UNIVERSITI PUTRA MALAYSIA

FEATURE EXTRACTION BASED ON WORD EMBEDDINGS AND OPINION LEXICALS FOR SENTIMENT ANALYSIS

EISSA MOHAMMED MOHSEN ALSHARI

FSKTM 2018 78

FEATURE EXTRACTION BASED ON WORD EMBEDDINGS AND OPINION LEXICALS FOR SENTIMENT ANALYSIS

By

EISSA MOHAMMED MOHSEN ALSHARI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Degree of Doctor of Philosophy

December 2018

DEDICATIONS

I would like to dedicate this thesis to my beloved motherland "Yemen".

&

To All whom I love.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

FEATURE EXTRACTION BASED ON WORD EMBEDDINGS AND OPINION LEXICALS FOR SENTIMENT ANALYSIS

By

EISSA MOHAMMED MOHSEN ALSHARI

December 2018

Chairman: Azreen Azman, PhD
Faculty: Computer Science and Information Technology

Sentiment Analysis has become one of the important researches in natural language processing due to the exponential increase of user reviews and comments online. The goal of sentiment analysis is to determine the polarity orientation of a review text to either positive or negative. Many techniques rely on generic opinion lexicons such as the SentiWordNet to construct features for the sentiment classification task. The lexicons consist of words with positive or negative polarity, and sometimes with assigned scores reflecting the degree of the sentiment polarity. The presence of the opinion lexicons in a text indicates the overall sentiment of the text. The lexical based sentiment analysis works by the summation of all polarity scores given by the opinion lexicons in the text to indicate its polarity, while feature vectors are constructed from the opinion lexicons and their scores to be used by the machine learning classifiers in the supervised learning task.

Firstly, in this context, the features to be used for classification are limited to only that opinion words presence in the text, while other non-opinion words in the text will be neglected (will be assigned zero values in the vector). It has become the limiting factor to the effectiveness of sentiment analysis. It is assumed that the collection of features should be enriched by including other non-opinion words in the text as features. In this thesis, the Dic2vec model is proposed to learn the polarity of non-opinion words based on the Word2vec. As such, the features for sentiment analysis are enriched by the combination of opinion words and non-opinion words.

Secondly, many feature extraction techniques have been proposed to alleviate the

data density and sparsity issue by mean of feature clustering. Such methods often result in the reduction of vector dimension and assign a more effective weighting scheme to improve the efficiency and effectiveness of sentiment analysis. One of the feature clustering methods used for sentiment analysis is based on computing semantic orientation of words in the labeled corpus and groups those words based on predefined ranges of semantic orientation scores. The score is measured based on the Pointwise Mutual Information (PMI) of words in the positive and negative reviews dataset. As a result, clusters of words are derived and used as features. The main disadvantage of this feature clustering method is that the strength in the polarity of words will be under represented in the vector. Two or more words with similar but high scores will only be represented by a binary value of 1, which is equals to any two or more words with similar but lower scores. As such, the effect of the significant words in the classification is diminished. In this thesis, the Senti2vec model is proposed to discover polarity clusters from the corpus to be used as features. The aim is to group non-opinion words around opinion words to produce more effective weighting scheme for the features in the sentiment analysis task.

Finally, the thesis focuses on the problem generating domain-dependent opinion lexicons through semi-supervised learning. It is based on the assumption that generic opinion lexicons such as the SentiWordNet is unable to capture the specific characteristics of the domain in order to discriminate among classes. The problem can be defined as assigning the polarity of target words based on a given set of opinion lexicons as the seed. The recent method proposed for this problem constructs a graph where nodes corresponds to subjective words and the edges reflect the similarity between those words. The similarity is measured by the co-occurrence of words pair within the same linguistic unit, such as an *n*-gram, phrase or sentence. Given that the polarity of a seed word is known, the polarity of target words is derived based on the strength of the edges between the seed word and the target word. It is argued that the Word2vec is much superior in representing the distributional semantics among words in a language. As such, in this thesis a semi-supervised learning method is proposed to learn the polarity of words from seeds opinion words by using the Word2vec.

All proposed methods and models in this thesis are evaluated by using a collection of movie reviews labeled dataset with 50,000 reviews. Based on the experiment, the performance of the Dic2vec model is about 2.5% to 6% better than the baseline. In addition, the Senti2vec model shows an improvement of up to 6.5% as compared to the baseline. Finally, the proposed semi-supervised method for learning opinion lexicons is better than the recent co-occurence graph method by more than 12%.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PENGEKSTRAKAN FITUR BERDASARKAN KEPADA PEMBENAMAN KATA DAN LEKSIKAL PENDAPAT UNTUK ANALISIS SENTIMEN

Oleh

EISSA MOHAMMED MOHSEN ALSHARI

Disember 2018

Pengerusi: Azreen Azman, PhD
Fakulti: Sains Komputer dan Teknolologi Maklumat

Analisis Sentimen telah menjadi salah satu penyelidikan penting dalam bidang pemprosesan bahasa tabii kerana peningkatan mendadak ulasan dan komen pengguna secara atas talian. Matlamat analisis sentimen adalah untuk menentukan orientasi kekutuban teks ulasan sama ada positif atau negatif. Kebanyakan teknik bergantung kepada leksikon pendapat yang generik seperti SentiWordNet untuk membina ciri dalam klasifikasi sentimen. Leksikon tersebut terdiri daripada kata-kata yang mempunyai kekutuban positif atau negatif, dan kadang-kadang berserta dengan skor yang diberikan untuk mencerminkan tahap kekutuban sentimen. Kehadiran leksikon pendapat dalam teks boleh menunjukkan sentimen untuk keseluruhan teks. Analisis sentimen berasaskan leksikal berfungsi dengan penjumlahan semua skor kekutuban yang diberikan oleh leksikon pendapat dalam teks untuk menunjukkan kekutubannya, sementara vektor ciri dibina dari leksikon pendapat dan markah mereka akan digunakan oleh kelas pembelajaran mesin dalam pembelajaran terselia.

Pertama, dalam konteks ini, ciri-ciri yang digunakan untuk klasifikasi adalah terhad kepada hanya kata-kata pendapat yang ada di dalam teks, sementara kata-kata bukan pendapat yang lain dalam teks akan diabaikan (akan diberi nilai kosong dalam vektor). Ia telah menjadi faktor yang membataskan keberkesanan analisis sentimen. Ia diandaikan bahawa koleksi ciri tersebut harus diperkaya dengan memasukkan kata-kata lain yang bukan pendapat dalam teks sebagai ciri-ciri. Dalam tesis ini, model Dic2vec dicadangkan untuk mempelajari kekutuban perkataan yang bukan pendapat berdasarkan kepada Word2vec. Oleh itu, ciri-ciri untuk analisis sentimen diperkaya dengan gabungan perkataan pendapat dan kata-kata bukan pendapat.

Kedua, banyak teknik pengekstrakan ciri telah dicadangkan untuk mengurangkan ketumpatan data dan masalah kejarangan melalui pengelompokan ciri-ciri. Kaedah sedemikian sering mengakibatkan pengurangan dimensi vektor dan memberikan skema pemberat yang lebih berkesan untuk meningkatkan kecekapan dan keberkesanan analisis sentimen. Salah satu kaedah pengelompokan ciri-ciri yang digunakan untuk analisis sentimen adalah berdasarkan pengiraan orientasi semantik kata-kata dalam korpus berlabel dan mengelompokkan kata-kata itu berdasarkan julat yang telah ditentukan sebelumnya daripada skor orientasi semantik. Skor diukur berdasarkan pada Pointwise Mutual Information (PMI) perkataan dalam dataset ulasan positif dan negatif. Akibatnya, kelompok perkataan diperoleh dan digunakan sebagai ciri-ciri. Kekurangan utama kaedah pengelompokkan ini ialah kekuatan dalam kekutuban kata-kata akan kurang diwakili dalam vektor. Dua atau lebih perkataan dengan markah yang serupa dan tinggi hanya akan diwakili oleh nilai binari 1, yang sama dengan mana-mana dua atau lebih perkataan dengan markah yang serupa tetapi rendah. Oleh itu, kesan kepada perkataan penting dalam klasifikasi itu akan berkurangan. Dalam tesis ini, model Senti2vec dicadangkan untuk mencari kluster kutub dari korpus untuk digunakan sebagai ciri-ciri. Matlamatnya adalah untuk mengelompokkan kata-kata yang bukan kata pendapat di sekitar kata-kata pendapat untuk menghasilkan skema pemberat yang lebih berkesan untuk ciri-ciri dalam analisis sentimen.

Akhirnya, tesis ini memberi tumpuan kepada masalah leksikon yang bergantung kepada domain melalui pembelajaran separuh diselia. Ia berasaskan kepada anggapan bahawa leksikon pendapat yang generik seperti SentiWordNet tidak dapat menangkap ciri-ciri khusus domain dalam mendiskriminasi antara kelas. Masalahnya boleh ditakrifkan sebagai mengenalpasti kutub sentimen bagi kata sasaran berdasarkan satu set pendapat leksikon sebagai benih. Kaedah terbaru yang dicadangkan untuk masalah ini ialah membina graf di mana nod bersesuaian dengan kata-kata subjektif dan keseluruhan mencerminkan kesamaan antara kata-kata itu. Kesamaan diukur oleh terjadinya pasangan perkataan dalam unit linguistik yang sama, seperti textit n -gram, frasa atau ayat. Memandangkan kekutuban perkataan asas diketahui, kekutuban kata sasaran diperolehi berdasarkan kekuatan keseluruhan di antara perkataan benih dan kata sasaran. Dikatakan bahawa Word2vec lebih unggul dalam perwakilan taburan semantik di antara kata-kata dalam sesuatu bahasa. Oleh itu, dalam tesis ini, kaedah pembelajaran yang separuh diselia dicadangkan untuk mempelajari kutub kata-kata dari kata-kata pendapat benih dengan menggunakan Word2vec.

Semua kaedah dan model yang dicadangkan dalam tesis ini dinilai dengan menggunakan koleksi ulasan filem dilabelkan dengan 50,000 ulasan. Berdasarkan eksperimen, prestasi model Dic2vec adalah lebih kurang 2.5 % hingga 6 % lebih baik daripada piawai. Di samping itu, model Senti2vec menunjukkan peningkatan sehingga 6.5% berbanding dengan piawai. Akhir sekali, cadangan kaedah separuh diselia untuk pembelajaran leksikon pendapat adalah lebih baik daripada kaedah geraf kewujudan terbaharu dengan lebih daripada 12%.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor Dr. Azreen Azman for the continuous support of my study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. His encouragement and help made me feel confident to overcome every difficulty I encountered in all the stages of this research. What I really learned from him, however, is his attitude to work and life - always aiming for excellence.

I would like to extend my gratitude and thanks to the distinguished committee member, Associated Professor Dr. Shyamala Doraisamy , and Associated Professor Dr. Norwati Mustapha for their encouragement and insightful comments.

I am very grateful to the Faculty of Computer Science and Information Technology and the staff of Postgraduate office, School of Graduate Studies, Library and Universiti Putra Malaysia, for providing me excellent research environment. Thanks to every person who has supported me to pursue and finish my Ph.D.

I am very grateful to my family my father my mother my brothers and my sisters for their unflagging love and support throughout my life. I have no suitable words that can fully describe my everlasting love to them except, I love you all.

Words fail me to express my appreciation to my lovely wife whose dedication, love and persistent confidence in me, has taken the load off my shoulder. I owe her for being unselfishly let her intelligence, passions, and ambitions collide with mine. Special thank goes to my sons and you are my joy and my guiding lights. Thanks for giving me your valuable time through all this long process. I promise I will never let you alone any more.

Last but not least, it gives me immense pleasure to express my deepest gratitude to my friends, colleagues and lab mates, especially Mostafa Alksher for their unlimited support and encouragement.

Finally, I would like to thank everybody who was important to the successful realization of this thesis, as well as I express my apology that I could not mention you all personally.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Azreen Azman, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Shyamala A/p C Doraisamy, PhD
Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(Member)

Norwati Mustapha, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

ROBIAH BINTI YUNUS, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: ______________________ Date: __________

Name and Matric No.: Eissa Mohammed Mohsen Alshari (GS42248)

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ____________________
Name of Chairman
of Supervisory
Committee: Assoc. Prof. Dr. Azreen Azman

Signature: ____________________
Name of Member
of Supervisory
Committee: Assoc. Prof. Dr. Shyamala A/p C Doraisamy

Signature: ____________________
Name of Member
of Supervisory
Committee: Assoc. Prof. Dr. Norwati Mustapha

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial intelligence
BoW	Bag-of-Words
CNET	Micropinion Generation Dataset
CNN	Conventional Neural Network
COW	Continuous Bag of Word
df	Document Frequency
Dic2vec	Proposed Dictionary to vector
Doc2vec	Document Vector on word embeddings
DOR	Document Occurrence Representation
FE	Feature Extraction
idf	Inverse Document Frequency
IG	Information Gathering
IMDB	Internet Movie Database
LDA	Latent Dirichlet Analysis
LR	Logistic Regression
LSA	Latent Semantic Analysis
ML	Machine Learning
MM	Markov Model
NB	Naive Bayes
NLP	Nature language Processing
PV-DBOW	Word Version of Paragraph Vector
RBF	Radial Basis Function
RNN	Recurrent Neural Network
SA	Sentiment Analysis
SOA	Second Order Attributes
SVM	Support Vector Machine
SWN	SentiWordNet
tf	Term Frequency
tf_idf	Term Frequency-Inverse Document Frequency
VSM	Vector Space Model

CHAPTER 1

INTRODUCTION

1.1 Background

The number of users of on-line shopping websites and the social media, (e.g., reviewers, tweeters, commenter) is continuously increasing. Such website usually provides facility for the users to give comments and ratings to the products being sold on the websites. The textual information can be useful as a recommendation for other users in making their purchase decision. These textual information are divided into two types: opinions and facts. Opinion statements are subjective in nature and mostly describe the people's sentiments around events or entities. Most of the current research have been emphasized on the factual data in various natural language processing tasks (NLP), e.g., information retrieval, text classification, etc. Research on sentiment analysis from sentences is still limited due to a huge number of challenges involved in the field (Turney, 2002; Cambria et al., 2010; Liu, 2012; Cambria et al., 2013). Several websites often provide facilities for the users to give comments and ratings to the products being sold on these websites. Therefore, understanding polarity orientation of text (positive or negative) through opinion mining techniques gives new chances for organizations to determine their wise development strategy (Berners-Lee et al., 2001).

Sentiment Analysis (SA) is a collection of opinion mining methods for analyzing people's opinion and sentiment (usually a sequence of words in a text) towards entities such as products. Sentiment analysis has three categories; document-level (Turney, 2002; Maas et al., 2011a; Lau and Baldwin, 2016), sentence-level (Meena and Prabhakar, 2007; Socher et al., 2011), and aspect-level (Mudinas et al., 2012; Agarwal et al., 2015c). The document-level SA considers a document as a single unit (such as review, document, or comment) and classifies it based on whether it has positive or negative sentiment polarity.

The sentence-level in SA processes a sentence to extract the opinion expressed in that sentence. Both the document and the sentence levels do not exactly detect what people liked and did not like in the entities. The aspect-level or feature-level identifies the people's opinion of entities described in the text (Hu and Liu, 2004a; Liu, 2012). The development of techniques for the document-level sentiment analysis is one of the significant components of this area and there is several research available in literature for detecting sentiment from the document text (Abbasi et al., 2008; Liu, 2012; Kaji and Kitsuregawa, 2007).

In this research, new methods are proposed to extract features from the unstructured text that can include semantic, and common sense knowledge from the embeddings of words. Techniques employed by sentiment analysis models can be broadly categorized into lexicon (Kaji and Kitsuregawa, 2007), semantic orientation (Dai et al., 2011; Turney and Littman, 2003), machine learning (Pang and Lee, 2008) approaches.

Semantic orientation and lexicon approaches detect the polarity of the words on the basis of the corpus or opinion dictionary, whereas Machine learning (ML) model requires large training dataset. There are three steps to construct the semantic orientation-based approach. At first, the features that contain rich opinion of the users are extracted from the untrusted text; for instance, 'good movie' expresses a positive orientation.

Further, semantic polarity orientation of (non-opinion) from rich opinion features are determined as a dictionary based, corpus based or word embeddings. At long last, the overall polarity of the document or comment is computed by summation the polarity of the feature. The polarity computed in two types in the semantic orientation based approach; (i) opinion lexicon or knowledge based (Dai et al., 2011) and (ii) corpus based.

In the dictionary-based (lexicon-based) or Knowledge-based approaches, polarity value is determined based on utilizing the pre-developed polarity lexicons, such as Bing lie (Hu and Liu, 2004a), SentiWordNet Baccianella et al. (2010), WordNet (Miller, 1995), Sentiment 140 (Go et al., 2016), (Das and Chen, 2007) and etc. Whereas, corpus-based approaches compute the polarity based on the co-occurrences of the term with other negative or positive seed words in the corpus.

The main inspiration behind this approach is that the semantic orientation of any feature is said to be negative if it has association with negative seed words (e.g., bad). Also, it is said to be positive semantic orientation in the event that it has relationship with positive seed words (e.g., good and excellent).

On the other hand, machine learning model is as follows; Initially, the keywords may not carry precise sentiment of the user and thus the intelligent features are extracted from the document that can incorporate the syntactic, semantic, knowledge and sentiment. Next, fitting weighting schemes are required to offer weight to features according to their importance. Further, an effective feature selection technique is required to extract only the important features (feature extraction) by removing the irrelevant features for better classification results. Finally, a significant machine learning method is required for the classification.

Feature extraction (FE) in SA is an emergent research field. This research is concentrated on related work performed in this area to investigate and address some issues of feature extraction on sentiment analysis. FE is facing several issues such as redundancy, large feature space problems, domain dependency, limited work on Lexicon-structural features, difficulty in implicit feature identification and word embeddings. The general challenges in FE, identified by (Beijing 2010, Zhang 2011, Abbasi 2011, liu 2015), are discussed as follows:

- Redundancy, such as N-grams, that are highly redundant causing redundancy problems in both multivariate methods and univariate. Therefore, ability of hybrid methods to overcome problems arising from redundancy needs further experiments (Joshi and Penstein-Rosé, 2009).
- Large feature sets (High dimensionality) causes performance retrogression due to computational problems, therefore selection the essential features are required.
- Domain dependency, performance of clustering based FE techniques is domain dependent, generalization problems and creating cross domain.
- Unlike semantic and syntactic features, limited work is carried out on lexicon structural features in feature extraction algorithms.

Consequently, there are several directions to overcome the feature extraction issues such as refine the lexicon and extend the sentiment feature-extraction procedure. Further, there are several techniques to represent the features for SA, as such the BoW , N-gram models and etc.. Mikolov et al. (2014a); Villegas et al. (2016a).

BoW is an approach that models text numerically in many text mining and information retrieval tasks. Several weighting schemes have been successfully used in the BoW such as the n-gram, Boolean, term co-occurrence and tf-idf. The SA is often based on deep learning and machine learning technologies that have been significantly developed and acquired widespread attention since 2010.

The high-performance computing and cloud computing expedite the development of word embedding technologies which are more easily be adopted in practical applications. The features that are used in the classification of text play an important role in polarity classification success. On the other hand, feature extraction methods can be divided to either discrete distribution (applicable to the scenarios where the set of possible outcomes is discrete, such as a roll of dice or a coin toss) like Document Occurrence Representation (DOR) (Lavelli et al., 2004) , Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997), Latent Dirichlet Analysis (LDA) (Hoffman et al., 2010), Second Order Attributes (SOA) (López-Monroy et al., 2013) and BoW (Le and Mikolov, 2014; Villegas et al.) or continuous distribution (applicable to the scenarios where the set of possible outcomes can take on values in a continuous range (real numbers), such as the temperature on a given day) like Word2vec (Mikolov et al. 2013d), Doc2Vec (Le and Mikolov, 2014), Glove (Pennington et al., 2014) and other neural network techniques.

In this research, a novel approach by incorporating semantic and word embeddings is proposed for sentiment analysis. The proposed concept extraction approach exploits the relationship between words; it obtains the semantic relationship between words based on word embeddings such as word2vec. The importance of word embeddings from domain distribution for the sentiment analysis model is investigated. Hence, various feature extraction techniques were employed to mine the prominent features for machine learning model.

1.2 Problems Statement

The problem of SA is to determine the polarity orientation of a text to either positive, negative or neutral (Hu and Liu, 2004a; Li et al., 2014; Chen et al., 2013). Over the years, researchers have developed different techniques for SA to classify the reviews or comments into their polarity classes (Graves et al., 2005; Lui and Croft, 2003; Shojaee et al., 2013; Mikolov et al., 2014a; Yazdani et al., 2017). Classification methods have been effective in text classification, where a corpus contains a lot of documents which are converted into document matrix as a numerical or binary vector related to the occurrence of its documents (Liu and Croft, 2005). The document matrix achieves good classification performances by assigning larger weights to discriminative feature and smaller weights to non-discriminative ones during training. However, it still has a weakness to classify the SA polarity orientation because the common unsupervised term weighting used in the document matrix is based on the co-occurrence models such as Term Frequency (tf), Term Frequency-inverse Document Frequency (tf_idf) and *n*-gram which is unable to draw a significant semantic weighting scheme of non-opinion words in the corpus (Sebastiani, 2002; Agarwal et al., 2015d; Yazdani et al., 2017).

Many SA techniques rely on generic opinion lexicons such as the SentiWordNet to construct features for SA task. Opinion lexicons consist of words with positive or negative polarity, and sometimes with assigned scores reflecting the degree of the sentiment polarity. In general, it is assumed that the presence of the opinion lexicons in a text indicates the overall sentiment of the text. Therefore, in the lexical based SA, the polarity scores given by the opinion lexicons in the text is added to indicate the overall polarity of the text. On the other hand, in the classification based SA, feature vectors are constructed from the opinion lexicons and in some cases the vectors are weighted based on the polarity scores. In general, there are several issues are investigated in this work as the following:

The first problem can be summarized as depicted in Figure 1.1. The features to be

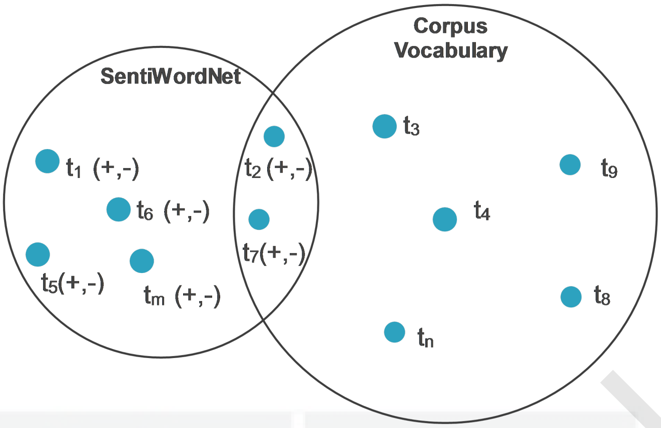

Figure 1.1 : The problem of combining vocabulary size and lexical size

used for classification are limited to only those words in the intersection between the *corpus vocabulary* set and *SentiWordNet* set. In addition, the SentiWordNet and other opinion lexicon does not include all terms in the corpus vocabulary. The terms that will be included as the features for sentiment classification reside within the intersection of the two sets. As such, this can be the limitation to the performance of any SA method.

The second problem investigates the effectiveness of feature extraction technique for SA. Many feature extraction techniques have been proposed to alleviate the data density and sparsity issue by mean of feature clustering. Here, supervised machine learning technique is capable of extracting semantic information among data to form clusters, which later used as feature vector. There are several clustering methods investigated to group the closest semantic features together (Agarwal and Mittal, 2014; Lin et al., 2014; Andrews and Fox, 2007). Such methods often result in the reduction of vector dimension and assign a more effective weighting scheme to improve the efficiency and effectiveness of SA method. One of the method for clustering features for SA is proposed by Agarwal and Mittal (2014). The method calculates semantic orientation of words in the labeled corpus and groups those words based on predefined ranges of semantic orientation scores. The semantic orientation score is measured based on the Pointwise Mutual Information (PMI) of words in the positive and negative reviews dataset. As a result, clusters of words are derived and used as features. Thus, the dimension of the feature vector is based to the number of clusters produced. The method used binary weighting scheme for the vector. The main disadvantage of this feature clustering method is that the strength in the polarity of words will be under represented in the vector. Two or more words with similar but high scores will only be represented by a binary value of 1, which is equals to any two or more words with similar but low scores. As such, the effect of the significant words in the classification is diminished.

The use of sentiment lexicons as features for sentiment analysis is based on the notion that the presence of those words in a text will give an indication of overall sentiment of the text. If there are many positive words appear in the text, it will indicate the positive sentiment of the text and vice versa. In machine learning approaches of sentiment analysis, the lexicons can be effective in discriminating text between positive and negative classes. However, many research have discovered that the performance of sentiment analysis model is better when all words are used as features as compared to only using the sentiment lexicons. Such discovery may be due to the fact that generic sentiment lexicon is unable to capture domain-dependent characteristics of the collection, such as movie reviews or tweets.

The third problem focuses on the problem of semi-supervised learning of opinion lexicons. It can be defined as the problem of assigning the polarity of target words based on a given set of opinion lexicons as the seed. In (Hatzivassiloglou and McKeown, 1997), the authors suggested that two adjectives conjoined by the word *and* should have the same sentiment polarity while conjoined by the word *but* should have different polarity. Based on that assumption, the polarity of target words can be derived by measuring the co-occurrence of two words conjoined by *and* or *but* within a corpus. In Turney algorithm (Turney, 2002), the polarity of a phrase is learnt by measuring its co-occurrence with a given opinion lexicon as the seed by using the Pointwise Mutual Information (PIM) score. In addition, the synonyms and antonyms relationship between a seed opinion lexicon and target words derived from the Wordnet have been used to infer the polarity of the words (Hu and Liu, 2004b; Kim and Lee, 2014). In Khan et al. (2016) proposed a semi-supervised method to learn the weight of features for sentiment analysis based on SentiWordNet.

More recently, Kim (Kim, 2018) constructed a graph where nodes corresponds to subjective words and the edges reflect the similarity between those words. The similarity is measured by the co-occurrence of words pair within the same linguistic unit, such as an *n*-gram, phrase or sentence. Given that the polarity of a seed word is known, the polarity of target words is derived based on the strength of the edges between the seed word and the target word. As such, the existing work on the semi-supervised learning of opinion lexicons are based on the co-occurrences of seed sentiment word with target words.In (Mikolov et al., 2013c), the authors have discovered that the Word2Vec is more effective representation of words in a continuous space to capture distributional semantics among words in a language. As such, a semi-supervised learning of opinion lexicons should take into consideration a more robust word embeddings techniques such as the Word2vec.

1.3 Research Objectives

The goal of this work is to build a lexical word embeddings model that allows flexible context analysis and generates the features from the text polarity. The specific objectives of this study were:

- To enrich features for sentiment analysis by learning the polarity of non-opinion words based on modeling distributional semantic of Word2vec.
- To propose an effective feature extraction method based on discovering polarity clusters by using the Word2vec.
- To develop effective semi-supervised learning of opinion lexicons from corpus based on Word2vec.

1.4 Research Contributions

The main contributions of this thesis are proposed to use Word2vec and sentiment lexical for feature extraction as following:

- The Dic2vec model is proposed by extracting the Word2vec features based on sentiment lexical to expand the BoWs representation model by used a significant non-opinion words as features. In Addition, Dic2vec is based on the assumption that the polarity of any words in the vocabulary can be learned from combining the terms vector in the SentiWordNet and word embeddings.
- The model, that is named as Senti2vec is based on the assumption that the dimensionality of the document matrix is decreased by selecting the centroid (best representative) of clustering which calculated from opinion lexicon distance and Word2vec distribution rather than all terms in the vocabulary.
- The internally semi-supervised lexical is developed from the labeled dataset instead of using external lexicon. As a result, the number of opinion words is increased which achieves a good polarity classification for Sentiment Analysis.

Consequently, the achieved results of our approaches are significant better than the other state-of-the-art Sentiment Analysis approaches.

1.5 Research Scope

There are three parts in the development of SA model (data collection and cleaning, the feature extracting, training model and evaluation). In this thesis, the focus on the feature extraction part because it is the most important and has a lot of challenges. This work aims to design the effective approaches to extract the features that will be learned from the word embeddings with a lexicon of sentiment words for the weighting scheme in documents matrix. The other tasks data collection, training and evaluation are involved to measure the performance of this research methods. For feature extraction, several scenarios of minimize and optimize the weight of feature are discussed in details chapter 4 to 6. Also, the different classification methods will be compared to measure the performance and calculate the accuracy.

1.6 Organization of the Thesis

This thesis consists of seven chapters and the details description of the sentiment analysis, related works, frameworks, and contributions experiments and analysis are presented for each chapter as follows:

Chapter 1 presents the introductory overview of the sentiment analysis and its approaches with brief of limits and drawbacks, problem statement, objectives, scope and contributions of this work.

Chapter 2 highlights and investigation of the research gap and motivation by giving overview of related work to position this work therein. The important concepts in sentiment analysis, feature extraction and word embeddings that complete the understanding of polarity extraction are reviewed in this chapter

Chapter 3 describes the composition of the sentiment analysis framework and the overall methodology details. In addition, all stage in the SA framework is discussed with the NLP processing and word embeddings techniques. Finally, the software, hardware and methodology are descried in this chapter.

Chapter 4 introduces the proposed Dic2vec model that learn the polarity of words in the vocabulary through measuring semantic relations between opinion words and non-opinion words based on the Word2vec and opinion lexical dictionary. The lexicons produced by the model is later combined with the existing opinion lexicons from dictionary to construct the set of features for sentiment analysis. In addition, the effectiveness of the proposed approach is evaluated using the IMDB Movie Review dataset (Maas et al., 2011b) with the SentiWordNet (Baccianella et al., 2010) as an opinion lexicon.

Chapter 5 proposes a feature extraction method based on modeling polarity clusters within the Word2vec vectors in order to improve the effectiveness of SA. It is assumed that each word in the vocabulary has its polarity alignment and will produce a better representation of text for SA. The method proposed in this study consists of three main components, which are; the learning of word embeddings based on Word2vec, the discovery of polarity clusters based on opinion lexical dictionary, and the construction of features matrix for classification based on cluster best representation.

Chapter 6 investigates and experiments the problem of semi-supervised learning of opinion lexicons that can be defined as the problem of assigning the polarity of target words based on a given set of opinion lexicons as the seed. Based on that assumption, the polarity of target words can be derived by measuring the co-occurrence of words on the Word2vec distribution and the distance of the seeds word in the labeled dataset. In addition, this chapter explains adaptive semi-supervised internal lexical.

Chapter 7 presents the main conclusions of the thesis and highlights future research work in the related areas. In this study, the interested issues of semantic relation between terms in the dataset to extract and predict the polarity of text for Sentiment Analysis were investigated.

REFERENCES

Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12.

Agarwal, B. and Mittal, N. (2014). Semantic feature clustering for sentiment analysis of english reviews. *IETE Journal of Research*, 60(6):414–422.

Agarwal, B., Mittal, N., Bansal, P., and Garg, S. (2015a). Sentiment Analysis Using Common-Sense and Context Information. *Computational Intelligence and Neuroscience*, 2015:1–9.

Agarwal, B., Mittal, N., Bansal, P., and Garg, S. (2015b). Sentiment analysis using common-sense and context information. *Computational intelligence and neuroscience*, 2015.

Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., and Hussain, A. (2015c). Concept-level sentiment analysis with dependency-based semantic parsing: A novel approach. *Cognitive Computation*, pp. 1–13.

Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., and Hussain, A. (2015d). Concept-level sentiment analysis with dependency-based semantic parsing: A novel approach. *Cognitive Computation*, pp. 1–13.

Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., and Hussain, A. (2015e). Concept-Level Sentiment Analysis with Dependency-Based Semantic Parsing: A Novel Approach. *Cognitive Computation*, 7(4):487–499.

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

Aggarwal, C. C. and Zhai, C. (2012b). A survey of text clustering algorithms. In *Mining text data*, pp. 77–128. Springer.

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, pp. 623–635.

Alksher, M. A., Azman, A., Yaakob, R., Kadir, R. A., Mohamed, A., and Alshari, E. M. (2016). A review of methods for mining idea from text. In *Information Retrieval and Knowledge Management (CAMP), 2016 Third International Conference on*, pp. 88–93. IEEE.

Alshari, E. M., Azman, A., Doraisamy, S., Mustapha, N., and Alkeshr, M. (2017). Improvement of sentiment analysis based on clustering of word2vec features. In *Database and Expert Systems Applications (DEXA), 2017 28th International Workshop on*, pp. 123–126. IEEE.

Alvim, L., Vilela, P., Motta, E., and Milidiú, R. L. (2010). Sentiment of financial news: a natural language processing approach. In *1st Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology, Buenos Aires*, p. 16.

Andrews, N. O. and Fox, E. A. (2007). Recent developments in document clustering.

Audi, R. (1999). The cambridge dictionary of philosophy.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pp. 2200–2204.

Bailey, P., Moffat, A., and Thomas, P. (2015). User Variability and IR System Evaluation.

Bansal, M., Gimpel, K., and Livescu, K. (2014). Tailoring continuous word representations for dependency parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Beineke, P., Hastie, T., and Vaithyanathan, S. (2004). The sentimental factor: Improving review classification via human-provided information. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, p. 263. Association for Computational Linguistics.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, 3:1137–1155.

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., and Reynar, J. (2008). Building a sentiment summarizer for local service reviews. In *WWW workshop on NLP in the information explosion era*, volume 14, pp. 339–348.

Blitzer, J., Dredze, M., and Pereira, F. (2009). Multi-domain sentiment dataset (version 2.0).

Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A. (2017a). Affective computing and sentiment analysis. In *A Practical Guide to Sentiment Analysis*, pp. 1–10. Springer.

Cambria, E., Poria, S., Gelbukh, A., and Thelwall, M. (2017b). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*, volume 10.

Chen, Y., Perozzi, B., Al-Rfou, R., and Skiena, S. (2013). The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226*, 28:2–9.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Da, Z., Engelberg, J., and Gao, P. (2015). The sum of all fears investor sentiment and asset prices. *Review of Financial Studies*, 28(1):1–32.

Dai, L., Chen, H., and Li, X. (2011). Improving sentiment classification using feature highlighting and feature bagging. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 61–66. IEEE.

Danielsson, P.-E. (1980). Euclidean distance mapping. *Computer Graphics and image processing*, 14(3):227–248.

Das, S. R. and Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management science*, 53(9):1375–1388.

Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pp. 519–528. ACM.

Deng, Z.-H., Luo, K.-H., and Yu, H.-L. (2014). A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications*, 41(7):3506–3513.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pp. 1370–1380.

Dey, L. and Haque, S. M. (2009). Opinion mining from noisy text data. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3):205–226.

Duwairi, R., Ahmed, N. A., and Al-Rifai, S. Y. (2015). Detecting sentiment embedded in arabic social media–a lexicon-based approach. *Journal of Intelligent & Fuzzy Systems*, 29(1):107–117.

Elabd, E., Alshari, E., and Abdulkader, H. (2015). Semantic boolean arabic information retrieval. *International Arab Journal of Information Technology (IAJIT)*, 12(3).

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Emad, E., Alshari, E. M., and Abdulkader, H. (2013). Arabic vector space model based on semantic. In *International journal of computer science (IJISI),*, volume 8, pp. 94–101. Ain Shams.

Emery, F. and Trist, E. (1973). The early detection of emergent processes. In *Towards A Social Ecology*, pp. 24–37. Springer.

Eneko Agirre, Oier López de Lacalle, A. S. (2014). Random walks for knowledge-based word sense disambiguation. *Dissertation Abstracts International, B: Sciences and Engineering*, 70(8):4943.

Enríquez, F., Troyano, J. A., and López-Solaz, T. (2016). An approach to the use of word embeddings in an opinion classification task. *Expert Systems with Applications*, 66:1–6.

Esuli, A. and Sebastiani, F. (2007). Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation*, pp. 1–26.

Fan, X., Li, X. X., Du, F., and Li, X. X. (2016). Apply Word Vectors for Sentiment Analysis of APP Reviews. (Icsai):1062–1066.

Filatova, E. (2017). Sarcasm detection using sentiment flow shifts.

Gamon, M. and Aue, A. (2005). Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pp. 57–64. Association for Computational Linguistics.

Ganesan, K. (2010). Opinrank review dataset.

Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 340–348.

Ganesan, K., Zhai, C., and Viegas, E. (2012). Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st international conference on World Wide Web*, pp. 869–878. ACM.

Ganu, G., Elhadad, N., and Marian, A. (2009). Beyond the stars: improving rating predictions using review text content. In *WebDB*, volume 9, pp. 1–6. Citeseer.

Giatsoglou, M., Vozalis, M. G., Diamantaras, K., Vakali, A., Sarigiannidis, G., and Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224.

Gionis, A., Indyk, P., Motwani, R., et al. (1999). Similarity search in high dimensions via hashing. In *Vldb*, volume 99, pp. 518–529.

Go, A., Bhayani, R., and Huang, L. (2016). Sentiment140. *Site Functionality, 2013c. URL http://help. sentiment140. com/site-functionality. Abruf am*, 20.

Goldberg, Y. and Levy, O. (2014). word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. *arXiv preprint arXiv:1402.3722*, (2):1–5.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Goyal, A. and Daumé III, H. (2011). Generating semantic orientation lexicon using large data and thesaurus. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 37–43. Association for Computational Linguistics.

Graves, A., Fernández, S., and Schmidhuber, J. (2005). Bidirectional lstm networks for improved phoneme classification and recognition. In *International Conference on Artificial Neural Networks*, pp. 799–804. Springer.

Haddi, E., Liu, X., and Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science*, 17:26–32.

Hamdan, H., Bellot, P., and Bechet, F. (2015). lsislif: Feature extraction and label weighting for sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 568–573.

Harper, F. M. and Konstan, J. A. (2016). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):19.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pp. 174–181. Association for Computational Linguistics.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pp. 856–864.

Hotho, A., Nürnberger, A., and Paaß, G. (2005). A brief survey of text mining. In *Ldv Forum*, volume 20, pp. 19–62.

Hu, M. and Liu, B. (2004a). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*, 04:168.

Hu, M. and Liu, B. (2004b). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177. ACM.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012a). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 873–882. Association for Computational Linguistics.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012b). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pp. 873–882. Association for Computational Linguistics.

Huang, M., Qian, Q., and Zhu, X. (2017). Encoding syntactic knowledge in neural networks for sentiment classification. *ACM Transactions on Information Systems*, 35(3). cited By 0.

Iskandar, B. S. (2017). Terrorism detection based on sentiment analysis using machine learning. *Journal of Engineering and Applied Sciences*, 12(3):691–698.

Jakkula, V. (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37.

Joshi, M. and Penstein-Rosé, C. (2009). Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pp. 313–316.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Joulin, A. and Mikolov, T. (2015a). Inferring algorithmic patterns with stack-augmented recurrent nets. *arXiv preprint arXiv:1503.01007*.

Joulin, A. and Mikolov, T. (2015b). Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets. pp. 1–10.

Jun, S., Park, S.-S., and Jang, D.-S. (2014). Document clustering method using dimension reduction and support vector clustering to overcome sparseness. *Expert Systems with Applications*, 41(7):3204–3212.

Kaji, N. and Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Karatzoglou, A., Meyer, D., and Hornik, K. (2005). Support vector machines in r.

Khan, F. H., Qamar, U., and Bashir, S. (2016). Swims: Semi-supervised subjective feature weighting and intelligent model selection for sentiment analysis. *Knowledge-Based Systems*, 100:97–111.

Kim, J., Yoo, J., Lim, H., Qiu, H., Kozareva, Z., and Galstyan, A. (2013). Sentiment Prediction using Collaborative Filtering. *Icwsm*.

Kim, K. (2018). An improved semi-supervised dimensionality reduction using feature weighting: Application to sentiment analysis. *Expert Systems with Applications*, 109:49–65.

Kim, K. and Lee, J. (2014). Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. *Pattern Recognition*, 47(2):758–768.

Kiritchenko, S., Zhu, X., Cherry, C., and Mohammad, S. (2014). Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 437–442.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Kumar, P., Prasad, R., Choudhary, A., Mishra, V. N., Gupta, D. K., and Srivastava, P. K. (2017). A statistical significance of differences in classification accuracy of crop types using different classification algorithms. *Geocarto International*, 32(2):206–224.

Kwon, N., Shulman, S. W., and Hovy, E. (2006). Multidimensional text analysis for erulemaking. In *Proceedings of the 2006 international conference on Digital government research*, pp. 157–166. Digital Government Society of North America.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lau, J. H. and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Lau, R. Y., Li, C., and Liao, S. S. (2014). Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. *Decision Support Systems*, 65:80–94.

Lavelli, A., Sebastiani, F., and Zanoli, R. (2004). Distributional term representations: an experimental comparison. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pp. 615–624. ACM.

Le, H. S. (2012). *Continuous space models with neural networks in natural language processing*. PhD thesis, Paris 11.

Le, Q. and Mikolov, T. (2014a). Distributed Representations of Sentences and Documents. *International Conference on Machine Learning - ICML 2014*, 32:1188–1196.

Le, Q. and Mikolov, T. (2014b). Distributed Representations of Sentences and Documents. *International Conference on Machine Learning - ICML 2014*, 32:1188–1196.

Lee, S., Jin, X., and Kim, W. (2015). Sentiment Classification for Unlabeled Dataset using Doc2Vec with JST. pp. 1–5.

Lee, S., Jin, X., and Kim, W. (2016). Sentiment classification for unlabeled dataset using doc2vec with jst. In *Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World*, p. 28. ACM.

Li, X., Xie, H., Chen, L., Wang, J., and Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23.

Lin, Y.-S., Jiang, J.-Y., and Lee, S.-J. (2014). A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering*, 26(7):1575–1590.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666.

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

Liu, X. and Croft, W. B. (2005). Statistical language modeling for information retrieval. *Annual Review of Information Science and Technology*, 39(1):1–31.

López-Monroy, A. P., Montes-Y-Gomez, M., Escalante, H. J., Pineda, L. V., and Villatoro-Tello, E. (2013). Inaoe's participation at pan'13: Author profiling task notebook for pan at clef 2013. In *CLEF (Working Notes)*.

Lu, Y., Hu, X., Wang, F., Kumar, S., Liu, H., and Maciejewski, R. (2015). Visualizing social media sentiment in disaster scenarios. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pp. 1211–1215. International World Wide Web Conferences Steering Committee.

Lui, X. and Croft, W. B. (2003). Statistical Language Modeling For Information Retrieval. *Annual Review of Information Science and Technology 2005 Volume 39*, 39:1.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011a). Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011b). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011c). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 142–150. Association for Computational Linguistics.

Mao, Y., Balasubramanian, K., and Lebanon, G. (2010). Dimensionality reduction for text using domain knowledge. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 801–809. Association for Computational Linguistics.

Martineau, J., Finin, T., et al. (2009). Delta tfidf: An improved feature space for sentiment analysis. *Icwsm*, 9:106.

Matsumoto, S., Takamura, H., and Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 301–311. Springer.

McAuley, J. and Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pp. 165–172. ACM.

McCormick, C. (2016). Word2vec tutorial-the skip-gram model.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

McKeown, K. (2009). N-Grams and Corpus Linguistics.

Meena, A. and Prabhakar, T. (2007). Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. pp. 573–580.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Distributed Representations of Words and Phrases and their Compositionality. *Nips*, pp. 1–9.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *Nips*, pp. 1–9.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013c). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Corrado, G., Chen, K., and Dean, J. (2013d). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12.

Mikolov, T., Joulin, A., Chopra, S., Mathieu, M., and Ranzato, M. (2014a). Learning longer memory in recurrent neural networks. *arXiv preprint arXiv:1412.7753*.

Mikolov, T., Joulin, A., Chopra, S., Mathieu, M., and Ranzato, M. (2014b). Learning Longer Memory in Recurrent Neural Networks. pp. 1–9.

Mikolov, T., Kopecky, J., Burget, L., Glembek, O., et al. (2009). Neural network based language models for highly inflective languages. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4725–4728. IEEE.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013e). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mishne, G., Glance, N. S., et al. (2006). Predicting movie sales from blogger sentiment. In *AAAI spring symposium: computational approaches to analyzing weblogs*, pp. 155–158.

Mittal, A. and Goel, A. (2012). Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15.

Mudinas, A., Zhang, D., and Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, p. 5. ACM.

Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. (2013). What yelp fake review filter might be doing? In *Seventh international AAAI conference on weblogs and social media*.

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

Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2015). Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. *arXiv preprint arXiv:1504.06654*.

Nikhil, R., Tikoo, N., Kurle, S., Pisupati, H. S., and Prasad, G. (2015). A survey on text mining and sentiment analysis for unstructured web data. In *Journal of Emerging Technologies and Innovative Research*, volume 2. JETIR.

O'Keefe, T. and Koprinska, I. (2009). Feature selection and weighting methods in sentiment analysis. In *Proceedings of the 14th Australasian document computing symposium, Sydney*, pp. 67–74. Citeseer.

Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines: an application to face detection. In *cvpr*, p. 130. IEEE.

Pandarachalil, R., Sendhilkumar, S., and Mahalakshmi, G. S. (2015). Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach. *Cognitive Computation*, 7(2):254–262.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*, pp. 271–278.

Pang, B. and Lee, L. (2005). Seeing stars. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, (1):115–124.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Passos, A., Kumar, V., and McCallum, A. (2014). Lexicon infused phrase embeddings for named entity resolution. *arXiv preprint arXiv:1404.5367*.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pp. 1532–1543.

Pletea, D., Vasilescu, B., and Serebrenik, A. (2014). Security and emotion: sentiment analysis of security discussions on github. In *Proceedings of the 11th working conference on mining software repositories*, pp. 348–351. ACM.

Poria, S., Cambria, E., and Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2539–2544.

Pouransari, H. and Ghili, S. (2014). Deep learning for sentiment analysis of movie reviews.

Pozzi, F. A., Fersini, E., Messina, E., and Liu, B. (2016). *Sentiment analysis in social networks*. Morgan Kaufmann.

Prollochs, N., Feuerriegel, S., and Neumann, D. (2015). Enhancing sentiment analysis of financial news by detecting negation scopes. *2015 48th Hawaii International Conference on System Sciences (HICSS). Proceedings*, pp. 959–968.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Republic, C. and Mikolov, T. (2012). Statistical Language Models Based on Neural Networks. *Wall Street Journal*, (April):1–129.

Rong, X. word2vec Parameter Learning Explained Continuous Bag-of-Word Model. pp. 1–21.

Rong, X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., and Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3):140–157.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5:3.

Saif, H., He, Y., Fernandez, M., and Alani, H. (2015a). Contextual semantics for sentiment analysis of twitter. *Information Processing & Management*.

Saif, H., He, Y., Fernandez, M., and Alani, H. (2015b). Contextual semantics for sentiment analysis of Twitter. *Information Processing and Management*, pp. 1–15.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Sharma, A. and Dey, S. (2012). A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM research in applied computation symposium*, pp. 1–7. ACM.

Shi, Y., Larson, M., Pelemans, J., Jonker, C. M., Wambacq, P., Wiggers, P., and Demuynck, K. (2015). Integrating meta-information into recurrent neural network language models. *Speech Communication*, 73:64–80.

Shojaee, S. and bin Azman, A. (2013). An evaluation of factors affecting brand awareness in the context of social media in malaysia. *Asian Social Science*, 9(17):72.

Shojaee, S., Murad, M. A. A., Azman, A. B., Sharef, N. M., and Nadali, S. (2013). Detecting deceptive reviews using lexical and syntactic features. In *Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on*, pp. 53–58. IEEE.

Smailović, J., Grčar, M., Lavrač, N., and Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285:181–203.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (ii):151–161.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., Potts, C., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, p. 1642. Citeseer.

Soleymani, M., Schuller, B., and Chang, S.-F. (2017). Guest editorial: Multimodal sentiment analysis and mining in the wild.

Steinberger, J., Brychcín, T., and Konkol, M. (2014). Aspect-level sentiment analysis in czech. In *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pp. 24–30.

Tabatabaei, N. (2011). Detecting weak signals by internet-based environmental scanning. Master's thesis, University of Waterloo.

Thorleuchter, D., Van den Poel, D., and Prinzie, A. (2010). Mining ideas from textual information. *Expert Systems with Applications*, 37(10):7182–7188.

Tomar, D. S. and Sharma, P. (2016). A text polarity analysis using sentiwordnet based an algorithm. *IJCSIT) International Journal of Computer Science and Information Technologies.*

Tripathy, A., Agrawal, A., and Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 57:117–126.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417–424. Association for Computational Linguistics.

Turney, P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Van de Kauter, M., Desmet, B., and Hoste, V. (2015). The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment. *Language Resources and Evaluation*, pp. 1–36.

Villegas, M. P., José, M., Ucelay, G., Fernández, J. P., Álvarez-Carmona, M. A., Errecalde, M. L., and Cagnina, L. C. (2016a). Vector-based word representations for sentiment analysis: a comparative study. pp. 785–793.

Villegas, M. P., José, M., Ucelay, G., Fernández, J. P., Álvarez-Carmona, M. A., Errecalde, M. L., and Cagnina, L. C. (2016b). Vector-based word representations for sentiment analysis: a comparative study. pp. 785–793.

Wang, S. and 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-Volume 2*, pp. 90–94. Association for Computational Linguistics.

Whitelaw, C., Garg, N., and Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 625–631. ACM.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 347–354. Association for Computational Linguistics.

Wu, D. and Chi, M. (2017). Long short-term memory with quadratic connections in recursive neural networks for representing compositional semantics. *IEEE Access*, 5:16077–16083. cited By 0.

Wu, H., Hu, Y., Li, H., and Chen, E. (2015). A New Approach to Query Segmentation for Relevance Ranking in Web Search. *Inf. Retr.*, 18(1):26–50.

Wu, S.-J., Chiang, R.-D., and Chang, W.-T. (2016). Extracting new opinion elements in the semi-automatic chinese opinion-mining system from internet forums. In *International Conference on Frontier Computing*, pp. 529–541. Springer.

Xue, B., Fu, C., and Shaobin, Z. (2014). A study on sentiment computing and classification of sina weibo with word2vec. In *2014 IEEE International Congress on Big Data*, pp. 358–363. IEEE.

Yang, C. C. and Ng, T. D. (2007). Terrorism and crime related weblog social network: Link, content analysis and information visualization. In *Intelligence and Security Informatics, 2007 IEEE*, pp. 55–58. IEEE.

Yazdani, S. F., Murad, M. A. A., Sharef, N. M., Singh, Y. P., and Latiff, A. R. A. (2017). Sentiment classification of financial news using statistical features. *International Journal of Pattern Recognition and Artificial Intelligence*, 31(03):1750006.

Yu, L.-C., Wang, J., Lai, K. R., and Zhang, X. (2015a). Predicting Valence-Arousal Ratings of Words using a Weighted Graph Method. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL-2015)*, pp. 788–793.

Yu, Z., Wang, H., Lin, X., and Wang, M. (2015b). Learning term embeddings for hypernymy identification. In *IJCAI*, pp. 1390–1397.

Zainuddin, N., Selamat, A., and Ibrahim, R. (2016). Twitter feature selection and classification using support vector machine for aspect-based sentiment analysis. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 269–279. Springer.

Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent Neural Network Regularization. *Iclr*, (2013):1–8.

Zhang, J., Zong, C., et al. (2015). Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, 15.

Zheng, L., Wang, H., and Gao, S. (2015). Sentimental feature selection for sentiment analysis of chinese online reviews. *International Journal of Machine Learning and Cybernetics*, pp. 1–10.

Zhu, J., Wang, H., and Mao, J. (2010). Sentiment classification using genetic algorithm and conditional random fields. In *Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on*, pp. 193–196. IEEE.

Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

BIODATA OF STUDENT

Eissa M. Alshari was born on 7th March, 1979 in Ibb - Yemen. He obtained his early education at his native town and finished higher secondary education from Khaled Bin Alwaled, Ibb, Yemen. Afterwards, he proceeded to get his degree from the National Institute of Administrative Sciences in Computer science in 2001. After completion of Higher diploma Degree, he worked at Ibb university for 8 years.

Then, he had been bachelor from National University, Taiz, Yemen on Computer science in 2009. During the year from 2009 to 2011 he has been working as a Lecturer in a few colleges in Yemen.

Next, he had been sponsored by the Ibb university in Yemen by offering a master and PhD scholarship. The Master in Information systems was obtained from Menofia university, Egypt during 2011-2014.

Finally, in the begining of 2015 he joined Universiti Putra Malaysia (UPM) for doing his Doctor of Philosophy (PhD) programme in the field of Intelligent Systems.

The Author is married in 1999 and is blessed with four kids.

LIST OF PUBLICATIONS/PATENTS

International Refereed Journals

Alshari, Eissa M.; Azman, Azreen; Mustapha, Norwati; Doraisamy, Shyamala C.; Alksher, Mostafa(2017). Senti2Vec: An Effective Feature Extraction Technique for Sentiment Analysis Based on Word2Vec. Malaysian Journal of Computer Science (MJCS) **(Acceptance 2018)**

Alshari, Eissa M.; Azman, Azreen; Mustapha, Norwati; Doraisamy, Shyamala C.; Alksher, Mostafa(2017). Effective Bag-of-Words Features Based on Enlarged Opinion Dictionary for Sentiment Analysis. IAENG International Journal of Computer Science **(Acceptance 2019)**

International Refereed Conferences

Alshari, Eissa M.; Azman, Azreen; Mustapha, Norwati; Doraisamy, Shyamala C.; Alksher, Mostafa;,Prediction of Rating from Comments based on Information Retrieval and Sentiment Analysis,2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP),32-36,2017,IEEE Conference Publications

Alshari, Eissa M; Azman, Azreen; Doraisamy, Shyamala; Mustapha, Norwati; Alksher, Mostafa;,Improvement of Sentiment Analysis Based on Clustering of Word2vec Features,"Database and Expert Systems Applications (DEXA), 2017 28th International Workshop on",123-126,2017,IEEE

Azman, Azreen; Alshari, Eissa M; Sulaiman, Puteri Suhaiza; Abdullah, Muhamad Taufik; Alksher, Mostafa; Kadir, Rabiah Abdul; ,Feasibility of Using Rating to Predict Sentiment for Online Reviews,2017 Asia Modelling Symposium (AMS),37-41,2017,IEEE

Alshari, Eissa M; Azman, Azreen; Mustapha, Norwati; Alksher, Mostafa; ,Effective Method for Sentiment Lexical Dictionary Enrichment based on Word2vec for Sentiment Analysis, (CAMP),2018,IEEE Conference Publications

UNIVERSITI PUTRA MALAYSIA

STATUS CONFIRMATION FOR THESIS / PROJECT REPORT AND COPYRIGHT

ACADEMIC SESSION : ____________________

TITLE OF THESIS / PROJECT REPORT :

FEATURE EXTRACTION BASED ON WORD EMBEDDINGS AND OPINION LEXICALS FOR SENTIMENT ANALYSIS

NAME OF STUDENT: EISSA MOHAMMED MOHSEN ALSHARI

I acknowledge that the copyright and other intellectual property in the thesis/project report belonged to Universiti Putra Malaysia and I agree to allow this thesis/project report to be placed at the library under the following terms:

1. This thesis/project report is the property of Universiti Putra Malaysia.

2. The library of Universiti Putra Malaysia has the right to make copies for educational purposes only.

3. The library of Universiti Putra Malaysia is allowed to make copies of this thesis for academic exchange.

I declare that this thesis is classified as :

*Please tick (√)

□ **CONFIDENTIAL** (Contain confidential information under Official Secret Act 1972).

□ **RESTRICTED** (Contains restricted information as specified by the organization/institution where research was done).

□ **OPEN ACCESS** I agree that my thesis/project report to be published as hard copy or online open access.

This thesis is submitted for :

□ **PATENT** Embargo from______________ until ______________
(date) (date)

Approved by:

____
(Signature of Student)
New IC No/ Passport No.:

Date :

____
(Signature of Chairman of Supervisory Committee)
Name:

Date :

[Note : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization/institution with period and reasons for confidentially or restricted.]