INTEGRATION OF FEATURE SUBSET SELECTION METHODS FOR SENTIMENT ANALYSIS

ALIREZA YOUSEFPOUR

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

NOVEMBER 2019

DEDICATION

“To my beloved wife and son”

ACKNOWLEDGEMENT

Thanks to Almighty ALLAH for providing me the knowledge, guidance, and patience to achieve this goal.

My sincere thanks to my project supervisor, Prof. Madya Dr. Roliana binti Ibrahim for motivation, encouragement and valuable comments. I also wish to appreciate my co-supervisor Dr. Haza Nuzly Abdull Hamed for giving me helpful guidelines and recommendations and developing my knowledge and understanding.

I would like to express my sincere appreciation to my mother and father, who have always supported me emotionally and dealt with difficulties while I was away for my studies. I would also thank Rahil, my beloved wife, and Meraj, my son, who helped me and encouraged me for my studies.

I would like to thank the staff of Universiti Teknologi Malaysia, and especially the Faculty of Computing, for their kind cooperation.

Lastly, I want to thank from all the people who have helped me during conducting my PhD project and the ones who provided such a great academic environment for research and education.

ABSTRACT

Feature selection is one of the main challenges in sentiment analysis to find an optimal feature subset from a real-world domain. The complexity of an optimal feature subset selection grows exponentially based on the number of features for analysing and organizing data in high-dimensional spaces that lead to the high-dimensional problems. To overcome the problem, this study attempted to enhance the feature subset selection in high-dimensional data by removing irrelevant and redundant features using filter and wrapper approaches. Initially, a filter method based on dispersion of samples on feature space known as mutual standard deviation method was developed to minimize intra-class and maximize inter-class distances. The filter-based methods have some advantages such as they are easily scaled to high-dimensional datasets and are computationally simple and fast. Besides, they only depend on feature selection space and ignore the hypothesis model space. Hence, the next step of this study developed a new feature ranking approach by integrating various filter methods. The ordinal-based and frequency-based integration of different filter methods were developed. Finally, a hybrid harmony search based on search strategy was developed and used to enhance the feature subset selection to overcome the problem of ignoring the dependency of feature selection on the classifier. Therefore, a search strategy on feature space using integration of filter and wrapper approaches was introduced to find a semantic relationship among the model selections and subsets of the search features. Comparative experiments were performed on five sentiment datasets, namely movie, music, book, electronics, and kitchen review dataset. A sizeable performance improvement was noted whereby the proposed integration-based feature subset selection method yielded a result of 98.32% accuracy in sentiment classification using POS-based features on movie reviews. Finally, a statistical test conducted based on the accuracy showed significant differences between the proposed methods and the baseline methods in almost all the comparisons in k-fold cross-validation. The findings of the study have shown the effectiveness of the mutual standard deviation and integration-based feature subset selection methods have outperformed the other baseline methods in terms of accuracy.

ABSTRAK

Pemilihan ciri merupakan salah satu cabaran utama dalam analisis sentimen untuk mencari subset ciri optimum dari domain dunia sebenar. Kerumitan pilihan subset ciri optimum berkembang pesat berdasarkan bilangan ciri-ciri untuk menganalisis dan menganjurkan data dalam ruang dimensi tinggi yang membawa kepada masalah dimensi tinggi. Untuk mengatasi masalah ini, kajian ini cuba untuk meningkatkan pemilihan subset ciri dalam data dimensi tinggi dengan membuang ciri-ciri tidak relevan dan berlebihan menggunakan pendekatan penapis dan bungkus. Pada mulanya, kaedah penapis berdasarkan penyebaran sampel pada ruang ciri yang dikenali sebagai kaedah sisihan piawai bersama telah dibangunkan untuk meminimumkan kelas intra dan memaksimumkan jarak antara kelas. Kaedah berasaskan penapis mempunyai beberapa kelebihan seperti mudah diperingkatkan kepada dataset berkepadatan tinggi dan dikira mudah dan cepat. Selain itu, ia hanya bergantung kepada ruang pemilihan ciri dan mengabaikan ruang model hipotesis. Oleh itu, langkah seterusnya dalam kajian ini adalah untuk membangunkan pendekatan skala ciri baru dengan mengintegrasikan pelbagai kaedah penapis. Penyepaduan berasaskan ordinal dan frekuensi berasaskan kaedah penapis yang berbeza telah dibangunkan. Akhirnya, pencarian harmoni hibrid berdasarkan strategi pencarian telah dibangunkan dan digunakan untuk meningkatkan pemilihan subset ciri untuk mengatasi masalah mengabaikan ketergantungan pemilihan ciri pada pengelas. Oleh itu, strategi carian pada ruang ciri menggunakan pendekatan penapis dan bungkus diperkenalkan untuk mencari hubungan semantik antara pilihan model dan subset ciri carian. Eksperimen perbandingan dilakukan pada lima kumpulan sentimen, iaitu filem, muzik, buku, elektronik, dan kajian semula peralatan kajian. Penambahbaikan prestasi yang besar telah diperhatikan di mana kaedah pemilihan subset ciri yang berasaskan integrasi yang dicadangkan menghasilkan hasil ketepatan 98.32% dalam klasifikasi sentimen menggunakan ciri berdasarkan POS pada ulasan filem. Akhir sekali, ujian statistik yang dijalankan berdasarkan ketepatan menunjukkan perbezaan yang ketara antara kaedah yang dicadangkan dan kaedah asas dalam hampir semua perbandingan dalam *k-fold cross-validation*. Dapatan kajian telah menunjukkan keberkesanan kaedah sisihan piawai bersama dan kaedah pemilihan subset ciri berasaskan integrasi telah mengatasi kaedah asas lain dari segi ketepatan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SA	-	Sentiment Analysis
NLP	-	Natural Language Processing
IR	-	Information Retrieval
MSD	-	Mutual Standard Deviation
CrSD	-	Cross Standard Deviation
SD	-	Standard Deviation
FCV	-	Fold-Cross Validation
OIFV	-	Ordinal-based Integration of Feature Vectors
FIFS	-	Frequency-based Integration of Feature Subsets
OFV	-	Ordinal-based Features Vector
FV	-	Features Vector
MFS	-	Middle Feature Subset
FFV	-	Frequency-based Features Vector
HS	-	Harmony Search
GA	-	Genetic Algorithm
BHS	-	Basic Harmony Search
DHS	-	Dynamic Harmony Search
HHS	-	Hybrid Harmony Search
SFS	-	Selected Feature Space
OSFS	-	Outer Selected Feature Space
HM	-	Harmony Memory
HMS	-	Harmony Memory Size
HMCR	-	Harmony Measuring Considering Rate
PAR	-	Pitch Adjusting Rate
BND	-	Bandwidth
maxIter	-	Maximum Iteration
NH	-	New Harmony
CNEF	-	Call Number of Evaluation Function
POS	-	Part-Of-Speech
TF-IDF	-	Term Frequency-Inverse Document Frequency

TDM	-	Term Document Matrix
NUM	-	Number
SW	-	Stop Words
TV	-	Term Variance
DF	-	Document Frequency
CHI	-	Chi-square
IG	-	Information Gain
WLLR	-	Weighted Log-Likelihood Ratio
SVM	-	Support Vector Machine
NB	-	Naive Bayes
ME	-	Maximum Entropy
LDF	-	Linear Discriminant Function

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

By the exponential growth of World Wide Web (WWW), many people are able to post their opinions or sentiments on a range of topics in different websites. These online posts can help people to observe and receive each other's opinions. Therefore, there is a large amount of data containing opinions generated from a variety of sources such as reviews, forum discussions, post from blogs, and other Carey-Simos (2015). Online reviews provide a significant information source to help customers and companies make decisions. In fact, these reviews can reassess their purchase decision and ultimately change their purchasing behaviour (Ye *et al.*, 2011).

Some studies reveal the interest that customers show in online reviews about product and services in term of relying on the online advice or recommendations to make purchase decision. For example, a study by Shrestha (2016) indicated that 92% of customers read online reviews before purchasing the product, whereas 66% of people rely on online product reviews (Stone, 2015). In addition, 63% of customers are more interested in purchasing products or services from website which has user reviews (iPerceptions, 2011). As a result, the request for sentiment analysis is important because of this surge of interest. In other words, it can be useful for making intelligent decisions by knowing the product's positive and negative sentiments.

Sentiment analysis is the type of a field in the computational study to process opinions, attitudes, sentiments and to assess people's comment about movies, events, products, topics, and their respective characteristics. There are different names for this area of study, such as sentiment mining, review mining, text mining, opinion extraction, subjectivity analysis, and emotion analysis. Opinion mining is a process for extracting subjective information from a text or a review while the main aim of

sentiment analysis is to identify and extract opinions, attitudes and comments in the overwhelming majority of generated content, whereas sentiment analysis is the evaluation of the extracted information. Recent studies have presented different techniques of sentiment extraction and analysis. Sentiment analysis is intended to identify and extract the opinions, attitudes, and sentiments in the overwhelming majority of generated online contents and classify them into polarity sentiment (positive and negative). The main task of sentiment analysis is categorized into two main steps: the first step involves selection or extraction of the relevant features from the textual reviews, and the last step covers the sentiment classification of the reviews into multi-classes (Ekbal and Saha, 2013; Pang and Lee, 2005).

Feature selection is one of the main challenges in sentiment analysis. More reviews on document-level have expressed a high-dimensional in feature space. The main task of feature selection is the reduction of dimension in feature space by removing irrelevant and redundant features in order to improve the performance of sentiment classification (Saeys *et al.*, 2007). To overcome this problem, using filter-based feature selection method can be helpful because of the advantages of filter-based methods. They are fast and simple in computation, easily scaled to a high-dimensional feature space, and independent from the classifier. As a result, to overcome the problem of optimal feature subset selection in high-dimensional space, this study makes a different view to filter methods based on the distribution of features on space and also the integration of different filter methods using heuristic algorithm. In this research, feature subset selection by using hybridizations of the filter and wrapper methods are proposed to improve the sentiment classification accuracy.

1.2 Background of the Study

Nowadays, reviews, are created by users, are very important in e-commerce and business. They can help companies to improve product quality and customers to select a better product. To this end, the science of sentiment analysis and opinion mining, which are a combination of information retrieval methods and natural language processing methods, have emerged to help to analyse opinions, emotions,

and attitudes of the user about products, to classify subjective text into negative and positive sentiment classes.

Two primary works attempting the sentiment analysis task were demonstrated by Pang *et al.* (2002) and Turney (2002) who introduced two different approaches. The unsupervised learning approach was used by Turney (2002) while the sentiment lexicon was used to identify and classify documents to sentiment polarity by calculating the word sentiment orientation using the POS patterns or a dictionary and using a search engine to estimate the association of words with a known polarity seed set. These works are known as unsupervised learning methods and are strongly dependent on sentiment lexicons. Furthermore, some of the researches have been undertaken based on unsupervised learning methods (Harb *et al.*, 2008; Hu *et al.*, 2013; Taboada *et al.*, 2011; Turney, 2002). On the other hand, Pang *et al.* (2002) exploited the corpus-based approach to sentiment classification in supervised learning. Some researchers used a set of several features to improve the accuracy of classification (Liu, 2012; Ortigosa-Hernández *et al.*, 2012; Zhou *et al.*, 2013; Zhu *et al.*, 2013). Other feature selection methods were employed to earn better performance such as Information Gain (Ye and Keogh, 2009). Further, some researches also attempted to ensemble several methods using hybrid classifiers (Prabowo and Thelwall, 2009). Sentiment classification using supervised learning is a popular approach in recent researches that attempts to train a classifier from a large amount of labelled data (Pang *et al.*, 2002; Zhang and Liu, 2011).

In addition, three levels of sentiment analysis based on the level of granularities, namely the document, sentence and aspect levels, have been investigated in recent researches (Liu and Zhang, 2012). Sentiment classification based on document-level granularity is performed to classify one whole document as showing either an overall negative or positive sentiment (Liu, 2012; Taboada *et al.*, 2011). A research by McDonald *et al.* (2007) and Nakagawa *et al.* (2010) used sentence-level granularity to determine whether each sentence presented a negative, positive or neutral sentiment. This task is often called subjective classification in the literature (Tang et al., 2009). Instead of investigating at the language structure-level, aspect-level or feature-based level, it looks directly at the opinions on the basis of the overall

opinion (that contains a polarity sentiment) and a target opinion (Jin *et al.*, 2015; Qiu *et al.*, 2011).

The main task of sentiment analysis is categorized into two main steps: the first step involves selection of the relevant features from the textual reviews, and the last step covers the sentiment classification of the reviews into multi-classes (Ekbal and Saha, 2013; Pang and Lee, 2005). Furthermore, feature selection methods in order to select a subset of most relevant features are classified according to three main aims: (1) techniques for overcoming of the overfitting problem and the improvement of the performance of sentiment classification, (2) techniques to provide a model with less time complexity and more cost-effectiveness, and (3) techniques to obtain the best understanding of the basic process for the data generated. As a result, The selection of an optimal subset of the features is the main task before applying a learning algorithm in designing systems based on machine learning and pattern recognition (Jin *et al.*, 2015; Sotoca and Pla, 2010). A survey by Saeys *et al.* (2007) investigated that feature selection techniques can be classified into three categories: filter, wrapper, and embedded.

1) The filter approach provides ways to assess the relevance of features by looking only at the properties of the data in order to find an optimal feature subset. In most cases, score is calculated for each feature and low-scoring features will be removed. The related advantages of the filter approach are easily scaling of high-dimensional data and offering a simple and quick computation method. However, it suffers from problems such as ignoring the interaction with the classifier and relinquishing relationships between features (Ekbal and Saha, 2013; Saeys *et al.*, 2007).

2) In the wrapper approach, the choice of an optimal feature subset is provided by generating and evaluating complete subsets on the feature space of states. To search the all possible features space, a search algorithm is wrapped around the classification model which has a time complexity of $O(2^N)$. In fact, as the feature space exponentially grows with the number of features, heuristic search techniques are employed to reduce the search time for finding an optimal subset.

For example, Meta-heuristic algorithms, which are inspired by behaviors in nature, have been used in optimization problems (Geem *et al.*, 2001b). In order to generate the subsets, meta-heuristic methods are used to solve the exponential time to find a candidate feature subset. Regarding the evaluation of candidate subsets, the classifier evaluates the effect of selecting a feature subset on the performance of sentiment classification to find an optimal feature subset. The advantage of the wrapper approach is finding a semantic relationship between the hypothesis model selections and subsets of the search features. These methods depend on classification algorithms. Nevertheless, they have a higher overfitting risk and significant complexity in computation and cost.

3) In the third approach, which is termed as embedded approach, the search comes with combining of the feature selection strategy into a classifier structure (hypothesis model). The advantages of embedded approach are that it reduces the computational time in comparison with wrapper methods, and it is able to interact with the classifier model to select most effective features, but they suffer from classifier dependent selection (Ekbal and Saha, 2013; Tabakhi *et al.*, 2014).

Furthermore, filter methods are a popular approach because of simple methods and low time computational, efficiency, scaling high-dimensional, and independence of the learning algorithms. In order to the high-dimensional problem, Rogati and Yang (2002) investigated feature selection using several filter methods to deal with high-dimensional data for text classification. They scored features by five methods: information gain (IG), mutual information (MI), chi-square (CHI), document frequency (DF), and term frequency (TF). Another work by Yang and Pedersen (1997) improved classification accuracy with the removal of up to 98% unique features through the IG and CHI methods. In some studies, feature selection methods such as the IG and CHI methods were found to achieve better accuracy than other methods (Uğuz, 2011; Ye and Keogh, 2009). Feature ranking using filter methods are divided into four categories based on information, distance, dependence, consistency measures. Information-based feature ranking method is a popular approach in recent research. For instance, many researchers have used filter methods based on information-based, such as document frequency and term frequency (Rogati and Yang,

2002; Yang and Pedersen, 1997), information gain, Chi-square (Rogati and Yang, 2002; Uğuz, 2011; Yang and Pedersen, 1997; Ye and Keogh, 2009), conditional mutual information (Peng and Fan, 2017), maximizing global information gain (Shang *et al.*, 2013), and dynamic mutual information (Hua *et al.*, 2009). Some filter methods based on distance-based ranking were employed, such as term variance (Tabakhi *et al.*, 2014), Euclidean distance-based (Li and Lu, 2009), Laplacian score (He *et al.*, 2006), fisher Markov selector (Cheng *et al.*, 2011). Also, other filter methods using the dependency of the features were introduced, such as covariance, correlation coefficient, and predominant correlation (Yu and Liu, 2003). The consistency-based filter as a consistency measures was investigated by (Bolón-Canedo *et al.*, 2014). Common drawbacks of these filter methods are that they suffer from problems like redundant features and the lack of interaction information between the feature selection and the classifier.

In wrapper method, meta-heuristic algorithms are the most efficient techniques to search for global-optimal solutions by overcoming the main problem of local-search techniques in large-scale problems, for example, getting trapped in local extremes in the space of the search (Oreski and Oreski, 2014). One of the meta-heuristic methods for optimization of feature subset selection problem is harmony search (HS) algorithm (Geem *et al.*, 2001a). Diao and Shen (2012) introduced a method for feature selection using HS which escapes from local-solutions and identifies multiple solutions with respect to the stochastic nature of HS and controls its parameters. They compared the HS algorithm with other meta-heuristic techniques (such as the genetic algorithm-GA) on the UCI benchmark datasets and showed that the HS was able to identify good-quality feature subsets for most of the test datasets. Moreover, in the wrapper method, feature selection using HS, was first manifested by Diao and Shen (2012). They applied the HS algorithm for feature selection on the UCI benchmark datasets. They showed that the HS algorithm is able to identify good-quality feature subsets. On the other hand, Wang *et al.* (2015) used the HS algorithm for feature selection in email classification. In a research related to HS, a few modified variants of the original HS algorithm were proposed by Geem *et al.* (2001b), to enhance the accuracy and convergence rate without targeting any specific application. Using dynamically adjusted parameters instead of constant parameters by mathematical techniques, Mahdavi *et al.* (2007) proposed an improved HS algorithm to increase the accuracy

and convergence rate of the original HS algorithm. Furthermore, Omran and Mahdavi (2008) introduced a global-based HS algorithm that exploits the concept of swarm intelligence. In addition, a self-adaptive HS algorithm was developed by Wang and Huang (2010) whereby the values of the HS parameters are adjusted automatically based on past experiences. Moreover, some applications of HS in engineering, medical, robotics, and control fields were investigated (Manjarres *et al.*, 2013).

In general, the sentence like "the best feature selection method" merely does not exist for each problem in the literature. Thus, choosing one method over another is a difficult decision to make. Some researchers introduced ensemble feature selection methods using the two most popular approaches. In order to incorporate the advantages filter and wrapper methods, the hybrid approach has also been introduced to solve low-accuracy sentiment classification in the filter methods and higher-computation burden problems respectively in wrapper approach. To solve these problems, this approach is employed by applying the methods to the selected feature subset using filtering method in the first step (Warren Liao, 2010). Moreover, a hybrid filter and wrapper methods for forecasting of short-term load has been introduced (Hu *et al.*, 2015).

Moreover, some researches, instead of using a single feature subset as a prediction model, used feature sets ensemble with aiming at combining these different feature sets while still taking advantage of their benefits, and improving their performance. As a result, bagging and boosting were the most popular approaches. For example, Xia *et al.* (2011) investigated the effectiveness of the ensemble technique on feature sets and sentiment classification in three main steps. First, they extracted feature sets using the part-of-speech-based (POS-based) and the word-relation-based. Second, they employed three base classifiers for each set with the use of the support vector machine (SVM), naive Bayes (NB) and maximum entropy (ME) classifiers. Last, they combined such methods based on the fixed, meta-classifier and weighted combination respectively as ensemble strategies. Research has shown ensemble methods result in higher accuracy and have demonstrated an efficient way to improve classification performance via a combination of different feature sets and classifiers.

In the next step of sentiment analysis, the sentiment classification task involves analysing and predicting opinions and sentiments in relation to the polarity sentiment classification. It presents an important role in social media and web. Sentiment classification using machine learning algorithms in a supervised approach was a popular approach in many recent researches, such as SVM (Badawi and Altınçay, 2014; Boiy and Moens, 2009; Pang *et al.*, 2002), NB (Pang and Lee, 2004; Tan and Zhang, 2008), ME (Pang *et al.*, 2002; Speriosu *et al.*, 2011), and linear discriminant function (LDF) (Guyon *et al.*, 2002).

Although the mentioned researches have attempted to overcome some problems in feature subset selection for sentiment analysis, there are still several problems in this area. These problems have rarely been considered in the literature such as, the high-dimensional data, time-complexity, difficult computational, diversity of classification result on the various domain, irrelevant and redundant features, risk of overfitting, dependent on a ranking method, dependent on the classifier. Filling of these gaps can potentially improve the performance of sentiment analysis in various domain using integration of different filter-based feature selection methods and wrapper methods. Figure 1.1 illustrates the problems and proposed desired solutions for feature subset selection in sentiment analysis.

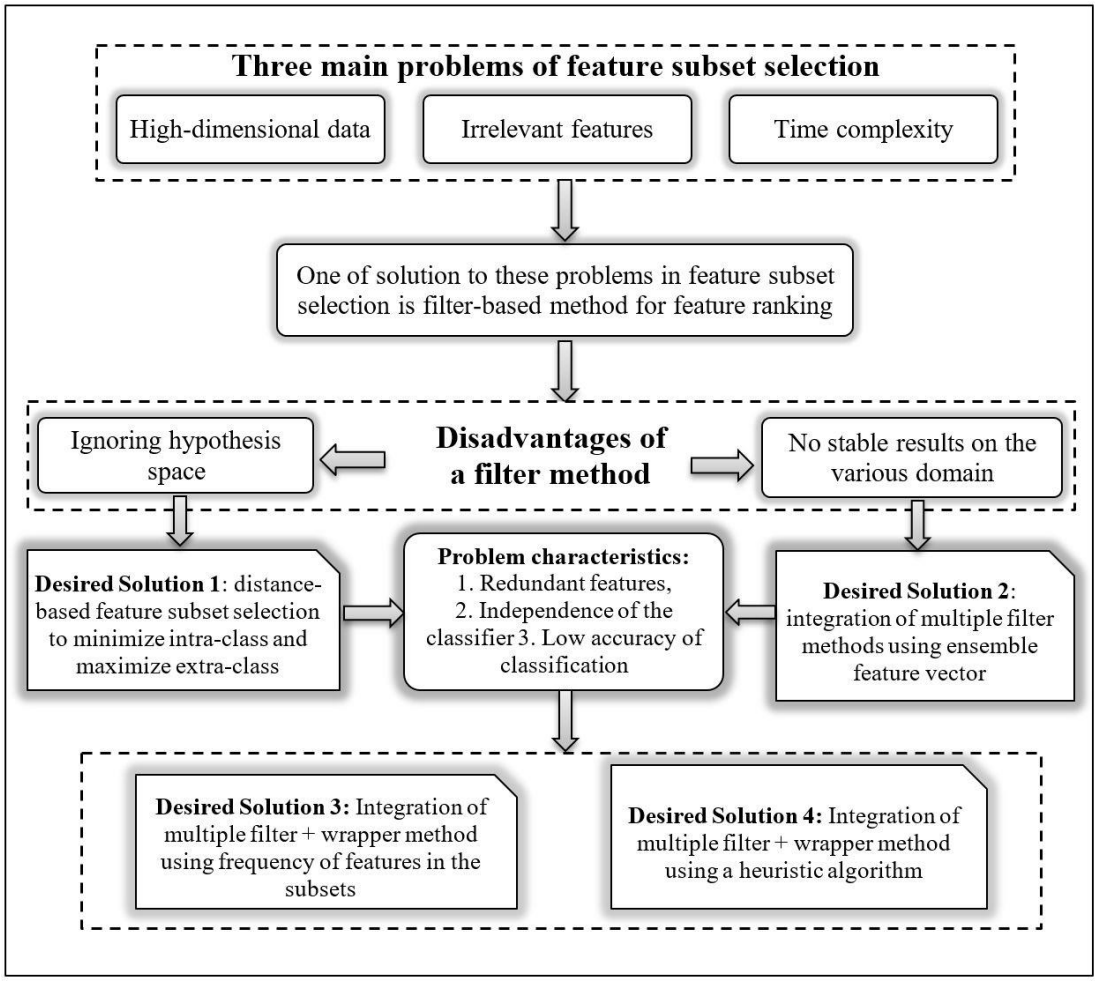

Figure 1.1 Problems and desired solutions of feature subset selection

1.3 Problem Statement

In this study, overcoming the problem of optimal feature subset selection for sentiment analysis is intended. The problem can be defined as follow:

Let a set of D = $\{d_1, d_2, .., d_M\}$ denotes the text documents set in polarity sentiment classification where d_i is a positive or negative text document and M is the number of the text documents. And also, a feature set F = $\{f_1, f_2, .., f_N\}$ denotes the features in the dataset with dimension N. The task of feature subset selection is a search to find an optimal subset of features S with dimension K on 2^N candidate feature subsets where $K \leq N$, and $S \subseteq F$. The subset *S* should make equal or better classification

accuracy compared to feature set F. The main three problems are defined by existing feature subset selection for sentiment analysis as follow:

1) **Lack of information between feature subset search and class discriminatory in filter methods.** More researches employed the information measure to weight and rank the features in literature.

2) **In filter approach, selected features are dependent on one feature ranking method.** It causes diversity in classification accuracy at the various domains. On the other hand, filter-based methods have some advantages which are easily scaling the high-dimensional problem and also they are a simple and quick computation technique.

3) **Ignoring interaction information between feature subset search and the hypothesis model search.** Whereas filter methods suffer from the problem of an optimal feature subset selection because of independency of the model selection step, wrapper methods embed the hypothesis model search within the feature subset search. Since filter methods rank each feature separately in which feature dependencies may be ignored, a poor classification performance may be obtained. In fact, these methods ignore dependencies between feature selection and classifier.

1.4 Research Question

By considering the defined problems for overcoming the feature subset selection in sentiment analysis, the research hypothesis is:

"How to select a subset of the most relevant features so as to remove irrelevant and redundant features with keeping the most class discriminatory information at the same time based on a given collection of review documents"

In order to answer the research hypothesis, the following research questions that address the problems in detail are defined:

1) How to improve filter-based feature selection methods with respect to dispersion of samples on feature space.

2) How to integrate feature subset selection methods based on ranked features.

3) How to hybrid the filter and wrapper methods in order to enhance the sentiment classification accuracy.

1.5 Research Goal

The aim of this thesis is to explore an effective way to select the most relevent features from raw reviews to improve the performance of sentiment classification. By addressing the existing problems in previous works, the research strives to propose enhanced methods for finding an optimal feature subset based on integration of filter and wrapper methods with the ultimate goal of improving the performance of sentiment analysis.

1.6 Research Objective

In order to attain the research goal, several research objectives have been identified and listed as follows:

1) To propose a weighting method of the features using the distance-based measure in order to minimize intra-class and maximize extra-class distances for more class discrimination.

2) To propose integration methods for different filter-based feature selection methods in order to reduce the problems such as dependency of the selected features on a feature ranking method and stability of classification accuracy on the various domains.

3) To propose a hybrid method with the ability of embedding the model hypothesis search within the feature subset search in order to find a semantic relationship among the model selection and features selection by integrating filter and wrapper methods.

1.7 Research Scope

To solve the feature subset selection problems in this research, the following constraints are considered:

1) This research focuses on classifying the movie and Amazon products reviews in English language on the overall sentiment of each review dataset.

2) This research uses some information retrieval tools to detect the useful features from raw review datasets to improve accuracy of sentiment classification. For example, the Stanford POS tagger tools is used in this research to annotate documents with the part-of-speech (POS).

3) This research employs some machine learning algorithms to increase the performance of sentiment classification such as SVM, NB, ME, and LDF algorithms.

4) The programming languages such as Microsoft visual C# 2012, MATLAB, and Excel are used for implementation and visualization of proposed methods.

1.8 Significance of Research

As mentioned in Pang and Lee (2008), 81% of internet users have performed online search on a product at least once and 73% to 87% of these users report that product reviews had a significant influence on their purchase. About 80% of these users expressed their opinions that reviews are of great importance in their decision making on the purchase. These statistics show that the sentiment classification of

reviews is very helpful to customers to select appropriate products which has motivated researchers to pay more attention to this area. Moreover, in many applications, companies want to analyze and compare the opinions of their customers on their services. This example can express that how reviews can be useful for selecting an ideal product by costumers. Analyzing and organizing these reviews leads to the high-dimensional problem. To deal with the problem, this study aims to propose some methods for selecting an optimal feature subset through the processing and understating information using information retrieval techniques and natural language processing algorithms to improve the sentiment classification.

1.9 Thesis Outlines

Seven chapters are organized in this research as follow:

Chapter 1, *Introduction*, starts with an introduction to the research topic. The research background and research problems are explained. After that, the research questions and objectives are introduced. Finally, the importance of research is expressed.

Chapter 2, *Literature review*, provides the background information and reviews the previous studies in this field that leads to find the research gaps and formulate the research problem.

Chapter 3, *Research methodology*, explained the methods and datasets, which are used in this research. The research flow is described systematically in this chapter. Evaluation metric and evaluation framework also are explained in this chapter.

Chapter 4, *Mutual standard deviation method based on distance-based feature ranking*, explains the development process of the first proposed method, which determine the relevance of features by respect to the distribution and dispersion of features on feature space. This method is evaluated and compared with some other baseline methods in this chapter.

Chapter 5, *Ordinal-based and frequency-based integration of filter feature selection methods*, addresses the design and development steps of the second proposed methods, which enhances the first proposed method through integration with other filter-based feature selection methods based on the sequence of the features of the vectors.

Chapter 6, *Integration of filter and wrapper methods using hybrid harmony search with controlled parameters*, describes the design and implementation process of the last proposed method, in which hybrid basic harmony search and dynamic harmony search algorithms integrated based on proposed search strategy and integration of several filter-based feature selection methods to find an optimal feature subset.

Chapter 7, *Conclusion and future works*, concludes the research, provides the list of contributions, states the limitations of proposed methods and expresses some recommendation for future study.

REFERENCES

Apolloni, J., Leguizamón, G., and Alba, E. (2016). Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Applied Soft Computing, 38, 922-932.

Arcilla, M., Esquivel, A., Quiros, C., Velasco, K., and Cheng, C. (2013). Felex builder: a semi-supervised lexical resource builder for opinion mining in product reviews. Paper presented at the The Second International Conference on Digital Enterprise and Information Systems (DEIS2013), 66-71.

Asur, S., and Huberman, B. A. (2010). Predicting the future with social media. Paper presented at the Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, 492-499.

Badawi, D., and Altınçay, H. (2014). A novel framework for termset selection and weighting in binary text classification. Engineering Applications of Artificial Intelligence, 35, 38-53.

Bennasar, M., Hicks, Y., and Setchi, R. (2015). Feature selection using Joint Mutual Information Maximisation. Expert Systems with Applications, 42(22), 8520-8532.

Bennasar, M., Setchi, R., and Hicks, Y. (2013). Feature Interaction Maximisation. Pattern Recognition Letters, 34(14), 1630-1635.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Paper presented at the ACL, 440-447.

Boiy, E., and Moens, M.-F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. Information retrieval, 12(5), 526-558.

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

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2014). Data classification using an ensemble of filters. Neurocomputing, 135, 13-20.

Brank, J., Mladenic, D., Grobelnik, M., and Milic-Frayling, N. (2008). Feature Selection for the Classification of Large Document Collections. J. UCS, 14(10), 1562-1596.

Bryll, R., Gutierrez-Osuna, R., and Quek, F. (2003). Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. Pattern Recognition, 36(6), 1291-1302.

George, S.(2015). 'How Much Data Is Generated Every Minute On Social Media', [Online]. Available at: http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/ (Accessed:21 Dec 2015).

Chen, L.-S., Liu, C.-H., and Chiu, H.-J. (2011). A neural network based approach for sentiment classification in the blogosphere. Journal of Informetrics, 5(2), 313-322.

Cheng, Q., Zhou, H., and Cheng, J. (2011). The fisher-markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data. IEEE transactions on pattern analysis and machine intelligence, 33(6), 1217-1233.

Dash, M., and Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(3), 131-156.

Diao, R., and Shen, Q. (2012). Feature selection with harmony search. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 42(6), 1509-1523.

Ekbal, A., and Saha, S. (2013). Combining feature selection and classifier ensemble using a multiobjective simulated annealing approach: application to named entity recognition. Soft Computing, 17(1), 1-16.

Geem, Z. W., Kim, J. H., and Loganathan, G. (2001a). A new heuristic optimization algorithm: harmony search. Simulation, 76(2), 60-68.

Geem, Z. W., Kim, J. H., and Loganathan, G. V. (2001b). A new heuristic optimization algorithm: harmony search. simulation, 76(2), 60-68.

Godbole, N., Srinivasaiah, M., and Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. ICWSM, 7.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3), 389-422.

Harb, A., Plantié, M., Dray, G., Roche, M., Trousset, F., and Poncelet, P. (2008). Web opinion mining: how to extract opinions from blogs? Paper presented at the Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, 211-217.

Hatzivassiloglou, V., and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. Paper presented at the 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, 174-181.

He, X., Cai, D., and Niyogi, P. (2006). Laplacian score for feature selection. Paper presented at the Advances in neural information processing systems, 507-514.

Hu, D., Kaza, S., and Chen, H. (2009). Identifying significant facilitators of dark network evolution. Journal of the American Society for Information Science and Technology, 60(4), 655-665.

Hu, M., and Liu, B. (2004). Mining and summarizing customer reviews. Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177.

Hu, X., Tang, J., Gao, H., and Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. Paper presented at the Proceedings of the 22nd international conference on World Wide Web, 607-618.

Hu, Z., Bao, Y., Xiong, T., and Chiong, R. (2015). Hybrid filter–wrapper feature selection for short-term load forecasting. Engineering Applications of Artificial Intelligence, 40(0), 17-27.

Hua, J., Tembe, W. D., and Dougherty, E. R. (2009). Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition, 42(3), 409-424.

Huang, Y.-F., Lin, S.-M., Wu, H.-Y., and Li, Y.-S. (2014). Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. Data & Knowledge Engineering, 92, 60-76.

iPerceptions, (2011) iPerceptions Releases Retail / E-Commerce Industry Report Q3 2011. [Online]. Available at: https://uk.finance.yahoo.com/news/iPerceptions-Releases-Retail-iw-1564944333.html (Accessed: 10 March 2013)

Jin, J., Ji, P., and Kwong, C. K. (2015). What makes consumers unsatisfied with your products: Review analysis at a fine-grained level. Engineering Applications of Artificial Intelligence(0).

Kechaou, Z., Wali, A., Ammar, M. B., Karray, H., and Alimi, A. M. (2013). A novel system for video news' sentiment analysis. Journal of Systems and Information Technology, 15(1), 24-44.

Kennedy, A., and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22(2), 110-125.

Kim, S.-M., and Hovy, E. (2004). Determining the sentiment of opinions. Paper presented at the Proceedings of the 20th international conference on Computational Linguistics, 1367.

Korkontzelos, I., Klapaftis, I. P., and Manandhar, S. (2008). Reviewing and evaluating automatic term recognition techniques. In Advances in Natural Language Processing (pp. 248-259): Springer.

Kwak, N., and Chong-Ho, C. (2002). Input feature selection by mutual information based on Parzen window. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(12), 1667-1671.

Li, Q., Jin, Z., Wang, C., and Zeng, D. D. (2016). Mining opinion summarizations using convolutional neural networks in Chinese microblogging systems. Knowledge-Based Systems, 107, 289-300.

Li, S., Wang, Z., Zhou, G., and Lee, S. Y. M. (2011). Semi-supervised learning for imbalanced sentiment classification. Paper presented at the Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three, 1826-1831.

Li, S., Xia, R., Zong, C., and Huang, C.-R. (2009). A framework of feature selection methods for text categorization. Paper presented at the Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, 692-700.

Li, Y., and Lu, B.-L. (2009). Feature selection based on loss-margin of nearest neighbor classification. Pattern Recognition, 42(9), 1914-1921.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

Liu, B., and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining Text Data (pp. 415-463): Springer.

Liu, H., Sun, J., Liu, L., and Zhang, H. (2009). Feature selection with dynamic mutual information. Pattern Recognition, 42(7), 1330-1339.

Liu, L., Kang, J., Yu, J., and Wang, Z. (2005). A comparative study on unsupervised feature selection methods for text clustering. Paper presented at the Natural

Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on, 597-601.

Liu, M., and Zhang, D. (2016). Feature selection with effective distance. Neurocomputing, 215, 100-109.

Liu, Y., Huang, X., An, A., and Yu, X. (2007). ARSA: a sentiment-aware model for predicting sales performance using blogs. Paper presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 607-614.

Mahdavi, M., Fesanghary, M., and Damangir, E. (2007). An improved harmony search algorithm for solving optimization problems. Applied Mathematics and Computation, 188(2), 1567-1579.

Manjarres, D., Landa-Torres, I., Gil-Lopez, S., Del Ser, J., Bilbao, M. N., Salcedo-Sanz, S., et al. (2013). A survey on applications of the harmony search algorithm. Engineering Applications of Artificial Intelligence, 26(8), 1818-1831.

McDonald, R., Hannan, K., Neylon, T., Wells, M., and Reynar, J. (2007). Structured models for fine-to-coarse sentiment analysis. Paper presented at the Annual Meeting-Association For Computational Linguistics, 432.

Moh'd Alia, O., and Mandava, R. (2011). The variants of the harmony search algorithm: an overview. Artificial Intelligence Review, 36(1), 49-68.

Nakagawa, T., Inui, K., and Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 786-794.

Ofek, N., Caragea, C., Rokach, L., Biyani, P., Mitra, P., Yen, J., et al. (2013). Improving Sentiment Analysis in an Online Cancer Survivor Community Using Dynamic Sentiment Lexicon. Paper presented at the Social Intelligence and Technology (SOCIETY), 2013 International Conference on, 109-113.

Omran, M. G. H., and Mahdavi, M. (2008). Global-best harmony search. Applied Mathematics and Computation, 198(2), 643-656.

Oreski, S., and Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications, 41(4), 2052-2064.

Ortigosa-Hernández, J., Rodríguez, J. D., Alzate, L., Lucania, M., Inza, I., and Lozano, J. A. (2012). Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers. Neurocomputing, 92, 98-115.

Padmaja, S., and Fatima, S. S. (2013). Opinion Mining and Sentiment Analysis–An Assessment of Peoples' Belief: A Survey. International Journal.

Pang, B., and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Paper presented at the Proceedings of the 42nd annual meeting on Association for Computational Linguistics, 271.

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., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, 79-86.

Peng, H., and Fan, Y. (2017). Feature selection by optimizing a lower bound of conditional mutual information. Information Sciences, 418-419, 652-667.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., et al. (2012). Sentiment analysis of suicide notes: A shared task. Biomedical Informatics Insights, 5(Suppl. 1), 3.

Prabowo, R., and Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2), 143-157.

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. Computational linguistics, 37(1), 9-27.

Qu, L., Ifrim, G., and Weikum, G. (2010). The bag-of-opinions method for review rating prediction from sparse text patterns. Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics, 913-921.

Rogati, M., and Yang, Y. (2002). High-performing feature selection for text classification. Paper presented at the Proceedings of the eleventh international conference on Information and knowledge management, 659-661.

Rustamov, S., Mustafayev, E., and Clements, M. A. (2013). Sentiment analysis using Neuro-Fuzzy and Hidden Markov models of text. Paper presented at the Southeastcon, 2013 Proceedings of IEEE, 1-6.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.

Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V., and Alonso-Betanzos, A. (2017). Ensemble feature selection: Homogeneous and heterogeneous approaches. Knowledge-Based Systems, 118, 124-139.

Shang, C., Li, M., Feng, S., Jiang, Q., and Fan, J. (2013). Feature selection via maximizing global information gain for text classification. Knowledge-Based Systems, 54, 298-309.

Sharma, A., and Dey, S. (2012). A document-level sentiment analysis approach using artificial neural network and sentiment lexicons. ACM SIGAPP Applied Computing Review, 12(4), 67-75.

Shrestha, K. (2016). '50 Stats You Need to Know About Online Reviews'[Online]. Available at: https://www.vendasta.com/blog/50-stats-you-need-to-know-about-online-reviews/ (Accessed: 5 September 2016).

Stone, Z.(2015). 'A Surprisingly Large Amount of Amazon Reviews Are Fake' [Online]. Available at: http://thehustle.co/a-surprisingly-large-number-of-amazon-reviews-are-scams-the-hustle-investigates (Accessed: 22 December 2015)

Solorio-Fernández, S., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. (2016). A new hybrid filter–wrapper feature selection method for clustering based on ranking. Neurocomputing, 214, 866-880.

Sotoca, M. J., and Pla, F. (2010). Supervised feature selection by clustering using conditional mutual information-based distances. Pattern Recognition, 43(6), 2068-2081.

Speriosu, M., Sudan, N., Upadhyay, S., and Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. Paper presented at the Proceedings of the First workshop on Unsupervised Learning in NLP, 53-63.

Su, Y., Zhang, Y., Ji, D., Wang, Y., and Wu, H. (2013). Ensemble learning for sentiment classification. In Chinese Lexical Semantics (pp. 84-93): Springer.

Tabakhi, S., Moradi, P., and Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence, 32, 112-123.

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

Tan, S., and Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. Expert Systems with Applications, 34(4), 2622-2629.

Tang, H., Tan, S., and Cheng, X. (2009). A survey on sentiment detection of reviews. Expert Systems with Applications, 36(7), 10760-10773.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178-185.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Paper presented at the Proceedings of the 40th annual meeting on association for computational linguistics, 417-424.

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24(7), 1024-1032.

Wang, C.-M., and Huang, Y.-F. (2010). Self-adaptive harmony search algorithm for optimization. Expert Systems with Applications, 37(4), 2826-2837.

Wang, Y., Liu, Y., Feng, L., and Zhu, X. (2015). Novel feature selection method based on harmony search for email classification. Knowledge-Based Systems, 73, 311-323.

Warren Liao, T. (2010). Feature extraction and selection from acoustic emission signals with an application in grinding wheel condition monitoring. Engineering Applications of Artificial Intelligence, 23(1), 74-84.

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

Williams, S. (2013). An Information Extraction System for English Ontology Identifier Names.

Xia, R., Zong, C., and Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6), 1138-1152.

Yang, Y., and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. Paper presented at the ICML, 412-420.

Yano, T., and Smith, N. A. (2010). What's Worthy of Comment? Content and Comment Volume in Political Blogs. Paper presented at the ICWSM.

Ye, L., and Keogh, E. (2009). Time series shapelets: a new primitive for data mining. Paper presented at the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 947-956.

Ye, Q., Law, R., Gu, B., and Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Computers in Human behavior, 27(2), 634-639.

Yu, L., and Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. Paper presented at the Proceedings of the 20th international conference on machine learning (ICML-03), 856-863.

Zhang, C., Wang, H., Cao, L., Wang, W., and Xu, F. (2016). A hybrid term–term relations analysis approach for topic detection. Knowledge-Based Systems, 93, 109-120.

Zhang, D., Chen, S., and Zhou, Z.-H. (2008). Constraint Score: A new filter method for feature selection with pairwise constraints. Pattern Recognition, 41(5), 1440-1451.

Zhang, L., and Liu, B. (2011). Identifying Noun Product Features that Imply Opinions. Paper presented at the ACL (Short Papers), 575-580.

Zhou, S., Chen, Q., and Wang, X. (2013). Active deep learning method for semi-supervised sentiment classification. Neurocomputing.

Zhu, S., Xu, B., Zheng, D., and Zhao, T. (2013). Chinese Microblog Sentiment Analysis Based on Semi-supervised Learning. In Semantic Web and Web Science (pp. 325-331): Springer.