict in freqDict_list:
                counter += 1
                for k in dict['freq_dict'].keys():
                        count = sum([k in tempDict['freq_dict'] for tempDict in freqDict_list])
                        temp = {'doc_id' : counter, 'IDF_score' : math.log(len(doc_info)/count), 'key' : k}
                        IDF_scores.append(temp)
        return IDF_scores

def computeTFIDF(TF_scores, IDF_scores):
        TFIDF_scores = []
        for j in IDF_scores:
                for i in TF_scores:
                        if j['key'] == i['key'] and j['doc_id'] == i['doc_id']:
                                temp = {'doc_id' : j['doc_id'],
                                        'TFIDF_score' : j['IDF_score'] * i['TF_score'],
                                        'key' : i['key']}
                TFIDF_scores.append(temp)
        return TFIDF_scores
```

Figure 4.5: Programming Codes for TF-IDF Score Calculation

Universiti Malaya

Figure 4.6 is shows partial results of TF-IDF score calculation for the analyzed sentences.

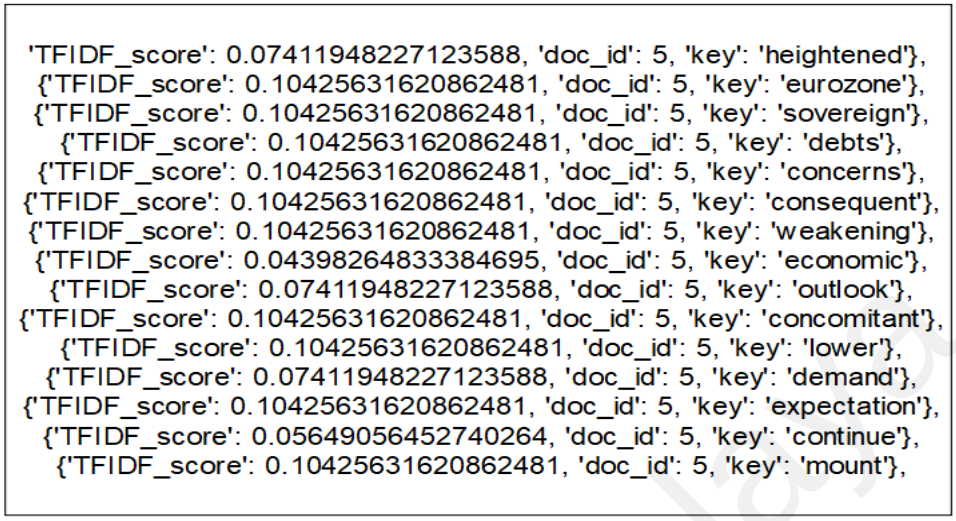

Figure 4.6: Example of TF-IDF Score Calculation

4.2.2 Testing Machine Learning Algorithms

To determine which machine learning algorithm-based classifier performs the best in the classification task, multiple supervised machine learning algorithms-based classifiers are tested separately. Table 4.3 shows the results of performance measure which named classification reports of each tested machine learning algorithms-based classifiers.

Table 4.3: Classification Reports of Tested Machine Learning Algorithms

Classifier	Sentiment (S)	Precision	Recall	F1-Score	Support	Accuracy Score (Rank)
Stochastic Gradient Descent Classifier	-1	0.71	0.72	0.71	803	0.70 (1)
	0	0.61	0.63	0.62	792	
	1	0.79	0.75	0.77	806	
Gaussian Naïve Bayes Classifier	-1	0.55	0.50	0.52	815	0.51 (6)
	0	0.43	0.52	0.47	780	
	1	0.56	0.51	0.53	806	

Universiti Malaya

Table 4.3, continued

Multinomial Naïve Bayes Classifier	-1	0.68	0.75	0.72	815	0.67 (2)
	0	0.65	0.42	0.51	780	
	1	0.67	0.83	0.74	806	
Complement Naïve Bayes Classifier	-1	0.61	0.71	0.66	803	0.61 (4)
	0	0.58	0.34	0.43	792	
	1	0.62	0.78	0.69	806	
Bernoulli Naïve Bayes Classifier	-1	0.74	0.25	0.37	803	0.46 (7)
	0	0.38	0.87	0.53	792	
	1	0.67	0.27	0.39	806	
Support Vector Classifier	-1	0.73	0.61	0.67	803	0.66 (3)
	0	0.53	0.71	0.61	792	
	1	0.80	0.68	0.74	806	
Random Forest Classifier	-1	0.71	0.38	0.50	803	0.56 (5)
	0	0.43	0.79	0.56	792	
	1	0.78	0.52	0.62	806	

4.2.3 Classifying OPEC news

As Stochastic Gradient Descent Classifier (SGDC) outperforms other tested supervised machine learning algorithms, it is used to classify the OPEC news sentiments. As mentioned in Chapter 3, all OPEC news analyzed in this research were collected from OPEC official website. SGDC analyzes each sentence in an OPEC news article, and then tags those sentences with its sentiment, -1, 0, or 1 representing negative, neutral or positive sentence, respectively. As mentioned in Chapter 3, the sentiment of an article is calculated by using relative proportional difference evaluating method (formula 3.2). The sentiment value of each analyzed article is ranging from -1 to 1.

Universiti Malaya

Following table 4.4 shows the sentiment value of OPEC news between the period 14-06-2012 to 21-12-2017.

Table 4.4: Sentiment of OPEC News from 2012 to 2017

Event Day	Senti-ment	Event Day	Senti-ment	Event Day	Senti-ment	Event Day	Senti-ment	Event Day	Senti-ment	Event Day	Senti-ment
2012-06-14	-0.5	2014-10-02	-0.75	2016-09-09	-0.14	2016-12-15	-0.25	2017-04-24	-0.2	2017-09-05	-0.25
2012-06-28	0.33	2014-11-27	-0.67	2016-09-26	1	2016-12-16	-0.25	2017-04-27	0.09	2017-09-06	0
2012-07-16	-0.2	2015-02-05	-0.33	2016-09-28	-0.33	2017-01-08	-0.5	2017-04-28	-0.2	2017-09-14	0.2
2012-09-25	0	2015-06-04	0.33	2016-10-14	-0.57	2017-01-11	-0.56	2017-05-22	0	2017-09-15	-0.33
2012-09-27	-0.42	2015-06-05	-0.08	2016-10-18	-0.43	2017-01-14	-0.41	2017-05-24	-0.5	2017-09-22	0.56
2012-10-04	0.43	2015-06-24	-0.4	2016-10-19	-0.33	2017-01-17	-0.23	2017-05-25	-0.42	2017-09-28	-1
2012-12-12	0.17	2015-07-30	-0.6	2016-10-24	0.33	2017-01-22	-0.22	2017-05-31	0.14	2017-09-29	-0.6
2013-03-21	0.2	2015-09-01	-0.33	2016-10-26	-0.65	2017-02-07	-0.14	2017-06-01	-0.75	2017-10-10	0.43
2013-05-31	-0.25	2015-09-08	-0.6	2016-10-29	0.6	2017-02-08	0	2017-06-09	0.5	2017-10-21	0.43
2013-07-29	-0.2	2015-12-04	-0.67	2016-11-02	-1	2017-02-14	-0.4	2017-06-13	-0.78	2017-10-24	0.2
2013-10-24	-0.33	2015-12-18	0.33	2016-11-05	0	2017-02-15	0.17	2017-06-22	0.2	2017-11-07	0.25
2013-11-08	0.33	2016-03-21	0	2016-11-07	-0.45	2017-02-21	0	2017-06-26	-0.2	2017-11-30	-0.27
2013-11-11	0.11	2016-06-02	-0.23	2016-11-08	0.4	2017-02-24	-0.67	2017-07-18	-0.2	2017-12-01	-0.05
2013-12-04	-0.17	2016-06-22	-0.4	2016-11-17	0.2	2017-03-06	0	2017-07-23	-1	2017-12-13	0.43
2014-03-31	0.5	2016-06-30	-0.6	2016-11-19	-0.14	2017-03-10	0.22	2017-07-24	0.11	2017-12-20	0.4
2014-04-30	0.33	2016-08-01	-0.5	2016-11-21	-0.43	2017-03-11	-0.4	2017-07-29	-0.52	2017-12-21	-0.33
2014-06-11	-0.56	2016-08-04	0.25	2016-11-30	-0.29	2017-03-16	0.14	2017-08-02	-0.56		
2014-06-24	-0.1	2016-08-08	0.33	2016-12-10	-0.33	2017-03-26	-0.3	2017-08-08	-0.6		
2014-07-18	-0.27	2016-09-05	-0.8	2016-12-13	-0.09	2017-04-04	0.54	2017-08-14	-0.72		
2014-09-16	0.6	2016-09-06	-0.75	2016-12-14	-0.27	2017-04-11	0.6	2017-08-24	-0.14		

4.3 Analysis of Energy Sector (Oil & Gas) Historical Stock Prices

Statistical data analysis in this research aims to analyze the fluctuation of stock prices of selected public listed energy sector (oil & gas) companies. As it has explained in Chapter 3, the Cumulative Abnormal Return (CAR) is used as the index to show the fluctuation of stock prices of the six selected companies during the event window time, based on each event date (news release date). It indicates whether the OPEC news

University of Malaya

announcements have any impact on the stock prices of six selected companies and how the influence happened using formulas mentioned in Chapter 3.

Based on the explanation provided in chapter 3, this research adopts the 30 trading days before the starting day of event window as estimate period, and the event window is set as the period of 5 trading days before and after event day.

The cumulative abnormal return of selected companies are calculated based on each event day's event window as well as each event's average cumulative abnormal return of all those six selected companies are shown in the Table 4.5 below.

Table 4.5: CAR and Average CAR of Six Companies on Each Event Day

Event Day	1201 CAR	3042 CAR	4324 CAR	5199 CAR	5210 CAR	5218 CAR	Average CAR
2012-06-14	0.017878	0.000696	0.004354	0.007515	0.004741	-0.00479	0.275%
2012-06-28	-0.15662	-0.00093	0.002489	0.006285	-0.00346	-0.00522	-0.149%
2012-07-16	-0.0047	-0.00015	-0.00156	0.00969	0.003104	0.006867	0.103%
2012-09-25	-0.16933	0.005061	0.003416	0.003455	0.00213	0.001936	-0.101%
2012-09-27	-0.0772	0.005612	0.001347	0.002548	0.001517	0.003484	0.160%
2012-10-04	0.013434	0.003167	0.001649	0.005682	0.000719	0.005488	-0.163%
2012-12-12	0.059761	-0.00347	-0.00428	0.00145	0.00067	0.007555	-0.089%
2013-03-21	-0.10686	-0.0005	-0.00367	0.005092	-0.00088	-0.00317	-0.107%
2013-05-31	0.066684	-0.00896	-0.00043	-0.00159	-0.0062	-0.00433	0.149%
2013-07-29	0.024748	-0.00068	-0.0001	-0.00224	0.001968	-0.00444	0.109%
2013-10-24	-0.11317	-0.00084	-0.00355	-0.01668	0.000178	0.000377	0.117%
2013-11-08	-0.04396	0.000248	0.001392	-0.0077	-0.00515	0.001574	-0.158%
2013-11-11	-0.07808	0.000459	0.003281	-0.0077	-0.0004	0.003566	-0.069%
2013-12-04	0.003706	0.000342	0.002247	0.016507	-0.00286	0.003676	0.069%
2014-03-31	-0.00364	0.001158	-0.00141	-0.00141	0.004962	0.010471	-0.214%
2014-04-30	0.00313	0.001505	0.000733	0.007446	-0.00076	-0.00177	-0.211%
2014-06-11	0.002771	-0.00972	1.61E-05	0.006066	0.00453	0.00693	0.348%
2014-06-24	0.017716	-0.00731	-0.00105	0.006126	-0.00037	0.004041	0.089%

Key: CAR – Cumulative Abnormal Return

Universiti Malaya

Table 4.5, continued

Event Day	1201 CAR	3042 CAR	4324 CAR	5199 CAR	5210 CAR	5218 CAR	Average CAR
2014-07-18	0.016524	0.00342	-0.00322	0.00187	0.007025	-0.00414	0.144%
2014-09-16	-0.01125	-0.00158	-0.00496	-0.00301	0.015471	-0.02804	-0.344%
2014-10-02	-0.01236	0.001063	0.002992	0.002461	-0.01277	-0.01005	0.369%
2014-11-27	-0.01271	-0.00796	0.001157	0.005933	-0.016	-0.01101	0.392%
2015-02-05	-0.00527	-0.00035	0.00486	0.000849	-0.0064	-0.00394	0.226%
2015-06-04	-0.00227	-0.00275	0.001735	0.009274	-0.00501	-0.00266	-0.255%
2015-06-05	-0.00227	-0.00122	0.000247	0.008258	-0.00603	-0.0037	0.031%
2015-06-24	0.004333	-0.00252	-0.00065	0.003934	0.008058	0.005597	0.279%
2015-07-30	-0.0085	0.007309	0.001643	0.00608	-0.00306	0.000992	0.245%
2015-09-01	0.020796	-0.00136	0.00115	-0.00123	0.003898	0.018048	0.178%
2015-09-08	0.022415	0.001679	0.004682	0.001951	0.012716	0.01774	0.301%
2015-12-04	-0.00904	0.022478	0.005041	0.026234	-0.01533	-0.01389	0.327%
2015-12-18	0.002058	0.005385	0.010402	0.022554	0.000154	0.003946	-0.199%
2016-03-21	-0.00062	0.007902	0.015104	-0.02132	0.00401	-0.01703	-0.181%
2016-06-02	-0.00054	0.005799	0.001301	0.011421	0.006907	0.010145	0.158%
2016-06-22	-0.00104	0.005824	-0.00056	-0.00288	0.004659	0.001533	0.134%
2016-06-30	0.004924	0.004537	0.000639	-0.0004	0.008956	0.004958	0.361%
2016-08-01	0.00582	0.002985	0.000982	-0.00585	-0.0015	0.006969	0.221%
2016-08-04	-0.01507	0.010567	0.001196	-0.00775	-0.0029	0.001086	-0.132%
2016-08-08	0.001007	0.01515	0.00123	-0.0023	0.001858	0.008449	-0.282%
2016-09-05	-0.00519	-0.00565	-0.00025	-0.00846	0.012375	-0.00517	0.300%
2016-09-06	-0.00857	-0.00589	-0.00129	-0.00782	0.009902	-0.00692	0.323%
2016-09-09	-0.00872	-0.00528	-0.00044	0.005505	-0.00945	-0.00311	0.066%
2016-09-26	-0.00305	-0.0028	0.001249	0.001474	-0.00838	0.000872	-0.493%
2016-09-28	-0.00091	-0.00186	0.000402	-0.0018	-0.00703	0.004162	0.190%
2016-10-14	0.014638	0.00684	-0.00011	0.028373	-0.00078	-0.00131	0.217%
2016-10-18	-6.2E-05	0.001486	0.000615	0.046613	-0.0027	-0.00167	0.195%
2016-10-19	-6.2E-05	-0.00101	0.000291	0.020263	-0.00538	-0.00583	0.172%
2016-10-24	-0.01521	-0.00168	0.000214	0.012695	-0.01231	-0.00087	-0.111%
2016-10-26	-0.03191	-0.00477	-0.0003	0.015108	0.003575	-0.00632	0.326%

Key: CAR – Cumulative Abnormal Return

Universiti Malaya

Table 4.5, continued

Event Day	1201 CAR	3042 CAR	4324 CAR	5199 CAR	5210 CAR	5218 CAR	Average CAR
2016-10-29	0.007629	0.035339	0.020334	0.049399	0.010225	-0.00817	-0.253%
2016-11-02	-0.02612	-0.00276	-0.0016	-0.02181	-0.0015	-0.00832	0.428%
2016-11-05	-0.00254	-0.00576	-0.00029	-0.02168	-0.00552	-0.01391	-0.254%
2016-11-07	-0.00254	-0.00576	-0.00029	-0.02168	-0.00552	-0.01391	0.266%
2016-11-08	0.005291	-0.00259	0.001017	-0.02214	-0.00705	-0.01348	-0.239%
2016-11-17	0.022626	1.43E-06	0.001172	-0.0217	-0.01565	0.000229	-0.131%
2016-11-19	0.014109	-0.00049	-0.00207	-0.0119	-0.01897	0.001125	0.067%
2016-11-21	0.014109	-0.00049	-0.00207	-0.0119	-0.01897	0.001125	0.290%
2016-11-30	-0.00999	-0.0063	-0.02977	0.00185	-0.01243	0.00596	0.166%
2016-12-10	0.021935	-0.00099	-0.01654	0.015432	0.016308	0.008055	0.176%
2016-12-13	0.021935	-0.00099	-0.01654	0.015432	0.016308	0.008055	0.042%
2016-12-14	0.025654	-0.00079	-0.01186	0.018369	0.01556	0.010988	0.176%
2016-12-15	0.023694	-0.00169	0.000775	0.016762	0.014956	0.008688	0.134%
2016-12-16	0.019534	-0.00106	0.002139	0.019336	0.013466	0.010079	0.197%
2017-01-08	-0.01291	0.000371	0.030583	-0.00447	0.002453	-0.00177	0.259%
2017-01-11	-0.03127	0.002807	0.033946	-0.00399	-0.00559	-0.00606	0.279%
2017-01-14	-0.03163	0.002064	0.019352	-0.00778	-0.00534	-0.00993	0.257%
2017-01-17	-0.03853	0.006234	0.009164	-0.01747	-0.0071	-0.00845	0.135%
2017-01-22	-0.02024	0.007943	0.01435	-0.02296	-0.00806	-0.00389	0.162%
2017-02-07	-0.00045	0.011701	0.007946	-0.00126	0.008344	0.005039	0.176%
2017-02-08	-0.00579	0.010741	0.004046	0.000619	0.005104	0.003844	-0.780%
2017-02-14	-0.01869	-0.00307	-0.00663	-0.01542	0.019848	0.004551	0.290%
2017-02-15	-0.02878	-0.0031	-0.00672	-0.01644	0.021553	8.05E-05	-0.123%
2017-02-21	-0.03713	0.047491	0.049192	-0.03285	0.003037	-0.00851	-0.003%
2017-02-24	-0.00934	0.014748	0.006372	-0.01699	0.003411	-0.00078	0.033%
2017-03-06	-0.01034	-0.01945	0.004042	-0.02087	-0.01199	-0.00552	-0.225%
2017-03-10	-0.0087	-0.01964	-0.01919	-0.01797	-0.0038	-0.00695	-0.235%
2017-03-11	-0.00632	-0.01979	-0.01965	-0.01503	-0.00415	-0.00946	0.176%
2017-03-16	0.008244	-0.01801	-0.01645	-0.01402	-0.00302	-0.01359	-0.519%
2017-03-26	0.002164	-0.00377	-0.0106	0.013021	-0.00498	-0.00495	0.284%

Key: CAR – Cumulative Abnormal Return

Universiti Malaya

Table 4.5, continued

Event Day	1201 CAR	3042 CAR	4324 CAR	5199 CAR	5210 CAR	5218 CAR	Average CAR
2017-04-04	0.004483	0.001294	0.00749	0.027446	0.005431	0.01073	-0.480%
2017-04-11	0.006669	-0.00964	0.000111	0.002844	0.005137	0.005709	-0.330%
2017-04-24	-0.00644	0.008901	0.000428	0.001141	-0.00105	0.000592	0.173%
2017-04-27	-0.00672	0.00459	-0.00492	-0.00164	0.000941	-0.00475	-0.235%
2017-04-28	0.000274	0.01117	-0.00531	-0.00015	0.000959	-0.00296	0.089%
2017-05-22	0.006207	-0.01561	0.017224	-0.00347	-0.00356	-0.00279	-0.082%
2017-05-24	-0.01034	-0.01495	0.004926	0.00845	-0.00249	-0.00785	0.387%
2017-05-25	-0.01127	-0.01259	0.009052	0.0108	0.001068	-0.00451	0.330%
2017-05-31	0.016847	0.060591	0.092551	0.029248	0.022477	-0.00684	-0.270%
2017-06-01	-0.00053	-0.01352	0.001727	0.005875	-0.00182	-0.00294	0.391%
2017-06-09	-0.0027	-0.01319	-0.01974	-0.01279	-0.00144	0.00558	-0.205%
2017-06-13	0.007653	-0.01418	-0.0189	-0.01123	-0.00264	-0.00776	0.328%
2017-06-22	0.015135	-0.01722	-0.02382	-0.00766	-5.5E-05	-0.01444	-0.118%
2017-06-26	-0.00358	-0.00966	-0.01558	0.002105	0.001553	-0.00529	0.111%
2017-07-18	-0.00506	0.009327	0.010092	0.002722	-0.00238	0.000774	0.163%
2017-07-23	-0.00515	0.010973	0.029847	0.004789	-0.00388	-0.00367	0.430%
2017-07-24	-0.00515	0.010973	0.029847	0.004789	-0.00388	-0.00367	-0.324%
2017-07-29	-0.00798	0.013618	0.0353	0.003635	-0.00201	-0.0016	0.273%
2017-08-02	0.01241	0.005866	0.024854	0.014403	0.002564	0.002842	0.214%
2017-08-08	0.003269	-0.00537	-0.01602	0.009174	0.00302	0.000988	0.254%
2017-08-14	0.01123	-0.01479	-0.00833	0.00869	0.006239	8.32E-06	0.348%
2017-08-24	0.006129	-0.00723	-0.01084	-0.00662	0.003912	-0.00274	0.067%
2017-09-05	0.001145	0.016997	0.005642	0.00323	-0.00111	0.012313	0.165%
2017-09-06	0.001145	0.01251	0.003627	0.00491	-0.00046	0.01005	0.232%
2017-09-14	0.021204	0.004918	-0.00409	0.036973	0.000253	0.013017	-0.125%
2017-09-15	0.021204	0.003704	-0.00558	0.035126	-0.00144	0.004986	0.141%
2017-09-22	0.01159	0.00068	-0.00094	0.036371	-0.00328	-0.0018	-0.209%
2017-09-28	-0.01296	0.002822	0.00367	-0.01139	-0.00671	-0.01342	0.469%
2017-09-29	-0.02227	0.002891	0.000542	-0.00053	-0.00777	-0.01209	0.276%
2017-10-10	-0.00952	0.003089	0.003031	0.002542	0.000608	0.001574	-0.258%

Key: CAR – Cumulative Abnormal Return

Universiti Malaya

Table 4.5, continued

Event Day	1201 CAR	3042 CAR	4324 CAR	5199 CAR	5210 CAR	5218 CAR	Average CAR
2017-10-21	-0.00082	-0.00718	-0.00421	-0.02284	0.001264	0.005173	-0.201%
2017-10-24	-0.00082	-0.00385	-0.00124	-0.02472	0.001891	0.003771	-0.099%
2017-11-07	0.017938	0.002222	0.019262	-0.01155	0.007733	-0.00673	-0.198%
2017-11-30	0.003215	0.00109	-0.00191	-0.00158	-0.00397	-0.0279	0.136%
2017-12-01	-0.00588	-0.00179	-0.00271	0.000573	-0.0082	-0.0388	0.026%
2017-12-13	-0.01298	0.012403	0.013253	0.00718	-0.0045	-0.04685	-0.371%
2017-12-20	-0.00056	0.004301	0.018724	0.002203	0.001917	0.011223	-0.251%
2017-12-21	0.015202	0.007214	0.021001	0.000526	0.0033	0.009911	0.723%

Key: CAR – Cumulative Abnormal Return

The average CAR of six companies on each event day (news release day) shows the fluctuation of six companies stock prices during the event window. Figure 4.7 shows the fluctuation of average CAR based on the event day starting from 2012 until 2017.

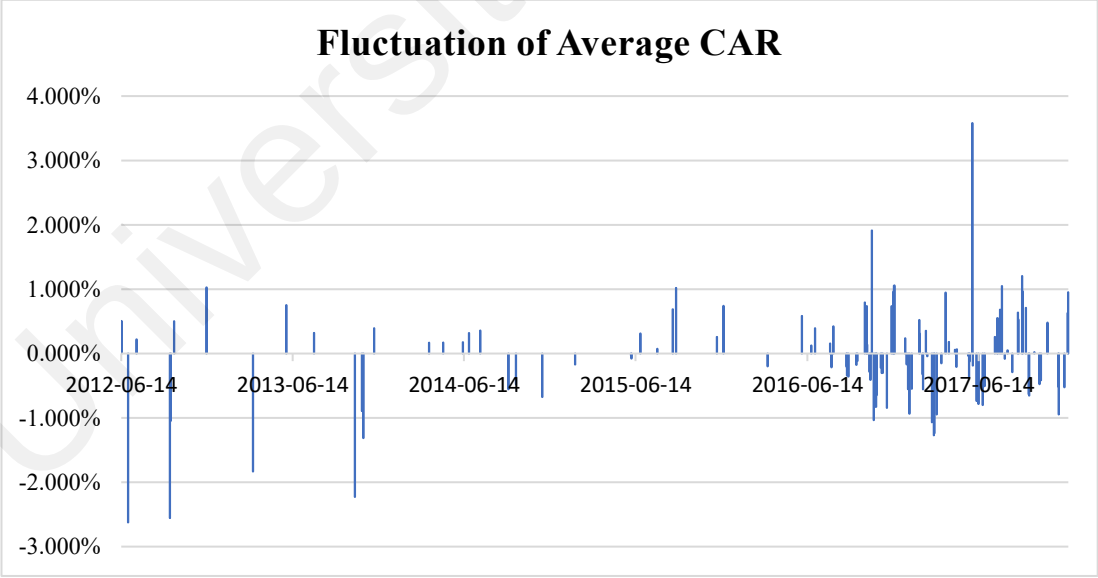

Figure 4.7 Fluctuation of the Average CAR

The relationship of average CAR and sentiments of OPEC news is further analyzed in the following section (Section 4.4).

Universiti Malaya

To study whether the stock prices of the six selected energy sector (oil & gas) companies were influenced by OPEC news sentiment on the event day (news release day), the abnormal return of these companies and its respective average abnormal return on each event day are analyzed. Table 4.6 shows the results of analysis.

Table 4.6: Abnormal Return of Six Companies on Event Day

Event Day	1201 EDAR	3042 EDAR	4324 EDAR	5199 EDAR	5210 EDAR	5218 EDAR	Average EDAR
2012-06-14	1.014115	-0.002058	0.0316229	0.0012462	-0.007664	-0.018635	0.197%
2012-06-28	-0.21078	-0.005426	0.0017668	0.0319448	-0.005772	-0.012212	-0.972%
2012-07-16	-0.1612	-0.005625	0.0105064	0.0248138	0.0005568	-0.00535	0.753%
2012-09-25	-0.20382	0.002744	0.0312452	-0.017429	-0.001305	-0.00773	-0.775%
2012-09-27	-0.18949	-0.00492	-0.009986	0.0209979	0.0012233	-0.007424	-0.174%
2012-10-04	1.017044	0.005091	0.025657	-0.015447	0.0142035	0.009876	0.777%
2012-12-12	-0.12049	-0.008937	-8.67E-05	0.0014296	-0.003164	0.034529	-0.065%
2013-03-21	-0.19008	0.007456	-0.016063	0.0013382	0.0073135	0.008564	0.646%
2013-05-31	-0.11176	-0.010244	-0.002264	-0.001031	-0.018517	0.086473	0.442%
2013-07-29	-0.64702	-0.018046	-0.005414	0.0007907	0.001837	0.000526	-1.088%
2013-10-24	-0.18625	-0.016155	-0.007501	-0.008892	-0.000298	-0.010341	0.089%
2013-11-08	0.084088	-4.92E-05	0.0018752	-0.021337	-0.010405	0.042241	-0.027%
2013-11-11	-0.02125	0.000162	0.0079629	0.0045972	-0.002588	-0.003016	-0.657%
2013-12-04	0.006453	0.007221	0.0069379	-0.004831	0.0042619	-0.005216	-0.559%
2014-03-31	-0.01634	0.000771	-0.002578	-0.007756	0.0193259	0.004478	-0.086%
2014-04-30	0.004405	0.002312	0.0011833	0.0011866	-0.000537	-0.004736	-0.188%
2014-06-11	0.000893	-0.002814	0.0003222	0.0095675	-0.008064	0.034435	0.540%
2014-06-24	0.128316	-0.010496	0.0003238	0.0094792	0.0008738	-0.001695	0.256%
2014-07-18	0.005783	0.00625	-0.0079	-0.005326	-0.002445	-0.009096	-0.403%
2014-09-16	-0.02946	0.000322	0.0046202	0.0078953	-0.02431	-0.001508	-0.908%
2014-10-02	0.001835	0.011599	0.0086135	0.0086008	-0.023228	-0.027916	-0.783%
2014-11-27	-0.04452	-0.010906	0.0014429	0.0102636	-0.04955	-0.036282	-1.630%
2015-02-05	-0.05028	0.026597	-0.013806	-0.008894	-0.038094	-0.023923	-1.656%
2015-06-04	0.002107	-0.006312	-0.000146	0.1072304	0.0026234	-0.017843	0.296%

Key: EDAR – Event Day Abnormal Return

Universiti Malaya

Table 4.6, continued

Event Day	1201 EDAR	3042 EDAR	4324 EDAR	5199 EDAR	5210 EDAR	5218 EDAR	Average EDAR
2015-06-05	-0.02289	-0.049947	-0.052355	0.0232729	-0.013094	-0.006869	0.364%
2015-06-24	0.080803	-0.002239	-0.002063	0.0075871	0.020046	0.008346	1.037%
2015-07-30	0.001332	-0.001324	-0.000169	-0.002221	0.0212118	0.032325	1.014%
2015-09-01	0.011368	-0.021571	0.0060867	-0.006051	-0.009733	0.05166	1.262%
2015-09-08	0.088181	0.000632	-0.002819	-0.004289	0.0069671	0.033175	0.583%
2015-12-04	0.004079	0.031915	-0.006562	0.1293083	-0.013852	-0.010905	-0.542%
2015-12-18	0.004479	-0.023732	0.0046812	-0.015821	-0.021847	-0.030406	-1.200%
2016-03-21	-0.00062	0.014382	0.0166574	-0.011079	0.0062615	-0.016989	-0.230%
2016-06-02	0.003233	-0.008135	0.0201293	-0.021226	0.0228108	0.022545	1.159%
2016-06-22	0.052641	0.011543	0.0032159	0.0272106	0.0386753	-0.004298	0.127%
2016-06-30	-0.04334	0.012005	0.000284	-0.000542	0.0247192	0.017059	0.717%
2016-08-01	0.004276	-0.006504	0.0104457	-0.001484	-0.000392	0.006226	0.726%
2016-08-04	-0.0025	0.001594	-0.006537	-0.003278	-0.002419	0.008189	-0.320%
2016-08-08	0.000528	0.041281	0.0002821	-0.002491	0.0042651	-0.000255	0.081%
2016-09-05	-0.00091	-0.008592	0.0030128	-0.004206	0.005442	0.016092	-0.154%
2016-09-06	0.000495	0.011919	0.003146	-0.001289	-0.004321	0.004484	0.385%
2016-09-09	0.000846	-0.010457	0.003121	-0.022007	-0.012374	0.003581	-0.320%
2016-09-26	0.060226	-0.004203	0.0009452	-0.026802	-0.006629	-0.022448	0.261%
2016-09-28	0.003254	-0.004584	-0.00316	-0.002038	-0.005247	-0.016043	-0.590%
2016-10-14	0.008288	0.010913	-0.00012	-0.019835	-0.002146	-0.013504	-0.309%
2016-10-18	0.067117	-0.001946	0.0035945	-0.020668	0.0038812	-0.005122	1.174%
2016-10-19	0.004617	0.02312	-7.81E-06	-0.027931	0.0176016	-0.008058	0.339%
2016-10-24	0.004617	-0.018578	0.0002115	0.1683041	-0.014303	-0.006419	-0.682%
2016-10-26	0.002828	-0.022724	-6.4E-06	0.0071364	0.0022298	-0.008608	-0.261%
2016-10-29	0.084284	-0.01835	0.0002107	0.0069153	0.0165663	-0.001996	-0.262%
2016-11-02	-0.08042	-0.010105	-0.003383	-0.048211	-0.004678	-0.006614	-0.138%
2016-11-05	0.011343	-0.003915	-0.003274	0.0348384	0.0089653	-0.00158	0.267%
2016-11-07	0.011343	-0.003915	-0.003274	0.0348384	0.0089653	-0.00158	-0.046%
2016-11-08	0.008565	-0.001362	0.0001036	-0.016584	0.0151236	0.011192	-0.059%

Key: EDAR – Event Day Abnormal Return

Universiti Malaya

Table 4.6, continued

Event Day	1201 EDAR	3042 EDAR	4324 EDAR	5199 EDAR	5210 EDAR	5218 EDAR	Average EDAR
2016-11-17	0.01202	-0.003933	0.0005415	-0.070692	-0.030375	-0.006221	-0.538%
2016-11-19	0.011768	0.001333	0.0069407	0.082834	0.0030376	0.009272	0.186%
2016-11-21	0.011768	0.001333	0.0069407	0.082834	0.0030376	0.009272	-0.079%
2016-11-30	0.095933	-0.002537	-0.003094	-0.005547	-0.021856	0.025817	-0.505%
2016-12-10	0.111192	-0.005781	0.0215899	0.0253843	0.0417325	0.040528	1.111%
2016-12-13	0.111192	-0.005781	0.0215899	0.0253843	0.0417325	0.040528	1.327%
2016-12-14	0.014222	0.010745	0.0099591	0.0760224	0.0152083	-0.01616	0.263%
2016-12-15	0.012261	-0.011181	0.0059491	0.0430245	-0.009391	0.001943	-0.177%
2016-12-16	-0.08198	0.007446	0.0144779	-0.013259	0.0153357	0.015037	0.522%
2017-01-08	0.045427	-0.011546	0.0868888	0.0776123	0.0014396	-0.009864	-0.802%
2017-01-11	-0.07562	-0.011504	0.0047275	0.0234743	-0.009949	-0.028871	-1.302%
2017-01-14	-0.14189	0.009816	-0.017423	-0.097222	-0.018735	-0.024049	-0.499%
2017-01-17	0.052684	0.015094	0.0082066	0.0579329	0.0030903	0.000146	-0.312%
2017-01-22	-0.01575	0.017386	-0.00078	-0.009313	-0.038016	-0.005586	-0.489%
2017-02-07	-0.0135	0.035348	0.0330739	-0.020266	0.0075415	-0.029778	-0.149%
2017-02-08	-0.12461	0.0042	0.0308902	-0.028607	-0.01742	0.01928	0.778%
2017-02-14	-0.01461	0.02531	-0.010445	-0.030646	0.0145233	0.018763	0.776%
2017-02-15	-0.01935	-0.007894	-0.014309	-0.040619	-0.022195	-0.008664	-0.897%
2017-02-21	-0.07024	-0.055544	0.0215682	0.0090715	0.0027841	-0.004173	-0.118%
2017-02-24	-0.06315	0.006467	-0.036376	0.0008183	-0.035932	0.010898	-1.091%
2017-03-06	-0.06613	-0.012712	-0.010598	0.0079506	-0.019808	0.006669	-0.126%
2017-03-10	0.067923	-0.039992	-0.043194	-0.058326	-0.012845	-0.052407	-1.628%
2017-03-11	-0.06779	-0.027987	-0.026709	-0.04662	-0.026607	0.005256	-1.628%
2017-03-16	0.000939	-0.001643	0.0002622	0.0106862	0.0233655	0.037944	0.977%
2017-03-26	0.000919	-0.01599	-0.009066	-0.030435	-0.032999	-0.008952	-0.845%
2017-04-04	0.003184	-0.004898	0.0020224	-0.012452	0.0107339	0.068925	1.183%
2017-04-11	0.00537	-0.022511	-0.00851	0.0034053	2.323E-05	0.029315	0.841%
2017-04-24	-0.00125	-0.001362	-0.000807	0.0019794	-0.009505	0.004282	0.593%
2017-04-27	-0.06776	-0.002569	-0.012017	-0.009074	-0.007807	-0.009237	-0.182%

Key: EDAR – Event Day Abnormal Return

Universiti Malaya

Table 4.6, continued

Event Day	1201 EDAR	3042 EDAR	4324 EDAR	5199 EDAR	5210 EDAR	5218 EDAR	Average EDAR
2017-04-28	0.070337	0.027537	-0.00805	-0.009884	-0.007873	0.004102	-0.411%
2017-05-22	0.080549	-0.013303	0.0315373	0.033242	0.0057952	0.019975	1.383%
2017-05-24	0.001062	-0.001952	8.442E-05	-0.033301	-0.007935	0.003016	-0.899%
2017-05-25	0.001221	-0.011633	0.0824571	-0.010072	0.0055591	0.00374	0.335%
2017-05-31	-0.07829	-0.032903	-0.062612	0.1341881	0.0476086	-0.05332	-0.325%
2017-06-01	0.005775	0.027194	0.0514412	0.0126463	-0.00614	0.018126	0.350%
2017-06-09	0.003498	-0.011343	-0.031945	-0.024487	0.0134242	0.041248	1.332%
2017-06-13	0.006276	-0.005424	-0.010179	-0.002039	0.0002636	0.00712	0.060%
2017-06-22	0.006182	-0.040705	-0.059848	-0.014513	-0.013372	-0.024094	-1.611%
2017-06-26	0.003311	-0.003527	-0.016966	-0.011893	-0.019689	0.002797	1.093%
2017-07-18	0.001554	0.026526	0.0113134	-0.022633	0.0019522	0.005169	0.241%
2017-07-23	0.001465	-0.010736	0.0189674	-0.007515	0.0011033	-0.003855	-1.130%
2017-07-24	0.001465	-0.010736	0.0189674	-0.007515	0.0011033	-0.003855	-1.130%
2017-07-29	-0.00136	0.015884	0.139932	0.0035667	0.0090175	-0.002166	0.036%
2017-08-02	0.001667	0.021327	0.0523676	0.0032486	0.0103885	0.012916	0.897%
2017-08-08	0.101616	-0.047645	-0.025278	0.1107161	0.0096027	0.012509	0.348%
2017-08-14	0.001313	-0.015099	-0.042925	0.0229508	0.0304359	0.014832	1.233%
2017-08-24	-0.09571	0.000174	-0.007748	0.0097465	-0.004932	-0.003153	-0.807%
2017-09-05	0.001145	0.032586	0.0110566	-0.024927	-0.000544	0.024739	0.646%
2017-09-06	0.001145	0.024561	0.0097557	0.047252	0.000123	0.046013	0.873%
2017-09-14	0.001751	0.006515	-0.005498	-0.012644	0.0119936	0.015152	0.934%
2017-09-15	0.101751	-0.014026	-0.008359	0.0313859	0.0181117	0.051228	1.102%
2017-09-22	-0.00098	-0.002091	-0.001248	-0.003494	-0.002183	-0.000649	-1.215%
2017-09-28	0.083569	0.002982	0.0011442	-0.021872	-0.01672	-0.077988	-2.232%
2017-09-29	-0.09114	-0.005291	0.0004159	-0.013117	0.0040027	0.016701	0.235%
2017-10-10	-0.00195	0.004668	0.0027193	0.030375	0.0075795	0.000239	0.405%
2017-10-21	-0.10183	0.025418	0.005761	-0.028537	0.0006061	-0.009066	-0.771%
2017-10-24	-0.00183	-0.023411	-0.016419	-0.071754	0.0076484	-0.023402	-1.058%
2017-11-07	-0.0019	0.00457	0.0155067	-0.00432	-0.005928	0.021093	0.639%

Key: EDAR – Event Day Abnormal Return

Universiti Malaya

Table 4.6, continued

Event Day	1201 EDAR	3042 EDAR	4324 EDAR	5199 EDAR	5210 EDAR	5218 EDAR	Average EDAR
2017-11-30	-0.0988	-0.030258	-0.026153	-0.023004	0.0112041	0.009607	1.143%
2017-12-01	0.001195	-0.00286	-0.00833	-0.002197	-0.002948	0.002029	1.143%
2017-12-13	-0.10591	0.015153	0.0366855	0.0151083	-0.008888	-0.004985	0.116%
2017-12-20	-0.00258	0.024698	0.0181719	-0.005002	0.0131458	0.000557	0.781%
2017-12-21	0.004091	-0.003773	-0.023083	0.0084504	0.0008139	0.014642	0.999%

Key: EDAR – Event Day Abnormal Return

Figure 4.8 shows the fluctuation of average event day abnormal return (EDAR) based on the OPEC news release date.

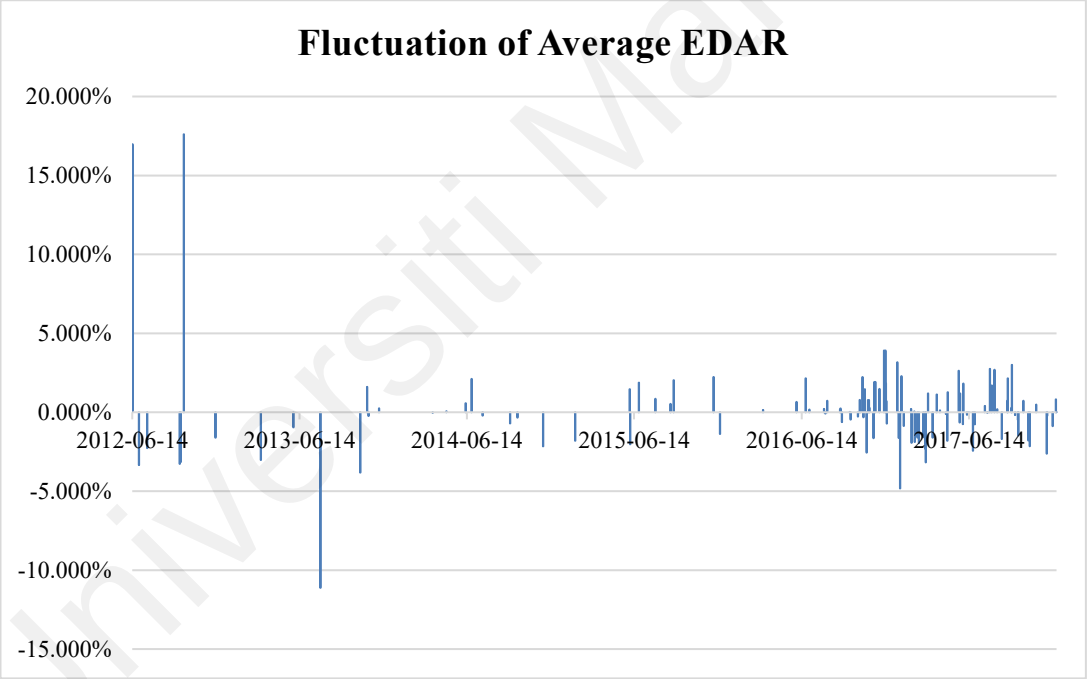

Figure 4.8 Fluctuation of Average EDAR

4.4 OPEC News Sentiment Impact

Before performing statistical analysis, proper hypotheses need to be defined and all the datasets are normally distributed (Das & Imon, 2016). The normality of each dataset: OPEC sentiments score, six companies average CAR during event window, and six companies average abnormal return on each event day (news release day) are determined

Universiti Malaya

before performing test of hypotheses. This research uses linear regression analysis to study the relationship between OPEC news sentiment and the movement of stock prices of the six selected companies. Before apply linear regression analysis method, datasets should be check for the five assumptions for linear regression analysis which are Linearity, Normality, Homoscedasticity, No or little multicollinearity, and No auto-correlation (Dalington & Hayes, 2016). As mentioned in Chapter 3, the statistical analysis of this research is conducted using IBM SPSS Statistics 22.

4.4.1 Hypotheses

To study whether OPEC news sentiments have impact on the stock prices of the six selected energy sector (oil & gas) companies, hypotheses are defined before performing statistical analysis. Null hypotheses (H_0) and alternative hypotheses (H_1) in this research are defined as follow:

Independent Variable: OPEC news sentiment

Dependent Variable: Six Companies Average CAR

H_0: There is no statistically significant relationship between OPEC news sentiment and the average cumulative abnormal return of stock prices of the six selected public listed energy sector companies.

H_1: There is statistically significant relationship between OPEC news sentiment and the average cumulative abnormal return of stock prices of the six selected public listed energy sector companies.

Independent Variable: OPEC news sentiment

Dependent Variable: Event Day Fluctuation of the analyzed companies

Universiti Malaya

H'$_0$: There is no statistically significant relationship between OPEC news sentiment and the average cumulative abnormal return of stock prices of the six selected public listed energy sector companies.

H'$_1$: There is statistically significant relationship between OPEC news sentiment and the average cumulative abnormal return of stock prices of the six selected public listed energy sector companies.

4.4.2 Assumptions for Linear Regression Analysis

1) Linearity: The first assumption for linear regression analysis is linearity. It is the primary assumption. It states that the independent variables in the regression have a straight-line relationship with the dependent variable (Dalington & Hayes, 2016). The linearity of the datasets were tested through SPSS (SPSS lnc, 1990). Following table shows the results.

Table 4.7: Linearity Analysis for Average CAR with OPEC Sentiment

			Sum of Squares	df	Mean Square	F	Sig.
Average CAR * OPEC Sentiment	Between Groups	(Combined)	5.450	46	.118	.828	.751
		Linearity	.031	1	.031	.214	.645
		Deviation from Linearity	5.419	45	.120	.841	.730
	Within Groups		9.877	69	.143		
	Total		15.327	115			

Table 4.7 above shows that the Sig. value of Deviation from Linearity is 0.730 which is greater than 0.05. This implies that there is a linear relationship between the variables of OPEC Sentiment and Average CAR.

Universiti Malaya

Table 4.8: Linearity Analysis for Average Event Day Fluctuation with OPEC Sentiment

			Sum of Squares	df	Mean Square	F	Sig.
Average Event Day Fluctuation * Sentiment	Between Groups	(Combined)	30.657	46	.666	1.045	.428
		Linearity	.102	1	.102	.160	.691
		Deviation from Linearity	30.555	45	.679	1.065	.402
	Within Groups		44.010	69	.638		
	Total		74.667	115			

As shown in Table 4.8 above, the Sig. value of Deviation from Linearity is 0.402 > 0.05. This implies that there is a linear relationship between the variables of OPEC Sentiment and the Average Event Day Fluctuation.

2) Normality: It is one of the most common assumptions made in the development and use of statistical procedures (Thode, 2002). In this research, SPSS (SPSS lnc, 1990) was used not only to do the normality test of datasets and but also to analyze the statistical relationship between these datasets.

Normality test can be divided into graphical or non-graphical tests (Stevens, 2012): Non-graphical tests include the chi-square goodness of fit test, the Kolmogorov–Smirnov test, the Shapiro–Wilks test, and the evaluation of kurtosis and skewness values. Graphical tests include the normality probability plot and the histogram. Non-graphical tests are preferred for small to moderate sample sizes, with the Shapiro–Wilks test and the evaluation of kurtosis and skewness values being preferred methods for sample sizes of less than 40 (Stevens, 2012). Since the sample size of these datasets are all 116 (the same as number of OPEC news collected), the graphical tests are applied to analyze normality of the datasets.

In this research the histogram and Quantile-Quantile (Q-Q) plot graphical tests were applied. Histogram is the easiest and simplest plot, in which the observed values are

Universiti Malaya

plotted against their frequency. Histogram states a visual indication whether the distribution is approximately bell shaped or not. Data that can be represented as bell-shaped curve has normal distribution. Otherwise, its non-normal data (Das & Imon, 2016). Quantile-Quantile (Q-Q) plot compares the quantiles of data distribution with the quantiles of a standardized theoretical distribution (normal distribution). When quantiles of two distributions are met, plotted dots are all approximately on the 45 degree rising straight line, which means that the analyzed data has normal distribution (Das & Imon, 2016). Figure 4.9 to Figure 4.14 are the histograms and Q-Q plots of the three datasets mentioned above.

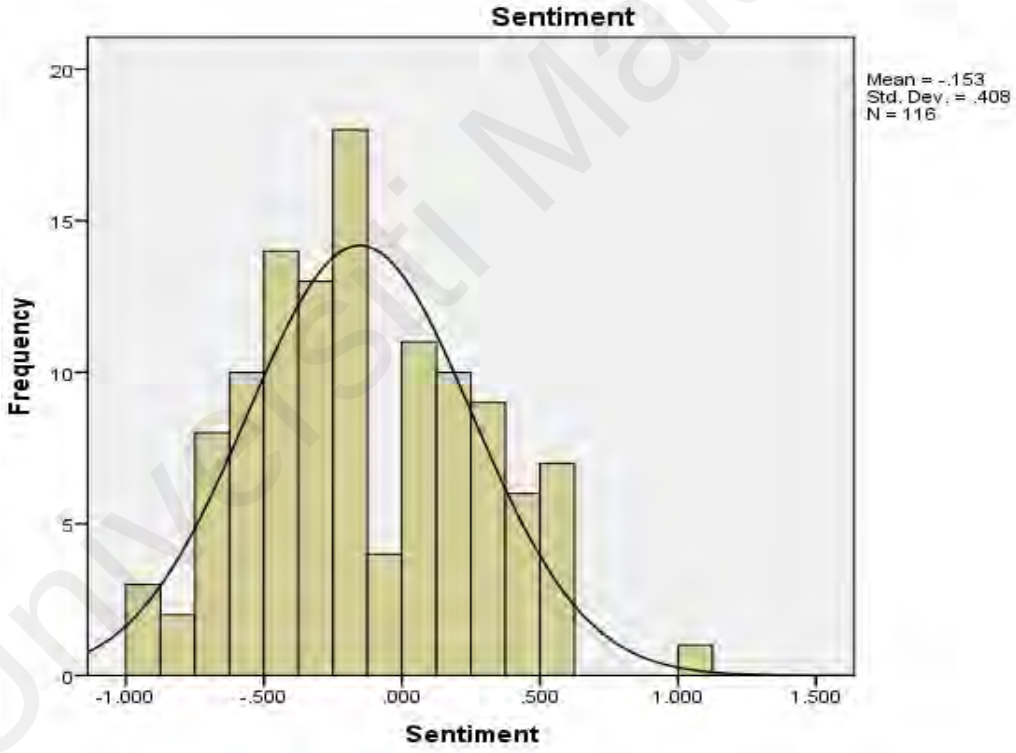

Figure 4.9: Histogram of Sentiment Score

Universiti Malaya

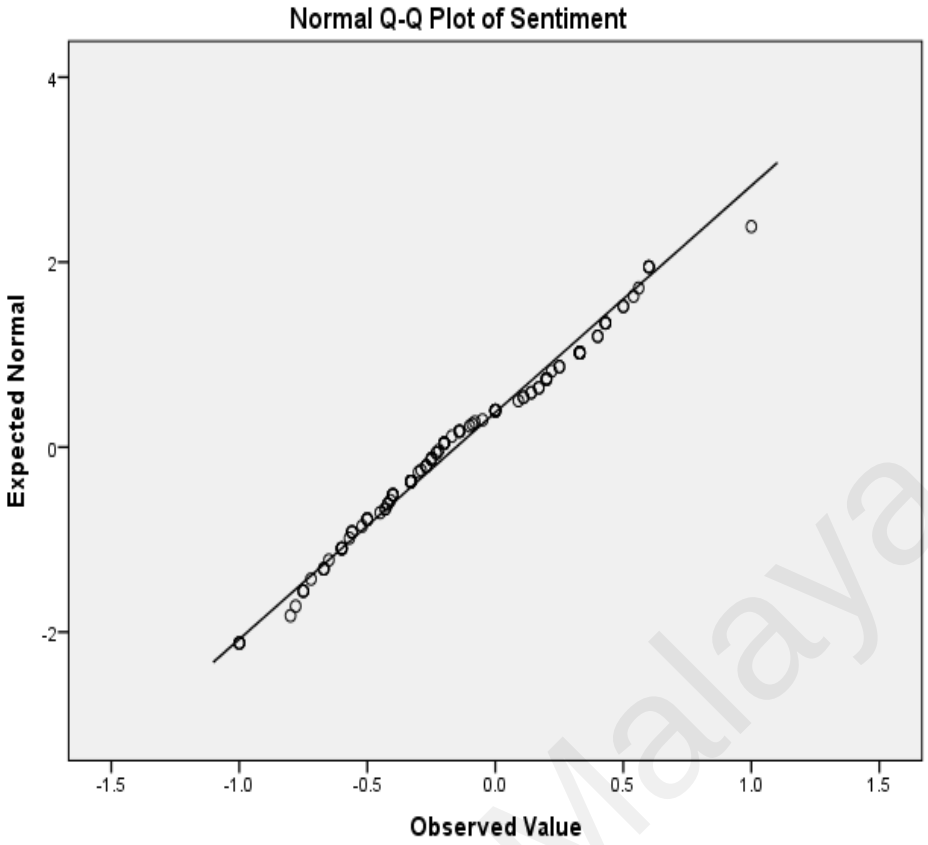

Figure 4.10: Normal Q-Q Plot of Sentiment

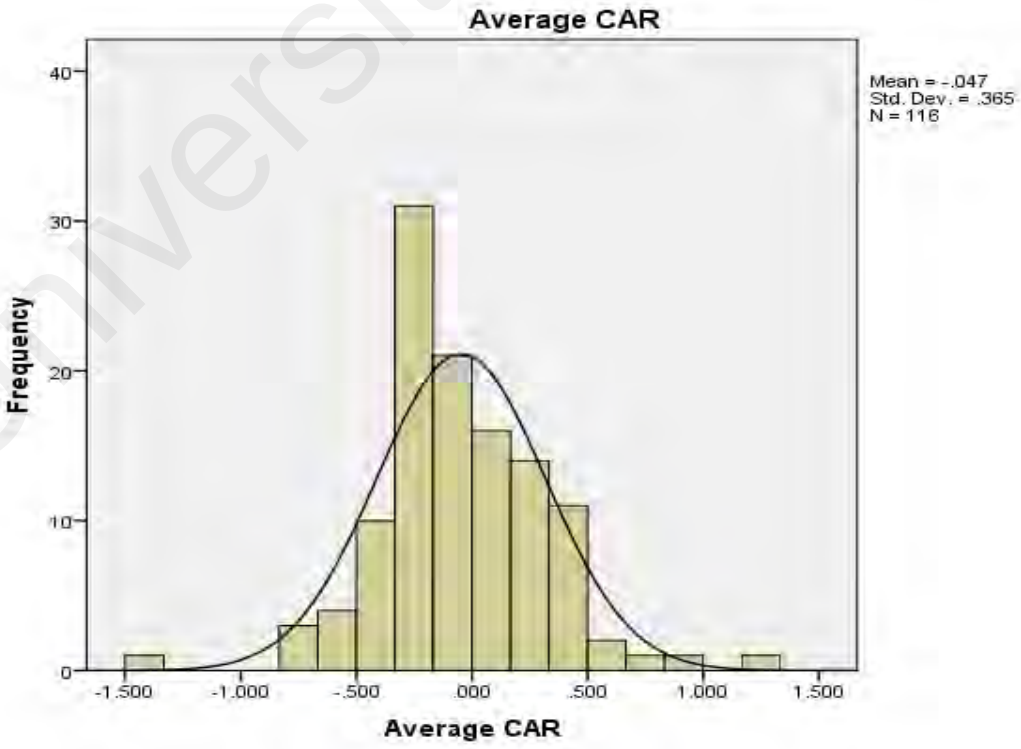

Figure 4.11: Histogram of Six Companies' Average CAR

Universiti Malaya

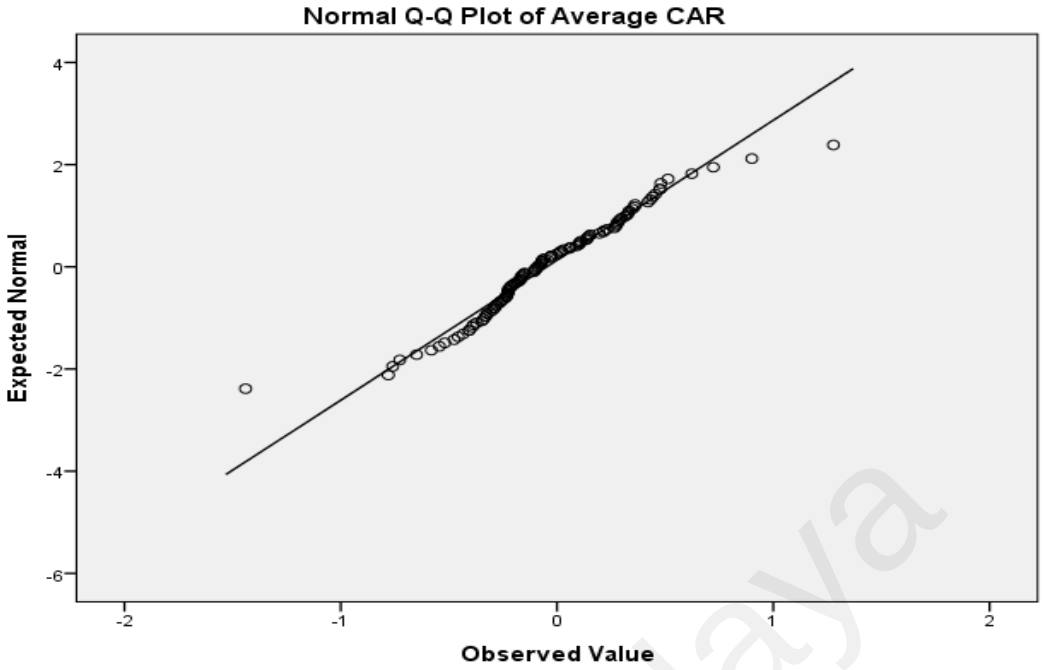

Figure 4.12: Normal Q-Q Plot of Average CAR

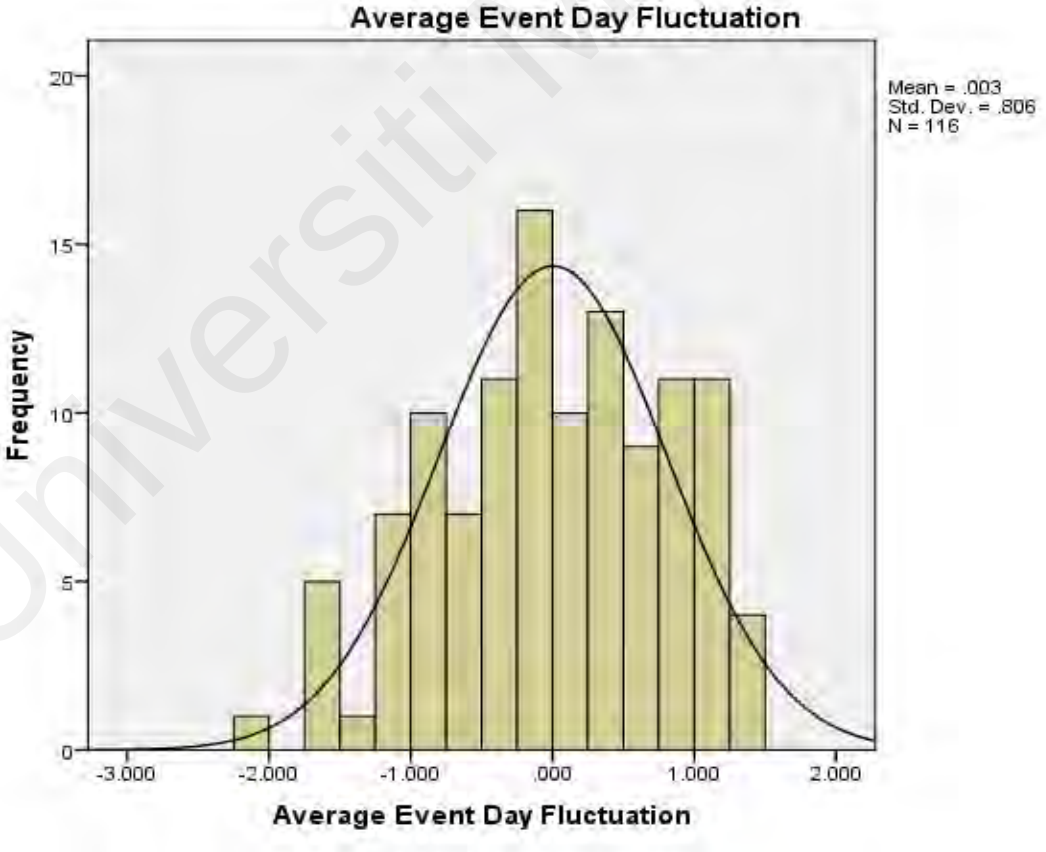

Figure 4.13: Histogram of Average Event Day Fluctuation of the Six Companies

University Malaya

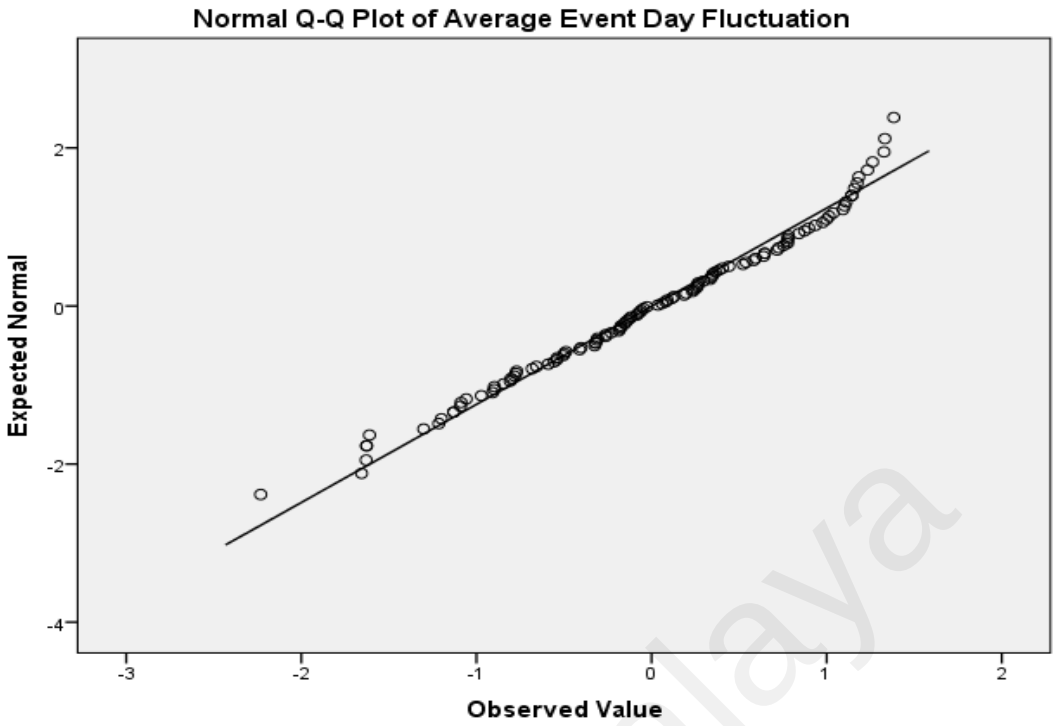

Figure 4.14: Normal Q-Q Plot of Six Companies Average Event Day Fluctuation

From the results of graphical tests above, three datasets are all approximately have a perfect bell-shape curve in histogram, and also fit the expected normal line in Q-Q plot. In the book titled "SPSS survival manual" mentioned that for large sample (> 40), the moderate departure of normality should not cause major problems (Pallant, 2013). It implies that we can apply parametric procedures even the data are not perfectly normally distributed (Elliott & Woodward, 2007). Hence, the results above show that these datasets are suitable for parametric analysis.

3) Homoscedasticity: It states that the conditional distribution of dependent variables have equal variance (Dalington & Hayes, 2016). If dependent variables meet homoscedasticity assumption, its scatterplot of residuals should be equally distributed above and below zero on the X axis, and to the left and right of zero on the Y axis without obvious pattern (StatisticsSolutions, 2019). The scatterplots of residuals are as follow.

Universiti Malaya

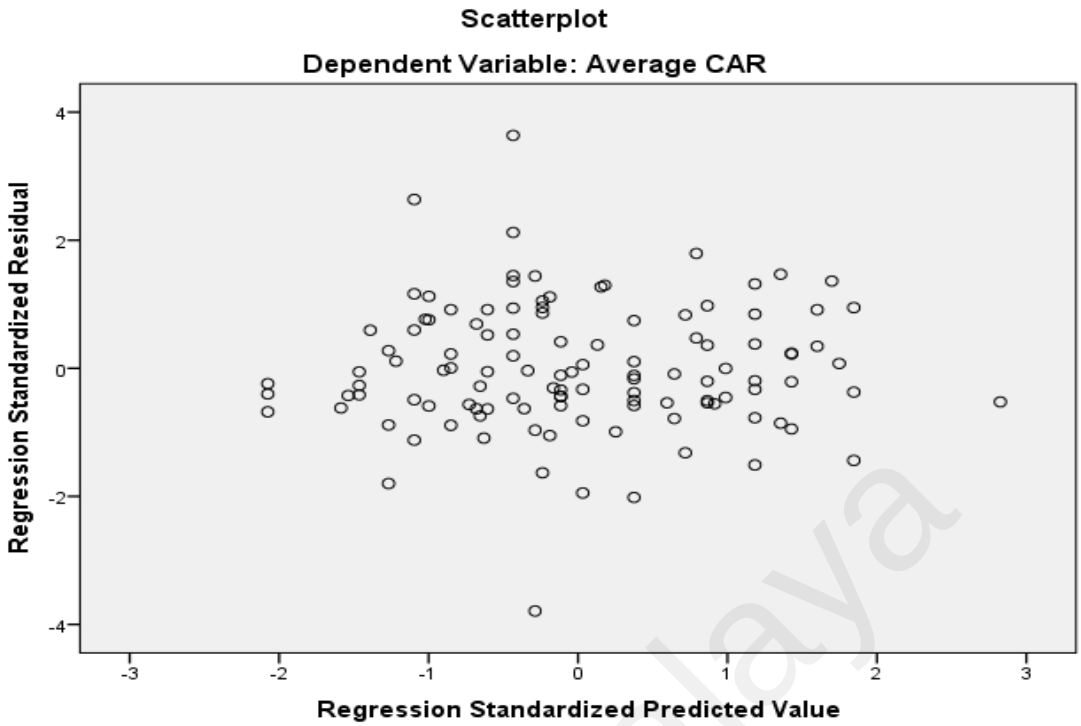

Figure 4.15: Scatterplot of Regression Standardized Residual (Average CAR)

Figure 4.15 above shows that the residuals are equally distributed above and below zero on the X axis, and to the left and right of zero on the Y axis without obvious pattern. Hence, Average CAR meets the assumption of homoscedasticity.

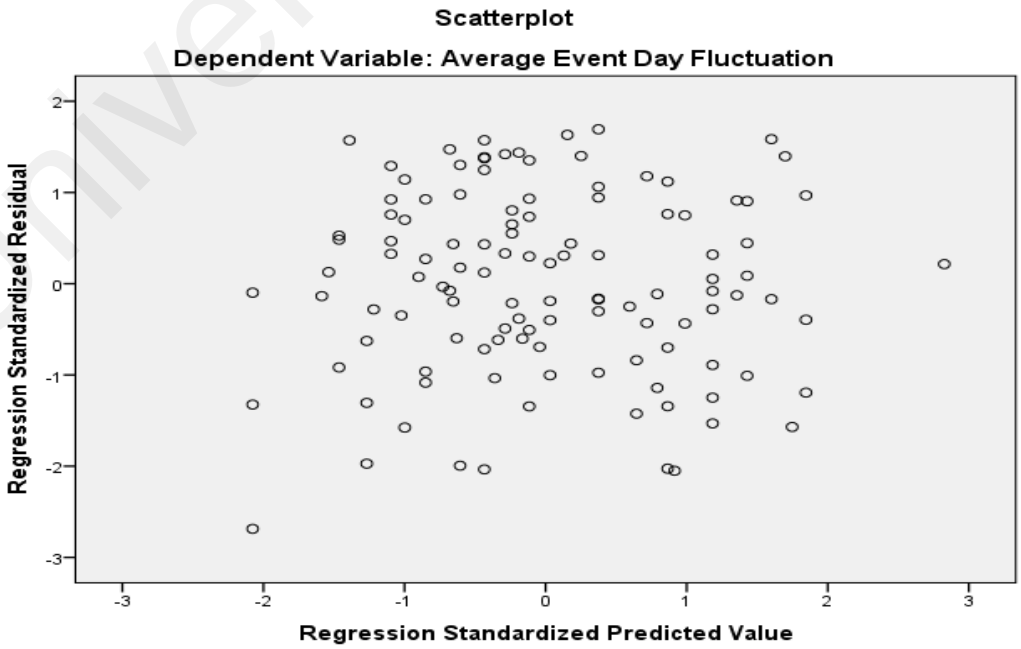

Figure 4.16: Scatterplot of Regression Standardized Residual (Average Event Day Fluctuation)

Universiti Malaya

Figure 4.16 above shows that the residuals are equally distributed above and below zero on the X axis, and to the left and right of zero on the Y axis without obvious pattern. Hence, Average Event Day Fluctuation meets the assumption of homoscedasticity.

4) No or little multicollinearity: It means there is no linear relationship between any two or more independent variables (Dalington & Hayes, 2016). This research only has one independent variable: OPEC news sentiment. Thus, it meets the assumption of no multicollinearity.

5) No auto-correlation: There is no correlation between the values of the same variables across different observations in the data (Dalington & Hayes, 2016). Auto-correlation can be checked by Durbin Watson test in IBM SPSS Statistics (Pallant, 2013). If the value of Durbin Watson test is between 1.5 to 2.5, it indicates that there is no auto-correlation in the data. The following tables present the results of Durbin Watson test.

Table 4.9: Durbin Watson Test (Average CAR)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.868[a]	.753	.750	.1267155	2.009

a. Predictors: (Constant), OPEC News Sentiment

b. Dependent Variable: Average Cumulative Abnormal Return

Table 4.10: Durbin Watson Test (Average Event Day Fluctuation)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.037[a]	.001	-.007	.808753	1.610

a. Predictors: (Constant), OPEC News Sentiment

b. Dependent Variable: Average Event Day Fluctuation

Universiti Malaya

As shown in Table 4.9 and 4.10 above, the values of Durbin Watson test are both between 1.5 to 2.5, respectively. Therefore, the analyzed datasets meet the assumption of no auto-correlation.

4.4.3 Linear Regression Analysis

Since the five assumptions of linear regression analysis are all meet. The linear regression analysis is used to study the relationship between OPEC news sentiment and the movement of stock prices of the six selected companies.

Regression analysis is a statistical modeling which contains a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables. It provides a conceptually simple method to investigate functional relationships among variables, regression analysis has become one of the most widely used statistical analysis tools for analyzing multifactor data. The standard approach in linear regression analysis is to take data, fit a model, and then evaluate the fit by using t statistic, F score, and R^2 (Samprit & Hadi, 2015). Since the datasets in this research meet the assumption of linearity. The linear regression analysis is chosen to test the hypotheses.

Following table 4.11 and table 4.12 present the results of linear regression analysis from SPSS that analyzed the relationship between OPEC news sentiments and Average Cumulative Abnormal Return (CAR) of selected companies which were observed during the event window (five days before and after the event day) based on each event day (news release day) of OPEC news.

Universiti Malaya

Table 4.11: Model Summary for Linear Regression Analysis of Average CAR with OPEC News Sentiment

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.868[a]	.753	.750	.1267155	2.009

a. Predictors: (Constant), OPEC News Sentiment

b. Dependent Variable: Average Cumulative Abnormal Return

From the table 4.11 above, it shows that there is strong correlation between the two variables, and more than 75 percent of the data of sentiment scores fit the regression line. Moreover, Table 4.12 below shows the results from analysis of variance (ANOVA) of two variables. It is a statistical method used to analyze the differences among group means in a sample (Blanca, Alarcón, Arnau, Bono, & Bendayan, 2017).

Table 4.12: ANOVA for Linear Regression Analysis of Average CAR with OPEC News sentiment

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.567	1	5.567	346.722	.000[b]
	Residual	1.830	114	.016		
	Total	7.398	115			

a. Dependent Variable: Average Cumulative Abnormal Return

b. Predictors: (Constant), OPEC News Sentiment

As shown in Table 4.12, due to the results With F = 346.7 and 115 degrees of freedom, and p-value is less than 0.05, it indicates strong evidence against the null hypothesis (Samprit & Hadi, 2015) which implies the test is highly significant. Thus, there is a statistically significant linear relationship exists between two analyzed variables.

Universiti Malaya

Table 4.13: Coefficients for Linear Regression Analysis of Average CAR and OPEC News sentiment

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-.026	.013		-2.089	.039		
	Sentiment	-.539	.029	-.868	-18.620	.000	1.000	1.000

a. Dependent Variable: Average Cumulative Abnormal Return

Table 4.13 shows the regression coefficients, the intercept and the significance of all coefficients in the model. The linear regression analysis estimates the linear regression function to be:

$$\text{Average CAR} = -0.26 - 0.539^{*} \text{ OPEC News Sentiment} \qquad (4.1)$$

The figure below shows the sentiment line fit plot.

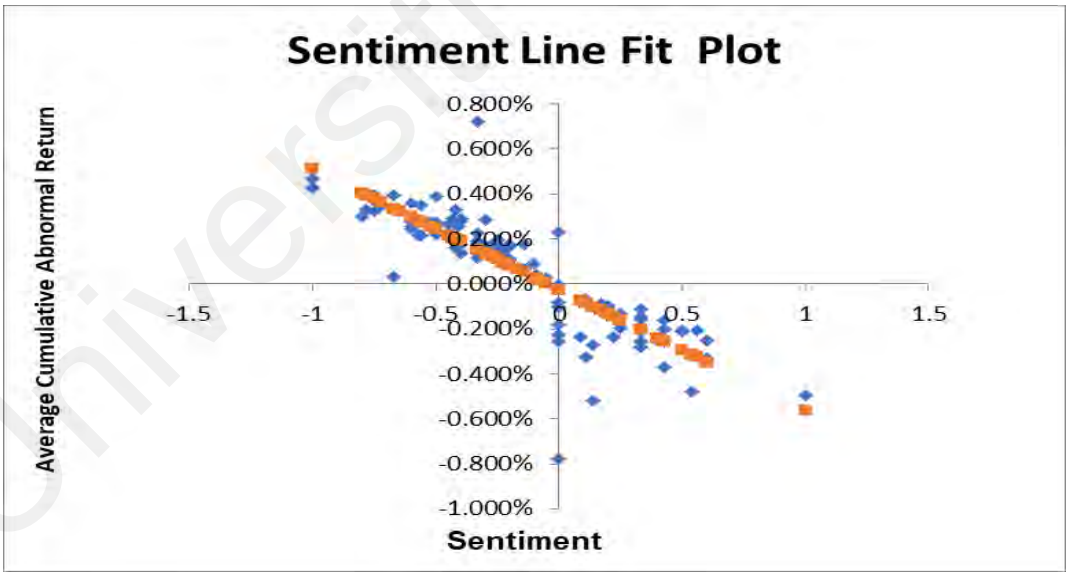

Figure 4.17: Sentiment Line Fit Plot (Average CAR)

Figure 4.17 above shows majority of the sentiment score and Average CAR fits a linear function line. It further proves the linear relationship between OPEC news sentiment and analyzed companies' Average CAR.

Universiti Malaya

Furthermore, since this research also collected data of each selected company's abnormal return on the OPEC news release date, the relationship between OPEC news sentiment and selected companies' event (news release) day's abnormal return is also studied. From the analysis results of relationship between OPEC news sentiment and selected energy sector (oil & gas) companies' event (news release) day's abnormal return, this research can find out how the OPEC news sentiment influence the energy sector (oil & gas) companies stock prices on event day (news release day). Analysis of their relationship is conducted by using linear regression analysis in IBM SPSS Statistics. The results of linear regression analysis for relationship between OPEC news sentiment and event day's fluctuation of analyzed companies stock prices are as follow.

Table 4.14: Model Summary for Linear Regression Analysis of Average Event Day Fluctuation with OPEC News Sentiment

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.037[a]	.001	-.007	.808753	1.610

a. Predictors: (Constant), OPEC News Sentiment

b. Dependent Variable: Average Event Day Fluctuation

As shown in Table 4.14 above, the value of R square is 0.001, which is close to 0. This means that there is no strong correlation between the two variables. Table 4.15 shows the analysis of variance of two variables (ANOVA).

Table 4.15: ANOVA for Linear Regression Analysis of Average Event Day Fluctuation with OPEC News sentiment

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.102	1	.102	.156	.694[b]
	Residual	74.565	114	.654		
	Total	74.667	115			

a. Dependent Variable: Average Event Day Fluctuation

b. Predictors: (Constant), OPEC News Sentiment

Universiti Malaya

Since the result shows that the significance P value is greater than 0.05, this implies that there is no significant statistical relationship between the OPEC news sentiment and the stock market fluctuation of the analyzed companies on the event day (news release day). Finally, the figure of sentiment fit line plot based on the analyzed companies' event day abnormal return is plotted (Figure 4.18).

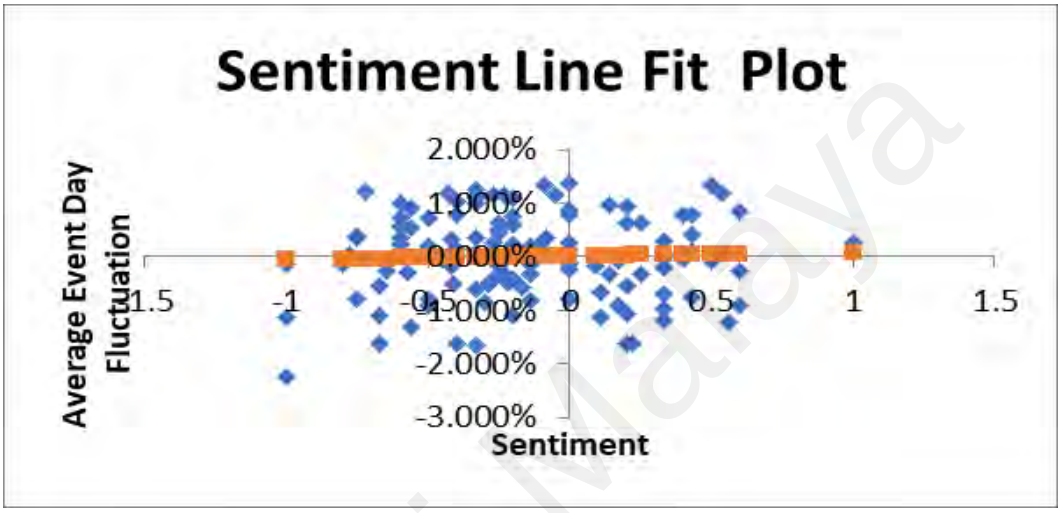

Figure 4.18: Sentiment Line Fit Plot (AEDF)

As shown in Figure 4.18, there is no obvious pattern can be observed in the sentiment line fit plot. This outcome shows that there is no significant statistical relationship between the OPEC news sentiment and the analyzed companies' event day fluctuation of stock prices.

4.5 Conclusion

In summary, the results of the data analysis in this research indicate there is negative correlation exists between OPEC news sentiment and selected energy sector (oil & gas) companies' average cumulative abnormal return during the event window based on each OPEC news release date. However, there is no significant statistical relationship between OPEC news sentiment and studied oil & gas companies abnormal return on event day. Hence, according to the results of data analysis, the hypotheses H_1 and H'_0 are supported and hypothesis H_0 and H'_1 are rejected

Universiti Malaya

Findings of this research provide useful guidelines to those investors who invest in those six energy sector companies especially near the OPEC news release day, which can help them make better investment decision.

Universiti Malaya

CHAPTER 5: CONCLUSION AND DISSCUSSION

This chapter summaries the findings of this research. It also explains the problems encountered when conducting this research. The weakness of the study and suggestions for future works are also presented in this chapter.

5.1 Research Findings

Based on the results of data analysis explained Chapter 4, the findings of this research can be summarized in relation to each research objectives

Objective 1: To build an innovative classifier to classify the OPEC news sentiment

The machine learning algorithm-based classifier is used to classify the OPEC news based on its sentiments. Stochastic Gradient Descent Classifier (SGDC) is found to outperform other tested supervised machine learning classifiers. The performance of the classifier is improved by generating the training feature of the machine learning classifier from financial news articles. The training data are prepared using lexicon-based labelling (financial dictionary) method which is better and much more efficient than manual labelling. Combing Stochastic Gradient Descent Classifier with financial dictionary labelling is innovation in related research field.

Objective 2: To improve the accuracy of the innovative classifier by using proper lexicon resource in labelling training data.

To improve the accuracy of the classifier, combination of Lexicon-based approach and machine learning algorithm are used. A sentiment dictionary from the finance domain is applied to label the training data, in this way, bias due non-financial sentiment words can be reduced. The machine learning algorithm applied in the classifier was selected after testing multiple supervised machine learning algorithms, which further ensures the performance of the classifier.

Universiti Malaya

Figure 5.1 below compares the difference of classification methods between this research and the most recent and related research conducted by Nasim (Nasim, 2018). It further proves the machine learning classifier is innovative and with better accuracy.

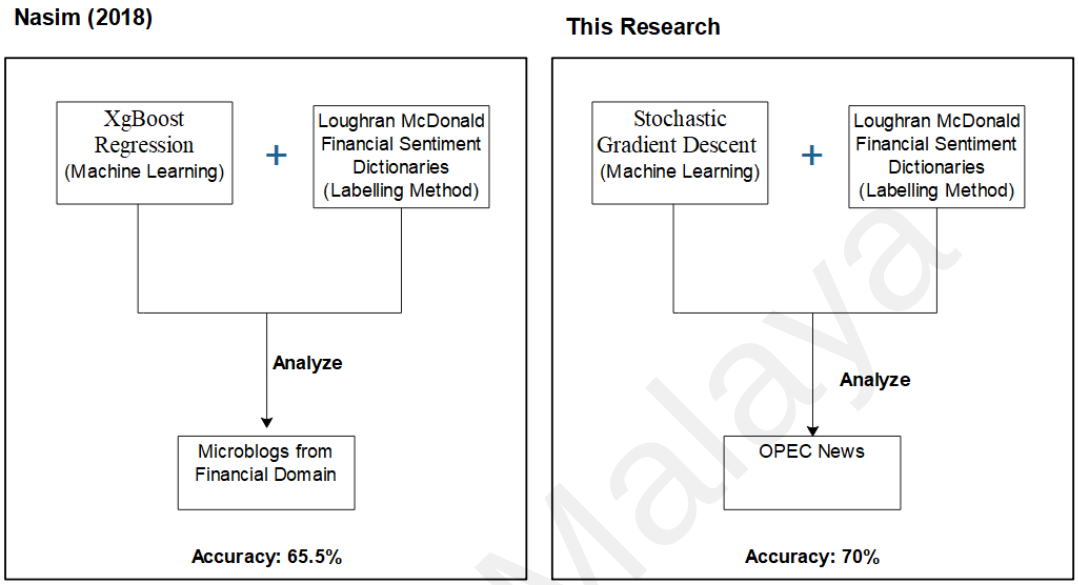

Figure 5.1: Comparison with Related Research

Objective 3: To find out how would the public listed Malaysian energy sector (oil & gas) companies' stock prices react to the OPEC news sentiments.

The results of data analysis show that there is a negative correlation between OPEC news sentiment and the average cumulative abnormal return of six energy sector (oil & gas) companies happened during the event window (five days before and after the news release day). However, the stock prices of these six Malaysian public listed (energy sector: oil & gas) companies do not react to OPEC news sentiments on the event day.

Event study method provides the timeline which helps to isolate the impact of OPEC news from other events. Statistical analysis of stock prices which were based on event study shows the fluctuation of stock prices occurred during the event window of OPEC news announcement. This research contributed to the finding of which classifier is the best performed machine learning classifier.

Universiti Malaya

5.2 Research Contribution

The contributions of this research are as follow:

Firstly, this research combined and extended the lexicon-based approach and machine learning algorithm to the study of news sentiments. Loughran-McDonald Master Dictionary from finance domain improves the accuracy of labelling training data. Stochastic Gradient Descent is a simple but efficient machine learning algorithm which can discriminatively learn linear classifier under convex loss functions such as Logistic Regression and linear Support Vector Machine. By combining Loughran-McDonald Master Dictionary with Stochastic Gradient Descent machine learning algorithm, an innovative financial news sentiment classifier with relatively higher accuracy (70%) is built.

Secondly, the findings of this research show the negative correlation between the sentiment of OPEC news and the average cumulative abnormal return of Malaysian energy sector (oil & gas) companies. This is a valuable information to investors in making better investment decision based on OPEC news sentiments.

5.3 Problems Encountered

Challenges in conducting this research comes from different aspects. Firstly, the data sets applied in this research should be processed correctly, especially the training data set. This research applies machine learning algorithm-based classifier to do the classification task. The training data has significant influence on the results of the classifications, and thus, generate a proper training dataset is a crucial task in this research. Good training data can result in an outperforming machine learning classifier. But there is no exact automatic procedure to measure how much of those training data are labelled correctly except manually checking. Since there are over 37,000 sentences in training dataset, and

Universiti Malaya

this research does not include any linguistic experts, manually checking the label of training data is difficult.

On the other hand, to find out which machine learning algorithm-based classifier has better performance, different machine learning algorithms need to be tested. Each of those machine learning algorithms has their own mathematical logic. But there is still no algorithm suitable for this research perfectly. Building a new algorithm specific for OPEC news sentiment classification based on its own characteristics is out of this research's scope.

Finally, the historical stock prices datasets are another critical part in this research. To analyze the fluctuation of stock prices of the six selected energy sector (oil & gas) companies during 2012 to 2017, historical stock prices data of these companies need to be collected. But there is no perfectly reliable source of these data. Although this research collected historical stock market data from Yahoo finance, which is chosen by most of the researchers from finance domain, it is just only a relatively more reliable resource.

5.4 Weakness of the Study

In this research, Stochastic Gradient Descent Classifier (SGDC) outperforms the other tested supervised machine learning algorithm-based classifiers. This is due to its one versus all (OVA) scheme which combines multiple binary classifiers that making SGDC to achieve 70 percent accuracy only. Many factors may influence the accuracy of the machine learning classifier. For example, the processing of textual data, the feature of training data and the parameters of the machine learning algorithms. The accuracy score of the machine learning classifier built in this research is still not significant enough.

The research scope and the size of datasets applied in this research are limited. Only six companies (21.4%) which were randomly chosen from the 28 energy sector (oil &

Universiti Malaya

gas) companies listed on the Main Market Board of Bursa Malaysia. Hence, the results are not representative due to the limited research sample.

Besides just using cumulative abnormal return as index of the fluctuation of stock prices, professional financial analysis model can also be applied to analysis the behavior of stock prices of selected energy sector (oil & gas) companies. The procedure may be more complicate, but the results can be better as well.

5.5 Future Works

Future works can focus on building a machine learning classifier with better performance. This can be done in two ways. Firstly, since the Loughran and McDonald Financial Sentiment Dictionaries only has limited number of financial text sentiment words, by building a better lexicon dictionary in financial domain, the accuracy of lexicon analysis of the financial news text can be improved. Consequently, the accuracy score of machine learning algorithm-based classifier can be also improved by using better training data which is processed by improved lexicon dictionary.

Secondly, the machine learning algorithm-based classifier can be improved by applying better parameters. This needs further study in mathematical domain. By understanding the mathematical theory behind the algorithm, the parameters of the machine learning algorithm can be chosen more appropriately. Thus, the performance of the machine learning algorithm-based classifier can also be enhanced.

Finally, future works can also focus on further analysis the impact of OPEC news sentiments on energy sector (oil & gas) companies stock prices by expanding the research scope.

Universiti Malaya

REFRENCES

Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. *Journal of Applied Mathematics*, *2014*, 1–6. https://doi.org/10.1155/2014/425731

Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016). Deep learning for stock prediction using numerical and textual information. *In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 1–6.

Albanese, D., Merler, G. S., & Jurman, R. V. (2008). MLPy: high-performance Python package for predictive modeling. *NIPS,MLOSS Workshop*.

Almalki, S. (2016). Integrating quantitative and qualitative data in mixed methods research—challenges and benefits. *Journal of Education and Learning*, *5*(3), 288. https://doi.org/10.5539/jel.v5n3p288

Atkins, A., Niranjan, M., & Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, *4*(2), 120–137. https://doi.org/10.1016/j.jfds.2018.02.002

Bali, T. G., & Zhou, H. (2016). Risk, uncertainty, and expected returns. *Journal of Financial and Quantitative Analysis*, *51*(3), 707–735. https://doi.org/10.2139/ssrn.1993304

Bangi. (2007). *Research methodology*. Selangor: Malaysian Nuclear Agency. https://doi.org/10.1007/978-94-007-4848-4_3

Belgiu, M., & Dragut, L. (2016). Random forest in remote sensing : a review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Universiti Malaya

Biltawi, M., Etaiwi, W., Tedmori, S., Hudaib, A., & Awajan, A. (2016). Sentiment classification techniques for Arabic language: A survey. *International Conference on Information and Communication Systems, ICICS 2016*, 339–346. https://doi.org/10.1109/IACS.2016.7476075

Bina, C., & Vo, M. (2007). OPEC in the epoch of globalization : An event study of global oil prices. *Global Economy Journal*, *7*(1).

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29*(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Bordino, I., Kourtellis, N., Laptev, N., & Billawala, Y. (2014). Stock trade volume prediction with Yahoo Finance user browsing behavior. *2014 IEEE 30th International Conference on Data Engineering*, 1168–1173.

Breiman, L., & Cutler, A. (2007). Random forests-classification description. *Department of Statistics*, *2*.

Bryan, K. (2018). How to calculate daily stock return. Retrieved from https://pocketsense.com/calculate-daily-stock-return-5138.html

Bursa Malaysia sectorial index series. (2018).

Campbell, J. Y., Champbell, J. W., Wen-Chuan Lo, A., & MacKinlay, A. C. (1997). *The econometrics of financial markets*. princeton University press.

Chan, S. W. K., & Chong, M. W. C. (2017). Sentiment analysis in financial texts. *Decision Support Systems*, *94*, 53–64. https://doi.org/10.1016/j.dss.2016.10.006

Chandrasheka, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers*

Universiti Malaya

and Electrical Engineering, *40*(1), 16–28.

Chau, M., & Chen, H. (2008). A machine learning approach to web page filtering using content and structure analysis, *44*, 482–494. https://doi.org/10.1016/j.dss.2007.06.002

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2008.06.054

Chen, L. (2017). News-Processed-Dataset. Retrieved from https://figshare.com/articles/News-Processed-Dataset/5296357

Colgan, J. D. (2014). *The emperor has no clothes: the limits of OPEC in the global oil market*. *International Organization* (Vol. 68). https://doi.org/10.1017/S0020818313000489

Cortes, C., & Vapnik, V. (1995). Support-Vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1109/64.163674

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., & Davis, B. (2018). SemEval-2017 task 5: fine-grained sentiment analysis on financial microblogs and news. In *The 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 847–851). https://doi.org/10.18653/v1/s17-2144

Croese, M., & Westerman, W. (2015). OPEC's influence on European oil stock returns. *Energy Technology and Valuation Issues*, 57–80. https://doi.org/10.1007/978-3-319-13746-9

D'Andrea, A., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and

Universiti Malaya

applications for sentiment analysis implementation. *International Journal of Computer Applications*, *125*(3), 26–33. https://doi.org/10.5120/ijca2015905866

Dalington, R. ., & Hayes, A. . (2016). *Regression analysis and linear models:Concepts,applications,and implementation*. Guilford Publications.

Das, K. R., & Imon, A. H. . R. (2016). A Brief Review of Tests for Normality. *American Journal of Theoretical and Applied Statistics*, *5*(1), 5. https://doi.org/10.11648/j.ajtas.20160501.12

Demirer, R., & Kutan, A. M. (2010). The behavior of crude oil spot and futures prices around OPEC and SPR announcements: An event study perspective. *Energy Economics*, *32*(6), 1467–1476. https://doi.org/10.1016/j.eneco.2010.06.006

Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, *340*, 250–261. https://doi.org/10.1016/j.ins.2016.01.033

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using Naïve Bayes ' and K -NN classifier. *ArXiv Preprint ArXiv:1610.09982*, *8*(4), 54–62.

Di Nunzio, G. M. (2014). A new decision to take for cost-sensitive Naïve Bayes classifiers. *Information Processing and Management*, *50*, 653–674.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. *Proceedings of the International Conference on Web Search and Web Data Mining*, 231–240. https://doi.org/10.1145/1341531.1341561

Universiti Malaya

Dutta, A. (2014). Parametric and nonparametric event study tests : a review. *International Business Research*, *7*(12), 136–142. https://doi.org/10.5539/ibr.v7n12p136

El-halees, A. M. (2007). Arabic text classification using maximum entropy. *The Islamic University Journal (Series of Natural Studies and Engineering)*, *15*(1), 157–167.

Elder, J., & Serletis, A. (2010). Oil price uncertainty. *Journal of Money, Credit and Banking*, *42*(6), 1137–1159. https://doi.org/10.1111/j.1538-4616.2010.00323.x

Elliott, A. ., & Woodward, W. . (2007). *Statistical analysis quick reference guidebook:With SPSS examples*. Sage.

Engelberg, J. E., & Parsons, C. A. (2011). The causal impact of media in financial markets. *The Journal of Finance*, *66*(1), 67–97.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A high-coverage lexical resource for opinion mining. *Evaluation*, *19*(9), 1–26. Retrieved from http://nmis.isti.cnr.it/sebastiani/Publications/2007TR02.pdf%0Ahttp://sentiwordnet.isti.cnr.it/

Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International Ecnomic Review*, *10*(1), 1–21.

Farid, D. M., Zhang, L., Rahman, C. M., Hossian, M. ., & Strachan, R. (2014). Hybrid decision tree and naïve bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, *41*, 1937–1946.

Gidófalvi, G. (2001). Using news articles to predict stock price movements. *Department of Computer Science and Engineering, University of California, San Diego*. https://doi.org/10.1111/j.1540-6261.1985.tb05004.x

Universiti Malaya

Girju, R., Giuglea, A.-M., Olteanu, M., Fortu, O., Bolohan, O., & Moldovan, D. (2004). Support Vector Machines applied to the classification of semantic relations in nominalized noun phrases. *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, (1), 68–75. Retrieved from http://dl.acm.org/citation.cfm?id=1596431.1596441

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(12), 1–6. https://doi.org/10.1016/j.sedgeo.2006.07.004

Groth, S. S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, *50*(4), 680–691. https://doi.org/10.1016/j.dss.2010.08.019

Guidi, M. G. D., Russell, A., & Tarbert, H. (2006). The effect of OPEC policy decisions on oil and stock prices. *OPEC Review*, *30*(1), 1–18. https://doi.org/10.1111/j.1468-0076.2006.00157.x

Gupta, K., & Banerjee, R. (2018). Does OPEC news sentiment influence stock returns of energy firms in the United States? *Energy Economics*, *77*, 34–45. https://doi.org/10.1016/j.eneco.2018.03.017

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, *55*(3), 685–697. https://doi.org/10.1016/j.dss.2013.02.006

Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2015). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. *Proceedings - 2014 6th*

Universiti Malaya

International Conference on Information Technology and Electrical Engineering: Leveraging Research and Technology Through University-Industry Collaboration, ICITEE 2014, 0–3. https://doi.org/10.1109/ICITEED.2014.7007894

Han, Z. (2012). *Data and Text Mining of Financial Markets Using News and Social Media*. *M.Sc. thesis*. University of Manchester.

Hanabusa, K. (2012). The effect of 107th OPEC Ordinary Meeting on oil prices and economic performances in Japan. *Renewable and Sustainable Energy Reviews*, *16*(3), 1666–1672. https://doi.org/10.1016/j.rser.2011.11.034

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, *7*(1), 37–53. https://doi.org/10.1007/s12021-008-9041-y

Hardeniya, N. (2015). *NLTK essentials*. Packt Publishing Ltd.

Harisinghaney, A., Dixit, A., Gupta, S., & Arora, A. (2014). Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm. *2014 International Conference on Reliability, Optimization and Information Technology*, 153–155. https://doi.org/10.1109/ICROIT.2014.6798302

Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, *2015*(1). https://doi.org/10.1155/2015/198363

Hirschberg, J., & Manning, C. D. (2015). Advances in Natural Language Processing. *Science (New York, N.Y.)*, *349(6245)*, 261–266. https://doi.org/10.1126/science.aaa8685

Universiti Malaya

Horan, S. M., Peterson, J. H., Mahar, J., Horan, S. M., Peterson, J. H., & Mahar, J. (2004). Implied volatility of oil futures options surrounding OPEC meetings. *The Energy Journal*, *25*(3), 103–125. Retrieved from http://www.jstor.org/stable/41323044

Hutto, C. ., & Gilbert, E. (2014). VADER : A parsimonious rule-based model for sentiment analysis of social media text. *In Eighth International AAAI Conference on Weblogs and Social Media.*

Ji, Q., & Guo, J. F. (2015). Oil price volatility and oil-related events: An Internet concern study perspective. *Applied Energy*, *137*, 256–264. https://doi.org/10.1016/j.apenergy.2014.10.002

Jiang, L., Li, C., Wang, S., & Zhang, L. (2016). Deep feature weighting for naive Bayes and its application to text classification. *Engineering Applications of Artificial Intelligence*, *52*, 26–39. https://doi.org/10.1016/j.engappai.2016.02.002

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kabir, F., Siddique, S., Kotwal, M. R. A., & Huda, M. N. (2015). Bangla text document categorization using stochastic gradient descent (SGD) classifier. *Proceedings - 2015 International Conference on Cognitive Computing and Information Processing, CCIP 2015*. https://doi.org/10.1109/CCIP.2015.7100687

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. *ArXiv Preprint ArXiv:1404.2188.* https://doi.org/10.3115/v1/P14-1062

Kaya, M. (2010). Stock price prediction using financial news articles. *Information and Financial Engineering ICIFE 2010 2nd IEEE International Conference On*, 478–

Universiti Malaya

482. https://doi.org/10.1109/ICIFE.2010.5609404

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *2014 Science and Information Conference.*, 372–378. https://doi.org/10.1109/SAI.2014.6918213

Ko, J., & Lee, C. (2015). International economic policy uncertainty and stock prices : *Economics Letters*, *134*, 118–122. https://doi.org/10.1016/j.econlet.2015.07.012

Kothari, S. P., & Warner, J. B. (1980). Measuring long-horizon security price performance. *Journal of Financial Economics*, *43*(3), 301–339. https://doi.org/10.1016/S0304-405X(96)00899-9

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering*, *3*(5), 1787–1797.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *In International Conference on Machine Learning*, *32*, 1188–1196. https://doi.org/10.1145/2740908.2742760

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2014). *IBM SPSS for intermediate statistics: Use and interpretation* (Vol. 53). https://doi.org/10.1017/CBO9781107415324.004

Li, J., Fong, S., Zhuang, Y., & Khoury, R. (2015). Hierarchical classification in text mining for sentiment analysis of online news. *Soft Computing*, *20*(9), 3411–3420.

Universiti Malaya

https://doi.org/10.1007/s00500-015-1812-4

Li, J., Xu, Z., Yu, L., & Tang, L. (2016). Forecasting oil price trends with sentiment of online news articles. *Procedia Computer Science*, *91*(71433001), 1081–1087. https://doi.org/10.1016/j.procs.2016.07.157

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, *278*, 826–840. https://doi.org/10.1016/j.ins.2014.03.096

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*(1), 14–23. https://doi.org/10.1016/j.knosys.2014.04.022

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Publishing Group*, *16*(6), 321–332. https://doi.org/10.1038/nrg3920

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. *2015 IEEE 14th Interational Conference on Cognitive Informatics & Cognitive Computing*, 136–140.

Lin, S. X., & Tamvakis, M. (2010). OPEC announcements and their effects on crude oil prices. *Energy Policy*, *38*(2), 1010–1016. https://doi.org/10.1016/j.enpol.2009.10.053

Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., & Dempsey, E. (2014). Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*, *3*.

Universiti Malaya

Loughran, T. I. M., & Mcdonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loutia, A., Mellios, C., & Andriosopoulos, K. (2016). Do OPEC announcements influence oil prices? *Energy Policy*, *90*, 262–272. https://doi.org/10.1016/j.enpol.2015.11.025

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, *16(1)*, 100–103.

McCusker, K., & Gunaydin, S. (2015). Research using qualitative, quantitative or mixed methods and choice based on the research, 537–542. Retrieved from https://doi.org/10.1177/0267659114559116

McDonald, D. M., Chen, H., & Schumaker, R. P. (2005). Transforming open-source documents to terror networks: The arizona terrornet. *AAAI Spring Symposium - Technical Report*, 62–69.

Mejova, Y. (2009). Sentiment analysis: An overview. *University of Iowa, Computer Science Department.* https://doi.org/10.1210/jc.2010-2284

Mensi, W., Hammoudeh, S., & Yoon, S. M. (2014). How do OPEC news and structural breaks impact returns and volatility in crude oil markets? Further evidence from a long memory process. *Energy Economics*, *42*, 343–354. https://doi.org/10.1016/j.eneco.2013.11.005

Meyer, D., & Wien, F. . (2015). Support vector machines. *The Interface to Libsvm in Package E1071*, 28.

Universiti Malaya

Mocormick, K., & Salcedo, J. (2017). *SPSS statistics for data analysis and visulization*. John Wiley & Sons.

Mukwazvure, A., & Supreethi, K. P. (2015). A hybrid approach to sentiment analysis of news comments. In *2015 4th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO* (pp. 1–6). IEEE. https://doi.org/10.1109/ICRITO.2015.7359282

Munot, N., & Govilkar, S. S. (2014). Comparative study of text summarization methods. *International Journal of Computer Applications*, *102*(12), 33–37. https://doi.org/10.5120/13115-0449

Narayan, S., & Narayan, P. K. (2017). Are oil price news headlines statistically and economically significant for investors? *Journal of Behavioral Finance*, *18*(3), 258–270. https://doi.org/10.1080/15427560.2017.1308942

Nasim, Z. (2018). NLG301 at SemEval-2017 task 5: fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 847–851). https://doi.org/10.18653/v1/s17-2144

Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. *2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013*. https://doi.org/10.1109/ICCCNT.2013.6726818

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

Universiti Malaya

Ontivero, M., Castellanos, A., Valente, G., & Goebel, R. (2017). Fast gaussian Naïve Bayes for searchlight classification analysis. *NeuroImage*, *163*, 471–479.

OPEC Secretariat. (2003). OPEC production agreements: a detailed listing. *OPEC Energy Review*, (March), 65–77.

Pallant, J. (2013). *SPSS survial manual*. McGraw-Hill Education(UK).

Pang, B., Lee, L., Rd, H., & Jose, S. (2002). Thumbs up ? Sentiment classification using machine learning techniques. *In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, *10*, 79–86.

Parikh, K. S., & Shah, T. P. (2016). Support vector machine – a large margin classifier to diagnose skin illnesses. *3rd International Conference on Innovations in Automation and Mechatronics Engineering*, *23*, 369–375. https://doi.org/10.1016/j.protcy.2016.03.039

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

Perikos, I., & Hatzilygeroudis, I. (2016). Recognizing emotions in text using ensemble of classifiers. *Engineering Applications of Artificial Intelligence*, *51*, 191–201.

Perkins, J. (2010). *Python text processing with NLTK 2.0 cookbook*. Packt Publishing Ltd.

Pham, M., Alse, S., Knoblock, C. A., & Szekely, P. (2016). Semantic labeling: A domain-independent approach. *The Semantic Web–ISWC 2016*, *9981*(PART 1), 33–48. https://doi.org/10.1007/978-3-319-46523-4

Phan, D. H. B., Sharma, S. S., & Narayan, P. K. (2015). Oil price and stock returns of

Universiti Malaya

consumers and producers of crude oil. *Journal of International Financial Markets, Institutions and Money*, *34*, 245–262. https://doi.org/10.1016/j.intfin.2014.11.010

Philipp, L., & Andre, K. S. (2016). Information processing in freight and freight forward markets: an event study on OPEC announcements. *Diskussionspapier*, *172*.

Plante, M. (2015). OPEC in the news: The effect of OPEC on oil price uncertainty. *Ssrn*, *75201*. https://doi.org/10.2139/ssrn.2627662

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of twitter sentiment on stock price returns. *PLoS ONE*, *10*(9), 1–21. https://doi.org/10.1371/journal.pone.0138441

Rao, Y., Lei, J., Wenyin, L., Li, Q., & Chen, M. (2014). Building emotional dictionary for sentiment analysis of online news. *World Wide Web*, *17*(4), 723–742. https://doi.org/10.1007/s11280-013-0221-9

Raschka, S. (2014). Naive bayes and text classification I: introduction and theory. *ArXiv Preprint ArXiv:1410.5329.*, 1–20.

Raschka, S. (2015). *Python machine learning*. Packt Publishing Ltd.

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. *Information Processing and Management*, *52*(1), 5–19. https://doi.org/10.1016/j.ipm.2015.01.005

Salwen, M. B., Garrison, B., & Discoll, P. D. (2004). *Online news and the public*.

Samprit, C., & Hadi, A. . (2015). *Regression analysis by example*. John Wiley & Sons.

Schaul, T., Bayer, J., Wierstra, D., & Sun, Y. (2010). PyBrain. *Journal of Machine*

Universiti Malaya

Learning Research, *11*, 743–746.

Schmidbauer, H., & Rösch, A. (2012). OPEC news announcements: Effects on oil price expectation and volatility. *Energy Economics*, *34*(5), 1656–1663. https://doi.org/10.1016/j.eneco.2012.01.006

Schumaker, R. P., & Chen, H. (2009a). A quantitative stock prediction system based on financial news. *Information Processing and Management*, *45*(5), 571–583. https://doi.org/10.1016/j.ipm.2009.05.001

Schumaker, R. P., & Chen, H. (2009b). Textual analysis of stock market prediction using breaking financial news. *ACM Transactions on Information Systems*, *27*(2), 1–19. https://doi.org/10.1145/1462198.1462204

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, *53*(3), 458–464. https://doi.org/10.1016/j.dss.2012.03.001

Scikit. (2011a). 1.5. Stochastic Gradient Descent. Retrieved from https://scikit-learn.org/stable/modules/sgd.html

Scikit. (2011b). 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#id1

Sekine, S., & Nobata, C. (2004). Definition , dictionaries and tagger for extended named entity hierarchy. *Lrec*, 1977–1980. https://doi.org/10.1.1.3.91

Seng, J.-L., & Yang, H.-F. (2017). The association between stock price volatility and

Universiti Malaya

financial news – a sentiment analysis approach. *Kybernetes*, *46*(8), 1341–1365. https://doi.org/10.1108/K-11-2016-0307

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., & Chanona-Hernández, L. (2014). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, *41*(3), 853–860. https://doi.org/10.1016/j.eswa.2013.08.015

Singh, P. (2019). Natural Language Processing. In *Machine Learning with PySpark* (pp. 191–218). https://doi.org/10.1007/978-1-4842-4131-8

Sinha, N. R. (2016). *Underreaction to news in the US stock market*. *Quarterly Journal of Finance* (Vol. 06). https://doi.org/10.1142/s2010139216500051

Sonnenburg, S., Rätsch, G., Henschel, S., Behr, J., Zien, A., Binder, A., … Franc, V. (2010). The SHOGUN machine learning toolbox. *Journal of Machine Learning Research*, *11*, 1799–1802. Retrieved from http://www.swig.org

Sorescu, A., Warren, N. L., & Ertekin, L. (2017). Event study methodology in the marketing literature: an overview. *Journal of the Academy of Marketing Science*, *45*(2), 186–207. https://doi.org/10.1007/s11747-017-0516-y

Soroka, S., Young, L., & Balmas, M. (2015). Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content. *Annals of the American Academy of Political and Social Science*, *659*(1), 108–121. https://doi.org/10.1177/0002716215569217

Spencer, S., & Bredin, D. (2019). Agreement matters: OPEC announcement effects on WTI term structure. *Energy Economics*, *80*, 589–609. https://doi.org/10.1016/j.eneco.2019.01.018

Universiti Malaya

SPSS lnc. (1990). *SPSS reference guide*. Spss.

StatisticsSolutions. (2019). Testing Assumptions of Linear Regression in SPSS. Retrieved October 12, 2019, from https://www.statisticssolutions.com/testing-assumptions-of-linear-regression-in-spss/

Stevens, J. . (2012). *Applied multivariate statistics for the social sciences*. Routledge.

Sun, A., Lim, E., & Liu, Y. (2009). On strategies for imbalanced text classi fi cation using SVM : A comparative study. *Decision Support Systems*, *48*(1), 191–201. https://doi.org/10.1016/j.dss.2009.07.011

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tan, S., Wang, Y., & Wu, G. (2011). Adapting centroid classifier for document categorization. *Expert Systems with Applications*, *38*(8), 10264–10273. https://doi.org/10.1016/j.eswa.2011.02.114

Tang, B., He, H., Baggenstoss, P. M., & Kay, S. (2016). A bayesian classification approach using class-specific features for text catergorization. *IEEE Transactions on Knowledge and Data Engineering*, *28*(6), 1602–1606. https://doi.org/10.1109/TKDE.2016.2522427

Tang, B., Member, S., Kay, S., He, H., & Member, S. (2016). Toward optimal feature selection in naive bayes for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, *28*(9), 2508–2521. https://doi.org/10.1109/TKDE.2016.2563436

Thode, H. C. (2002). *Testing for normality*. CRC press.

Tolle, K. M., & Chen, H. (2000). Comparing noun phrasing techniques for use with

Universiti Malaya

medical digital library tools. *Journal of the American Society for Information Science*, *51*(4), 352–370.

Tong, Z. (2004). Sloving large scale linear prediction problems using stochastic gradient descent algorithms. *Twenty-First International Conference on Machine Learning*, 116. https://doi.org/10.1145/1015330.1015332

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117–126. https://doi.org/10.1016/j.eswa.2016.03.028

Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based framework for text categorization. *Procedia Engineering*, *69*, 1356–1364. https://doi.org/10.1016/j.proeng.2014.03.129

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 417–424. https://doi.org/10.3115/1073083.1073153

Uysal, A. K. (2015). An improved global feature selection scheme for text classification. *Expert Systems With Applications*, *43*, 82–92. https://doi.org/10.1016/j.eswa.2015.08.050

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction.

Universiti Malaya

Simulation Modelling Practice and Theory, *55*, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining - an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Wang, S., Jiang, L., & Li, C. (2014). Adapting naive Bayes tree for text classification. *Knowledge and Information Syatem*, *44*(1), 77–89. https://doi.org/10.1007/s10115-014-0746-y

Wang, S., & Manning, C. D. (2012). Baselines and bigrams : simple , good sentiment and topic classification. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, *2*, 90–94.

Will, L., Benoit, K., Slava, M., & Laver, M. (2011). Scaling policy preferences from coded political texts. *Legislative Studies Quarterly*, *36*(1), 123–155. https://doi.org/10.1111/j.1939-9162.2010.00006.x

Wisniewski, T. P., & Lambe, B. (2013). The role of media in the credit crunch: The case of the banking sector. *Journal of Ecnomic Behavior and Organization*, *85*(1), 163–175. https://doi.org/10.1016/j.jebo.2011.10.012

Witten, I. H., Eibe, F., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

Xia, T., & Chai, Y. (2011). An improvement to TF-IDF: Term distribution based term

Universiti Malaya

weight algorithm. *Journal of Software*, *6*(3), 413–420. https://doi.org/10.4304/jsw.6.3.413-420

Xu, Y. S. (2014). Stock Price Forecasting Using Information from Yahoo Finance and Google Trend. *UC Brekley.*

Zhang, H. (2006). On the optimality of Naïve Bayes. *Pattern Recognition Letters*, *27*(7), 830–837. https://doi.org/10.1016/j.patrec.2005.12.001

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). *Combining lexicon-based and learning-based methods for twitter sentiment analysis*. *HP Laboratories, Technical Report*.

Zhang, W., & Gao, F. (2011). An improvement to naive bayes for text classification. *Procedia Engineering*, *15*, 2160–2164. https://doi.org/10.1016/j.proeng.2011.08.404

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *In Advances in Neural Information Processing Systems*, 649–657. https://doi.org/10.1093/ndt/gfv436

Zhang, Y., Dang, Y., Chen, H., Thurmond, M., & Larson, C. (2009). Automatic online news monitoring and classi fi cation for syndromic surveillance. *Decision Support Systems*, *47*(4), 508–517. https://doi.org/10.1016/j.dss.2009.04.016

Zhu, S., Ji, X., Xu, W., & Gong, Y. (2005). Multi-labelled classification using maximum entropy method. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 274–281.

Zhu, Y., & Cai, L. (2015). The challenges of data quality and data quality assessment in

Universiti Malaya

the big data era. *Data Science Journal*, *14*(2), 1–10. https://doi.org/http://doi.org/10.5334/dsj-2015-002

Zito, T., Wilbert, N., Wiskott, L., & Berkes, P. (2009). Modular toolkit for data processing (MDP): a Python data processing framework. *Frontiers in Neuroinformatics*, *2*(8), 1–7. https://doi.org/10.3389/neuro.11.008.2008

Universiti Malaya

Publication: Paper Presented and Published in e-Proceedings

A paper has been accepted and presented in ICIT 2019 conference, and published in the e-Proceedings of the conference:

Wu, L., & Ow, S. H. (2019). Financial News Sentiment Analysis Using Lexicon-Based Labelling and Machine Learning-Based Algorithm, *e-Proceeding of the 5th International Conference on Information Technology & Society (ICITS 2019): The Innovation of ICT & New Media*, August 20th, 2019, University of Malaya, Kuala Lumpur, pp. 46-54 (e-ISBN:978-967-2122-xx-x).

The first page of the published paper is given as reference below.

Universiti Malaya