IMPROVING EXPLICIT ASPECTS EXTRACTION IN SENTIMENT ANALYSIS USING OPTIMIZED RULESET

MOHAMMAD AHMAD JOMAH TUBISHAT

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

2019

IMPROVING EXPLICIT ASPECTS EXTRACTION IN SENTIMENT ANALYSIS USING OPTIMIZED RULESET

MOHAMMAD AHMAD JOMAH TUBISHAT

THESIS SUBMITTED AS FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

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

2019

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Mohammad Ahmad Jomah Tubishat

Matric No: WHA160006

Name of Degree: **Doctor of Philosophy (PhD)**

Title of Thesis ("this Work"): **IMPROVING EXPLICIT ASPECTS EXTRACTION IN SENTIMENT ANALYSIS USING OPTIMIZED RULESET**

[Main Supervisor Dr. Norisma Binti Idris (University of Malaya) & Co-Supervisor Dr. Mohammad Abushariah (University of Jordan)]

Field of Study: **ARTIFICIAL INTELLIGENCE**

I do solemnly and sincerely declare that:

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

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

IMPROVING EXPLICIT ASPECTS EXTRACTION IN SENTIMENT ANALYSIS USING OPTIMIZED RULESET

ABSTRACT

Aspect extraction, also known as opinion target extraction, is the fine-grained identification of users' opinion targets, such as the extraction of opinionated product aspects from customer reviews. Aspect extraction is considered as the core task in aspect-based sentiment analysis and other applications. Currently, many studies were conducted using dependency relation rules which give promising results. However, these dependency-based extraction approaches perform better on formal text as its accuracy is based on the dependency parser which gives correct results if the text follow the English rules and grammars. On the other hand, there are also many studies were conducted using sequential syntactic patterns which mimic and follow the ways users expressed their opinion without giving attention to the language rules but give better results on informal text. However, customer reviews normally are a mixed of both types of reviews including formal and informal text. In addition, extraction rules including either pattern-based or dependency-based rules should be selected in a correct way to remove the irrelevant rules and minimize the extraction errors Thus, in this study, to select the most effective extraction rules, an improved version of Whale Optimization Algorithm (IWOA) is developed and applied to a full set of rules. This set of rules includes combination of new created extraction rules with dependency-based rules and pattern-based rules from the previous studies. In addition, the improved WOA is developed by using Cauchy mutation and local search algorithm to solve its local optima problem and improve population diversity. The algorithm was then applied to the full set of 126 rules. Finally, after the aspects list was obtained from the selected rules, a pruning algorithm (PA) is developed to remove the incorrect aspects and retain the correct aspects. Our results from the conducted experiments revealed that the proposed algorithm outperform the state-of-the-art aspect extraction algorithms and optimization algorithms. The IWOA algorithm outperforms other optimization algorithms includes native WOA, PSO, MFO, FFA, GWO, MVO, SSA, and SCA and achieved 86% precision, 94% recall, and 90% F-measure respectively. IWOA superiority resulted because of its ability to escape from local optima and balance between exploitation and exploration. In addition, after application of PA, IWOA+PA outperforms other state-of-the-art aspect extraction works and achieved 92% precision, 93% recall, and 92% F-measure respectively.

Keywords: Sentiment Analysis; Improved Whale Optimization Algorithm; Aspects Extraction. Rules Selection, Pruning Algorithm

PENAMBAHBAIKAN PENGEKSTRAKAN ASPEK NYATA DALAM ANALISA SENTIMEN MENGGUNAKAN SET PERATURAN YANG OPTIMA

ABSTRAK

Pengekstrakan aspek, yang juga dikenali sebagai pengekstrakan sasaran pendapat adalah identifikasi sasaran pendapat pengguna halus , seperti pengekstrakan aspek produk yang dinilai dari tinjauan pelanggan. Pengekstrakan aspek dianggap sebagai tugas utama dalam analisis sentimen berasaskan aspek dan aplikasi-aplikasi lain. Pada masa ini, banyak kajian telah dijalankan menggunakan kaedah hubungan ketergantungan yang memberikan hasil yang menjanjikan. Walau bagaimanapun, pendekatan pengekstrakan berasaskan ketergantungan ini berfungsi dengan lebih baik pada teks formal kerana ketepatannya adalah bergantung kepada penghurai yang memberikan hasil yang betul jika teks mengikuti peraturan dan tatabahasa Bahasa Inggeris.Selain daripada itu, terdapat juga banyak kajian yang dijalankan menggunakan corak sintaksik berurutan yang meniru dan mengikuti cara pengguna menyatakan pendapat mereka tanpa memberi perhatian kepada peraturan bahasa tetapi memberi hasil yang lebih baik pada teks tidak formal. Walau bagaimanapun, ulasan pelanggan biasanya adalah campuran antara kedua-dua jenis ulasan termasuk teks formal dan tidak formal. Di samping itu, peraturan pengekstrakan termasuk sama ada kaedah berasaskan corak atau berasaskan ketergantungan harus dipilih dengan cara yang betul untuk menghapuskan peraturan yang tidak relevan dan meminimakan ralat pengestrakan. Oleh itu, dalam kajian ini, untuk memilih peraturan pengekstrakan yang paling berkesan, versi Algoritma *Whale Optimization* yang telah ditambahbaik *(IWOA)* dibangunkan dan diapplikasikan pada set peraturan penuh. Set peraturan ini termasuk kombinasi kaedah pengekstrakan baru yang dibuat dengan peraturan berdasarkan ketergantungan dan peraturan berasaskan corak dari kajian terdahulu. Di samping itu, WOA yang telah ditambahbaik ini dibangunkan dengan menggunakan mutasi Cauchy dan algoritma carian tempatan untuk menyelesaikan masalah optima tempatan dan meningkatkan kepelbagaian populasi. Algoritma ini kemudiannya digunakan pada set lengkap yang mengandungi 126 peraturan. Akhir sekali, selepas senarai aspek diperolehi dari peraturan yang dipilih, Algoritma Pemangkasan (PA) dibangunkan untuk menghapuskan aspek yang tidak betul dan mengekalkan aspek yang betul. Hasil daripada eksperimen yang dijalankan menunjukkan bahawa algoritma yang dicadangkan mengatasi algoritma pengekstrakan aspek terkini dan algoritma pengoptimuman. Algoritma IWOA mengatasi algoritma pengoptimuman lain termasuk WOA, PSO, MFO, FFA, GWO, MVO, SSA, dan SCA asli dan mencapai ketepatan 86%, 94% bagi *recall*, dan 90% bagi ukuran-F. Keunggulan IWOA adalah disebabkan keupayaannya untuk mengelak daripada optima tempatan dan keseimbangan antara eksploitasi dan eksplorasi. Sebagai tambahan, selepas penggunaan PA, IWOA + PA mengatasi prestasi pengekstrakan aspek terkini dan mencapai ketepatan 92%, 93% bagi recall, dan 92% ukuran-F.

Kata kunci: Analisis Sentimen; Algoritma Whale Optimization; Pengekstrakan Aspek. Pemilihan peraturan, Algoritma Pemangkasan

ACKNOWLEDGEMENTS

Firstly, I would like to thank and express my deep appreciation to my supervisors Dr. Norisma Binti Idris and Dr. Mohammad Abushariah. I really appreciate Dr. Norisma's help, support, and patience during my PhD years. Moreover, her continuous comments, support, encouragement, and guidance have made my PhD journey become easy for me to finish my PhD. In addition, I should thank Dr. Mohammad Abushariah, for helping and supporting me during my PhD study to be in the correct way.

Finally, I should say thanks to my beloved wife Zainab Rawashdeh for her endless encouragement, caring, time, efforts, and hope which she gave to me in difficult days to accomplish my PhD journey. In addition, my heartiest thanks to my beloved daughter Lara who gave me the happiness and cute support during my research journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABSA	:	Aspect Based Sentiment Analysis
ACO	:	Ant Colony Optimization
ANN	:	Artificial Neural Networks
BA	:	Bat Algorithm
BBO	:	Biogeography-Based Optimization
CFO	:	Central Force Optimization
CM	:	Cauchy Mutation
CNN	:	Convolutional Neural Network
CNN+LP	:	Convolutional Neural Networks with Linguistic Patterns
CRF	:	Conditional Random Fields
CSO	:	Cat Swarm Optimization
CSS	:	Charged System Search
DE	:	Differential Evolution
DP	:	Double Propagation
EDR	:	Extrinsic-Domain Relevance
FA	:	Firefly Algorithm
FOA	:	Fruit fly Optimization Algorithm
GRU	:	Gated Recurrent Unit
GSA	:	Gravitational Search Algorithm
GWO	:	Grey Wolf Optimizer
HAC	:	High Adjective Count
HITS	:	Hyperlink-Induced Topic Search
HMM	:	Hidden Markov Models

IDR	:	Intrinsic-Domain Relevance
ILDA	:	Interdependent Latent Dirichlet Allocation
IN	:	Preposition or Subordinating Conjunction
IWOA	:	Improved Whale Optimization Algorithm
IWOA+PA	:	Improved Whale Optimization Algorithm + Pruning Algorithm
KH	:	Krill Herd
LDA	:	Latent Dirichlet Allocation
LRT	:	Likelihood Ratio Tests
LSA	:	Latent Semantic Analysis
LSA	:	Local Search Algorithm
ME	:	Maximum Entropy
MFO	:	Moth-Flame Optimization
MVO	:	Multi-Verse Optimizer
NFL	:	No Free Lunch
NGD	:	Normalized Google Distance
NLP	:	Natural Language Processing
NPs	:	Noun Phrases
OM	:	Opinion Mining
PA	:	Pruning Algorithm
PMI	:	Pointwise Mutual Information
POS	:	Part-of-Speech
PSO	:	Particle Swarm Optimization
PSWAM	:	Partially Supervised Word Alignment Model
RO	:	Ray Optimization

RSLS	:	Rule Selection using a Local Search Algorithm
SA	:	Sentiment Analysis
SA	:	Simulating Annealing
SCA	:	Sine Cosine Algorithm
SSA	:	Salp Swarm Algorithm
SVM	:	Support Vector Machines
TF-RBM	:	Two-Fold Rule-Based Model
VPs	:	Verb Phrases
WOA	:	Whale Optimization Algorithm
WTM	:	Word Based Translation Model

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Introduction

The emergence of different online shopping websites have changed our daily life habits by using e-commerce websites. These websites such as Amazon, Taobao, and many more allow people to post and share their reviews and feedbacks about the purchased products (Liu et al., 2017). Many customers refer to these reviews before making their buying decision, where about 70% of the customers find these reviews as one of the major trusted sources to make decisions. On the other hand, manufactures use these reviews to analyze the defects of their products, as the pre-discovery of product defects can maximize the reputation of the manufacturers, which in turn increase their sales (Law et al., 2017). In addition, manufacturers used these reviews to explore the most important product aspects by customers. According to Akalamkam and Mitra (2018), the pre-purchase decision of huge numbers of customers are based on conducting pre-search through online reviews provided by previous customers. In another study by Zhao et al. (2018), about 90% of consumers depend on online reviews for making purchase decision. This habit is now considered as one major step by customers before deciding to buy any products online. However, the process of reviewing and retrieving the reviews by the previous customers are very time consuming and cumbersome due to the huge volume of information provided online.

Sentiment analysis (SA) or also known as opinion mining (OM) is a computerized process of extracting user opinion and determine their semantic orientation into either negative, positive, or neutral opinion (Hu & Liu, 2004a). SA can be performed at different levels such as document level, sentence level, or aspect (feature) level. In document and sentence levels, the task is based on extracting all opinion words from a sentence or document, then it gives the opinion of the whole document or sentence, whereas, aspect level SA, concerns with the extraction of product aspects and the specific opinion words mentioned in the reviews toward each aspect (Rana & Cheah, 2016). As a conclusion, the analysis for document and sentence levels are based on the overall document or sentence without considering specific analysis of each aspect while the analysis in aspect level SA considers each aspect, where it associates each aspect with their related opinion words.

One of the most important tasks in aspect level SA is the aspect or feature extraction step (Rana & Cheah, 2016) where the aspect can be either explicit or implicit. The aspect is considered as explicit if it is mentioned explicitly in the text. For example, in the following review, "*the **screen** is very nice*", the aspect ***screen*** is mentioned explicitly. On the other hand, if the aspect is mentioned implicitly in a text, it is known as implicit aspect. For example, a review about a handphone, "*the phone is **expensive***", where the opinion word ***expensive*** is an indicator for implicit aspect, ***price*** (Hu & Liu, 2004a).

Several studies have been conducted for explicit aspect extraction. The extraction techniques used in the previous studies can be classified as either unsupervised (Hu & Liu, 2004a; Popescu & Etzioni, 2007; Moghaddam & Ester, 2010; Htay & Lynn, 2013; Hai et al., 2014; Poria et al., 2014), semi-supervised (Wu et al., 2009; Wei et al., 2010; Qiu et al., 2011; Yu et al., 2011; Liu, Xu, & Zhao, 2013; Xu et al., 2013; Samha et al., 2014; Yan et al., 2015), or supervised (Jin et al., 2009; Choi & Cardie, 2010; Jiang et al., 2010; Chen et al., 2012; Li et al., 2012; Cruz et al., 2013). In unsupervised approach, no training data is required for the aspect extraction. In the semi-supervised approach, a little training data is required to accomplish the extraction task. Finally, the supervised approach is based on the availability of labeled data for training the used extraction algorithm (Rana & Cheah, 2016).

This research mainly focuses on improving explicit aspect extraction algorithm, which will be used for extracting the opinionated explicit aspects from customer reviews of electronics products. This explicit aspect extraction algorithm can be used in sentiment analysis systems for doing this task at aspect level.

1.2 Research Motivation

This research focus on explicit aspect extraction in SA of customer reviews about electronic products due to the following reasons:

- **Customers**: From customers point of view, most online shopping websites such as Amazon provides a space for customers to write and share their reviews about their experiences with the products they bought. These reviews become an interest of coming customers. For customers, before deciding to buy a specific product, they normally would read these reviews to minimize the risks related to a decision of purchase (Maslowska et al., 2017). When customers search for information about the items they want to buy, they focus on the previous customers' reviews rather than the information provided by the product firm or manufacture, as they consider the information provided by previous customers are more reliable and enjoyable (Blazevic et al., 2013).
- **Manufacturers**: From the manufacturers side, these online reviews have considerable impact on customer's decision for either to purchase a product or not, and on product sales (Zhu & Zhang, 2010). Another important factor for manufacturers is the customer's satisfaction, which is considered as a measure of how much the products or services of the manufacturer or company meet the customer requirements and expectation. This satisfaction measure can be used to forecast customer's tendency to return and buy from the same company or

manufacturer (Moghaddam, 2015). In addition, manufacturers can analyze these online reviews to find the defects and weaknesses in their products according to customers experiences and feedbacks with their products, and to improve the quality of their products and build better business plan (Zhang et al., 2012).

Previous studies showed that the aspect extraction phase plays an important role in many applications, such as in recommender system to recommend a specific aspect of product or service according to users' needs (Bauman et al., 2017), product and feature ranking applications (Yu et al., 2011; Liu et al., 2017; Xue et al., 2017; Zhou & Zhang, 2017), and in aspect retrieval system (Caputo et al., 2017).

Furthermore, the extraction of various products or services aspects required by customers to match with products or services according to their aspect preferences (Li, Zhou, et al., 2015). In addition, a study conducted by Du et al. (2015), found that customers search trends about specific features of the product will give the manufacturers the opportunity of monitoring customers search trends for specific features, which allow the manufacturers to come with the best decisions. Moreover, aspect extraction task plays an important role in aspect-based sentiment analysis (ABSA) (Marrese-Taylor et al., 2014; Poria et al., 2016; Akhtar et al., 2017).

As a conclusion, aspect extraction plays an important role in many applications including SA. Thus, it motivates this research to be conducted as a contribution and improvement to these research fields.

1.3 Problem Statement

The rapid and vast growth of internet platforms allowed customers to share their reviews with other customers. However, the huge volume of available reviews have changed the life styles of many customers (Asghar et al., 2017; Musto et al., 2017; Rana & Cheah, 2017). High percentage of customers spend long time searching for the available online reviews about electronic product before deciding to buy it. A study conducted by Akhtar et al. (2017) found that customers spend on average eight hours daily on different types of social media applications and internet.

The online reviews shared by customers based on their previous experiences of different products or services has become an important source of information for prospective customers. These reviews were used by either individual customers or manufacturers to help them in decision making (Rana & Cheah, 2017). However, huge volume of available reviews for each product or service result in the difficulty and impossibility of reading all of them by individual customer or manager to make their decisions (Ravi & Ravi, 2015). As highlighted previously, SA was developed to solve these problems and it can be performed at different levels including aspect level. One of the most important type is ABSA as it gives users deeper insights at fine grained level. According to a survey conducted by Rana and Cheah (2016), the major part of this SA is the aspect extraction phase. Recently, several research have been conducted for aspect extraction, and they can be organized into three approaches which are unsupervised, semi-supervised, or supervised approach (Rana & Cheah, 2016).

As mentioned before, different works have been conducted in literature to extract explicit aspects. Many of the conducted methods used either dependency-based approach such as (Qiu et al., 2011; Kumar & Raghuveer, 2013; Poria et al., 2014; Liu, Gao, et al., 2016; Kang & Zhou, 2017) or patterns-based approach as in (Htay & Lynn, 2013; Maharani et al., 2015; Asghar et al., 2017; Rana & Cheah, 2017). The extraction types which based on pattern-based or dependency-based rules give promising results.

The advantages and weaknesses of dependency-based extraction approaches can be summarized as follows.

The main advantage of dependency-based extraction approaches is its ability to give accurate results if the reviews follow the English rules. Another advantage of dependency-based approach is that it is better for long sentences, as we cannot make specific pattern for each long sentence existed in the reviews. The main weakness in the previous works which used dependency-based extraction is that many extraction rules were not explored. In addition, some extraction rules used by the previous studies, may lead to extraction of many non-aspect words and these extracted non-aspect words can decrease the performance of the SA.

Besides than that, dependency-based extraction has the problem of generating error results since the dependency relations accuracy is based on the grammatical correctness of the reviews. However, not all online reviews follow the English grammars rules and these reviews usually are mixed types of informal and formal styles (Zhang et al., 2010; Hai et al., 2012; Liu, Xu, & Zhao, 2013; Liu, Xu, et al., 2015; Rana & Cheah, 2017). In addition, there is no restriction from the website on customers to follow language rules when they post their reviews. Customers may follow the language rules in writing and sometimes may violates some of them (Rana & Cheah, 2017). In some cases, dependency-based approach cannot be used because none of the existing dependency rules match with the exist sentence. The other weakness of dependency-based extraction is that it worked on extracting relations between single words, therefore it ignores the syntactic categories and local structure (Wu et al., 2009).

One of the most famous approach used for aspect extraction which based on dependency-based rules is double propagation (DP). DP is based on bootstrapping approach which use the known aspect or opinion words to extract other mentioned aspects or opinion words. However, DP works well for a dataset of medium size, but for large size datasets, it will extract many incorrect features. The main reason of generating many incorrect features is extracting many non-opinionated adjectives as opinion words. These non-opinionated adjectives will be used for extracting aspects, which result in extracting many noisy aspects and vice versa (Zhang et al., 2010). According to Liu, Gao, et al. (2016), not all dependency relations rules are of equal quality, where some rules combinations are good and other rules are not effective for extraction. In addition, not all dependency relation rules always provide good results and some of these rules are irrelevant (Liu, Gao, et al., 2016).

On the other hand, pattern-based extraction approach has been used in many studies for explicit aspects extraction. One of the main advantage of pattern-based approach is that it mimics the ways users write their reviews, which make it a good choice for informal text (Rana & Cheah, 2017).

However, the disadvantages of pattern-based approach include the following. Pattern-based approach is similar to dependency-based approach where it cannot be used in some cases as no pattern match with the review which contains the aspects to extract. However, these works do not cover all possible extraction rules which cover all extraction cases. In the previous studies, some pattern-based rules were effective for aspects extraction, but other patterns were not effective in aspect extraction as it resulted in extracting many non-aspect words (Maharani et al., 2015).

As a conclusion, dependency-based approach gives good results in formal written reviews, but pattern-based approach gives better results in informal written reviews. However, some reviews are written formally, some are written informally, and many reviews may contain a combination of formal writing and informal writing. Based on the previous discussion, pattern-based approach has some advantages that cannot be found in dependency-based approach, and vise-versa. However, each approach also has several weaknesses. To take the advantages of both approaches at the same time a combination of pattern-based approach and dependency-based approach is proposed. Which can increase the recall but may also decrease the precision since there are many irrelevant rules in each approach. Thus, proper selection of these rules' combinations must be carried out to select the optimal combinations.

Therefore, to make a proper selection from these aspects' extraction rules, an optimization algorithm can be used. One of the most recent developed optimization algorithms which proved its ability in solving number of applications is the Whale Optimization Algorithm (WOA). WOA mimics the hunting mechanism used by humpback whales to catch their preys (Mirjalili & Lewis, 2016). Based on a number of studies, WOA has been applied to different application areas and gives promising results in various fields as energy applications (Zhao et al., 2017; Chen et al., 2018; Saha & Saikia, 2018), solving computer networks problems (Ahmed et al., 2017; Jadhav & Shankar, 2017; Parambanchary & Rao, 2018), feature selection problem (Sayed et al., 2016; Zamani & Nadimi-Shahraki, 2016; Canayaz & Demir, 2017; Sharawi et al., 2017; Eid, 2018; Hegazy et al., 2018; Mafarja & Mirjalili, 2018; Tubishat, Abushariah, et al., 2018; Hussien et al., 2019), image processing application (El Aziz et al., 2017; Mostafa et al., 2017; El Aziz et al., 2018; Hassan & Hassanien, 2018), to find the optimal weights of the neural network (Aljarah, Faris, et al., 2018), spammer identification problem (Ala'M et al., 2018), clustering application (Nasiri & Khiyabani, 2018), neural network training (Bhesdadiya et al., 2016), classification application (Tharwat et al., 2017; Karlekar & Gomathi, 2018), for improving power system (Sahu et al., 2018), for solving constrained economic load dispatch problems (Kumar et al., 2018), requirements prioritization problem (Alzaqebah et al., 2018), finding the minimal cost of network (Ghahremani-Nahr et al., 2019), finding the concrete columns compressive strength (Moodi et al., 2018), design problem (Mukherjee et al., 2017), feature extraction in modulation signal (Miao et al., 2019), parameter optimization (Sai & Huajing, 2017), antenna design (Yuan et al., 2017), speed prediction of wind (Osama et al., 2017), link prediction (Barham & Aljarah, 2017), and for finding the optimal planning for robot path (Dao et al., 2016).

Based on these previous works, WOA was successfully applied in several fields and has proved its suitability and superiority to solve these problems. In addition, WOA is featured with its efficiency and simplicity (Mirjalili & Lewis, 2016). This motivated the current research to apply WOA for rules selection.

Despite that the standard WOA show competitive results as shown in previous works in comparison to other optimization algorithms. WOA as other optimization algorithms may stuck in local optima and has a problem of solutions diversity based on the problem to be solved. In addition, based on No Free Lunch (NFL) theorem which states that a specific optimization algorithm cannot solve all types of problems and at the same time outperforms all other optimization algorithms in all these problems (Wolpert & Macready, 1997).

To further improve WOA performance, the current study provided two major improvements into the original WOA to solve its weaknesses and make it suitable for the current research problem. The first improvement is to use Cauchy mutation (CM) to improve the solutions diversity (the rules selection) and make balance between its exploration and exploitation. The second improvement is to use local search algorithm (LSA) to solve its local optima problem. LSA will be used to check if there is a better solution than the current best solution by select and deselect rules in the solution based on precision and recall values of included rules in the solution. In addition, these included improvements are used to further improve its performance in aspect extraction.

Besides than that, many previous works are only based on frequency of the terms in the corpus (Qiu et al., 2011; Kang & Zhou, 2017; Rana & Cheah, 2017) that may pruned low frequent aspects. In a study by Kang and Zhou (2017), they used pruning based on semantic meaning using similarity based on Wordnet. However, not all words can be found in Wordnet or sometimes an aspect is in reality part of the product but not semantically related. In addition, pruning in a work conducted by Rana and Cheah (2017), used normalized google distance (NGD) between the candidate aspect and main entity, but the cooccurrence returned by google is not reliable as the two words occurred in same page may be very far from each other and not related. Which shows the unavailability of good pruning algorithm.

As a summary, the available customer reviews on internet shopping websites are composed of mixed types of structured and unstructured text. For structured text the dependency-based rules give better results, while for unstructured text the pattern-based approach gives better results. Thus, to cover both text types a combination of dependency and pattern rules is proposed by using rules from the previous studies with some new rules. However, not all these combined rules are of equal quality because using all these rules together will increase recall but decrease precision. Therefore, a proper selection of these rules will be required to select optimal rules combination. The proper selection of these rules can be accomplished via IWOA. In addition, the unavailability of suitable pruning algorithm in literature to remove incorrect aspects and keep correct pruned aspects. Thus, to improve the current pruning algorithm, a three-phase pruning is proposed. This three-phase pruning algorithm is based on using frequency, product manual, and direct opinion association.

1.4 Aim and Objectives

The main aim of this research work is to improve explicit aspect extraction algorithm, which can be used to extract the opinionated explicit aspects in products reviews. To fulfill this aim, the primary objectives of this research are specified as follows:

1. To develop a set of aspect extraction rules based on dependency-based or pattern-based rules.
2. To improve WOA using LSA (Local Search Algorithm) and CM (Cauchy Mutation) for rules selection.
3. To improve pruning algorithm using product manual and direct opinion association.
4. To evaluate the developed algorithms using performance metrices including precision, recall, and f-measure.

1.5 Research Questions

RQ1) How to overcome the problems of the current extraction rules?

RQ2) How to select the best rules for explicit feature extraction and discard worst rules?

RQ3) How can incorrect extracted aspects be discarded while retaining the correct aspects?

RQ4) How is the performance of the proposed algorithms compared to baseline algorithms?

1.6 Research Methodology

To carry out the research work, the research methodology process is divided into four phases, as shown in Figure 1.1 and the details of the phases are as follows:

1) **Phase 1 - Identification of the Problem:** In this phase, the research starts by conducting a comprehensive literature review of the previous works on explicit aspect extraction algorithms to identify the current problems. Then, the problem statement will be identified and formulated before the research aim, research objectives, and research questions will be specified.

2) **Phase 2 – Rules and Algorithms Formulation and Development:** This phase starts with analyzing the problem statement to determine all factors that lead to the problem. Then, the proposed solution to accomplish each of the main objectives will be formulated as algorithms. To fulfil the main aim of this research work, three algorithms are proposed. The first algorithm can be achieved by the combination of different rules types which are pattern-based rules from the previous studies, dependency-based rules from the previous studies and the new developed rules. These full set of rules are combined to take their advantages. In addition, some new rules are formulated as there are some cases that cannot be extracted by the old

rules. In addition, the new rules are developed to improve the precision of aspect extraction.

Furthermore, there are restrictions which are combined with the new rules to extract only the target opinionated aspect and not any aspect in sentence. There are many rules in literature which can extract aspect, but sometimes cannot extract the opinion target aspects. In addition, in majority of the studies, they consider opinion words either adjective only or the opinion word from opinion lexicon, but did not consider both at the same time. However, in this research to take advantage of both concurrently, the opinion word is considered opinion word if its type is adjective or found in opinion lexicon as sometimes there are opinion words that are not adjective, but can be found in opinion lexicon. The second algorithm is accomplished by the development of the rule selection algorithm, which will be used to select the useful extraction rules from the whole set of rules and discard the worst rules. This algorithm is used to improve aspect extraction performance by selecting the best rules based on precision and recall combinations using F-measure as objective function to make balance between recall and precision. Thus, after the application of optimization algorithm on full set of rules it will select optimal rules combination and discard rules that give many incorrect aspects and decrease the extraction performance. Finally, the third algorithm called aspect pruning algorithm, will be used to filter and remove incorrect extracted aspects.

3) **Phase 3 – Evaluation of the Algorithms:** To evaluate the proposed algorithms, several experiments will be conducted on customer review datasets created by Hu and Liu (2004a), which used by most studies on explicit aspect extraction. In addition, these experiments will also be conducted to determine the effectiveness

of the proposed algorithms. In the first experiment, the full sets of rules will be used for aspects extraction. In the second experiment, the proposed Improved whale optimization algorithm (IWOA) will be used to select the optimal rules combination and apply these rules for aspect extraction. In addition, IWOA results will be compared by original WOA and other recent and famous optimization algorithms. Furthermore, several experiments will also be carried out to compare the performance of the proposed algorithms with the most recent and famous aspect extraction. The performance metrices used in the experiments include precision, recall, and f-measure.

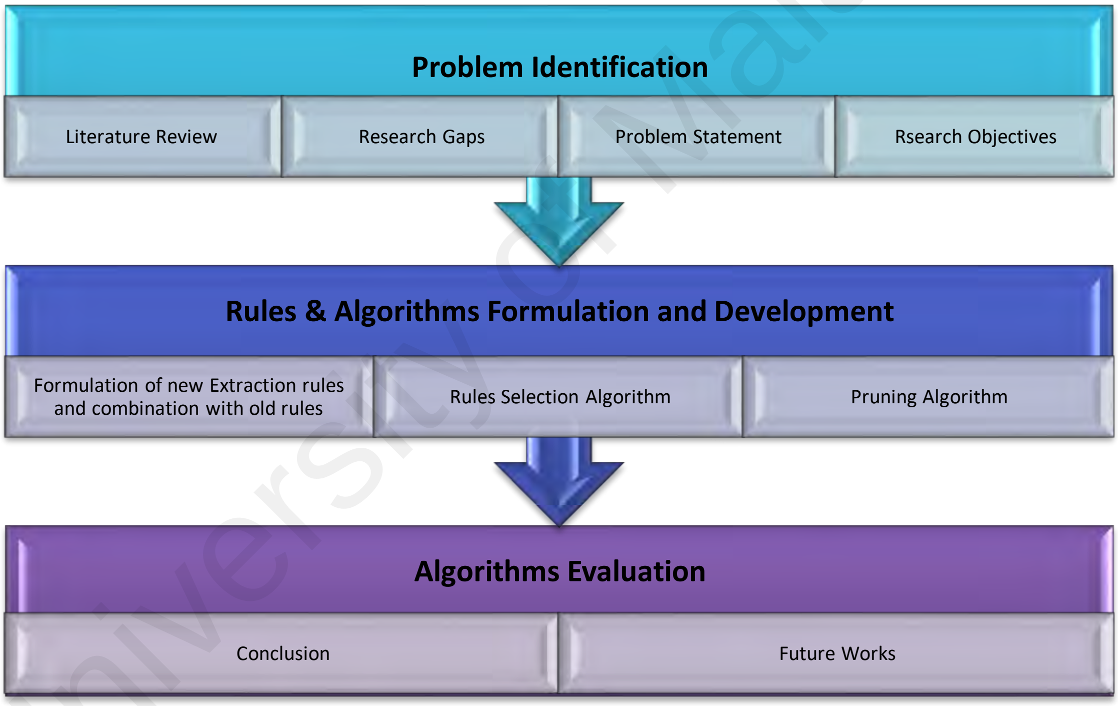

Figure 1.1: Research Methodology

1.7 Thesis overview

This thesis is organized as follows:

- Chapter 1: This chapter gives a brief introduction about the research topic, research motivation, details about the problem statement, the objectives of this work, the research methodology on how to conduct this research work, and the research questions.

- Chapter 2: This chapter starts with a basic introduction about SA and its commonly used terminologies. In addition, it gives some examples of SA applications and details about opinion types. Furthermore, it provides a brief overview about different SA levels. It describes the three types of aspect extraction methods and the details of works conducted in each type. It also provides the limitations of each type. It presents a brief introduction about implicit aspect extraction methods types. Finally, a review on related dependency and pattern-based approaches.

- Chapter 3: This chapter provides the details about the aspect extraction rules and the preprocessing employed in this work. For each rule details of how it works with an example from the used benchmark dataset were discussed.
- Chapter 4: Chapter 4 provides details about WOA algorithm, LSA algorithm, and CM. It provides details about the improvements carried out on standard WOA which is called IWOA. Furthermore, it gives a detail about the pruning algorithm with two improvements which are pruning based on product manual and direct opinion association.
- Chapter 5: This chapter presents the details about the used datasets and the baseline methods. In addition, it provides details about different experiments conducted and the results of each experiment.
- Chapter 6: In this last chapter, it gives an overall conclusion about the thesis. In addition, it provides a summary about thesis contribution and also suggests possible future works.

CHAPTER 2: ASPECTS EXTRACTION IN SENTIMENT ANALYSIS

2.1 Introduction to Sentiment Analysis

With the wide spreading of social network applications as (Facebook, Twitter and Instagram), online discussion forums, and ecommerce websites as (Amazon and Agoda) that allows consumers to post their reviews. Theses reviews include users' opinions or experiences towards a product, hotel, services, events, issues, etc. Furthermore, there are increase demands on the analysis of these different types of opinions either by individual customers or companies. However, the process of manually searching and extracting the relevant piece of information from these reviews are cumbersome and very time consuming. Moreover, these difficulties faced by either individuals or companies to get the valuable information, result in ambiguity for customers or businesses to make their decisions. Thus, an automatic system to extract required information from these huge reviews, process it, and analyze its opinion orientation is needed. These requirements direct the researches community to a new research area called Sentiment Analysis (SA) or Opinion Mining (OM). SA is defined as the automatic analysis of people's opinions, emotions, stances, and sentiments from the written reviews(Liu, 2012). SA involves applying natural language processing (NLP) tasks such as feature extraction, opinion expressions extraction, and opinion expressions classification to find writer attitude (Yadollahi et al., 2017).

Opinion (or Sentiment): opinion or sentiment are synonyms of each other in SA. In SA, opinion is judgement expression used by opinion holder towards an entity or aspect which represents his/her stance (Liu, 2012). According to Liu (2012) the opinion can be composed of two main components, which are the opinion target and the opinion (or sentiment) itself. The representation of opinion as (g, s), where g represents aspect, feature, or entity, while *s* represents the opinion (sentiment) expressed toward *g*. The opinion *s* can be either numeric rating score, negative, positive, or neutral. These values of *s* which represented by numeric rating score, negative, positive, or neutral have many synonyms which can be called sentiment orientations, opinion orientations, sentiment polarities, or opinion polarities. For example, the opinion word in Example 1 is *'large',* while the opinion target is *'screen'*.

Example 1: *'The screen is large'*

2.2 Sentiment Analysis Applications

With the wide coverage of internet and the dramatic increase of its use, different type of websites or social media applications are available that allow user to post their reviews to the public. Thus, the huge volume of publicly available reviews with different interest types motivates many researches to invest SA in their applications. The following discussions give some examples of various applications of SA.

2.2.1 Product Improvement

In a work conducted by Das et al. (2014) , SA was utilized to monitor user reviews on Samsung Galaxy, they collect the reviews from Twitter and analyze it using SA to help the management. Companies use the results of SA to make an informed decision toward enhancing the product according to customer reviews.

2.2.2 Stock Market Price Prediction

In a research by Nguyen et al. (2015) , SA was applied for the problem of stock price prediction by combining the users' sentiments posted on different social media to monitor and predict stock price changes. In their study they only extracted sentiments which expressed to topics related to the company services or products.

2.2.3 Transportation Traffic Problem

Another application of SA was to solve transportation problem by proposing traffic sentiment analysis system, which able to extract and process the user reviews about traffic problems (Cao et al., 2014). The results of SA on traffic information can support safety and information sharing among users which provided by the intelligent transportation system.

2.2.4 Express Industry

A study by Zhang et al. (2017) applied SA to express industry reviews by applying the proposed technique to a data collected from reviews about five express enterprises. Then classifying the reviews according to four basic aspects in this field including price, service, speed, and security. After that, classifying the reviews about each aspect using SA to help these enterprises improve their services and to reduce their weakness.

2.2.5 Learning Applications

In a work proposed by Barrón-Estrada et al. (2017), SA was adopted in Intelligent Tutoring System. To solve the problems of using questioners' surveys they applied SA on the students reviews about the course contents regularly. From the results they look at negative feedbacks and use it to improve the course materials. In Ortigosa et al. (2014) study, SA was also employed in the e-learning field. The implemented application was called SentBuk which retrieved students posted messages on Facebook, and from these messages, only topics related to teaching materials were selected. Then, SA was used to classify these messages into either positive or negative messages. This analysis was used to help students and teachers. In case of students, the system can adaptively change their practices and activities provided to each student according to SA results of his/her feedbacks, while the teachers can monitor these SA analysis results and improve the contents of eLearning courses or change their teaching methodology based on the SA results.

2.2.6 Public Service

Citymis OpTree, a system which was developed by Gamboa et al. (2016), where SA was employed in the system. In this system, the main idea is to collect citizens' reviews about services provided by municipalities, then analyze these reviews and response quickly to citizens request. The advantages of using this technique were to reduce reply time and improve communication quality with citizens. In the constructed system, the reviews about each specific service are organized as a tree of opinions. For example, if the management sent a query to the system about street light system, the system search for opinion tree associated with the received query. In the next step, the system uses SA to analyze the negative and positive reviews associated with each node using naïve Bayes classifier. Finally, the result of SA was used by the government for decision making.

2.2.7 Medical Applications

In Ceyhan et al. (2017) study, SA was applied in medical domain to measure patients satisfaction about health services in private hospitals in Turkey. They collected the negative and positive reviews posted by patients about health services provided by hospitals. Four classifiers were used to classify these reviews into either positive or negative. In addition, the results of SA analysis can be used by medical experts to fulfil patients' requirements according to their feedbacks and improve the services by building development strategy according to the reviews analysis results. In another work by Gopalakrishnan and Ramaswamy (2017), they invested SA idea to analyze patients' satisfaction about the drugs they received. The proposed work mainly based on collecting reviews about patients' experiences with drugs. They analyzed the effects of these drugs side effects using SA by applying neural network on these reviews to classify it as negative or positive reviews. In a study conducted by Zucco et al. (2017), they take the benefits of using SA in patients' depression detection and monitoring. The main target of the developed system was to support medical experts for diagnosing patients' depression. Furthermore, it can be used to help the doctors to monitor the progress of patients' treatments and at the same time the patients can be updated with his/her medical status. They hybridized the idea of affective computing with SA to develop the system. Affective computing can be used to detect and collect required data from patients through their mobile phones. From the collected data they used SA to analyze patient mood. Based on the results of mood analysis the doctor can decide which treatment should be given to the patient and whether to follow the progress of medication on that patient or to change the given medication.

2.2.8 Political Election Prediction

Almatrafi et al. (2015) study applied SA to predict the election results using location-based SA. 650,000 tweets about political teams in India were collected and Naive Bayes (NB) classifier was used to classify these tweets into positive and negative reviews. Based on the number of positive and negative reviews directed toward each team, the system can determine which team is closer to success in the election.

2.3 Opinion Types

Opinion types can be either regular opinion or comparative opinion based on the type of opinion words used. Details of each type are discussed as follows:

2.3.1 Regular Opinion

In literature, regular opinion refers to the term opinions and it has two types which are direct opinion (explicit opinion) or indirect opinion (implicit opinion). In direct opinion, the opinion is mentioned and expressed towards the aspect or feature directly. For example, in a sentence "*The camera resolution is clear*", the opinion word *clear* is a positive opinion which referred to the aspect *camera resolution* directly. On the other hand, in the indirect opinion, the opinion is mentioned and expressed towards the aspect (or feature) indirectly. An example of indirect opinion is as in the sentence "*After installing this program, my phone get much slower*". In this example, *slower* is an opinion word which is directed towards *phone* directly, but in this example the *program* cause the *phone* to become *slower* which represents an example of indirect opinion (Liu et al., 2011).

2.3.2 Comparative Opinion

Comparative opinion is expressed by opinion holder as a type of difference or similarity relation between compared entities based on a given aspect (or feature) shared between these entities according to opinion holder preferences. For example, in the following sentence "*Nikon is cheaper than Canon*", two entities are compared which are *Nikon* and *Canon*. The entities are compared based on the implicit aspect *price* which indicated by the opinion word *cheaper* (Jindal & Liu, 2006b, 2006a).

2.4 Sentiment Analysis (SA) Levels

The task of SA can be carried out at different levels based on the degree of granularities. In general, SA levels include document, sentence, and aspect levels.

2.4.1 Document Level

At document level, the task of SA is to consider the whole document as a single unit and to classify it into either negative or positive sentiments based on the opinion words contained in that document. For example, if the document contains a review about Nokia phone, in this case SA will find the overall polarity of that document regardless the individual sentences or aspects. Document level SA is based on the idea that the document expressed opinion about only one entity. Moreover, as the main idea behind document level SA is based on one entity which an opinion expressed inside the document and the result of SA in document level can be either negative, positive, or neutral.

However, the same document may contain a variety of different opinions words which are expressed towards a number of different products, thus applying document level SA in this case is not realistic and biased. The main problem of this type of SA is that it does not consider details of a review, as the review may contains positive opinion towards specific features of the product and at the same time it may contain negative opinions towards other features of the product. In addition, it does not give detail explanation about fine grained details because it considers the document as whole.

For example, in Moraes et al. (2013)work they compared between Support Vector Machines (SVM) and Artificial Neural Networks (ANN) in finding SA at document level over datasets of different domains such as Movies, Cameras, Books, and GPS. Another example was a work by Tripathy et al. (2017), where they applied SVM for feature selection, then they used these features for ANN classifier to perform SA at document level on reviews about IMDb datasets.

2.4.2 Sentence Level

Sentence level considers the sentence as a single unit like a small document and try to determine its orientation as either negative, positive, or neutral. Sentence level SA mainly involves two important tasks including subjectivity classification and opinion classification. Subjectivity classification used to classify the sentence as either opinionated or not-opinionated sentence. Opinion classification as a next step after subjectivity classification, which mean if the sentence is opinionated (subjective) it will determine its polarity orientation. Furthermore, sentence level SA also shares the same problem as document level SA as it does not consider each individual feature mentioned in the sentence as a separate case and does not give more detail about each aspect. Similar to document SA, it determines the overall sentiment based on the opinion words found in the sentence. There are a few studies conducted on the area of sentence level SA such as (Arulmurugan et al., 2017; Fu et al., 2017; Wu et al., 2017).

Sentiment Orientation: is the measure of opinion word polarity as either positive or negative according to the word direction evaluation. Which includes the determination of word polarity orientation whether the extracted opinion word imply positive or negative attitude. For example, the opinion word "*bad*" implies negative opinion, while the opinion word "*good*" implies positive opinion (Hatzivassiloglou & McKeown, 1997; Turney, 2002; Turney & Littman, 2003; Esuli & Sebastiani, 2005).

Subjective sentence: is a sentence that presents evaluations and opinion expressions (Hatzivassiloglou & Wiebe, 2000). For example, the sentence '*the car is expensive*' is an example of subjective sentence.

Objective sentence is a sentence that presents objective (factual) information (Hatzivassiloglou & Wiebe, 2000). For example, the sentence '*the car color is white*' is an example of objective sentence as it represents a fact with no opinion words mentioned or implied.

2.4.3 Aspect or Feature Level

As mentioned previously in sections 2.4.1 and 2.4.2, the main problem of both sentence and document levels SA is that they are coarse-grained and give result as one word either negative, positive, or neutral for the overall sentence or document. In addition, both types consider the whole document or sentence as expressing opinion about one object. However, the document or sentence may be about specific product like digital camera, but this document may contain details about this camera and different opinions towards its features such as its picture, screen, resolution, and many more. In these days, customers not only concern about general opinion of a product but they also need deeper insight at fine-grained level (aspect level) of the product. For example, a customer is looking for the picture quality of the camera to be the best, while another customer is looking for huge memory size. Thus, if document or sentence level are applied on this case, it will not give these customers the required details as it gives an opinion about the camera as overall, but not the opinion about each feature separately. In addition, if the opinion obtained from applying the document or sentence level is positive, this does not mean that all features are also positive. There may be some negative opinions words expressed towards some features and also positive opinion expressed towards other features. In addition, at manufacturer and administration levels, using either sentence or document levels is not enough as they do not know what aspects (features) that customers like or do not like in order to improve their products or services.

Therefore, ABSA can be used to solve these problems where it can perform a level of analysis lower than document and sentence levels by analyzing customers opinions expressed to each individual aspect of the product. Furthermore, the main aim of ABSA is to extract all relevant aspects of the given products together with the opinion words expressed towards each aspect. After the extraction, ABSA will determine the sentiment polarity of each aspect. In general, ABSA comprise two major tasks, which are aspect extraction and aspect opinion identification (Liu, 2012).

- **Aspect Extraction**: It is the most important task of ABSA which involves extracting the aspects (or features) of the given entity such as camera to which the customer expresses his/her opinion. For example, in the sentence "*The screen of Canon camera is good, but its battery is bad*" the ***screen*** is a feature of ***camera*** entity with an opinion word ***good*** while the other ***camera*** feature is ***battery*** with ***bad*** opinion word. In literature, approaches for conducting explicit aspect extraction techniques are classified into supervised, unsupervised, and semi-supervised (Rana & Cheah, 2016). The importance of aspect extraction not only important for decision makers, but also can be used in other systems such as recommender systems, search engine systems, ranking systems, review helpfulness ranking, and many more.

  Aspect (or Feature): is sub parts, attributes, or function of a given entity (or object) (e.g., battery aspect of phone entity) and this aspect can be explicit or implicit(Liu, 2012).

  1. **Explicit Aspect**: The aspect that mentioned explicitly in the reviews is called explicit aspect and normally can be either noun or noun phrase (Liu, 2012). (e.g, the aspect

button is mentioned explicitly in the sentence *'the camera button is small'* and in this case *button* is an example of explicit aspect).

2. **Implicit Aspect**: The opposite of explicit aspect is called implicit aspect in which the aspect is not mentioned explicitly in the reviews, but normally implied by an opinion word (Liu, 2012). (e.g, the aspect *size* is not mentioned explicitly in the sentence '*the camera button is small'*, but the implicit aspect *size* was implied by *small* opinion word).

- **Aspect Opinion Identification**: In this task, ABSA finds the sentiment orientation for each extracted feature as negative, positive, or neutral based on the opinion words related to each feature. In addition, ABSA will give sentiment to each unique feature. Furthermore, according to ABSA when opinion identification will be applied to previous example, then it will show a positive opinion expressed towards the ***screen*** feature of ***camera*** entity, also a negative opinion expressed towards the ***battery*** feature (Liu, 2012).

2.5 Aspects Extraction Methods

Many studies have been conducted for aspect extraction and these studies are either for explicit aspect extraction, implicit aspect extraction, or both aspects together.

2.5.1 Explicit Aspect Extraction

Several works have been conducted for explicit aspect extraction which were grouped into three approaches: supervised, unsupervised, and semi-supervised. In supervised method, it requires training data and semi-supervised requires little training data, while unsupervised approach does not require any training data (Rana & Cheah, 2016).

2.5.1.1 Unsupervised Extraction Methods

The unsupervised method means that the adopted technique does not require training data to accomplish extraction task. The following are some examples of previous works conducted using unsupervised approach.

(a) ***Frequency based extraction***

Hu and Liu (2004b) used unsupervised extraction for aspect extraction where they considered noun and noun phrases as a candidate features and the nearest adjective as its opinion word. The proposed technique worked by first labelling part of speech tags for all words, stop words removal, stemming, and fuzzy matching for misspellings corrections and matching of words variants. In the following step, it will extract the frequent aspects using Apriori algorithm through finding the frequent nouns and noun phrases. The candidate feature will be considered frequent if its frequency is higher than 1% of number of sentences in the review. After that, it will consider the nearest adjective as its opinion word. If a given sentence has no frequent feature but has opinion word, then it will extract the noun that is closest to the opinion word as infrequent feature.

For pruning purpose, they employed two techniques: the first one is called compactness pruning which is based on the distance between words to check if the words normally not appeared together for features with at least two words. The second one is called redundancy pruning which is based on discarding single words features that has support less than the specified threshold and it is subset of another feature phrase. The reported performance of the proposed method was 80% recall and 72% precision.

However, the proposed technique extracted many non-aspects words and pruned low frequent features.

Popescu and Etzioni (2007) extended the previous work of Hu and Liu (2004a) where web similarity measure was used to improve precision of feature extraction performance based on Pointwise Mutual Information (PMI). The experiments conducted on customer review dataset with 77% recall and 94% precision (Hu & Liu, 2004a).

(Eirinaki et al., 2012) study represented an improvement over (Hu & Liu, 2004a) as it is based on frequency method. A new algorithm called High Adjective Count (HAC) was developed to extract features. HAC is mainly based on the concept that the noun which frequently opinionated by reviewer can be a feature. The HAC starts by extracting all nouns and adjectives and set a counter for each noun which initialized by zero. If it finds the closest adjective to the corresponding noun on its left or right side the counter will be incremented by 1. In the end, the nouns with a score higher than a prespecified threshold will be considered as an approved features and other features will be discarded. The best results on customer review datasets (Hu & Liu, 2004a) on DVD player was 80% precision. However, not all adjectives normally are considered as opinion words.

(b) ***Frequency with patterns based***

Li et al. (2009) research, hybridized NLP technique with frequency for feature extraction. In addition, it utilized two types of dictionaries one for aspects and another one for opinion words. And used four pattern rules to extract aspects. The feature approved based on its frequency on both the reviews and the background corpus. The proposed technique was applied on Chinese dataset about mobile reviews with the best achievement of f-measure was 74.04%. However, for the approaches which based on statistics the results are not reliable when the used dataset was small (Qiu et al., 2011).

In another study by Moghaddam and Ester (2010), they improved on (Hu & Liu, 2004a) work by combining frequency-based approach with several POS (part of speech) patterns to extract and filter non-aspect words. To extract explicit aspects, they built a list of product known aspects together with product reviews as the input to the system. The patterns were constructed by matching known aspects with review text. Then, the closest adjective was considered as its opinion word. The pattern represents the POS tags of the words between the matched aspect and the found adjective word. The best results were 80% precision and 87% recall on customer review dataset about the number of electronic products (Hu & Liu, 2004a). However, not all aspects could be extracted as it is based on the previous knowledge of list of aspects.

(c) ***Frequency with dependency based***

Zhuang et al. (2006) improved the work of (Hu & Liu, 2004a). Features and their related opinion words were extracted using statistical approach which based on word frequencies and syntactic method using four dependency relation rules. The experiments were conducted on movie reviews where the best result was 52.9 % F-measure. However, not all rules were of equal efficiency in the extraction. In addition, for approaches which based on statistics, the results were not reliable when the used dataset is small (Qiu et al., 2011).

Hai et al. (2014) used one domain specific corpus and another independent domain corpus with three relation rules to extract explicit features from Chinese reviews. The technique started by extracting all candidate features using three syntactic dependency rules. In addition, two relevance measures were defined which are extrinsic-domain relevance (EDR) and intrinsic-domain relevance (IDR). EDR finds how relevant the candidate feature to the independent corpus, while IDR finds how relevant the candidate feature is to the domain specific corpus. If the EDR value of a given feature is less than a specified EDR threshold and IDR is greater than a specified IDR threshold, then the given feature will be approved as a correct feature. The experiments were conducted on two Chinese datasets about hotels and cellphones were the performance are 52.26% F-measure on hotel dataset and 63.6% F-measure on cellphone dataset. However, this work also used statistic approach and the results were not reliable when the used dataset is small (Qiu et al., 2011).

(d) ***Bootstrapping with pattern based***

In a work conducted by Bagheri et al. (2013), aspects were extracted by using bootstrapping method and the process started by POS tagging all sentences in the reviews. From the POS tagged sentence, the parts that matched one of the four patterns are extracted as candidate aspects. For each candidate aspect, the stem will be determined and the multi words aspects are chosen for further processing. In the following step, heuristics rules were applied for removing aspects with no opinion word in the sentence that contain that aspect. In addition, if the given aspect contains stop word it was removed. A new score was proposed called A-Score which based on the mutual relation between words and the frequency of aspect, that give a score for each aspect in the bootstrapping phase. The bootstrapping started by seed aspects and at each iteration from the candidate extracted aspects, the algorithm selected the aspect with the highest A-score and added it to the seed sets. From the final seed list of aspects obtained by bootstrapping, two pruning methods including subset-support pruning by removing meaningless word from multi words aspects, and superset-support pruning which based on removing single word aspects that is a part of multiword aspects were applied. The technique was applied on customer review dataset (Hu & Liu, 2004a) and the reported performance using F-measure was 72.9%. However, bootstrapping approach has the problem of error propagation (Zhang et al., 2010).

(e) ***Bootstrapping with dependency based***

In a study by Li, Qin, et al. (2015), bootstrapping approach based on number of dependency relation rules for aspect extraction and its related opinion words was used. They used six dependency relations patterns combined with sentiment information for extracting explicit features, where these relations were based on the grammatical relation between opinion words and features. The new features were used to generate new relations patterns through the used bootstrapping approach. Moreover, two measures were defined including Prevalence and Reliability to evaluate the confidence of the extracted features and the relations pattern used for extraction. The extracted features were grouped into number of clusters with a weight allocated for each cluster. A feature with low confidence was saved if it is grouped in the cluster which contains features with high confidence values. The experiment was conducted on customer review datasets (Hu & Liu, 2004a) with 89% using F1-measure. However, bootstrapping based approach has the problem of error propagation (Zhang et al., 2010) and not all possible extraction rules were used.

(f) ***Word alignment based***

A study by Liu et al. (2012), proposed a new technique called Word-based Translation Model (WTM) for features extraction. They found the association between candidate feature and their opinion words by formulating the extraction as word alignment. They then used a graph algorithm for finding relations between features and opinion words and considered the feature and opinion word with confidence higher than the threshold as an approved feature. In addition, noun and noun phrase in WTM were considered as possible features and adjective as opinion word. The experiments were conducted on three datasets in Chinese and English languages where the achievements were 74.5% F-measure for COAE2008 dataset, 77.66% F-measure for Large dataset, and 85.8% F-measure for customer review dataset (Hu & Liu, 2004a).

(g) ***Pattern based rules***

Htay and Lynn (2013) extracted features and their related opinion words using pattern-based approach. They defined eight pattern rules and considered the features as noun and noun phrase. The experiments were conducted on customer reviews datasets from Hu and Liu (2004a) with F-measure was 79%. However, from their findings, not all possible patterns were explored, and not all used patterns are effective for extraction.

Maharani et al. (2015) work based on using pattern rules from the previous studies (Turney, 2002; Htay & Lynn, 2013). They defined new patterns for feature extraction and used patterns from the previous studies including patterns from (Turney, 2002; Htay & Lynn, 2013). They tried to use different combination of these patterns rules and the best result was achieved when all patterns were used. The experiments were conducted on customer reviews datasets by (Hu & Liu, 2004a; Ding et al., 2008) with the best achievement for F-measure was 67.2%. However, many of the used patterns were not effective in feature extraction and many of the patterns were not explored.

In a recent study by Asghar et al. (2017), pattern-based approach was also used to extract features and their related opinion words where ten extraction rules were defined.

In addition, preprocessing was also conducted on the datasets such as removal of stop words, lemmatization, and POS tagging before applying the rules for extraction. The experiments were conducted on customer reviews datasets by Hu and Liu (2004a) where the F-measure was 77.16%. However, many pattern rules were not explored, and some of the used rules were not effective in extraction.

(h) ***Dependency based rules***

Samha (2016) defined 5 new dependency relation rules for feature extraction based on observation of the used datasets and combined these rules with 11 rules from previous studies (Qiu et al., 2011; Kumar & Raghuveer, 2013; Agarwal et al., 2015; Chinsha & Joseph, 2015). She used some preprocessing such as removing symbols, lemmatization, POS tagging, and finding of dependency relations using Stanford dependency parser. The experiments conducted on customer reviews datasets by Hu and Liu (2004a) with best achieved result was 71.8% F-measure. However, not all used rules were effective for feature extraction and many of the rules are still unexplored. In addition, there was also no pruning applied in this work.

(i) ***Topic modeling***

In a work by Brody and Elhadad (2010), each sentence in the reviews was considered as one separate document to solve the problem of low frequent features. They applied local LDA (Latent Dirichlet Allocation) on these collections of documents for feature extraction. Moreover, the extracted topics from each document were considered as the approved features. Moghaddam and Ester (2011) extended the standard LDA to Interdependent Latent Dirichlet Allocation (ILDA) model for extracting features and their sentiments.

Chen et al. (2014) Combined topic modeling with prior knowledge for extracting features. The prior knowledge was learned by applying LDA on multiple reviews domains, learnt the features in each domain, and learned the shared features among domains as prior knowledge. However, topic modeling approaches are not suitable for fine grained features extraction since they only able to extract general features (Zhang et al., 2010; Li, Qin, et al., 2015; Liu, Gao, et al., 2016; Wu et al., 2018).

2.5.1.2 Semi-supervised Extraction Methods

The semi-supervised approach means that the adopted technique requires little training data to accomplish extraction task. The following are examples of some of the previous works conducted using semi-supervised approach

(a) ***Bootstrapping with dependency based***

In a work by (Qiu et al., 2009, 2011), new technique known as DP was proposed, which based on the idea that there is syntactic relation exist between features and opinion words. They defined eight rules based on dependency relations and relations between feature and opinion word. DP considered only nouns and noun phrases as possible features and adjectives as possible opinion words. DP requires seed opinion words to start the propagation, DP also uses these opinion words with dependency rules to extract new features and vice versa. Besides than that, several pruning techniques were also proposed including clause pruning which based on the assumption that the clause contains only one feature unless there are conjunction words. Another used pruning was based on frequency of the feature over the used dataset. Also, one used pruning was pruning of other dealers and products related words. The experiments were conducted over customer reviews datasets (Hu & Liu, 2004a) with 86% F-measure. However, DP has a problem of error propagation and it is good only for medium size datasets (Zhang et al., 2010). For large datasets, it will extract many no-aspects words because through the propagation process, many non-opinionated adjectives will be extracted as opinion words which in turn will be used to extract new aspects that are not correct (Zhang et al., 2010).

(Zhang et al., 2010) improved DP technique by adding more rules including these two rules: 1) no-pattern rule with list of words to use with; and 2) part-whole rule for identifying more features which require to find the class concepts of the corpus. In addition, to find feature relevance they ranked these features using Hyperlink-induced topic search (HITS) algorithm and frequency-based approach. The best achieved performance was on mattress dataset with 77% precision and 64% recall. However, not all used rules were efficient in extraction and many rules are still unexplored.

In another previous study by Hai et al. (2012), proposed bootstrapping approach based on utilizing three types of dependency relations which exist between the feature and opinion word. The process started with feature seed list to start the bootstrapping process. Then, all noun/noun phrase were extracted as candidate features and adjectives and verbs as possible candidate opinion words. They also combined the bootstrapping approach with two new types of association models including Likelihood Ratio Tests (LRT) and Latent Semantic Analysis (LSA) which are used to find the degree of association between features and opinion words. The two proposed bootstrapping models were called LRTBOOT and LSABOOT. The experiments were conducted on two Chinese datasets with 61.9% F-measure on hotel reviews and 73.25% F-measure on cellphone reviews. However, this approach cannot extract non-frequent features and bootstrapping based approach has the problem of error propagation (Zhang et al., 2010).

In a work by Liu, Liu, et al. (2016), they improved the DP approach by employing aspect associations and semantic similarity. The experiments were conducted on customer reviews datasets by Hu and Liu (2004a) with 87% F-measure and electronics products datasets from Liu, Gao, et al. (2015) with 82.3% F-measure.

In another work by Kang and Zhou (2017), they improved DP method (Qiu et al., 2011). Features were extracted by extending DP with some new rules as comparative rules, and indirect-dependency rules. Besides than that, part–whole relation rules were also defined, and all these rules were applied for feature extraction. In addition, two pruning techniques were used for prune incorrect features based on two strategies which were, self-filtering by removing the feature if its not part of any multi-words candidate features and mutual exclusion (Qiu et al., 2011). Other pruning was based on term frequency in the document. They also defined semantic similarity filtering based on using WordNet. The experiments were conducted on customer reviews datasets where the F-measure was 87% (Hu & Liu, 2004a). However, some rules are irrelevant and produce many incorrect aspects. Furthermore, they pruned features based on WordNet which is not reliable since not all words contained in WordNet.

(b) ***Dependency based rules***

A work by (Wu et al., 2009), extracted features expression and their opinion expression using dependency parsing approach at phrase level. They considered only noun phrases (NPs) and verb phrases (VPs) as possible candidate features expressions. In addition, they considered opinion words as the words surrounded candidate feature expression based on existed opinion words dictionary. In addition, they used tree kernel with SVM for extracting the relations between the candidate features and opinion words. The experiments were conducted customer reviews datasets (Hu & Liu, 2004a) and (Jindal &

Liu, 2008) datasets with 57% F-measure. However, this approach extracted many non-aspect features.

In a work by Kumar and Raghuveer (2013), they extracted explicit aspects based on opinion seed lists and dependency relations. They defined 11 rules using a combination of different types of dependency relations. The experiments were conducted on customer reviews datasets (Hu & Liu, 2004a) with best achievements was 82% recall and 73% precision. However, not all used rules were effective for feature extraction and many rules still unexplored and no pruning was applied.

In a study by Yan et al. (2015), a technique of using four dependency relation to extract features and opinion words was proposed. In addition, the extracted features were expanded using synonym lexicon. The technique started by preprocessing the reviews, then parsed the reviews and extracted possible features and opinion words based on the four dependency relations. Then, they created a network based on the extracted feature-opinion pairs. Moreover, the NodeRank algorithm which is an extension of PageRank algorithm was used to rank all extracted pairs in the network. In the constructed network any aspect opinion pair with a NodeRank greater than the given threshold was approved, and the given feature added to the final list of features. The final list of features was expanded by finding the synonyms of each feature from synonym lexicon. The experiments were carried on Chinese reviews about three different electronic products with average F-measure was 73.96%. However, not all used rules were efficient in extraction and many rules still unexplored.

(c) ***Pattern based rules***

In study by Samha et al. (2014), frequent syntactic patterns with a total of 11 frequent patterns were used for aspect extraction. In the beginning opinion lexicon created by Hu and Liu (2004a) was used to match any opinion word in the review to one of the 11 patterns, before the corresponding feature was extracted. In addition, they built Feature-Dictionary from product specification which was found on product website and improved the dictionary by its synonyms. From the extracted feature, they approved only the feature which was found in the Feature-Dictionary. The experiment was conducted on customer review dataset (Hu & Liu, 2004a) with 77% F-measure. However, not all features could be found in Feature-Dictionary and not all possible extraction patterns are explored. In addition, many used rules were not efficient for extraction.

In a work by Rana and Cheah (2017), they used a hybrid approach for aspect extraction by extracting noun/noun phrase as feature and using opinion lexicon (Hu & Liu, 2004a). The features extracted by defining number of pattern rules based on noun and noun phrase as feature and search for opinion word in sentence based on opinion lexicon. Next, two pruning methods were applied one pruning is based on frequency by prune all feature with frequency less than threshold with 2 for noun phrase and 3 for nouns are set as frequencies thresholds, while the other pruning was based on Google results by using Normalized Google Distance (NGD). They applied NGD between the entity and pruned nouns form frequency-based pruning and keep only nouns with NGD value less than threshold. In Last step, they searched further for other aspects which not extracted by using SenticNet4 lexicon. The experiments were conducted on customer reviews datasets (Hu & Liu, 2004a) with F-measure is 89%. However, they used many patterns which produced many irrelevant features and used opinion words which found in opinion lexicon only, but there were many opinion words not exist in opinion lexicon (Hu & Liu, 2004a) and in this case they miss many correct features. In addition, long used patterns may introduce many noisy features. Moreover, pruning based on NGD not reliable as Google considered any two words cooccurred in the same page as semantically related words regardless the words are close together or very far from each other in the web page.

(d) ***Bootstrapping***

In a work by Wang and Wang (2008), they extracted features and their related opinion words using bootstrapping iterative learning technique based on the assumption that feature is surrounded by opinion word context-dependency. Thus, these opinion words can be used to extract features and vice versa. The iterative process started based on seed opinion list. In addition, they modified the mutual information formula and used it for finding the association between the extracted feature and opinion word. They also used one linguistic rule to find low frequent features. The experiments were conducted on Chinese reviews about three different electronic products with 75.05% F-measure. However, bootstrapping based approach has the problem of error propagation (Zhang et al., 2010).

(e) ***Association rule mining with lexicon***

In Wei et al. (2010), they developed a technique based on opinion words seed list which obtained from General Inquirer and by using semantic based approach which similar to the work of (Hu & Liu, 2004a, 2004b) . The proposed approach used preprocessing such as POS tagging, stemming, and noun phrase chunking. From the preprocessed reviews, association rule mining algorithm was applied for extracting features. Besides than that, redundancy and compactness pruning techniques were also employed. Based on the opinion words seed list from the General Inquirer, they semantically pruned frequent non-product features and features that are opinion irrelevant. In addition, these opinion words also extracted infrequent features. The experiments were conducted on customer review datasets (Hu & Liu, 2004a) with precision 46.8% and recall 75.7%.

(f) ***Topic modeling***

A study conducted by Ma et al. (2013), integrated both synonym lexicon and Latent Dirichlet Allocation (LDA) to extract products features. The technique started by extracting noun/noun phrase and combined with LDA for extracting candidate features. They then extended these candidate features with their synonyms based on lexicon synonyms. The experiments were conducted on Chinese reviews about electronic products with 69.19% F-measure. However, topic modeling approaches are not suitable for fine grained features as they are only able to extract general features (Zhang et al., 2010; Li, Qin, et al., 2015; Liu, Gao, et al., 2016; Wu et al., 2018).

(g) ***Word alignment based***

Liu, Xu, Liu, et al. (2013) used Partially supervised word alignment model (PSWAM) for extracting features. PSWAM based on the idea that noun/ noun phrases as the candidate features and adjective as the opinion word. Furthermore, PSWAM with some syntactic patterns was applied on the reviews to find possible relations between candidate features and opinion words, then determine degree of association in each pair. In addition, a graph algorithm was applied to find candidate feature confidence. Moreover, the candidate feature with confidence value greater than threshold were selected. The experiment was conducted on customer review dataset (Hu & Liu, 2004a) with 86% F-measure. However, low frequent features are pruned.

Another example of work which applied word alignment the work by Liu, Xu, et al. (2015) proposed alignment word model to find features and their related opinion words by identifying relations between opinion words and features based on the alignment model. In addition, they considered noun/noun phrase as candidate features and adjectives/verbs as possible opinion words candidates. Then, they estimated the confidence of each extracted feature or opinion word using graph based ranking algorithm. Lastly, candidate features or opinion words with confidence greater than threshold were approved as correct features and opinion words. The experiments were conducted on three datasets with 86.6% F-measure on customer review dataset (Hu & Liu, 2004a), COAE 2008 Chinese datasets with 77.5% F-measure, and Large dataset which contains English and Chinese reviews with 82% F-measure.

2.5.1.3 Supervised Extraction Methods

The supervised approach requires training data to accomplish extraction task. The following are examples of some of the previous works conducted using supervised approach

(a) ***Deep learning***

Previous study by Poria et al. (2016) utilized deep learning for explicit aspect extraction and used 7 layers deep neural network. In addition, the deep neural network was also combined with 5 linguistic rules and word embedding model which was created for sentiment analysis task. The experiments were conducted on two datasets including customer review datasets (Hu & Liu, 2004a) with 88% F-measure and SemEval 2014 dataset with 84.74% F-measure. However, not all possible extraction rules were explored.

In a work of Xu et al. (2018), deep learning used for aspect extraction by utilizing Convolutional Neural Network (CNN) with two types of embedding including domain specific and general embedding. Furthermore, this combined embedding with CNN applied on two datasets about laptop and restaurants with F1-score achieved as 81.59% for laptop dataset and 74.37% for restaurant dataset. However, the approach still has problems such as unable to extract the aspects if there is a conjunction.

(b) ***Conditional Random Fields***

In a work of Li et al. (2010), they proposed feature extraction based on using Conditional Random Fields (CRF). To model the sequential dependency relation which exist between words, they used linear-chain CRFs. In addition, they also used Skip-chain CRFs for dependency over long distance and Tree CRFs used to exploit the syntactic structure. The experiments were conducted over two datasets about movie and products reviews with the reported performance was 83.7% F-measure on movie reviews and 80.1% F-measure on products reviews.

Another example of study which applied CRF for aspects extraction a work of Jakob and Gurevych (2010). CRF was applied for feature extraction using a combination of different features to learn CRF such as token, POS, word distance, short dependency path, and opinion distance. The model was applied on different datasets about cars, cameras, movies, and web-services. The best achieved performance was on movies reviews with 70.2% F-measure. Further example of CRF use the study by Choi and Cardie (2010). They used hierarchical parameter sharing by CRFs for extracting features and their related opinion words.

Chen et al. (2012) also applied CRFs on feature extraction. At first step, they preprocessed the data by removing special symbols, misspelling words correction, and nouns stemming. At second step, they prepared the datasets by label it with the corresponding tags. At third step, they trained the CRFs model using the labelled data. Finally, they applied the trained CRFs model on the testing data for label the features. The experiments were conducted on customer reviews datasets (Hu & Liu, 2004a) and other data collected from Amazon about digital camera with 86.2% F-measure. Another example of CRF use is the work conducted by Huang et al. (2012). They utilized CRFs for feature extraction by combining it with three types of features including word feature, POS feature, and sentence structure feature. The experiments were conducted on (Hu & Liu, 2004a) dataset and another part was collected from Amazon with 75.6% F-measure. However, CRFs based approach is not suitable for long range patterns (Zhang et al., 2010).

To create ensemble classifier based on using particle swarm optimization (PSO), (Akhtar et al., 2017), used three type of classifiers for feature extraction including Maximum Entropy (ME), CRF and Support Vector Machine (SVM). Before applying these classifiers, they applied PSO for selecting best combinations of features for training these classifiers. The PSO was applied on number of features such as WordNet, word cluster, dependency relation, lexicons, and other types of features which were all syntactic and lexical features. The experiments were conducted on SemEval-2014 datasets with 84.52% F-measure on restaurant dataset and 74.93% F-measure on Laptop dataset.

(c) ***Pattern based rules***

In a work by Kobayashi et al. (2007), at first the opinion words were extracted through supervised dictionary-based approach. Furthermore, syntactic patterns used together with identified opinion words for extracting features. Moreover, machine learning classifier was used to evaluate the exist relations between extracted pairs. The features with scores higher than threshold were selected as correct features. The approach was conducted on Japanese dataset about restaurant with precision 72% and 62% recall.

(d) ***Bootstrapping with dependency based***

In a work by Liu, Gao, et al. (2016), they extended the extraction rules used in DP by adding more dependency relations than the standard DP algorithm. Furthermore, two algorithms were compared for selecting the best rules from the set of rules using either greedy algorithm or Simulating Annealing (SA) algorithm. Based on reported results the best achieved results obtained when SA was applied with 87.9% F-measure on customer reviews datasets (Hu & Liu, 2004a). However, according to Rana and Cheah (2017), the rules which were used based on dependency relations only, have problems when dealing with unstructured reviews. In addition, according to Zhang et al. (2010), DP have a problem of propagation errors and cannot work well in large datasets. Moreover, SA algorithm is a local search algorithm which is not suitable for global search and not give the global optimal solution. Besides than that, there was no pruning applied.

2.5.2 Implicit Aspect Extraction

According to Tubishat, Idris, et al. (2018) previous works on implicit aspect extraction techniques are classified into three main categories including unsupervised, semi-supervised, and supervised.

2.5.2.1 Unsupervised Extraction Methods

Unsupervised implicit aspects extraction uses unlabeled data to extract implicit aspects from the corpus and does not use any algorithm that require some training. The most commonly used unsupervised implicit aspects extraction methods are dependency parsing, association rule mining, mutual or association, hierarchy, ontology, topic modeling, co-occurrence, rule-based, and clustering (Tubishat, Idris, et al., 2018).

2.5.2.2 Semi-supervised Extraction Methods

Semi-supervised implicit aspects extraction utilizes both labeled and unlabeled data to extract implicit aspects from the corpus or require little training. The most commonly used semi-supervised implicit aspects extraction methods can be based on association rule mining, clustering, and topic modeling (Tubishat, Idris, et al., 2018).

2.5.2.3 Supervised Extraction Methods

Supervised implicit aspects extraction uses labeled data to extract implicit aspects from the corpus or use any algorithm that requires training. The most commonly used supervised implicit aspects extraction methods can be association rule mining, hierarchy with Co-occurrence, co-occurrence, classification, fuzzy, rule-based, and conditional random fields (Tubishat, Idris, et al., 2018).

2.6 Dependency-based and Pattern-based methods

Many studies conducted in literature used dependency relations rules for extracting aspect and its related opinion words such as (Qiu et al., 2011; Kumar & Raghuveer, 2013; Poria et al., 2014; Liu, Gao, et al., 2016; Kang & Zhou, 2017) in which Stanford dependency relation represents a binary grammatical relationship between words in the sentence will be used for aspects extraction. Also, Patterns-based approach was used in the previous studies (Htay & Lynn, 2013; Maharani et al., 2015; Asghar et al., 2017; Rana & Cheah, 2017) in which based on the idea that a pattern contain sequence of tags . These tags described the aspect, opinion word, and relation between these tags. Dependency relations accuracy is based on the grammatical correctness of the reviews, however, not all online reviews normally follow the English grammars rules and these reviews are usually mixed types of informal and formal styles (Zhang et al., 2010; Hai et al., 2012; Liu, Xu, & Zhao, 2013; Liu, Xu, et al., 2015; Rana & Cheah, 2017). In addition, there is no restriction from the website on customers to follow language rules when they post their review. Customers may follow the language rules in writing and sometimes may violates some of it (Rana & Cheah, 2017). Hence, these types of rules are better for formal text. On the other hand, pattern-based approach is good for informal text since the patterns mimic the ways users write their reviews (Rana & Cheah, 2017). Pattern-based approach normally follows the ways user express their opinion on the opinionated aspect in informal writing.

From reviewing the literature, it was found that most of the methods whether its supervised, semi-supervised, or unsupervised. Commonly used dependency-based, or pattern-based method for aspects extraction. However, there are some limitations still in these two methods types as discussed before in subsection 2.5.3.

Table 2.1 shows the limitations of related dependency-based approaches from the previous works discussed in the previous subsections. Table 2.2 summarized the limitations related to the pattern-based approaches from the previous works.

Table 2.1: Dependency-based approaches

References	Limitations
(Qiu et al., 2011)	• Not all used rules are efficient for extraction. • Not all possible extraction rules were explored. • Bootstrapping based approach has the problem of error propagation (Zhang et al., 2010)

(Kumar & Raghuveer, 2013)	• Not all used rules are efficient for extraction. • Not all possible extraction rules were used. • No refinements were made after extraction.
(Li, Qin, et al., 2015)	• Not all used rules are efficient for extraction. • Not all possible extraction rules were explored. • No refinements were made after extraction. • bootstrapping based approach has the problem of error propagation (Zhang et al., 2010).
(Samha, 2016)	• Not all used rules are efficient for extraction. • Not all possible extraction rules were used. • No refinements were made after extraction.
(Liu, Liu, et al., 2016)	• Not all used rules are efficient for extraction. • Not all possible extraction rules were used. • No refinements were made after extraction. • Bootstrapping based approach has the problem of error propagation(Zhang et al., 2010)
(Liu, Gao, et al., 2016)	• All The rules used were based on dependency relations only which have problems when dealing with informal written reviews according to (Rana & Cheah, 2017). • Double propagation have a problem of propagation errors and does not work well in large datasets according to (Zhang et al., 2010). • Simulating Annealing algorithm is a local search algorithm which is not suitable for global search and does not give the global optimal solution. • No refinements were made after extraction.
(Kang & Zhou, 2017)	• Not all used rules are efficient for extraction. • Not all possible extraction rules were explored. • Bootstrapping based approach has the problem of error propagation (Zhang et al., 2010). • they pruned features based on WordNet which is not reliable as not all words are contained in WordNet.

Table 2.2: Pattern-based approaches

References	Limitations
(Htay & Lynn, 2013)	• Not all possible patterns were explored. • Not all used patterns were effective for extraction. • No refinements were made after extraction.

(Samha et al., 2014)	• Not all possible patterns were explored. • Not all used patterns were effective for extraction. • No refinements were made after extraction.
(Maharani et al., 2015)	• Not all possible patterns were explored. • Not all used patterns were effective for extraction. • No refinements were made after extraction.
(Rana & Cheah, 2017)	• Not all used patterns were effective for extraction. • they used normalized google distance (NGD), but the cooccurrence returned by google is not reliable, as if the two words occurred in same page might be very far from each other and not related.
(Asghar et al., 2017)	• Not all possible patterns were explored. • Not all used patterns were effective for extraction. • No refinements were made after extraction.

As shown in Table 2.1, methods which applied dependency rules for aspects extraction, share that not all used rules were efficient for extraction and Not all possible dependency extraction rules were explored. In addition, extraction of aspects based on dependency rules has a problem when dealing with informal written reviews. Dependency rules give better results if the reviews follow English rules and grammar. However, reviews available online are mixed combination of formal and informal text.

On the other side as shown in Table 2.2, methods which applied pattern rules for aspects extraction, also share that not all used rules were efficient for extraction and Not all possible extraction pattern rules were explored. Pattern-based rules are better than dependency-based rules for informal text, but for long sentence dependency-based rules is better. In addition, Tables 2.1 and 2.2 showed that most of methods were either not applied aspects pruning or they did aspect pruning which need some improvements.

Thus, a combination of pattern and dependency rules from previous studies is proposed to overcome the problems of both methods. In addition, a development of new extraction rules is carried out to solve the problems of unexplored rules. However, not all rules are relevant. An algorithm to select the proper combinations of these rules is required. To select a proper subset of rules from the full set of rules, IWOA is proposed for rules selection. Finally, to discard incorrect aspects while keeping correct aspects, a three-phase pruning is proposed which is based on frequency, product manual, and direct opinion association.

2.7 Optimization Algorithms

In recent years, metaheuristic optimization algorithms had been widely used to solve different types of problems such as feature selection, engineering problems, energy, time tabling, parameters optimization, and many more as mentioned in Section 1.3. The wide use of these metaheuristic algorithms resulted from their characteristics which include the simplicity, ease of implementation, no need of gradient information, can avoid local optima problem, and can be applied to solve different types of problems (Mirjalili, 2016; Mirjalili & Lewis, 2016; Mirjalili et al., 2017). These metaheuristic optimization algorithms can be mainly divided into three groups which are evolution-based algorithms, physics-based algorithms, and swarm-based algorithms (Mirjalili & Lewis, 2016). These evolution algorithms are inspired from the natural evolution laws. In evolution algorithms, the common search procedure among these algorithms that they start by randomly generating number of solutions which represent the population to be used by the algorithm. These solutions then will be evolved over the algorithm iterations and the result will be the new generations. In evolution algorithm, the best solutions will be selected to be combined and generate new generation. This feature allows the algorithm to improve the population solutions over the iterations of algorithm (Mirjalili & Lewis, 2016; Mirjalili et al., 2017). There are a number of evolution algorithms, and one of the most famous evolution algorithms is the Genetic Algorithms (GA) (Holland, 1992). Other examples of evolutionary algorithms are Differential Evolution (DE) (Storn & Price, 1997), and Biogeography-Based Optimization (BBO) algorithm (Simon, 2008). The second group of metaheuristic optimization algorithms is physics-based algorithms, which mimic the physical laws in our world (Mirjalili & Lewis, 2016). Some examples of physics-based algorithms, are Simulated Annealing (SA) (Kirkpatrick, 1984), Ray Optimization (RO) (Kaveh & Khayatazad, 2012), Gravitational Search Algorithm (GSA) (Rashedi, 2009), Central Force Optimization (CFO) (Formato, 2007), and Charged System Search (CSS) (Kaveh & Talatahari, 2010), The last type is swarm-based algorithms, in which the algorithm mimic the animals' intelligence and social behavior in their nature (Mirjalili & Lewis, 2016; Mirjalili et al., 2017). One of the most famous swarm algorithms is Particle Swarm Optimization (PSO). PSO was developed by Eberhart and Kennedy (1995). The idea of PSO is originally inspired from imitating the social behavior that used by bird flocking. Additional examples of available swarm algorithms including Ant Colony Optimization (ACO) (Colorni et al., 1992), Grey Wolf Optimizer (GWO) (Mirjalili et al., 2014), Firefly Algorithm (FA), Whale Optimization Algorithm (WOA) (Mirjalili & Lewis, 2016), Salp Swarm Algorithm (SSA) (Mirjalili et al., 2017), Fruit fly Optimization Algorithm (FOA) (Pan, 2012), and Bat Algorithm (BA) (Yang, 2010).

WOA represents one of the new optimization algorithms which proved its ability and outperforms other well-known optimization algorithms in solving different type of real-world problems as mentioned in Chapter 1. This superiority of WOA and application in different type of problems motivated this research to use it for rules selection problem in aspect extraction. However, based on NFL theorem which states that none of the available optimization algorithms can solve all problems and at the same time outperform all other algorithms. In addition, based on NFL theorem, all optimization algorithms have weaknesses in some cases (Wolpert & Macready, 1997). Therefore, two main improvements are included in the original WOA to make it suitable for rules selection problem, solve WOA weaknesses, and further improve WOA performance. The first improvement includes the development of a new equation based on Cauchy Mutation (CM) to improve WOA solutions diversity and make balance between its exploitation and exploration. The second improvement is the development of a new Local Search Algorithm (LSA) and use it at the end of each WOA iteration to improve its exploitation and avoid it from stuck in local optima. The first improvement is important as it make diversity among the selection of aspect extraction rules, whereas the second improvement is also important to improve the current best solution at the end of each WOA iteration. LSA will set and rest the rules which are included in the current best solution based on the rules' performances (precision, recall).

2.8 Summary

This chapter contains comprehensive reviews about different aspect extraction methods, where these methods are classified into unsupervised, semi-supervised, or supervised. From the review, it was found that in most of the works, many non-aspects words were extracted, and low frequent features were pruned. For works which are based on statistics, the results were not reliable when the size of the used dataset is small. In the studies where they used either dependency-based or pattern-based rules, subset of these rules are irrelevant since these rules will extract many incorrect aspects. Besides than that, there are many extraction rules still unexplored. For works which only applied dependency rules for aspects extraction, these rules are good for structured reviews while for unstructured reviews, pattern rules are better. In other works which used bootstrapping approach, they had the problem of error propagation.

Based on reviews of the previous works, many of these works were conducted without applying aspects pruning. In addition, for works which were developed based on topic modeling for aspects extraction, topic modeling is not suitable for fine grained aspects extraction because it only able to extract general aspects. For works which were developed by using DP or through improving DP, these works had the problem of error propagation. In addition, DP will extract many incorrect aspects for large size datasets. In other works which were developed using WordNet or aspects dictionary for aspects pruning, many correct aspects cannot be found in these resources. There are works which were developed using CRF for aspects extraction. However, CRF is not suitable for long range patterns.

Although studies are still lacking on implicit aspect extraction, but there are also issues on explicit aspect extraction that need to be resolved as mentioned in this chapter. Thus, the focus of this work is only to resolve the issues in explicit aspects extraction.

As reviewed in this chapter, there are many studies which were developed using either dependency or pattern rules. However, as mentioned previously each rules type has its advantages and disadvantages. In addition, many rules are still unexplored, and a combination of all rules will improve the recall, but the selection of rules is also required to improve the precision.

Hence, a combination of different rules types is required to take the advantages of these rules together. In addition, a development of a set of new rules is required to overcome the weakness of the existing rules. However, not all these rules can be used together. Thus, a proper selection of rules combinations is required. This can be achieved using optimization algorithm and also pruning algorithm to remove the incorrect aspects and retain the correct ones.

CHAPTER 3: FORMULATION OF ASPECT EXTRACTION RULES

This chapter describes the details about the full set of extraction rules used for aspects extraction. At the beginning, it gives introduction in section 3.1. Then, some details about the required preprocessing are provided in section 3.2. Section 3.3 introduces the selected dependency-based rules that have been used in the previous studies (Zhuang et al., 2006; Qiu et al., 2011; Huang et al., 2012; Kumar & Raghuveer, 2013; Samha, 2016), while section 3.4 presents pattern-based rules which have been used in the previous studies (Htay & Lynn, 2013; Samha et al., 2014; Maharani et al., 2015; Asghar et al., 2017; Rana & Cheah, 2017, 2018a). In addition, as an improvement of these rules selected from the previous studies, an OP/JJ combination is used in these rules to cover majority of opinion words. Finally, section 3.5 describes the new formulated rules.

3.1 Introduction

Online reviews normally contain mix types of reviews including structured and unstructured text. To take the advantages of dependency-based and pattern-based rules, this research uses multiple types of rules. In addition to the existing rules, some new rules are also developed to overcome the weaknesses of the rules. There are also cases where two or more rules can extract the same aspects but one rule from these rules is better in precision as it has the minimal number of wrong extracted aspects. As a result, the total number of rules used in this work is 126 rules which consist of 17 dependency-based rules, 16 pattern-based rules, and 93 new rules. These rules are based on the idea that an aspects and opinion words exist in online reviews with an association relation between opinion words and aspects. However, this association relation between aspect and opinion word sometimes follow English language constraints and grammars, while sometimes it does not follow these constraints and grammars. Thus, in the case of association follow the language rules, dependency relation rules can be used and give accurate results. On other hand, in the case it where does not follow the language rules, the pattern rule can follow the ways user frequently expressed their opinion towards aspects and use these frequent patterns for extraction. In addition, sometimes we cannot use either pattern-based or dependency-based rules as no rules match. Therefore, new type of rules which based on regular expression can be used. In conclusion, the main aim of this chapter is to explore the various types of rules that are used in this work for aspects extraction.

The 126 used rules are presented on the following sections, where the ('nsubj', 'amod', 'prep', 'csubj', 'xsubj', 'dobj', 'iobj', 'conj', 'xcomp', 'compound', 'nmod', 'neg', 'det', 'nsubjpass', 'nummod', 'appos', and 'cop') are dependency relations from Stanford dependency parser which describe a relation between two words. The decsription of these relations are presented in appendix A and B. Furthermore, OP means any opinion word from opinion lexicon (Hu & Liu, 2004a) and the decsription of each POS tag is presented in Table 3.1. As shown in Figure 3.1, the general process of aspect extraction consists of a number of phases. In the first phase, data preprocessing is carried out on the customer reviews datasets to prepare the dataset for aspect extraction. The reviews datasets are splitted into sentences. Then, for each sentence, the dependency relations in that sentence will be extracted by Stanford dependency parser. For each sentence, the sequence of POS tags will be determined by Stanford POS tagger. The set of POS tags and dependency relations of each sentence will be used in the next phases. In the second phase, there are a combination of 126 rules including dependency and pattern rules from the previous studies (Zhuang et al., 2006; Qiu et al., 2011; Huang et al., 2012; Htay & Lynn, 2013; Kumar & Raghuveer, 2013; Samha et al., 2014; Maharani et al., 2015; Samha, 2016; Asghar et al., 2017; Rana & Cheah, 2017, 2018a) together with the new developed rules.

Table 3.1: POS tags description

POS Tag	Description
NN	Any type of noun
VB	Any type of verb
NNS	Plural noun
JJ	Any type of adjective
RB	Any type of Adverb
VBP	Non3rd person singular present
VBZ	3rd person singular present
VBD	Past tense verb
IN	Preposition or subordinating conjunction
DT	Determiner
TO	to
PRP	Personal pronoun
CC	Coordinating conjunction
VBN	past participle verb

These rules will be used by IWOA to select optimal combination rules subset from the full set of rules. The selected rules will be applied on the testing dataset for extracting aspects. Finally, the extracted aspects will be passed to the pruning phase where the correct aspects will be saved as final approved aspect, whereas the incorrect aspects will be discarded.

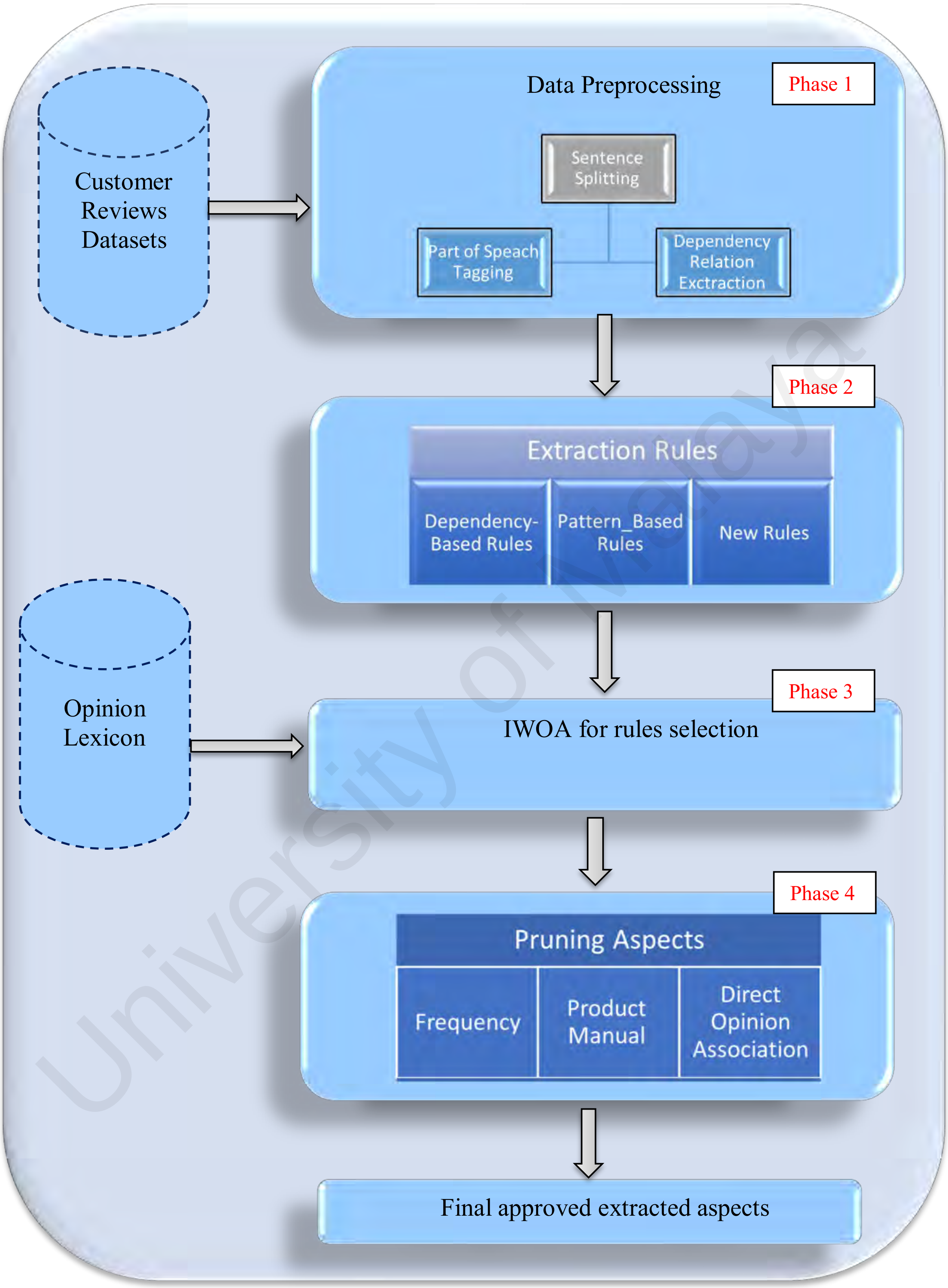

Figure 3.1 General process of aspect extraction

3.2 Preprocessing

The first step before applying any rule for aspect extraction is to do preprocessing on the datasets including the removal of unnecessary symbols in the datasets. Then, each review is tokenized using sentence tokenizer. After getting each sentence, each sentence will be parsed using Stanford tagger to obtain the part-of-speech (POS) type of each word in the sentence. In addition, for each sentence, Stanford dependency parser will be used to extract the list of dependency relations exist within the sentence in Dependency Relation Extraction step.

As mentioned in the previous section, there are groups of rules used in this work. The first set of rules which have been used in the previous studies, applied dependency relations for extraction. Many rules have been used in the previous studies, however, not all of these rules are relevant. Some of these rules extratct correct aspects, but some extract many incorrect aspects. These rules were selected based on the achieved results in their study as mentioned in Chapter 2, and manual observation to see which rules can be further used in the final set of 126 rules. Therefore, the selected dependency rules are presented in section 3.3

3.3 Dependency-based extraction rules

The following rules were used by the previous studies (Zhuang et al., 2006; Qiu et al., 2011; Huang et al., 2012; Kumar & Raghuveer, 2013; Samha, 2016). At the beginning, every tokenized sentence from the datasets will be parsed using Stanford dependency parser to extract all relations in the sentence. Then, any matched rule from these dependency rules will be used to extract the aspect from the sentence. Each rule was explained with an example from customer reviews datasets (Hu & Liu, 2004a) and some samples of these datasets are presented in Appendix C.

The Stanford dependency relation represents a binary grammatical relationship between words in the sentence. Furthermore, in this binary relation one word is called governor (head) and another word called dependent (De Marneffe & Manning, 2008). For example, Figure 3.2 shows the dependency relations on sentence "*The phone is nice*" which resulted from Stanford parser. In this nsub('nice','phone') relation, the **governor** is **nice,** and the **dependent** is **phone**.

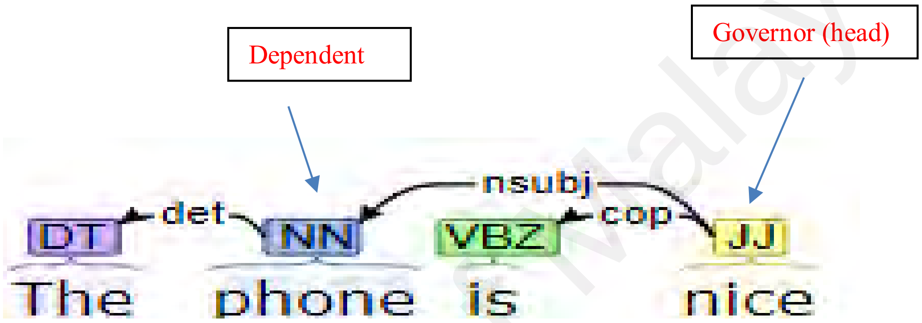

Figure 3.2 Dependency parsing of "*The phone is nice*" sentence

As shown on Figure 3.2, the word "***nice***" is an adjective word represents an opinion word and "***phone***" is a noun which represents an aspect. These two words are related via nsub('nice','phone') relation, where "***phone***" will be extracted as an aspect and "***nice***" as the opinion word expressed towards "***phone***" aspect. Thus, these types of dependency relations are used according to the fact that opinion words always have opinion targets (aspects) and these opinion words and opinion targets are connected via one of the dependency relations (Qiu et al., 2011).

The following rules represent the dependency rules which are selected from the previous studies.

Rule #1 (nsubj(JJ/OP,**NN**)**):** if 'nsubj' relation found in the sentence with first argument is either JJ or OP and the second argument is NN, then the second argument will be extracted as an aspect as the following example *"**video** was poor"*. There is "nsubj (poor-3, video-1)" relation between "***poor***" opinion word and "***video***" aspect, then "***video***" will be extracted as an aspect (Zhuang et al., 2006).

Rule #2 (ReL1(H1,**NN1**) and ReL2(H1,**NN2**)) **:** means that ReL1 and ReL2 represents any dependency relations from ['nsubj','amod','prep','csubj','xsubj','dobj','iobj']. In addition, both relations ReL1 and ReL2 have NN as a second argument and share the same first argument H1 (represent any word), then **NN1** and **NN2** represent the extracted aspects as the following example *" Canon **G3** has a great **lens**"*. There are "nsubj(has-3, G3-2) , dobj(has-3, lens-6)" relations and the extracted aspects are "***lens***" and "***G3***" (Qiu et al., 2011).

Rule #3 (nsubj(VB1,H1) and dobj(VB1,**NN**)) **:** means that two relations exist in the sentence with first relation of type "nsubj" and second relation of type "dobj". Furthermore, both relations share same first argument VB1 and the second argument of "dobj" relation is NN, then **NN** represents the extracted aspect as the following example *"honestly , i love this **player** "*. There are " nsubj(love-4, i-3) , dobj(love-4, player-6) " relations and the extracted aspect is "***player***" (Zhuang et al., 2006).

Rule #4 (nsubj(H1,**NN**) and xcomp(H1,JJ/OP)) **:** means that two relations exist in the sentence with first relation of type "nsubj" and second relation of type "xcomp". Furthermore, both relations share same first argument H1 and second argument of "xcomp" relation is either JJ or OP and second argument of "nsubj" relation is NN, then **NN** represents the extracted aspect as the following example *" it 's **size** also makes it ideal* *for travel "*. There are "nsubj(makes-5, size-3) , xcomp(makes-5, ideal-7)" relations and the extracted aspect is "***size***", and "***ideal***" is an opinion word in OP (Samha, 2016).

Rule #5 (amod(NN1,OP/JJ) and conj(**NN1**,**NN2**)) **:** means that two relations exist in the sentence with first relation of type "amod" with first argument is NN1 and second argument is either JJ or OP. In addition, the second relation of type "conj" with first argument is NN1 and equal the first argument of "amod" relation and the second argument is also NN2. Furthermore, the extracted aspects are **NN1** and **NN2** as the following example *" it plays original **dvds** and **cds"***. There are "amod(dvds-4, original-3), conj(dvds-4, cds-6)" relations and the extracted aspects are "***dvds***" and "***cds***" (Kumar & Raghuveer, 2013).

Rule #6 (nmod(OP/JJ,**NNS**)) **:** if 'nmod' relation found in the sentence with first argument is either JJ or OP and the second argument is NNS, then the second argument will be extracted as an aspect as the following example *"i find the lack of entertaining **games** on this phone quite disturbing"*. There is "nmod(lack-4, games-7)" relation between "***lack***" opinion word and "***games***" aspect, then "***games***" will be extracted as an aspect (Samha, 2016).

Rule #7 (amod(**NN**,OP)) **:** if 'amod' relation found in the sentence with first argument is NN and the second argument is OP, then the first argument will be extracted as an aspect as the following example *"the poor **manual"***. There is " amod(manual-3, poor-2)" relation between "***poor***" opinion word and "***manual***" aspect, then "***manual***" will be extracted as an aspect (Zhuang et al., 2006).

Rule #8 (ReL1(H1,**NN**) and ReL2(H1,OP/JJ)) **:** means that ReL1 and ReL2 represents any dependency relations from ['nsubj','amod','prep','csubj','xsubj','dobj','iobj']. In addition, both relations ReL1 and ReL2 share the same first argument H1 (represent any word). In addition, the second argument of ReL1 is NN and the second argument of ReL2 is either OP or JJ, then **NN** represents the extracted aspect as the following example *" this **camera** has a major design flaw ."*. There are " nsubj(has-3, camera-2) , dobj(has-3, flaw-7)" relations and the extracted aspect is "***camera***", and "***flaw***" represent an opinion word from OP (Qiu et al., 2011).

Rule #9 (nsubj(**NN** ,JJ/OP)) **:** if 'nsubj' relation found in the sentence with first argument is NN and the second argument is either JJ or OP, then the first argument NN will be extracted as an aspect as the following example *"my only gripe about the hardware is the **buttons**"*. There is " nsubj(buttons-9, gripe-3)" relation between "***gripe***" opinion word and "***buttons***" aspect, then "***buttons***" will be extracted as an aspect, and "***gripe***" represents the opinon word which exists in OP (Zhuang et al., 2006).

Rule #10 (dobj(JJ/OP, **NN**)) **:** if 'dobj' relation found in the sentence with first argument is either JJ or OP and the second argument is NN, then the second argument will be extracted as an aspect as the following example *"i especially like the more commonly used **buttons** ."*. There is " dobj(like-3, buttons-8)" relation between "***like***" opinion word and "***buttons***" aspect, then "***buttons***" will be extracted as an aspect (Qiu et al., 2011).

Rule #11 (nsubj(OP/JJ,NN1) and compound(NN1,NN2)) **:** means that two relations exist in the sentence with first relation of type "nsubj" with first argument is either JJ or

OP and second argument is NN1. In addition, the second relation of type "compound" with first argument is NN1 and equal the second argument of "nsubj" relation and the second argument is also NN2. Furthermore, the extracted aspect is **NN2 NN1** as the following example *"i find **onscreen displays** annoying ."*. There are "compound(displays-4, onscreen-3), nsubj(annoying-5, displays-4)" relations and the extracted aspect is "***onscreen displays***", and "***annoying***" represents an opinion word in OP (Kumar & Raghuveer, 2013).

Rule #12 (conj(**NN**1, **NN2**)) **:** if 'conj' relation found in the sentence with first argument is NN1 and the second argument is NN2, then the extracted aspects are NN1 and NN2 as the following example *"proven canon built **quality** and **lens**"*. There is " conj(quality-4, lens-6)" relation between "***quality***" and "***lens***" aspects, then the extracted aspects are "***quality*** " and "***lens***" (Qiu et al., 2011).

Rule #13 (amod(NN1, OP/JJ) and compound(**NN1**,**NN2**)) **:** means that two relations exist in the sentence with first relation of type "amod" with first argument is NN1 and second argument is either JJ or OP. In addition, the second relation of type "compound" with first argument is NN1 and equal the first argument of "amod" relation and the second argument is also NN2. Furthermore, the extracted aspect is **NN2 NN1** as the following example *"overall , the g3 delivers what must be considered the best **image quality**."*. There are " amod(quality-13, best-11), compound(quality-13, image-12)" relations and the extracted aspect is " ***image quality*** ", and "***best***" represents an opinion word (Kumar & Raghuveer, 2013).

Rule #14 (neg(OP/JJ, H1) and nsubj(OP/JJ,**NN**))**:** means that two relations exist in the sentence with first relation of type "neg" and second relation of type "nsubj". Furthermore, both relations share same first argument which can be either JJ or OP. In addition, the second argument of "nsubj " relation is NN, then **NN** represents the extracted aspect as the following example *"the **colors** on the screen are not as crisp as i 'd have liked them to be ."*. There are " nsubj(crisp-9, colors-2), neg(crisp-9, not-7)" relations and the extracted aspect is "***colors***", and "***crisp***" represents the opinion word (Kumar & Raghuveer, 2013).

Rule#15 (ReL(**NN,**JJ/OP)**):** means 'ReL any relation from ['nsubj', 'amod', 'prep', 'csubj', 'xsubj', 'dobj', 'iobj'] found in the sentence with first argument is NN and the second argument is either JJ or OP, then the first argument will be extracted as an aspect as the following example *"definitely a great **camera**"*. There is "amod(camera-4, great-3)" relation between "***great***" opinion word and "***camera***" aspect, then " ***camera*** " will be extracted as an aspect(Qiu et al., 2011).

Rule#16 (ReL(JJ/OP,**NN**)**):** means 'ReL any relation from ['nsubj', 'amod', 'prep', 'csubj', 'xsubj', 'dobj', 'iobj'] found in the sentence with first argument is either JJ or OP and the second argument is NN, then the second argument will be extracted as an aspect as the following example *"the **manual** is relatively clear"*. There is "nsubj(clear-5, manual-2)" relation between "***clear***" opinion word and "***manual***" aspect, then "***manual***" will be extracted as an aspect (Qiu et al., 2011).

Rule #17 (nsubj(OP1/JJ1,**NN**) and cop (OP1/JJ1,H1)) **:** means that two relations exist in the sentence with first relation of type "nsubj" and second relation of type "cop".

Furthermore, both relations share same first argument which can be either JJ or OP. In addition, the second argument of "nsubj" relation is NN, then **NN** represents the extracted aspect as the following example *"the **menus** are easy to navigate."*. There are "nsubj(easy-4, menus-2), cop(easy-4, are-3)" relations and the extracted aspect is "***menus***" (Huang et al., 2012).

As discussed before, the dependency rules are better for formal text and long sentences. However, for informal text like blogs, pattern rules are better as it follows the ways customers write their reviews. The following rules represent the pattern rules which are selected from the previous studies. These rules were selected based on their good performance in the previous study as mentioned in Chapter 2, and also through manual observation. In addition, from these rules, the frequent suitable rules according to the datasets were included. Thus, to overcome the weaknesses of dependency rules, pattern-based rules are also considered.

3.4 Pattern-based extraction rules

In general, customers tend to write the sentences that express an opinion towards an aspect by frequently sharing similar sentence structure based on the similar alignment of POS tags in the sentence. Thus, the association between aspects and opinion words may violates English language constraints. These frequent patterns contain sequence of tags which describe the aspect, opinion word, and relation between these tags. The following rules were used in (Htay & Lynn, 2013; Samha et al., 2014; Maharani et al., 2015; Asghar et al., 2017; Rana & Cheah, 2017, 2018a). At the beginning, every tokenized sentence will be tagged using Stanford POS tagger to define the POS tag of every word in the sentence. After that, any matched rule from these pattern rules will be used to extract the aspect from the sentence.

Rule #18 (NNS VBP OP /JJ) **:** For example, the sentence *"**controls** are **poorly** designed"* match with this rule as the tagged sentence as the following "controls/NNS are/VBP poorly/RB designed/VBN", where **NNS** represent the extracted aspect "***controls***", and "***poorly***" represents an opinion word from OP (Samha et al., 2014).

Rule #19 (NN VBZ JJ/OP) **:** For example, the sentence "***audio** is excellent**."*** match with this rule as the tagged sentence as the following "audio/NN is/VBZ excellent/JJ", where **NN** represents the extracted aspect "***audio***", and "***excellent***" represents an opinion word from OP (Samha et al., 2014).

Rule #20 (JJ/OP **NN1 NN2) :** For example, the sentence "*it 's a very nice **dvd player."*** match with this rule as the tagged sentence as the following " nice/JJ dvd/NN player/NN", where **NN1 NN2** represents the extracted aspect "***dvd player"*** and "***nice***" represents an opinion word from OP (Htay & Lynn, 2013).

Rule #21 (RB JJ/OP **NN) :** For example, the sentence *" it is a very amazing **product.."*** match with this rule as the tagged sentence as the following "it/PRP is/VBZ a/DT very/RB amazing/JJ product/NN", where **NN** represents the extracted aspect "***product***", and "***amazing***" represents an opinion word from OP (Htay & Lynn, 2013).

Rule #22 (OP **NN) :** For example, the sentence "*poor **reliability**."* match with this rule as the tagged sentence as the following " poor/JJ reliability/NN", where "***poor***" is an OP, and **NN** represents the extracted aspect "***reliability***" (Htay & Lynn, 2013).

Rule #23 (**NN** OP)**:** For example, the sentence *" creative* ***software*** *stinks.***"** match with this rule as the tagged sentence as the following " creative/JJ software/NN stinks/VBZ", where **NN** represents the extracted aspect "***software***", and "***stinks***" is represents an opinion word from OP (Maharani et al., 2015).

Rule #24 (**NN1** IN NN2)**:** For example, the sentence "***audio*** *on video also lacking."* match with this rule as the tagged sentence as following "audio/NN on/IN video/NN also/RB lacking/VBG", where **NN1** represents the extracted aspect "***audio****"* (Maharani et al., 2015).

Rule #25 (**NN1** IN DT NN2)**:** For example, the sentence "*the* ***construction*** *of the player is the cheesiest i have ever seen."* match with this rule as the tagged sentence as the following "construction/NN of/IN the/DT player/NN", where **NN1** represents the extracted aspect "***construction****"* (Maharani et al., 2015).

Rule #26 (NN1 IN DT **NN2**)**:** For example, the sentence "*overall , a good buy for the* ***price*** ***.****"* match with this rule as the tagged sentence as the following "good/JJ buy/NN for/IN the/DT price/NN", where **NN2** represents the extracted aspect "***price****"* (Maharani et al., 2015).

Rule #27 (OP to **VB**)**:** For example, the sentence "*it has refused to* ***read*** *second discs."* match with this rule as the tagged sentence as the following "refused/VBN to read/VB", where "***refused***" represents an opinion word from OP and **VB** represents the extracted aspect "***read***" (Asghar et al., 2017).

Rule #28 (JJ to **VB** such that JJ not in OP)**:** For example, the sentence "*it is exceedingly simple to* ***navigate****."* match with this rule as the tagged sentence as the following "simple/JJ to navigate/VB", where "***simple***" is not in OP and represent the opinion word and **VB** represents the extracted aspect "***navigate***" (Asghar et al., 2017).

Rule #29 (JJ1 **JJ2 NN** (such that JJ2 not in OP))**:** For example, the sentence " *the g3 has much sharper **white offsets"*** match with this rule as the tagged sentence as the following " sharper/JJR white/JJ offsets/NNS", where "***white***" does not exist in OP and **JJ2 NN** represents the extracted aspect "***white offsets"***(Rana & Cheah, 2017).

Rule #30 (**NN**1 VBZ/VBP DT OP/JJ NN2)**:** For example, the sentence "*the **manual** does a fine job."* match with this rule as the tagged sentence as the following "the/DT manual/NN does/VBZ a/DT fine/JJ job/NN", where **NN1** represents the extracted aspect "***manual***", and "***fine***" represents an opinion word from OP (Rana & Cheah, 2018a).

Rule #31 (**NN** * PRP/DT OP where * means any pattern found but no NN in the pattern)**:** For example, the sentence "***apex*** *is the best cheap quality brand for dvd players ."* match with this rule as the tagged sentence as the following "apex/NN is/VBZ the/DT best/JJS", where **NN** represents the extracted aspect "***apex***", and "***best***" represents an opinion word from OP (Rana & Cheah, 2018a).

Rule #32 (VB OP/JJ **NN**)**:** For example, the sentence "*get great **reception."*** match with this rule as the tagged sentence as the following "get/VB great/JJ reception/NN", where **NN** represents the extracted aspect "***reception***", and "***great***" represents an opinion word from OP (Rana & Cheah, 2018a).

Rule #33 (**NN** VB OP/JJ)**:** For example, the sentence "***audio*** *is excellent"* match with this rule as the tagged sentence as the following "audio/NN is/VBZ excellent/JJ", where

NN represents the extracted aspect "***audio***", and "***excellent***" represents an opinion word from OP (Rana & Cheah, 2018a).

As discussed before, the dependency rules are better for formal text and pattern rules are better for informal text. However, these rules still have weaknesses, and in some cases, these rules cannot be used to extract aspects as in the sentence *"my favorite being the* ***games*** *and the* **pim***, and the* **radio***."*, because this sentence contains more than one aspect. Thus, the following new rules were developed.

3.5 Formulation of new extraction rules

To complement the existing rules of the aspect extraction, we also took the initiative to come out with some new rules to overcome the weaknesses of the existing rules. These new rules were developed based on manual observation, and by finding frequent occurring rules which have not been explored by the previous studies. In addition, some of the new rules were developed to extract aspects which cannot be extracted by the previous rules. Furthermore, some new rules can extract same aspects which can be extracted by the rules used from previous studies. However, the new rules are characterized by minimizing the number of incorrect extracted aspects which can be obtained by the previous rules by adding more restrictions on the new developed rules to minimize the extraction of incorrect aspects.

Rule #34 ((Any POS but not NN) **NN** VBZ RB OP/JJ)**:** This is a five-pattern rule and the first word can be any word type but not NN. For example, the sentence *"the* ***manual*** *is relatively clear."* match with this rule as the tagged sentence as the following " the/DT manual/NN is/VBZ relatively/RB clear/JJ", where **NN** represents the extracted aspect "***manual***", and *"**clear**"* represents an opinion word from OP.

Rule #35 (amod (**NN**, Any POS) such that OP found on left or right with any pattern in between but no NN exists in the pattern)**:** if 'amod' relation found in the sentence with the first argument is NN and the second argument is any POS type. In addition, there is an OP opinion word in the left or right of the extracted NN, but with no NN between the extracted aspect and OP, then **NN** will be extracted as an aspect as the following example *"in addition it comes with a sleek and powerful* ***headset****."*. There is "amod(headset-10, sleek-7)" relation and there are "***powerful***" and "***sleek***" opinion words on left of "***headset***" with no NN in between, then "***headset***" will be extracted as an aspect**.**

Rule #36 (nmod (**NN**, Any POS) such that OP found on left or right with any pattern in between but no NN exists in the pattern)**:** if 'nmod' relation found in the sentence with the first argument is NN and the second argument is any POS type. In addition, there is an OP opinion word in the left or right of the extracted NN, but with no NN between the extracted aspect and OP, then **NN** will be extracted as an aspect as the following example "*overall it is the best* ***camera*** *on the market* ." There is "nmod(camera-6, market-9)" relation and there is "***best***" opinion word on left of " ***camera***" with no NN in between the aspect and the opinion word, then "***camera***" will be extracted as an aspect**.**

Rule #37 (nummod (**NN**, Any POS) such that OP found on left or right with any pattern in between but no NN exists in the pattern)**:** if 'nummod' relation found in the sentence with the first argument is NN and the second argument is any POS type. In addition, there is an OP opinion word in the left or right of the extracted NN, but with no NN between the extracted aspect and OP, then **NN** will be extracted as an aspect as the following example *"usb 2.0* ***transfer*** *is insanely fast ."*. There is " nummod(transfer-3, 2.0-2)"relation and there are "***fast***" opinion words on right of "***transfer***" with no NN in between, then "***transfer***" will be extracted as an aspect**.**

Rule #38 (appos (**NN**, Any POS) such that OP found on left or right with any pattern in between but no NN exists in the pattern)**:** if 'appos' relation found in the sentence with the first argument is NN and the second argument is any POS type. In addition, there is an OP opinion word in the left or right of the extracted NN, but with no NN between the extracted aspect and OP, then **NN** will be extracted as an aspect as the following example *"i love my new **nomad**, its great !."*. There is "appos (nomad-5, !-9)" relation and there are "***love***" and "***great***" opinion words on right and left of " ***nomad*** " with no NN in between, then "***nomad***" will be extracted as an aspect**.**

Rule #39 (det (**NN**, Any POS) such that OP found on left or right with any pattern in between but no NN exists in the pattern)**:** if 'det' relation found in the sentence with the first argument is NN and the second argument is any POS type. In addition, there is an OP opinion word in the left or right of the extracted NN, but with no NN between the extracted aspect and OP, then **NN** will be extracted as an aspect as the following example *"the **price** was right."*. There is "det (price-2, the-1)" relation and there is ***"right"*** opinion word on right of "***price***" with no NN in between, then "***price***" will be extracted as an aspect**.**

Rule #40 (nsubjpass (Any POS, **NN**) such that OP found on left or right with any pattern in between but no NN exists in the pattern)**:** if 'nsubjpass' relation found in the sentence with the second argument is NN and the first argument is any POS type. In addition, there is an OP opinion word in the left or right of the extracted NN, but with no NN between the extracted aspect and OP, then **NN** will be extracted as an aspect as the following example *"the i**nterfac**e used could be better designed."*. There is " nsubjpass (designed-7, interface-2)" relation and there is "***better***" opinion word on right of "***interface***" with no NN in between, then "***interface***" will be extracted as an aspect**.**

Rule #41 (nmod (Any POS, **NN**) such that OP found on left or right with any pattern in between but no NN exists in the pattern)**:** if ' nmod ' relation found in the sentence with the second argument is NN and the first argument is any POS type. In addition, there is an OP opinion word in the left or right of the extracted NN, but with no NN between the extracted aspect and OP, then **NN** will be extracted as an aspect as the following example *"everything else about the* ***camera*** *is great."*. There is "nmod (everything-1, camera-5)" relation and there is "***great***" opinion word on right of "***camera***" with no NN in between, then "***camera***" will be extracted as an aspect**.**

Rule #42 (cop (**NN**, Any POS) such that OP found on left or right with any pattern in between but no NN exists in the pattern)**:** if 'cop' relation found in the sentence with the first argument is NN and the second argument is any POS type. In addition, there is an OP opinion word in the left or right of the extracted NN, but with no NN between the extracted aspect and OP, then **NN** will be extracted as an aspect as the following example *"this is a wonderful* ***camera****."*. There is " cop (camera-5, is-2)" relation and there is "***wonderful***" opinion word on left of "***camera***" with no NN in between, then "***camera***" will be extracted as an aspect**.**

Rule #43 (**NN NN * OP** where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"their* ***customer service*** *is very poor."* match with this rule as the tagged sentence as following " their/PRP$ customer/NN service/NN is/VBZ very/RB poor/JJ", where **NN NN** represents the extracted aspect "***customer service***", *and* ***"poor"*** represents the opinion word in OP**.**

Rule #44 (**NN** RB (RB not in OP) * OP where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence "***interface*** *practically seamless ."* match with this rule as the tagged sentence as the following "interface/NN practically/RB seamless/VBZ", where **NN** represents the extracted aspect "***interface***", *and* ***"seamless"*** represents the opinion word in OP**.**

Rule #45 (**NN** VBD (VBD not in OP) * OP where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence "*the included* ***earbuds*** *were uncomfortable****.***" match with this rule as the tagged sentence as the following "the/DT included/JJ earbuds/NNS were/VBD uncomfortable/JJ", where **NN** represents the extracted aspect "***earbuds***", *and* ***"uncomfortable*"** represents the opinion word in OP**.**

Rule #46 ((Any POS but not NN) OP **NN**)**:** This a three-pattern rule and the first word can be any word type but not NN. For example, the sentence *"very comfortable* ***camera"*** match with this rule as the tagged as following "very/RB comfortable/JJ camera/NN", where **NN** represents the extracted aspect "***camera***", and "***comfortable***" is an opinion word in OP.

Rule #47 (NN1 of **NN2**)**:** For example, the sentence *"it does play a wide range of* ***formats****."* match with this rule as the tagged sentence as the following "range/NN of/IN formats/NNS", where **NN2** represents the extracted aspect "***formats***".

Rule #48 (DT (Any POS) **NN** VBZ JJ/OP)**:** This a five-pattern rule and the second word can be any POS word type. For example, the sentence *"the* ***sound quality*** *is okay"* match with this rule as the tagged sentence as the following "The/DT sound/JJ quality/NN is/VBZ okay/JJ ", where **(Any POS) NN** represents the extracted aspect "***sound quality***", and "***okay***" is a JJ represents the opinion word in OP.

Rule #49 (DT **NN** IN * JJ/OP where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"you can see the **interface** as modern or classical ."* match with this rule as the tagged sentence as the following "The/DT interface/NN as/IN modern/JJ or/CC classical/JJ", where **NN** represents the extracted aspect "***interface***", while ***"modern"** and **"classical"*** represent the opinion words in OP**.**

Rule #50 (DT **NN NN** * JJ/OP where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"the **software interface** supplied was very easy to use"* match with this rule as the tagged sentence as the following "The/DT software/NN interface/NN supplied/VBD was/VBD very/RB easy/JJ", where **NN NN** represents the extracted aspect "***software interface***", *and* "***easy"*** represents the opinion word in OP**.**

Rule #51 (**NN**1 NN2 VBZ/VBP RB RB)**:** For example, the sentence "*the **mms** technology is very well integrated with this phone , which you will enjoy ."* match with this rule as the tagged sentence as the following "mms/NNS technology/NN is/VBZ very/RB well/RB", where **NN1** represents the extracted aspect "***mms***".

Rule #52 (**NN NN** VB OP/JJ)**:** For example, the sentence "***replacement battery*** *is expensive."* match with this rule as the tagged sentence as the following "replacement/NN battery/NN is/VBZ expensive/JJ", where **NN NN** represents the extracted aspect "***replacement battery***", *and* "***expensive***" represents the opinion word in OP**.**

Rule #53 (**NN** (any POS) RB JJ/OP)**:** This a four-pattern rule and the second word can be any POS word type. For example, the sentence *"the **controls** are very intuitive"* match with this rule as the tagged sentence as the following "The/DT controls/NNS are/VBP very/RB intuitive/JJ", where **NN** represents the extracted aspect "***controls***", and "***intuitive***" represents the opinion word in OP.

Rule #54 (OP (any POS) OP/JJ **NN**)**:** This a four-pattern rule and the second word can be any word type. For example, the sentence *"like the smart **volume**"* match with this rule as the tagged sentence as the following "like/IN the/DT smart/JJ volume/NN", where **NN** represents the extracted aspect "***volume***", and "***like***" represents the first opinion word in OP and "***smart***" represents the second opinion word in OP.

Rule #55 (DT **NN NN** VB RB)**:** For example, the sentence *"the **lens cover** is surely loose"* match with this rule as the tagged sentence as the following "The/DT lens/NN cover/NN is/VBZ surely/RB", where **NN NN** represents the extracted aspect "***lens cover***".

Rule #56 (DT **NN NN** VB OP/JJ)**:** For example, the sentence *"the **storage capacity** is great for me"* match with this rule as the tagged sentence as the following "the/DT storage/NN capacity/NN is/VBZ great/JJ ", where **NN NN** represents the extracted aspect "***storage capacity***"**,** and "***great***" represents the opinion word in OP.

Rule #57 (DT **NN** VB RB * OP [such that RB not in OP] where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"the **headphones** are n't the best"* match with this rule as the tagged sentence as the following "the/DT headphones/NNS are/VBP n't/RB the/DT best/JJS", where **NN** represents the extracted aspect "***headphones***", and "***best***" represents the opinion word in OP**.**

Rule #58 (DT **NN** VB JJ/OP)**:** For example, the sentence "*the **grain** was terrible"* match with this rule as the tagged sentence as the following "the/DT grain/NN was/VBD terrible/JJ", where **NN** represents the extracted aspect "***grain***", and "***terrible"*** represents the opinion word in OP**.**

Rule #59 (OP * **NNS** where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"lack of good **accessories**."* match with this rule as the tagged sentence as the following "lack/NN of/IN good/JJ accessories/NNS", where **NNS** represents the extracted aspect "***accessories***", and "***lack"*** represents the opinion word in OP**.**

Rule #60 (DT **NN** * OP where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"this **camera** is closest to perfect."* match with this rule as the tagged sentence as the following "this/DT camera/NN is/VBZ closest/JJS to/TO perfect/JJ", where **NN** represents the extracted aspect "***camera***", and "***perfect"*** represents the opinion word in OP**.**

Rule #61 (RB JJ/OP * **NN** where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"very flexible and powerful **features.**"* match with this rule as the tagged sentence as the following "very/RB flexible/JJ and/CC powerful/JJ features/NNS", where **NN** represents the extracted aspect "***features***", and "***flexible"*** represents the opinion word in OP**.**

Rule #62 (RB JJ/OP * **NN NN** where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *" i am really impressed by this **dvd player** ."* match with this rule as the tagged sentence as the following "really/RB impressed/VBN by/IN this/DT dvd/NN player/NN", where **NN NN** represents the extracted aspect "***dvd player***", and *"**impressed**"* represents the opinion word in OP**.**

Rule #63 (OP/JJ IN DT **NN**)**:** For example, the sentence *"either way , can't go wrong with this **price**"* match with this rule as the tagged sentence as the following "wrong/JJ with/IN this/DT price/NN", where **NN** represents the extracted aspect "***price***", and ***"wrong"*** represents the opinion word in OP**.**

Rule #64 (OP * **NN NN** where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"i knew this before hand , and it is not that bad there is no **tiff format** ."* match with this rule as the tagged sentence as following "Bad/JJ there/EX is/VBZ no/DT tiff/NN format/NN", where **NN NN** represents the extracted aspect " ***tiff format***", and ***"bad "*** represents the opinion word in OP**.**

Rule #65 (DT JJ/OP **NN**)**:** For example, the sentence *"overall it is a great **unit**."* match with this rule as the tagged sentence as the following "a/DT great/JJ unit/NN", where **NN** represents the extracted aspect "***unit***", and **"*great"*** represents the opinion word in OP**.**

Rule #66 (DT **NN** VBZ/VBP RB JJ/OP)**:** For example, the sentence *"the **vibration** is not top"* match with this rule as the tagged sentence as the following "The/DT vibration/NN is/VBZ not/RB top/JJ", where **NN** represents the extracted aspect "***vibration***", and "***top"*** represents the opinion word in OP**.**

Rule #67 (**NN NN** VBZ JJ/OP)**:** For example, the sentence *"**color screen** is good"* match with this rule as the tagged sentence as the following "Color/NN screen/NN is/VBZ good/JJ", where **NN NN** represents the extracted aspect "***color screen***", and "***good"*** represents the opinion word in OP**.**

Rule #68 (DT RB JJ/OP **NN**)**:** For example, the sentence *"it 's a very intuitive **program**"* match with this rule as the tagged sentence as the following " a/DT very/RB intuitive/JJ program/NN", where **NN** represents the extracted aspect "***program***", and "***intuitive"*** represents the opinion word in OP**.**

Rule #69 (DT VBZ/VBP DT JJ/OP **NN**)**:** For example, the sentence *"this is a good **deal**"* match with this rule as the tagged sentence as the following "this/DT is/VBZ a/DT good/JJ deal/NN", where **NN** represents the extracted aspect "***deal***", and "***good"*** represents the opinion word in OP**.**

Rule #70 (RB VB JJ/OP **NN**)**:** For example, the sentence *"does not provide enough **volume"*** match with this rule as the tagged sentence as the following "not/RB provide/VB enough/JJ volume/NN", where **NN** represents the extracted aspect "***volume***", and "***enough"*** represents the opinion word in OP**.**

Rule #71 (OP/JJ DT **NN**)**:** For example, the sentence *"i would recommend this **product** to anyone."* match with this rule as the tagged sentence as the following "recommend/VB this/DT product/NN", where **NN** represents the extracted aspect "***product***", and "***recommend***" represents the opinion word in OP**.**

Rule #72 (DT JJ/OP JJ/OP **NN**)**:** For example, the sentence *"the nomad zen could use a little sturdier **construction"*** match with this rule as the tagged sentence as the following "a/DT little/JJ sturdier/JJR construction/NN", where **NN** represents the extracted aspect "***construction***", and "***little"*** is a JJ represents an opinion word which not exists in OP *and* "***sturdier***" represents the opinion word in OP**.**

Rule #73 (**NN** RB OP/JJ)**:** For example, the sentence "***interface** practically seamless*" match with this rule as the tagged sentence as the following "interface/NN practically/RB seamless/JJ", where **NN** represents the extracted aspect "***interface***", and "***seamless***" represents the opinion word in OP**.**

Rule #74 (VBZ/VBP RB OP/JJ **NN**)**:** For example, the sentence "*the phone is very light* ***weight***" match with this rule as the tagged sentence as the following "the/DT phone/NN is/VBZ very/RB light/JJ weight/NN", where **NN** represents the extracted aspect "***weight***", and ***"light"*** is a JJ which not exists in OP**.**

Rule #75 (NN1 VB JJ/OP **NN**2)**:** For example, the sentence *"the zen has minimal* ***stoppage*** *between tracks"* match with this rule as the tagged sentence as the following "the/DT zen/NN has/VBZ minimal/JJ stoppage/NN", where **NN2** represents the extracted aspect "***stoppage***", and "***minimal***" is an JJ which not exists in OP**.**

Rule #76 (OP/JJ TO VB (Any POS) **NN**)**:** This a five-pattern rule and the fourth word can be any POS word type. For example, the sentence *"very convenient to scroll in* ***menu"*** match with this rule as the tagged sentence as the following "convenient/JJ to/TO scroll/VB in/IN menu/NN", where **NN** represents the extracted aspect "***menu***", and "***convenient***" represents the opinion word in OP.

Rule #77 (JJ/OP JJ/OP **NN**)**:** For example, the sentence *"polyphonic sweet* ***tunes***" match with this rule as the tagged sentence as the following "polyphonic/JJ sweet/JJ tunes/NNS", where **NN** represents the extracted aspect "***tunes***", and "***polyphonic"*** is an JJ which not exists in OP and "***sweet***" represents the opinion word in OP**.**

Rule #78 (JJ/OP **NN** RB)**:** For example, the sentence *"bad* **picture** *yet*" match with this rule as the tagged sentence as the following "bad/JJ picture/NN yet/RB", where **NN** represents the extracted aspect "***picture***", and "***bad***" represents the opinion word in OP**.**

Rule #79 (DT **NN** VBZ/VBP VBN)**:** For example, the sentence *"the **display** is hinged"* match with this rule as the tagged sentence as following " the/DT display/NN is/VBZ hinged/VBN", where **NN** represents the extracted aspect "***display***", and "***hinged***" represents the opinion word which not exists in OP**.**

Rule #80 (**NN1** CC **NN2** VBP JJ/OP)**:** For example, the sentence *"**option**s and **control**s are easy."* match with this rule as the tagged sentence as the following "options/NNS and/CC controls/NNS are/VBP easy/JJ", where **NN1** and **NN2** represent the extracted aspects which are "***option**s*" *and* "***control**s*", while "***easy***" represents the opinion word in OP**.**

Rule #81 (PRP VBZ JJ/OP **NN**)**:** For example, the sentence *"it takes great **pictures**"* match with this rule as the tagged sentence as the following "it/PRP takes/VBZ great/JJ pictures/NNS", where **NN** represent the extracted aspect "***pictures***", and "***great***" represents the opinion word in OP**.**

Rule #82 (RB OP/JJ DT **NN**)**:** For example, the sentence *"just received this camera two days ago and already love the **features** it has."* match with this rule as the tagged sentence as the following "already/RB love/VB the/DT features/NNS", where **NN** represents the extracted aspect "***features***", and "***love***" represents the opinion word in OP**.**

Rule #83 (DT **NN NN** VBZ/VBP JJ/OP)**:** For example, the sentence *"the **menu options** are uncreative"* match with this rule as the tagged sentence as the following "the/DT menu/NN options/NNS are/VBP uncreative/JJ", where **NN NN** represent the extracted aspect "***menu options***", and "***uncreative***" is a JJ represents the opinion word which not exists in OP**.**

Rule #84 ((Any POS but not NN) **NN NN** VBZ JJ/OP)**:** This a five-pattern rule and the first word can be any POS word type but not NN. For example, the sentence "*the little **jog dial** seems weak and quirky.*" match with this rule as the tagged sentence as the following "little/JJ jog/NN dial/NN seems/VBZ weak/JJ", where **NN NN** represents the extracted aspect " ***jog dial*** ", and ***"weak"*** represents the opinion word in OP.

Rule #85 (DT **NN** VBZ/VBP JJ/OP)**:** For example, the sentence *" kind of bulky and the **wheel** is awkward , but i can deal with that."* match with this rule as the tagged sentence as the following "the/DT wheel/NN is/VBZ awkward/JJ", where **NN** represent the extracted aspect "***wheel***", and "***awkward***" represents the opinion word in OP**.**

Rule #86 (**NN NN** VB (Any POS) JJ/OP)**:** This a five-pattern rule and the fourth word can be any POS word type. For example, the sentence *"the **picture quality** are so great."* match with this rule as the tagged sentence as the following "the/DT picture/NN quality/NN are/VBP so/RB great/JJ", where **NN NN** represents the extracted aspect "***picture quality***", and ***"great"*** represents the opinion word in OP.

Rule #87 (DT **NN** VBZ/VBP DT JJ/OP)**:** For example, the sentence "*this **phone** is a winner.*" match with this rule as the tagged sentence as the following "this/DT phone/NN is/VBZ a/DT winner/NN", where **NN** represents the extracted aspect "***phone***", and "***winner***" represents the opinion word in OP.

Rule #88 (PRP JJ/OP * DT **NN** where * means any pattern found)**:** For example, the sentence *"my favorite being the **games** and the **pim**, and the **radio**."* match with this rule as the tagged sentence as the following "my/PRP favorite/NN being/VBG the/DT games/NNS and/CC the/DT pim/NN and/CC the/DT radio/NN", where **NN** represents the extracted aspects which are "***games***", "***pim***", and "***radio***". Whereas, "***favorite***" represents the opinion word in OP**.**

Rule #89 (DT NN1 VB * JJ/OP **NN**2 where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"the g3 is loaded with many useful **features** ."* match with this rule as the tagged sentence as the following "the/DT g3/NN is/VBZ loaded/VBN with/IN many/JJ useful/JJ features/NNS", where **NN2** represents the extracted aspect "***features***", and "***useful***" represents the opinion word in OP**.**

Rule #90 (DT **NN** * DT NNS * JJ/OP where * means any pattern found)**:** For example, the sentence *"the **player** itself has all sorts of problems ."* match with this rule as the tagged sentence as the following "the/DT player/NN itself/PRP has/VBZ all/DT sorts/NNS of/IN problems/NNS", where **NN** represents the extracted aspect "***player***", and "***problems***" represents the opinion word in OP**.**

Rule #91 (JJ/OP NN * **NN NN** where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"my favorite features , although there are many , are the **speaker phone** , the radio and the infrared ."* match with this rule as the tagged sentence as the following "favorite/JJ features/NNS although/IN there/EX although/IN there/EX are/VBP many/JJ are/VBP the/DT speaker/NN phone/NN", where **NN NN** represents the extracted aspect "***speaker phone***", and "***favorite***" represents the opinion word in OP**.**

Rule #92 (DT **NN** RB VB)**:** For example, the sentence *"the **calls** constantly drop in my area and i experince mega-static"* match with this rule as the tagged sentence as the following "the/DT calls/NNS constantly/RB drop/VBP", where **NN** represents the extracted aspect "***calls***".

Rule #93 (**NN** VB RB OP/JJ TO VB)**:** For example, the sentence *"the **menus** are very easy to navigate."* match with this rule as the tagged sentence as the following "menus/NNS are/VBP very/RB easy/JJ to/TO navigate/VB", where **NN** represents the extracted aspect "***menus***", and ***"easy"*** represents the opinion word in OP.

Rule #94 (DT RB JJ/OP **NN** (Any POS but not NN))**:** This a five-pattern rule and the fifth word can be any POS word type but not NN. For example, the sentence *"but at least youre starting with the most photorealistic **images** ive ever seen from a camera ."* match with this rule as the tagged sentence as the following "the/DT most/RBS photorealistic/JJ images/NNS", where **NN** represents the extracted aspect ***"images***", and ***"photorealistic"*** is a JJ and does not exists in OP.

Rule #95 (JJ/OP IN DT (Any POS) **NN**)**:** This a five-pattern rule and the fourth word can be any POS word type. For example, the sentence *"i am bored with the silver **look**."* match with this rule as the tagged sentence as the following "bored/VBN with/IN the/DT silver/NN look/NN", where **NN** represents the extracted aspect "***look***", and "***bored***" represents the opinion word in OP.

Rule #96 (nsubj (NN1, **NN2**) and amod (NN1, JJ/OP))**:** means that two relations exist in the sentence with first relation of type "nsubj" and second relation of type "amod". Furthermore, both relations share the same first argument which is NN1. In addition, the second argument of "nsubj" relation is NN2 and the second argument of "amod" can be either JJ or OP, then **NN2** represents the extracted aspect as the following example "***t-***

mobile *was a pretty good server.*" There are "nsubj (server-6, t-mobile-1), amod (server-6, good-5)" relations and the extracted aspect is "***t-mobile***"

Rule #97 (PRP VB (Any POS) **NN**)**:** This a four-pattern rule and the third word can be any POS type. For example, the sentence *"it takes wonderful* **pictures** *very easily in auto mode."* match with this rule as the tagged sentence as the following "it/PRP takes/VBZ wonderful/JJ pictures/NNS", where **NN** represents the extracted aspect "***pictures***", and "***wonderful***" represents the opinion word in OP.

Rule #98 (JJ/OP NN1 (any pos) DT **NN2**)**:** This a five-pattern rule and the third word can be any POS type. For example, the sentence *"the only drawback is the* ***viewfinder*** *is slightly blocked by the lens."* match with this rule as the tagged sentence as the following "only/JJ drawback/NN is/VBZ the/DT viewfinder/NN", where **NN2** represents the extracted aspect "***viewfinder***".

Rule #99 (DT **NN** * RB RB OP/JJ where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"when talking the* **voice** *is not very clear"* match with this rule as the tagged sentence as the following "the/DT voice/NN is/VBZ not/RB very/RB clear/JJ" where **NN** represents the extracted aspect "***voice***", and "***clear***" represents the opinion word in OP**.**

Rule #100 (DT **NN NN** * DT NN * JJ/OP where * means any pattern found)**:** For example, the sentence *"the* **battery life** *on this phone is surreal."* match with this rule as the tagged sentence as the following "the/DT battery/NN life/NN on/IN this/DT phone/NN is/VBZ surreal/JJ", where **NN NN** represents the extracted aspect "***battery life***" *and "**surreal**"* represents the opinion words in OP**.**

Rule #101 (PRP VB DT JJ/OP **NN**)**:** For example, the sentence *"it is a perfect* ***phone****."* match with this rule as the tagged sentence as following " it/PRP is/VBZ a/DT perfect/JJ phone/NN", where **NN** represents the extracted aspect ***"phone***", and ***"perfect"*** represents the opinion word in OP.

Rule #102 (OP (Not OP) **NN NN**)**:** This a four-pattern rule and the second word can be any word type but not in OP. For example, the sentence *"excellent polyphonic* ***ringing tones."*** match with this rule as the tagged sentence as the following "excellent/JJ polyphonic/JJ ringing/NN tones/NNS", where **NN NN** represents the extracted aspect "***ringing tones***", and "***excellent***" represents the opinion word in OP.

Rule #103 (**NN NN** * JJ/OP NN where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence ***"sunset feature*** *takes incredible pics in the morning , and the evening !."* match with this rule as the tagged sentence as following "sunset/NN feature/NN takes/VBZ incredible/JJ pics/NNS in/IN the/DT morning/NN", where **NN NN** represents the extracted aspect "***sunset feature***", and "***incredible***" represents the opinion word in OP**.**

Rule #104 (OP/JJ (Any POS) **NN1** CC **NN2**)**:** This a five-pattern rule and the second word can be any POS word type. For example, the sentence *"the system is terrific in* ***size*** *and* ***design"*** match with this rule as the tagged sentence as the following "the/DT system/NN is/VBZ terrific/JJ in/IN size/NN and/CC design/NN", where **NN1** and **NN2** represent the extracted aspects as "***size***" and "***design***", and "***terrific***" represents the opinion word in OP**.**

Rule #105 (RB JJ/OP (Any POS) **NN**)**:** This a four-pattern rule and the third word can be any POS type. For example, the sentence "*that is solved with the very comfortable* *handsfree* ***ear-piece*** *which is included.*" match with this rule as the tagged sentence as the following "very/RB comfortable/JJ handsfree/NN ear-piece/NN", where **NN** represents the extracted aspect "***ear-piece***", and "***comfortable***" represents the opinion word in OP.

Rule #106 (**NN NN** (such that no JJ/OP found in the sentence))**:** For example, the sentence "*the zen does not have a* ***stop button*** *!*" match with this rule as the tagged sentence as the following "the/DT zen/NN does/VBZ not/RB have/VB a/DT stop/NN button/NN", where **NN NN** represents the extracted aspect "***stop button***" and no opinion word exists in the sentence.

Rule #107 (DT JJ/OP **NN NN**)**:** For example, the sentence "*the 2nd dvd player had a faulty* ***power supply*** *which caused to occasionally not turn on.*" match with this rule as the tagged sentence as the following "a/DT faulty/JJ power/NN supply/NN", where **NN NN** represents the extracted aspect "***power supply***", and "***faulty***" represents the opinion word in OP.

Rule #108 (DT (Any POS) **NN** * OP where * means any pattern found but no NN exists in the pattern and second word can be any POS type)**:** For example, the sentence *"the included* ***earbuds*** *work quite well."* match with this rule as the tagged sentence as the following "the/DT included/JJ earbuds/NNS work/VBP quite/RB well/RB", where **NN** represents the extracted aspect "***earbuds***"**.** In addition, "***work***" *and* "***well"*** represent the opinion words in OP**.**

Rule #109 (compound(NN1,NN2) such that "OP", "no", or "not" found on left or "OP" found on right with any pattern in between but no NN exists in the pattern)) **:** if 'compound' relation found in the sentence with first argument is NN1 and the second argument is NN2, then **NN2 NN1** will be extracted as an aspect as the following example *"well **flash photos** are never great.".* There is " compound (photos-4, flash-3)" relation between "***photos***" and "***flash***" aspect, then " ***flash photos*** " will be extracted as an aspect. In addition, there is an opinion word on the right of the aspect and this opinion word is "***great***" which exists in OP.

Rule #110 (**JJ NN** * OP (such that JJ not in OP and * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"**manual functionality** is excellent."* match with this rule as the tagged sentence as the following "manual/JJ functionality/NN is/VBZ excellent/JJ", where **JJ NN** represents the extracted aspect "***manual functionality***", and "***excellent***" represents the opinion word in OP**.**

Rule #111 (RB JJ/OP IN DT **NN**)**:** For example, the sentence *"i am extremely pleased with this **camera**."* match with this rule as the tagged sentence as the following "extremely/RB pleased/JJ with/IN this/DT camera/NN", where **NN** represents the extracted aspect "***camera***", and "***pleased***" represents the opinion word in OP.

Rule #112 (OP * **JJ NN** (such that JJ not in OP and * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"the radio feature has superb **sound quality** ."* match with this rule as the tagged sentence as the following "the/DT radio/NN feature/NN has/VBZ superb/JJ sound/JJ quality/NN", where **JJ NN** represents the extracted aspect "***sound quality***", and "***superb***" represents the opinion word in OP**.**

Rule #113 (DT **NN1** * DT NN2 * JJ/OP where * means any pattern found)**:** For example, the sentence *"the **sound** of the player is pretty good."* match with this rule as the tagged sentence as the following "the/DT sound/NN of/IN the/DT player/NN is/VBZ pretty/RB good/JJ", where **NN1** represents the extracted aspect "***sound***"**.** In addition**,** "***pretty***", *and* "***good***" represent the opinion words in OP**.**

Rule #114 (JJ/VBN/OP * DT **NN** where * means any pattern found but no NN exists in the pattern)**:** For example, the sentence *"i would not be inclined to purchase an **apex** product again ."* match with this rule as the tagged sentence as the following "inclined/VBN to/TO purchase/VB an/DT apex/NN", where **NN** represents the extracted aspect "***apex***", and ***"inclined"*** represents VBN which does not exists in OP**.**

Rule #115 (JJ/OP NN1 (Any POS) **NN2**)**:** This a four-pattern rule and the third word can be any POS type. For example, the sentence *"this phone has a very cool and useful feature the **speakerphone**"* match with this rule as the tagged sentence as the following "useful/JJ feature/NN the/DT speakerphone/NN", where **NN2** represents the extracted aspect "***speakerphone***", and "***useful***" represents the opinion word in OP.

Rule #116 (RB VB * **NN** where * means any pattern that contain OP/JJ and no NN exists in between)**:** For example, the sentence *"i will never buy a creative **product** again."* match with this rule as the tagged sentence as the following "never/RB buy/VB a/DT creative/JJ product/NN", where **NN** represents the extracted aspect "***product***"**.**

Rule #117 (DT **NN NN** VBZ JJ/OP)**:** For example, the sentence "*the **auto mode** is good.*" match with this rule as the tagged sentence as the following "the/DT auto/NN mode/NN is/VBZ good/JJ", where **NN NN** represents the extracted aspect "***auto mode***", and "***good***" represents the opinion word in OP.

Rule #118 (nsubj (OP/JJ, NN1) and conj (**NN1**, **NN2**))**:** means that two relations exist in the sentence with the first relation of type "nsubj" with the first argument is either JJ or OP and the second argument is NN1. In addition, the second relation of type "conj" with the first argument is NN1 and equal the second argument of "nsubj" relation and the second argument is also NN2. Furthermore, the extracted aspects are **NN1** and **NN2** as the following example *"the **options** and **controls** are easy to use and logically laid out ."*. There are "nsubj (easy-6, options-2), conj(options-2, controls-4)" relations and the extracted aspects are "***options***" and "***controls***".

Rule #119 (xcomp (OP, **VB**)) **:** if ' xcomp' relation found in the sentence with the first argument is in OP and the second argument is VB, then the second argument will be extracted as an aspect as the following example *"easy to **use**"*. There is "xcomp(easy-1, use-3)" relation between "***easy***" opinion word and "***use***" aspect, then "***use***" will be extracted as an aspect.

Rule #120 (amod (NN1, JJ/OP) and amod (**NN1, JJ**2) such that JJ2 not in OP)**:** means that two relations exist in the sentence with type "amod" with the first argument is NN1 and the second argument is either JJ or OP in first relation. In addition, the second relation of type "amod" with the first argument is NN1 and equal the first argument of first "amod" relation and the second argument is JJ2 but not exist in OP. Furthermore, the extracted aspect is **JJ2 NN1** as the following example *"simply , the canon g3 is the best **digital camera**"* . There are "amod (camera-10, best-8), amod (camera-10, digital-9)" relations and the extracted aspect is "***digital camera***", and "***best***" represents the opinion word in OP.

Rule #121 (nmod (JJ1/OP1, NN1) and nsubj (JJ1/OP1, **NN2**))**:** means that two relations exist in the sentence with first relation of type "nmod" with the first argument is either JJ1 or OP1 and the second argument is NN1. In addition, the second relation of type "nsubj" with the first argument is either JJ1 or OP1 and equal the first argument of "nmod" relation and the second argument is NN2. Furthermore, the extracted aspect is **NN2** as the following example "***door*** *broke after a month ."*. There are "nmod (broke-2, month-5) nsubj (broke-2, door-1)" relations and the extracted aspect is "***door***", and "***broke***" represents the opinion word in OP.

Rule #122 (nmod (OP/JJ, NN1) and compound (**NN1**, **NN2**))**:** means that two relations exist in the sentence with first relation of type "nmod" with the first argument is either JJ or OP and the second argument is NN1. In addition, the second relation of type "compound" with first argument is NN1 and equal the second argument of "nmod" relation and the second argument is NN2. Furthermore, the extracted aspect is **NN2 NN1** as the following example *"i 've been pleased with the* ***picture quality."***. There are " nmod (pleased-5, quality-9), compound (quality-9, picture-8)" relations and the extracted aspect is "***picture quality***", and "***pleased***" represents the opinion word in OP.

Rule #123 (dobj (H1, **NN1**) and amod (**NN1**, JJ/OP))**:** means that two relations exist in the sentence with first relation type "dobj" with the first argument can be any POS word type and the second argument is NN1. In addition, the second relation of type "amod" with the first argument is NN1 and equal the second argument of "dobj" relation and the second argument is either JJ or OP. Furthermore, the extracted aspect is **NN1** as the following example *"you can also assign special* **rings** *to special people when they call"*. There are "dobj (assign-5, rings-7), amod (rings-7, special-6)" relations and the extracted aspect is "***rings***", and "***special***" represents a JJ which does not exists in OP.

Rule #124 (nsubj (JJ/OP, **NN1**) and det (**NN1**, H1))**:** means that two relations exist in the sentence with first relation type "nsubj" with the first argument can be either JJ or OP and the second argument is NN1. In addition, the second relation of type "det" with the first argument is NN1 and equal the second argument of "nsubj" relation and the second argument can be any POS word type. Furthermore, the extracted aspect is **NN1** as the following example *"the **prints** are beautiful."*. There are "nsubj (beautiful-4, prints-2), det (prints-2, the-1)" relations and the extracted aspect is "***prints***", and "***beautiful***" represents opinion word in OP.

Rule #125 (nmod (JJ/OP, NN1) and conj (**NN1**, **NN2**))**:** means that two relations exist in the sentence with first relation of type "nmod" with the first argument is either JJ or OP and the second argument is NN1. In addition, the second relation of type "conj" with first argument is NN1 and equal the second argument of "nmod" relation and the second argument is also NN2. Furthermore, the extracted aspects are **NN1** and **NN2** as the following example *"i am very pleased with its **quality** and **durability**"*. There are "nmod (pleased-4, quality-7), conj (quality-7, durability-9)" relations and the extracted aspects are "***quality***" and "***durability*** ". In addition, ***"pleased"*** represents opinion word in OP.

Rule #126 (nsubj (JJ/OP, **NN1**) and amod (**NN1**, H1))**:** means that two relations exist in the sentence with first relation type "nsubj" with the first argument can be either JJ or OP and the second argument is NN1. In addition, the second relation of type "amod" with first argument is NN1 and equal to the second argument of "nsubj" relation and the second argument can be any POS word type. Furthermore, the extracted aspect is **NN1** as the following example *"main **dial** is not backlit"*. There are "nsubj (backlit-5, dial-2), amod (dial-2, main-1)" relations and the extracted aspect is "***dial***", and ***"backlit"*** represents a JJ which does not exists in OP.

These new rules are developed to extract aspects that have not been extracted by the rules from the previous studies. For example, a sentence "i am really **impressed** by this **dvd player**", the extracted aspect is "***dvd player***" which can be extracted by the new developed Rule #62, and "***impressed***" is not an adjective but can be found in opinion lexicon. However, this aspect cannot be extracted by the existing rules which based on adjective as opinion word only. In addition, a new improvement was added to the rules from the previous studies (Rule #1 to Rule #33) using OP/JJ combinations. This combination will cover all possible types of opinion words, as the previous studies only consider the opinion word as either adjective or they used only the opinion lexicon, but not both. In the case of considering opinion lexicon or adjective only as an opinion word, the extraction algorithm will miss many correct aspects and cannot extract these aspects. For example, in the sentence *"main **dial** is not **backlit**"*, "***backlit***" represents an adjective which is used in the sentence to represents the opinion words, but the word "***backlit***" was not found in opinions lexicon. So, the previous studies which based on opinion lexicon only for defining an opinion word cannot extract the aspect "***dial***". In addition, the new rules are developed to improve low precision problem rules of the previous studies. As there are new rules which can extract the same aspects which were extracted by the previous rules, but with better precision by these new rules based on the added restrictions to new rules. Another example, taken from Rule# 66, in the sentence *"the **vibration** is not **top**"*, the previous studies extract "***vibration***" as the correct aspect but when they extracted it, they consider "***top***" as the opinion word. However, in this example before "***top***" opinion word there is "***not***" word, this restriction was added to Rule# 66 to extract accurate and reliable results. An example by Rule# 71 in the sentence *"i would **recommend** this **product** to anyone."*, it cannot be extracted by the previous studies which consider an adjective as only the opinion word because the opinion word "***recommend***" is a verb and was found in the opinion lexicon. Therefore, this problem was resolved by Rule# 71 using combination OP/JJ. Further example in Rule# 88 in the sentence *" my favorite being the* ***games*** *and the* **pim***, and the* **radio***."*, where in this case regular expression can extract all these aspects "***games*****",** "***pim*****",** and "***radio***", which were not extracted by the previous studies. An example of the advantage of using restrictions in the new developed rules as in Rule# 89 (DT NN1 VB * JJ/OP **NN**2 where * means any pattern found but no NN exist in the pattern). For example in sentence *"the g3 is loaded with many useful* **features** *."*, some of the previous studies extracted "***g3***" as the correct aspect, but the opinionated target aspect in this sentence is ***features***. Thus, Rule# 89 solved this problem and extract ***features*** as the correct aspect. These are some examples which show the importance of the new developed rules. Table 3.2 presents summary of all aspect extraction rules which are used in this research:

Table 3.2: Aspect Extraction Rules

Rule#	Rule	Example (Hu & Liu, 2004a)
1	nsubj(JJ/OP,**NN**)	***video** was poor*
2	ReL1(H1,**NN1**) and ReL2(H1,**NN2**)	*proven canon built **quality** and **lens***
3	nsubj(VB1,H1) and dobj(VB1,**NN**)	*honestly , i love this **player** .*
4	nsubj(H1,**NN**) and xcomp(H1,JJ/OP)	*it 's **size** also makes it ideal for travel*
5	amod(NN1,OP/JJ) and conj(**NN1**,**NN2**)	*it plays original **dvds** and **cds***
6	nmod(OP/JJ,**NNS**)	*i find the lack of entertaining **games** on this phone quite disturbing .*
7	amod(**NN**,OP)	*the poor **manual** .*
8	ReL1(H1,**NN**) and ReL2(H1, OP/JJ)	*this **camera** has a major design flaw .*
9	nsubj(**NN**,OP/JJ)	*my only gripe about the hardware is the **buttons***
10	dobj(OP/JJ,**NN**)	*i especially like the more commonly used **buttons** .*
11	nsubj(OP/JJ,NN1) and compound(**NN1**,**NN2**)	*i find **onscreen displays** annoying .*
12	conj(**NN1**,**NN2**)	*proven canon built **quality** and **lens***
13	amod(NN1,OP/JJ) and compound(**NN1**,**NN2**)	*overall , the g3 delivers what must be considered the best **image quality***
14	neg(OP/JJ, H1) and nsubj(OP/JJ,**NN**)	*the **colors** on the screen are not as crisp as i 'd have liked them to be .*
15	ReL(**NN**, OP/JJ)	*definetely a great **camera***
16	ReL(OP/JJ ,**NN**)	*the **manual** is relatively clear*
17	nsubj(OP1/JJ1,**NN**) and cop (OP1/JJ1,H1)	*the **menus** are easy to navigate.*
18	**NNS** VBP OP /JJ	***controls** are poorly designed*
19	**NN** VBZ JJ/OP	***audio** is excellent.*
20	JJ/OP **NN NN**	*it 's a very nice **dvd player**.*
21	RB JJ/OP **NN**	*it is a very amazing **product**.*
22	OP **NN**	*poor **reliability**.*
23	**NN** OP	*creative **software** stinks.*
24	**NN** IN NN	***audio** on video also lacking .*
25	**NN** IN DT NN	*the **construction** of the player is the cheesiest i have ever seen.*
26	NN IN DT **NN**	*overall , a good buy for the **price** .*
27	OP to **VB**	*it has refused to **read** second discs .*
28	JJ to **VB** such that JJ not in OP	*it is exceedingly simple to **navigate***
29	JJ1 **JJ2 NN** such that JJ2 not in OP	*the g3 has much sharper **white offsets** .*
30	**NN** VBZ/VBP DT OP/JJ NN	*the **manual** does a fine job.*
31	**NN** * PRP/DT OP (* any pattern found but no NN)	***apex** is the best cheap quality brand for dvd players .*
32	VB OP/JJ **NN**	*get great **reception**.*

Table 3.2, continued

Rule#	Rule	Example (Hu & Liu, 2004a)
33	NN VB OP/**JJ**	***audio** is excellent*
34	(Any POS but not NN) **NN** VBZ RB OP/JJ	*the **manual** is relatively clear.*
35	amod(**NN**, Any POS) (such that OP found on left or right with any pattern in between but no NN)	*in addition it comes with a sleek and powerful **headset** .*
36	nmod(**NN**, Any POS) (such that OP found on left or right with any pattern in between but no NN)	*overall it is the best **camera** on the market .*
37	nummod(**NN**, Any POS) (such that OP found on left or right with any pattern in between but no NN)	*usb 2.0 **transfer** is insanely fast .*
38	appos(**NN**, Any POS) (such that OP found on left or right with any pattern in between but no NN)	*i love my new **nomad** , its great !.*
39	det(**NN**, Any POS) (such that OP found on left or right with any pattern in between but no NN)	*the **price** was right .*
40	nsubjpass(Any POS, **NN**) (such that OP found on left or right with any pattern in between but no NN)	*the **interface** used could be better designed .*
41	nmod(Any POS, **NN**) (such that OP found on left or right with any pattern in between but no NN)	*everything else about the **camera** is great .*
42	cop(**NN**, Any POS) (such that OP found on left or right with any pattern but no NN)	*this is a wonderful **camera** .*
43	**NN NN** * OP (* any pattern found but no NN)	*their **customer service** is very poor .*
44	**NN** RB (RB not in OP) * OP (* any pattern found but no NN)	***interface** practically seamless*
45	**NN** VBD (VBD not in OP) * OP (* any pattern found but no NN)	*the included **earbuds** were uncomfortable*
46	(Any POS but not NN) OP **NN**	*very comfortable **camera***
47	NN of **NN**	*it does play a wide range of **formats**.*
48	DT **(Any POS) NN** VBZ JJ/OP	*the **sound quality** is okay*
49	DT **NN** IN * JJ/OP (* any pattern found but no NN)	*you can see the **interface** as modern or classical .*
50	DT **NN NN** * JJ/OP (* any pattern found but no NN)	*the **software interface** supplied was very easy to use*
51	**NN** NN VBZ/VBP RB RB	*the **mms** technology is very well*
52	**NN NN** VB OP/JJ	***replacement battery** is expensive.*
53	**NN** (any POS) RB JJ/**OP**	*the **controls** are very intuitive*
54	OP (any POS) OP/JJ **NN**	*like the smart **volume***
55	DT **NN NN** VB RB	*the **lens cover** is surely loose*
56	DT **NN NN** VB OP/JJ	*the **storage capacity** is great for me –*
57	DT **NN** VB RB * OP ([such that RB not in OP] and * any pattern found but no NN)	*the **headphones** are n't the best*
58	DT **NN** VB JJ/OP	*the **grain** was terrible*
59	OP * **NNS** (* any pattern found but no NN)	*lack of good **accessories** .*
60	DT **NN** * OP (* any pattern found but no NN)	*"this **camera** is closest to perfect*
61	RB JJ/OP * **NN** (* any pattern found but no NN)	*very flexible and powerful **features***
62	RB OP/JJ * **NN NN** (* any pattern found but no NN)	*i am really impressed by this **dvd player** .*

Table 3.2, continued

Rule#	Rule	Example (Hu & Liu, 2004a)
63	OP/JJ IN DT **NN**	*can't go wrong with this* ***price***
64	OP * **NN NN** (* any pattern found but no NN)	*it is not that bad there is no* ***tiff format*** *.*
65	DT JJ/OP **NN**	*overall it is a great* ***unit****.*
66	DT **NN** VBZ/VBP RB JJ/OP	*the* ***vibration*** *is not top.*
67	**NN NN** VBZ JJ/OP	***color screen*** *is good*
68	DT RB JJ/OP **NN**	*it 's a very intuitive* ***program***
69	DT VBZ/VBP DT JJ/OP **NN**	*this is a good* ***deal****.*
70	RB VB JJ/OP **NN**	*does not provide enough* ***volume***
71	OP/JJ DT **NN**	*i would recommend this* ***product*** *to anyone.*
72	DT JJ/OP JJ/OP **NN**	*the nomad zen could use a little sturdier* ***construction***
73	**NN** RB OP/JJ	***interface*** *practically seamless*
74	VBZ/VBP RB OP/JJ **NN**	*the phone is very light* ***weight*** *.*
75	NN VB JJ/OP **NN**	*zen has minimal* ***stoppage*** *between tracks*
76	OP/JJ TO VB (Any POS) **NN**	*very convenient to scroll in* ***menu***
77	JJ/OP JJ/OP **NN**	*polyphonic sweet* ***tunes***
78	JJ/OP **NN** RB	*bad* ***picture*** *yet .*
79	DT **NN** VBZ/VBP VBN	*the* ***display*** *is hinged*
80	**NN** CC **NN** VBP JJ/OP	***option****s and* ***control****s are easy.*
81	PRP VBZ JJ/OP **NN**	*it takes great* ***pictures***
82	RB OP/JJ DT **NN**	*already love the* ***features*** *it has .*
83	DT **NN NN** VBZ/VBP JJ/OP	*the* ***menu options*** *are uncreative*
84	(Any POS but not NN) **NN NN** VBZ JJ/OP	*the little* ***jog dial*** *seems weak and quirky*
85	DT **NN** VBZ/VBP JJ/OP	*kind of bulky and the* ***wheel*** *is awkward*
86	**NN NN** VB (Any POS) JJ/OP	*the* ***picture quality*** *are so great.*
87	DT **NN** VBZ/VBP DT JJ/OP	*this* ***phone*** *is a winner.*
88	PRP JJ/OP * DT **NN** (* any pattern)	*my favorite being the* ***games*** *and the* ***pim****, and the* ***radio****.*
89	DT NN VB * JJ/OP **NN** (* any pattern found but no NN)	*the g3 is loaded with many useful* ***features*** *.*
90	DT **NN** * DT NNS * JJ/OP (* any pattern)	*the* ***player*** *itself has all sorts of problems*
91	JJ/OP NN * NN NN (* any pattern found but no NN)	*my favorite features , although there are many , are the* ***speaker phone*** *, the radio and the infrared .*
92	DT **NN** RB VB	*the* ***calls*** *constantly drop in my area and i experince mega-static*
93	**NN** VB RB OP/JJ TO VB	*the* ***menus*** *are very easy to navigate.*
94	DT RB JJ/OP **NN** (Any POS but not NN)	*but at least youre starting with the most photorealistic* ***images*** *ive ever seen from a camera .*

Table 3.2, continued

Rule#	Rule	Example (Hu & Liu, 2004a)
95	JJ/OP IN DT (Any POS) **NN**	*i am bored with the silver **look**.*
96	nsubj(NN1,**NN2**) and amod(NN1,JJ/OP)	***t-mobile** was a pretty good server.*
97	PRP VB (Any POS) **NN**	*it takes wonderful **pictures** very easily in auto mode**.***
98	JJ/OP NN (any pos) DT **NN**	*the only drawback is the **viewfinder** is slightly blocked by the lens .*
99	DT **NN** * RB RB OP/JJ (* any pattern found but no NN)	*when talking the **voice** is not very clear*
100	DT **NN NN** * DT NN * JJ/OP (* any pattern)	*the **battery life** on this phone is surreal.*
101	PRP VB DT JJ/OP **NN**	*it is a perfect **phone**.*
102	OP (Not OP) **NN NN**	*excellent polyphonic **ringing tones.***
103	**NN NN** * JJ/OP NN (* any pattern found but no NN)	***sunset feature** takes incredible pics in the morning , and the evening !.*
104	OP/JJ (Any POS) **NN** CC **NN**	*the system is terrific in **size** and **design***
105	RB JJ/OP (Any POS) **NN**	*that is solved with the very comfortable handsfree **ear-piece** which is included .*
106	**NN NN** (such that no JJ/OP found in sentence)	*the zen does not have a **stop button** !*
107	DT JJ/OP **NN NN**	*the 2nd dvd player had a faulty **power supply**.*
108	DT (Any POS) **NN** * OP (* any pattern found but no NN)	*the included **earbuds** work quite well .*
109	compound (**NN1,NN2**) (such that "OP", "no", or "not" found on left or "OP" on right with any pattern in between but no NN)	*well **flash photos** are never great.* *the zen does not have a **stop button** !*
110	**JJ NN** * OP (such that JJ not in OP and * any pattern found but no NN)	***manual functionality** is excellent.*
111	RB JJ/OP IN DT **NN**	*i am extremely pleased with this **camera**.*
112	OP * **JJ NN** (such that JJ not in OP and * any pattern found but no NN)	*the radio feature has superb **sound quality** .*
113	DT **NN** * DT NN * OP/JJ (* any pattern)	*the **sound** of the player is pretty good .*
114	JJ/VBN/OP * DT **NN** (* any pattern found but no NN)	*i would not be inclined to purchase an **apex** product again .*
115	JJ/OP NN (Any POS) **NN**	*this phone has a very cool and useful feature the **speakerphone***
116	RB VB * **NN** (* mean any pattern that contain OP/JJ and no NN in between)	*i will never buy a creative **product** again .*
117	DT **NN NN** VBZ JJ/OP	*the **auto mode** is good.*

Table 3.2, continued

Rule#	Rule	Example (Hu & Liu, 2004a)
118	nsubj(OP/JJ, NN1) and conj(**NN1**, **NN2**)	*the **options** and **controls** are easy to use and logically laid out .*
119	xcomp(OP,**VB**)	*easy to **use***
120	amod(NN1,JJ/OP) and amod(**NN1,JJ**) (such that JJ not in OP)	*simply , the canon g3 is the best **digital camera***
121	nmod(JJ1/OP1,NN1) and nsubj(JJ1/OP1,**NN2**)	***door** broke after a month .*
122	nmod(OP/JJ,NN1) and compound(**NN1**,**NN2**)	*i 've been pleased with the **picture quality.***
123	dobj(H1,**NN1**) and amod(**NN1**,JJ/OP)	*you can also assign special **rings** to special people when they call .*
124	nsubj(JJ/OP,**NN1**) and det(**NN1**,H1)	*the **prints** are beautiful.*
125	nmod(JJ/OP,NN1) and conj(**NN1**,**NN2**)	*i am very pleased with its **quality** and **durability** .*
126	nsubj(JJ/OP,**NN1**) and amod(**NN1**,H1)	*main **dial** is not backlit*

* where ReL means any dependency relation from['nsubj','amod','prep','csubj','xsubj','dobj','iobj']

3.6 Summary

This chapter gives the details about the rules for aspect extraction by adapting some rules from the previous studies and formulating new developed rules based on the analysis of datasets. There are 126 rules in total and some of these rules can extract the same aspects, but the clear difference between these rules that one rule is better than other in term of precision as it extracts less non-aspects words. The rules including 17 dependency-based rules from previous studies, 16 pattern-based rules from previous studies, and 93 new created rules. However, there are also irrelevant rules which will result in extracting non-aspect words. Thus, a proper selection of an optimal rules' combination from the full set of rules is required to select relevant rules and discard irrelevant rules. In this work, an improved version of WOA is proposed to be used for rules selection and the details about this algorithm is presented in Chapter 4.

CHAPTER 4: DEVELOPMENT OF IMPROVED WHALE OPTIMIZATION ALGORITHM AND ASPECT PRUNING ALGORITHM

As discussed before in Chapter 3 many rules are available for aspect extraction. However, some rules are good in extraction, but on other hand some rules may result on extracting many non-aspects words. Thus, proper selection of optimal rules combination is required to select the efficient rules and discard the irrelevant ones. To select optimal rules combination IWOA (Improved Whale Optimization Algorithm) is used. The details are presented on section 4.1.

4.1 Introduction

Recently WOA has been applied on different problems. For example, a study by Prakash and Lakshminarayana (2017) utilized WOA to find the optimal size and placement of capacitors in radial system. In other work conducted by El Aziz et al. (2017), WOA was used for multilevel thresholding image segmentation. In a study by Tharwat et al. (2017), WOA was used to optimize SVM parameters to classify toxicity effects of bio transformed hepatic drugs. WOA was also used for the problem of optimal power flow (Bentouati et al., 2016), feature selection problem (Sharawi et al., 2017), and for scheduling problem in cloud computing (Sreenu & Sreelatha, 2017). (Ala'M et al., 2018) utilized WOA with SVM classifier for the identification of spammers in social networks, while in a work by Nazari-Heris et al. (2017), WOA was used to power and heat economic dispatch problem.

All these works and other examples which were mentioned in Chapter 1 show a noticeable performance which is competitive with other well-known optimization algorithms. This outperformance of WOA motivated this research to use it for rules selection. As the optimization algorithms will build a number of solutions from 126 rules, in which each solution contains a sequence of 0 or 1. In each solution, the value of 0 means that rule is not selected, whereas the value of 1 mean the given rule is selected. However, based on NFL theorem, one optimization algorithm cannot be superior to all optimization to solve all problems. In addition, optimization algorithms sometimes may be stuck in local optima or have problems of population diversity. Thus, to improve WOA and make it suitable for the current research problem, two improvements were added to WOA and the new improved algorithm is called Improved WOA (IWOA). The first improvement is using CM to improve solutions diversity and make balance between exploration and exploitation. In this way, the solution will not be concentrated in one area and will be diversified. Hence, in rules selection problems, not all rules will be selected or not selected, but it makes diversity of selected and non-selected rules. The second improvement includes the use of LSA in WOA to improve its exploitation. This will be achieved by setting or resetting the selection of rules in the current best solution based on recall and precision of these rules. Therefore, the benefits of using of LSA are to check if there is better solution to improve the extraction performance and to avoid WOA from being stuck at local optima.

The improved WOA (IWOA) is applied on the 126 rules for selecting the optimal rules combination and discarding low quality rules. In addition, an improved algorithm called pruning algorithm is also developed based on aspect frequency, product manual, and opinion direct association between the candidate aspect and opinion word.

The process flow of IWOA+PA is shown in Figure 4.1. In the first step, the optimal rules combinations from the full set of rules will be selected using a training dataset. In the next step, the selected rules will be applied on the testing dataset for extracting candidate aspects.

Finally, pruning algorithm will be applied on the list of candidate aspects to remove the incorrect aspect words and retain the correct aspects words.

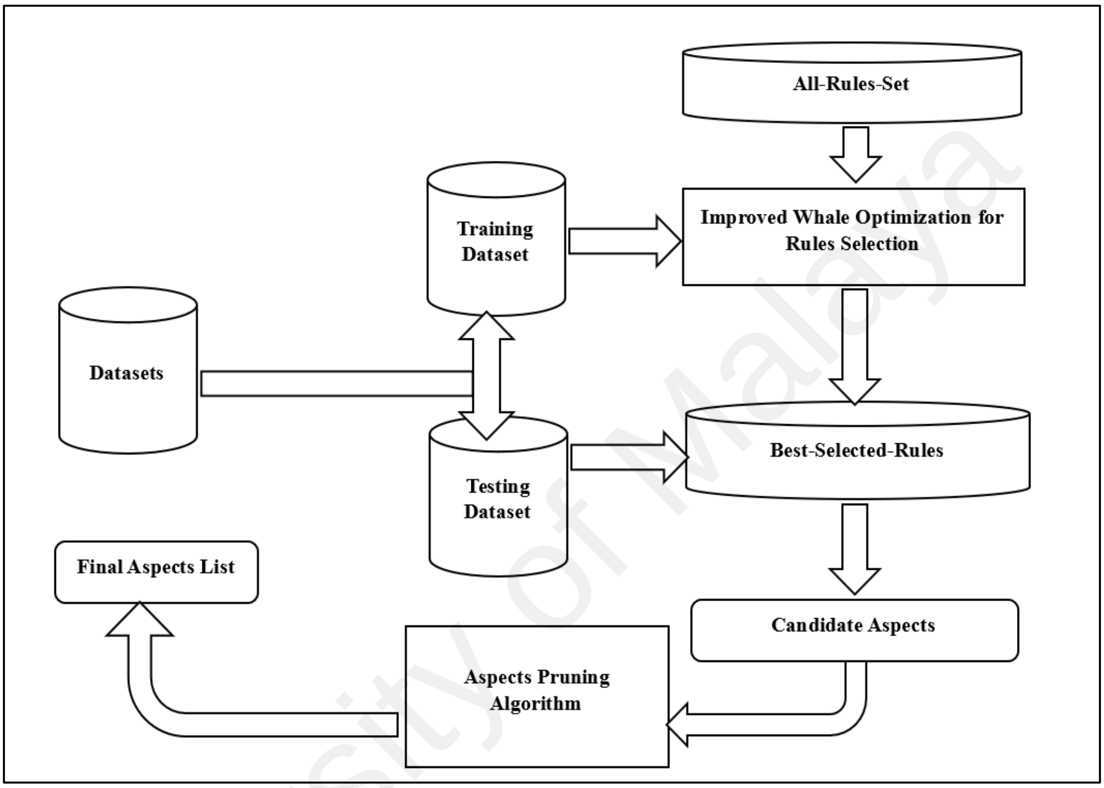

Figure 4.1 Process flow of IWOA+PA

4.2 Development of the Improved Whale Optimization Algorithm (IWOA)

4.2.1 Whale Optimization Algorithm (WOA)

WOA represents one of the recent metaheuristic optimization algorithms which gives promising results to a number of different applications. The idea of WOA originally initiated from formulating the humpback whale hunting behavior mathematically using number of equations for exploration and exploitation. The humpback whales use the bubble-net hunting technique to encircle and catches their preys, and this bubble-net hunting strategy represents the intelligent technique utilized in WOA. In addition, WOA represents and formulate how whales in nature communicate and live (Mirjalili & Lewis, 2016). In nature, the whales hunt and catch its preys through hunting the groups of small fishes that are swimming near the water surface. At first, the hunting technique used by whales begins by diving down deeper than the small fishes' preys around 12 meters, then it starts creating and sending huge numbers of '9' or circles shapes bubbles. These bubbles can be a trap for the small fishes by encircling the group of small fishes inside these bubbles. Because of this, all whales in the group start to swim towards the surface and begin hunting these small fishes (Mirjalili & Lewis, 2016). The WOA algorithm can be represented by three phases that include 1) encircling prey, 2) bubble-net attacking, and 3) searching for a prey (Mirjalili & Lewis, 2016):

1) **Encircling prey phase**: At this phase, first the search-agents (whales) will specify and identify the possible locations of prey. After that the whales will encircle them. In the following step, WOA will find the fitness value of each search agent and consider the search agent with best fitness value as the best candidate search-agents so far. Then, the remaining search-agents (whales) will move and keep updating their positions based on the reference with the initial best solution so far. The main equations for representing and formulating WOA search process are represented by the following Equations (1) and (2) (Mirjalili & Lewis, 2016):

$$\vec{D} = \left| \vec{C} \cdot \vec{X}^{*}(t) - \vec{X}(t) \right| \tag{1}$$

$$\vec{X}(t+1) = \vec{X}^{*}(t) - \vec{A} \cdot \vec{D} \tag{2}$$

where $\vec{A}$ and $\vec{C}$ variables are used as coefficient vectors, t represents the current iteration, the vector $\vec{X}^{*}$ represents the best solution so far, and the vector $\vec{X}$ represents the position vector of each search agent. The equations used for calculating $\vec{A}$ and $\vec{C}$ vectors values are Equations (3) and (4) (Mirjalili & Lewis, 2016):

$$\vec{A} = 2\vec{a} \cdot \vec{r} - \vec{a} \tag{3}$$

$$\vec{C} = 2 \cdot \vec{r} \tag{4}$$

In equation (3) $\vec{a}$ value is decreased linearly from 2 to 0 over the WOA algorithm iterations. In addition, $\vec{r}$ is vector with random values over [0,1].

2) **Bubble-net attacking Phase (Exploitation phase):** This phase is composed of two mechanisms including shrinking encircling and updating position using upward spiral mechanism. Where the Shrinking encircling mechanism is achieved using Equation (3). And the upward spiral mechanism was formulated using the spiral equation (5) as follows:

$$\vec{X}(t+1) = \overrightarrow{D'} \cdot e^{bl} \cdot cos(2\pi l) + \overrightarrow{X^*}(t) \tag{5}$$

where $\overrightarrow{D'} = \left|\vec{X}^*(t) - \vec{X}(t)\right|$ is the distance between the whale position X and the prey. Furthermore, b constant value defines the shape of upward spiral movement used by whales and l is a random number over [-1,1]. In WOA, they assumed that the whales have 50% probability to switch between shrinking encircling and upward spiral mechanism for movements around the prey. Thus, these two mechanisms are modeled using equation (6) (Mirjalili & Lewis, 2016):

$$\vec{X}(t+1) = \begin{cases} \vec{X}^*(t) - \vec{A} \cdot \vec{D} & if\ p < 0.5 \quad \text{(Encircling)} \\ \overrightarrow{D'} \cdot e^{bl} \cdot cos(2\pi l) + \overrightarrow{X^*}(t) & if\ p \geq 0.5 \quad \text{(upward spiral)} \end{cases} \tag{6}$$

where p value defines a random number over [0,1].

3) **Searching for a prey Phase (Exploration phase)**: The works in this phase are based on the value of $\vec{A}$ vector in which its variation values are considered for selecting to update the whale position either based on reference to the current best solution so far or with reference to randomly selected solution. Therefore, equations (7) and (8) are used for updating the search agent position based on $\vec{A}$ vector value (Mirjalili & Lewis, 2016).

$$\vec{D} = \left|\vec{C} \cdot \overrightarrow{X_{rand}} - \vec{X}\,\right| \tag{7}$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A} \cdot \vec{D} \tag{8}$$

where X_{rand} represents the value of randomly selected search agent (whale) which is randomly selected from the set of available whales, Figure 4.2 presents the WOA algorithm (Mirjalili & Lewis, 2016).

```
Initialize the search agents (whales) population X_i (i = 1, 2, ..., n)
Calculate the fitness of each search agent
X*=the best search agent
while (t < maximum_iterations)
   for each search agent
  Update a, A, C, l, and p values
     if (p<0.5)
        if (|A| < 1)
              Update the position of the current solution position using Eq. (2)
        else if (|A|≥1)
              Select a solution randomly search agent (X_rand) from list of solutions
              Update the position of the current search agent by the Eq. (8)
        end if
     else if (p≥0.5)
        Update the current search position using Eq. (5)
     end if
  end for
  Check if any search agent (whale) that goes outer the search space and amend it
  Calculate the objective value of each search agent
  Update X* if there is a better solution
  t=t+1
end while
 return X*
```

Figure 4.2 WOA algorithm (Mirjalili & Lewis, 2016)

4.2.2 The Cauchy Mutation Operator

The definition of the one-dimensional Cauchy density function is modeled using equation (9) (Zhang et al., 2019):

$$f(x, m, \mu) = \frac{1}{\pi} \frac{m}{m^2 + (x - \mu)^2} \text{ , } -\infty < x < \infty \tag{9}$$

When the values of $m = 1 \ and \ \mu = 0$, it is called the standard Cauchy distribution and it is defined using equation (10) (Zhang et al., 2019):

$$f(x) = \frac{1}{\pi} \frac{1}{1 + x^2} \text{ , } -\infty < x < \infty \tag{10}$$

Then, the value of standard Cauchy distribution function can be determined using equation (11) (Zhang et al., 2019):

$$CM(0,1) = \tan[(\xi - 0.5)\pi] \quad such \ that \ \xi \epsilon \ U[0,1] \tag{11}$$

In equation (11), the value of ξ is a random value over [0,1]. In general, Cauchy Mutation (CM) has been used in many studies to improve different types of optimization algorithms such as work conducted in (Pappula & Ghosh, 2017), in which CM was used with cat swarm optimization (CSO) algorithm to avoid the algorithm from being stuck in local optima and to avoid premature convergence problem. In a work conducted by Wang, Wang, et al. (2016), Firefly algorithm (FFA) was improved by utilizing the idea of CM with FA algorithm to improve its global search ability. Zou et al. (2017) research improved the balance between the speed and accuracy of Particle Swarm Optimization (PSO) algorithm and improved its population diversity by combining PSO algorithm with CM. Whereas Wang, Deb, et al. (2016) improved the Krill herd (KH) algorithm by combining it with CM to avoid it from trap into local optima and improve its population diversity. Furthermore, Ali and Pant (2011)study incorporated CM into Differential Evolution (DE) to improve its performance and avoid it from fall into local optima. In another study by Li et al. (2017), they incorporated CM into gravitational search algorithm (GSA) to improve its exploration ability. Wu and Law (2011) improved PSO algorithm by incorporation of CM in PSO to enhance its exploration ability. In the work by (Zhang et al., 2019), they used CM to improve the diversity of solution in Salp Swarm Algorithm (SSA). Thus, the idea to embed CM into WOA was motivated by these previous researches which incorporated CM into different optimization algorithms and improved the performance of these algorithms. In addition, the idea of CM is incorporated into standard WOA algorithm is to improve the population diversity and avoid it from being stuck at local optima.

4.2.3 Local Search Algorithm (LSA)

To improve the best solution, the new Local Search Algorithm (LSA) is used for adding and removing rules from the best solution received from WOA at the end of each iteration. The other reasons for incorporating LSA at the end of WOA is to improve WOA exploitation ability and to avoid it from being trapped into local optima. LSA algorithm iterates *MaxLIter* times trying to find a better solution than the current one received by WOA at the end of each iteration. The LSA algorithm is shown in Figure 4.3.

The input to LSA algorithm represents a list of tuples contains 126 rules (with each tuple contains rule number and rule precision and another list which contains each rule with its recall value). At the first line in LSA, the selected rules in current *NSol* will be ranked in ascending order based on their precision and saved into ***P1*** list. In addition, the non-selected rules in current *NSol* will be ranked in descending order based on their precision and saved into ***P0*** list. Then, the best selected rules in *NSol* will be in ***P1*** upper half, while the worst selected rules in in *NSol* will be in ***P1*** lower half.

```
Input: all 126 rules are ranked based on their precision, all 126 rules are ranked based on their recall
R=1
MaxLIter=T   (T represents the number of times LSA will repeat)
NSol=Leader_pos (Leader_pos represents the current best solution at end of WOA iteration)
  While  R < MaxLIter

        P1= Sort All Selected Rules in NSol in ascending order based on precision
        P0= Sort All Non-Selected Rules in NSol in descending order based on precision

        P1_Worst_half = P1 half which contains the worst selected rules based on precision
        P1_Best_half = P1 half which contains the best selected rules based on precision
        P0_Worst_half = P0 half which contains the worst Non-selected rules based on precision
        P0_Best_half = P0 half which contains the best Non-selected rules based on precision

        Select 5 rules randomly from P1_Worst_half and reset its values to 0 in NSol
        Select 5 rules randomly from P0_Best_half and set its values to 1 in NSol

        R1= Sort All Selected Rules in NSol in ascending order based on recall
        R0= Sort All Non-Selected Rules in NSol in descending order based on recall

        R1_Worst_half = R1 half which contains the worst selected rules based on recall
        R1_Best_half = R1 half which contains the best selected rules based on recall
        R0_Worst_half = R0 half which contains the worst Non-selected rules based on recall
        R0_Best_half = R0 half which contains the best Non-selected rules based on recall

        Select 5 rules randomly from R1_Worst_half and reset its values to 0 in NSol
        Select 5 rules randomly from R0_Best_half and set its values to 1 in NSol

        NSolVal =Evaluate Fitness value of NSol based on F-measure

        IF (NSolVal > Leader_score):
            Leader_pos= NSol
            Leader_score= NSolVal
        EndIF
        R=R+1
 End While
```

Figure 4.3 LSA algorithm

On the other hand, the best non-selected rules in *NSol* will be in ***P0*** lower half, while the worst non-selected rules in in *NSol* will be in ***P0*** upper half. The same steps apply, based on recall values of each rule, where the selected rules in current *NSol* will be ranked in ascending order based on their recall and saved into ***R1*** list. The non-selected rules in current *NSol* will be ranked in descending order based on their recall and saved into ***R0*** list. Then, the best selected rules in *NSol* will be in ***R1*** upper half, and the worst selected rules in in *NSol* will be in ***R1*** lower half while the best non-selected rules in *NSol* will be in ***R0*** lower half, and the worst non-selected rules in in *NSol* will be in ***R0*** upper half. In the following steps, five rules will be randomly selected from ***P1*** lower half and reset to 0 as non-selected rules in the *NSol* as these rules represents some rules that are selected with low precision. On the other hand, five rules will be randomly selected from ***P0*** lower half and reset to 1 as selected rule in the *NSol* as these rules represents some rules that are not selected and have high precision. In addition, five rules will be randomly selected from ***R1*** lower half and reset to 0 as non-selected rules in the *NSol* as these rules represents some rules that are selected with low recall. Five rules will also be randomly selected from ***R0*** lower half and reset to 1 as selected rule in the *NSol* as these rules represents some rules that are not selected and have high recall. Finally, the fitness value of *NSol* will be evaluated on the training dataset using F-measure and if it is better than the current *Leader_pos* fitness value it will replace the current *Leader_pos* with *NSol* value. This LSA algorithm will repeats *MaxLIter* times.

4.2.4 Improved WOA based on using Cauchy mutation and LSA for rules selection

This section presents the details of improvements to standard WOA, as displayed in Figure. 4.4 where CM is used with WOA to improve the population diversity and to avoid it from being stuck at local optima. In addition, LSA is also used at the end of WOA iteration to improve its exploitation ability and to avoid it from being stuck at local optima. As presented in Figure 4.4, the first improvement to standard WOA is based on introducing new mutation rate, that if MR greater than or equal 0.5 then the new CM equation will be used to update the position of the current solution using Equation (12); otherwise Equation 2 will be used to update the position of the current solution.

$$\vec{X}(t+1) = \vec{X}(t) \;\; + CM * \overrightarrow{D2} \tag{12}$$

where $\quad \overrightarrow{D2} = \left|\vec{X}^{*}(t) - \vec{X}(t)\right|$

```
Initialize the search agents (whales) population Xi (i = 1, 2, ..., n)
Calculate the fitness of each search agent
X*=the best search agent
while (t < maximum_iterations)
  for each search agent
  Update a, A, C, l, and p values
     if (p<0.5)
          if (|A| < 1)
                Mrate=rand(0,1)
                If   (Mrate < 0.5)
                     Update the position of the current solution position using Eq. (2)
                else
                     Update the position of the current solution position using Eq. (12) using
                     Cauchy mutation
                end if
          else if (|A|≥1)
                Select a solution randomly search agent (Xrand) from list of solutions
                Update the position of the current search agent by the Eq. (8)
         end if
     else if (p≥0.5)
          Update the current search position using Eq. (5)
     end if
  end for
  Check if any search agent (whale) that goes outer the search space and amend it
  Apply LSA algorithm on X* to check if there is a better solution
  Update X* if there is a better solution
  t=t+1
end while
return X*
```

Figure 4.4 Proposed IWOA algorithm based on CM and LSA algorithm

The other improvement to WOA as presented in Figure 4.4, include applying LSA algorithm at the end of each WOA iteration to find better solution than the current best solution. LSA works by adding and dropping rules from the current best solution and evaluates the fitness of the new solution, if the new solution fitness is better than the current best solution, then it will replace its value with the new solution value. The new improved IWOA algorithm will be used for rules selection which will be used for aspect extraction. At first, WOA will take the full set of 126 extraction rules and generate initial solutions and each solution will contain a random number of rules from the full set of 126 rules.

Rules selection by IWOA: IWOA is applied on full rules set to find the optimal rules combinations by taking best rules and discarding irrelevant rules. In results of IWOA, the selected rules in optimal solution are represented as a sequence of 1 and 0, such that the existence of "1" value in the solution means that the corresponding rule is selected; otherwise the existence of "0" value in the solution means that the corresponding rule is not selected. The following steps outlines the details of IWOA algorithm for rules selection:

1. **IWOA Initialization:** In the first step, IWOA randomly generates number of search-agents based on the population size (number of whales). Each whale (solution) contains number of randomly selected rules from the full set of rules. Then after generating the solutions, IWOA will evaluates the objective values of each solution based on F-measure and the best one will be set as best solution so far X*. Moreover, one of the datasets used by IWOA for training for rules selection and another dataset will be used for testing.
2. **Updates Solutions (Whales) Positions:** At this step, IWOA updates each solution position based on reference to one of the solutions selected randomly using equation (8) or updates the solution position with reference to X* using either equation (2) or (12) or by using equation (5). This is a new improvement, based on adding random variable rate, if the random variable rate is less than

0.5 then equation (2) will be used to update current solution position; otherwise equation (12) will be used to update current solution position based on CM.

3. **Apply LSA:** At this phase, LSA will be applied at the end of current IWOA iteration on current best solution so far X*. Furthermore, if LSA algorithm find better solution than the current X* solution, then it will replace the X* with the new solution and this step will repeat ***MaxLIter*** times.
4. **IWOA stop execution:** IWOA will iterate and repeats steps 2 and 3 *t* times (where the variable *t* used to specify the number of iterations IWOA must iterate over all solutions). In addition, IWOA will update the current best solution X* if it finds better solution at the end of each iteration.
5. **Best Solution:** At the end of IWOA execution, best solution will be obtained by IWOA which represents the set of optimal rules combination yielded by training IWOA on training dataset. The selected rules which were indicated by bit 1 as selected rules will be used on the testing datasets to evaluate the IWOA performance on the testing dataset.
6. **Dataset Testing:** The best rules which are selected in the best solution X* will be applied on testing dataset.

4.3 Pruning Algorithm (PA)

Aspects pruning is important because sometimes the extraction rules will extract incorrect aspects that are not frequent in the datasets. To remove these incorrect aspects, a frequency pruning can be used, since the aspect word that is occur frequently is normally considered as the correct aspect (Hu & Liu, 2004a). However, sometimes number of pruned aspects based on frequency are correct aspects but do not occur frequently in the datasets. Therefore, manual pruning will be used to look for these candidate aspects. Furthermore, sometimes an aspect which is a pruned based on frequency and manual can be a correct aspect but does not frequently occur or is not mentioned in the manual. Thus, to solve this issue, a pruning based on direct opinion association is used to look for these candidate aspects.

Aspect pruning algorithm is required to improve the precision of extracted aspects and will be applied on the aspects extracted from the application of the best selected rules which were resulted from IWOA. The importance of aspects pruning emerged as the extraction rules might extract many non-aspects words which might affect the precision of the extraction system. Thus, to filter the non-aspects words and retain the correct aspect words a pruning algorithm is required. The pruning algorithm is developed based on three phases that include frequency-based pruning, manual based pruning, and pruning based on direct opinion association. The pruning algorithm is shown in shown in Figure 4.5. Details of each phase are as the following subsections:

4.3.1 Pruning based on frequency

The frequency pruning was used by many studies such as (Hu & Liu, 2004a; Qiu et al., 2011; Kang & Zhou, 2017; Rana & Cheah, 2017). In this first phase, the frequency of each extracted aspect will be calculated and the aspects with frequency less than the specified threshold will be pruned and kept for next pruning phases. Then, the thresholds for pruning are 2 for single words aspects and 1 for multi words aspects. If the aspect is single word and its frequency greater than 2, it will be approved as a correct aspect; otherwise it will be pruned out and kept for the next two pruning phases. In addition, if the aspect is multi-words aspect with frequency greater than 1, it will be approved as a correct aspect; otherwise it will be pruned out and kept for the next two pruning phases. However, not all pruned aspects which are pruned based on frequency are non-aspect words, but some pruned words are correct aspects but not frequently mentioned in reviews and not meet with the frequency pruning thresholds. Therefore, to overcome this problem pruning based on product manual will be used.

4.3.2 Pruning based on Product Manual

In general, each electronic product bought by a customer will be accompanied with a product manual (also called user guide, instruction manual or owner's manual) that is normally stored as pdf file. The product manual is normally a small book that contains all details about the product such as safety instruction, product technical specification, installation instructions, setup instructions, operations instructions, maintenance instructions, service locations, and warranty information. It is a comprehensive book which include many information about the product[1].

Therefore, from the first phase, the aspects that do not meet the specified frequency thresholds will be pruned out. Now, for each pruned aspect, if its frequency in product manual is greater than a specified manual threshold, it will be approved as the correct aspect and will be added to the list of correct aspects; otherwise it will be pruned out and will be kept for the last pruning phase. Thus, two thresholds are specified based on manual one for single word aspects and another one for multi-words aspects. For single word aspect, if its frequency in the manual is greater than 2, it will be approved and added to the list of approved aspects; otherwise it will be pruned out and kept for the last pruning phases. In addition, for multi-words aspect, if its frequency in the product manual is greater than 0, it will be added to the list of approved aspects; otherwise it will be pruned out and kept for the last pruning phases. However, not all pruned aspects which are pruned based on manual are non-aspect words, but some pruned words are correct aspects and

[1] https://en.wikipedia.org/w/index.php?title=User_guide&oldid=872157779

can be used frequently by people but not written in the product manual. So, to overcome this problem a pruning based on single word in sub sentence, sentence or word that has direct relation with opinion word will be used.

4.3.3 Pruning based on direct relation with opinion word

In this phase, the aspects that do not meet with the specified frequency thresholds based on manual will be pruned out and checked in this phase. Now, for each aspect in the pruned aspects, the first step is to check if there is only one aspect in the sentence (Qiu et al., 2011). In this case the aspect will be added to the list of the approved aspects; otherwise if two or more aspects exist in the sentence that contains the current pruned aspect. Then, check if the pruned aspect has direct opinion relation such as 'amod', 'nsubj' with an opinion word in the current sentence or only one aspect exists in the sub-sentence which contains this pruned aspect then approve it; otherwise discard it. Figure. 4.4 shows the aspect pruning algorithm, where CA represents the aspects obtained after the application of selected rules by IWOA. In addition, FA represents the final approved aspects, and NFA is the non-frequent aspects (pruned aspects). Therefore, pruning algorithm works in phase order by applying frequency-based pruning at first, then pruning based on product manual and the last phase which is pruning based on one aspect in the sentence or aspect has direct opinion relation with opinion word or one aspect exist in the sub-sentence. The pruned aspect that do not met by all phases will be discard.

The following are examples of application of aspect pruning algorithm and they are presented in Table 4.1. In a sentence "*it's great to switch to* ***spot metering*** *and actually see it working on the lcd screen* ". The aspect "***spot metering***" has frequency of 1 in Canon dataset, then based on product manual of Canon Camera it will be approved as it was found in the manual. Another example, in a sentence " *this camera also has a great feel* *and* ***weight*** *to it*", "***weight***" aspect does not satisfy frequency threshold in Canon dataset, then based on manual it will be approved. Now, the following are examples of pruning based on direct opinion association, sub-sentence, or sentence. For example, in a sentence "*and for those that are interested the* **recharger** *works anywhere in the world and is quite small*", it contains two nouns include "***recharger***" and "***world***", but "***recharger***" has direct association with opinion word "***work***" via "***nsubj***" relation and only one aspect exists in the sub-sentence, then the aspect "***recharger***" will be approved as a correct aspect.

Furthermore, in example " *basic* **usage** *is easy , but the remote has a lot of buttons that i have n't used*", the sentence has more than one aspect, but the sub-sentence which contains "***usage***" aspect has only one aspect and "***usage***" aspect has direct opinion association with "***easy***" opinion word via "***nsubj***" relation, then "***usage***" aspect will be added to the correct aspects. In a sentence " *the little* **jog dial** *seems weak and quirky and i hope i do n't figure out a way to break it*." it will approve "***jog dial***" aspect as a correct aspect because it exists in a sub-sentence which contains only one aspect and the end of this sub-sentence indicated by "***and***". This is another example in which only one aspect exists in the sub-sentence, where "**,**" indicates the end of sub-sentence " *simple* **click buttons** *, back buttons volume and display are very easy to read , access and use*". In addition, the aspect "***click buttons***" has "***amod***" relation with opinion word ***"simple"***. Therefore, "***click buttons***" aspect will be added to the correct aspects. An example "*the* **grain** *was terrible*", the aspect "***grain***" will be added to the correct aspects because only one aspect exists in the sentence. Another example which is based on one aspect exist in the sub-sentence is "*solid, high-quality* **construction**" and in this case "***construction***" will be added to the correct aspects. In a sentence "*i am very pleased with its quality and* **durability**", the aspect ***"durability"*** will be added to the correct aspects because there is only one aspect in the sub-sentence where the end of this sub-sentence is indicated by ***"and"***. In example, " *the* ***voice quality*** *is very good , and it gets great reception*", the aspect ***"voice quality"*** will be added to the correct aspects because only one aspect exists in the sub-sentence where the end of sub-sentence is indicated by ***","***. In addition, aspect "***voice quality***" has direct opinion association with opinion word ***"good"*** via "***nsubj***" relation. In sentence " the **vibrate setting** is loud", the aspect ***"vibrate setting"*** will be added to the correct aspects because there is only one aspect in the sentence. In a sentence "***touchups****, redeye, and so on are very easy to alter, and correct*", the aspect ***"touchups"*** will be added to the correct aspects because there is only one aspect exists in the sub-sentence where the end of the sub-sentence is indicated by ***","***. In a sentence "*the design and* ***construction*** *are excellent*", the "***construction***" aspect will be added to the correct aspects because the sub-sentence contains only one aspect. In a sentence "*awesome camera with huge* ***print quality*****",** it has two aspects "***camera***" and "**print quality**". Therefore, based on direct opinion association, the aspect "***print quality***" will be added to the correct aspects because "***print quality***" aspect has direct association to opinion word ***"huge"*** via "***amod***" relation.

In a sentence "*great feature list , poor* ***reliability*****",** the aspect ***"reliability"*** will be added to the correct aspects because the sub-sentence only contains one aspect where the end of sub-sentence is indicated by ***","***. In addition, the aspect ***"reliability"*** has direct opinion relation to ***"poor"*** opinion word via "***amod***" relation.

As shown in Figure. 4.5, the pruning algorithm starts at the first phase by finding the frequency of each extracted candidate aspect. The next step, after getting the frequency of each extracted aspect, it will prune each aspect that does not meet the prespecified thresholds and approve the aspects that meet the frequency thresholds.

```
CA= Aspects received from application rules selected by IWOA
// Frequency Pruning Phase (First Phase)

for each aspect in CA do:
        If frequency(aspect) in dataset > threshold1: (threshold1 =2 for single word aspects and 1 for
        multiword aspects)
            Add aspect to the list of final correct aspects FA // Final Aspects
            NFA=CA-FA          //NFA is Non-frequent aspects (pruned aspects)
        end if
end for

// Product Manual Pruning Phase (Second Phase)
for each aspect in NFA do:
        If frequency(aspect) in manual > threshold2: (threshold2 =2 for single word aspects and 0 for
        multiword aspects)
           Add aspect to the list of final correct aspects FA // Final Aspects
           NFA=CA-FA          //NFA is Non-frequent aspects (pruned aspects)
        end if
end for

// Direct opinion association, sub-sentence, or sentence Pruning Phase (Third Phase)
for each aspect in NFA do:
        If only one aspect found in current sentence
            Add aspect to the list of final correct aspects FA // Final Aspects
           NFA=CA-FA          //NFA is Non-frequent aspects (pruned aspects)
        else if more than one aspect found in sentence:
              if the pruned aspect has direct opinion association with an opinion word with relation
              such as 'amod','nsubj' or the subsentence has only one aspect: // sub-sentence can be
              indicated  by and,or,but,",",".":

                    Add aspect to the list of final correct aspects FA // Final Aspects
                    NFA=CA-FA          //NFA is Non-frequent aspects (pruned aspects)
              else
                    Discard aspect
        end if
 end for
 Return FA as the final aspect list
```

Figure 4.5 Proposed aspects pruning algorithm

After that, from these pruned aspects from first phase, the aspects which meet with product manual thresholds it will be approved as correct aspects by the algorithm; otherwise it will be pruned and be kept for the last phase. Finally, phase three will be applied on left pruned aspects from phase two. In phase three, if one aspect exists only in the whole sentence then it will approve it; otherwise if two or more aspect exist in the sentence, then if one aspect exists in the sub-sentence or the aspect has direct association with the opinion word then approve it. If all phases were not satisfied with current pruned aspect, then the aspect will be pruned.

Table 4.1: Pruning examples by type

Example (Hu & Liu, 2004a)	Pruning Type
"*it 's great to switch to* ***spot metering*** ".	Product Manual
" *this camera also has a great feel and* ***weight*** *to it*"	Product Manual
"*the included* ***memory card*** *is too small*"	Product Manual
"*the* ***macro*** *works great for medical photographs* "	Product Manual
, "*the* ***lens cover*** *is surely loose*"	Product Manual
, "*the* ***service*** *from the supplier was exceptional*"	Product Manual
, "*the* ***volume key*** *can be hard to press*"	Product Manual
"*the* ***pc sync*** *feature is superb* "	Product Manual
"*audio* *on video also lacking*"	Product Manual
" *and for those that are interested the* ***recharger*** *works anywhere in the world and is quite small*"	Direct Opinion Association
" *basic* ***usage*** *is easy , but the remote has a lot of buttons that i have n't used*"	Direct Opinion Association
" *the little* ***jog dial*** *seems weak and quirky and i hope i do n't figure out a way to break it*."	Direct Opinion Association
" *simple* ***click buttons*** *, back buttons volume and display are very easy to read , access and use*"	Direct Opinion Association
"*the* ***grain*** *was terrible*"	Direct Opinion Association
"*solid , high-quality* ***construction***"	Direct Opinion Association
"*i am very pleased with its quality and* ***durability"***	Direct Opinion Association
" *the* ***voice quality*** *is very good , and it gets great reception*"	Direct Opinion Association
" the **vibrate setting** is loud"	Direct Opinion Association
"***touchups****, redeye, and so on are very easy to alter, and correct*"	Direct Opinion Association
"*the design and* ***construction*** *are excellent*"	Direct Opinion Association
"*awesome camera with huge* ***print quality"***	Direct Opinion Association
"*great feature list , poor* ***reliability"***	Direct Opinion Association

The proposed aspect extraction algorithm is shown in Figure 4.6, which consists of three phases. In phase 1, a combination of different extraction rules types which resulted in total of 126 rules were formulated. In the second phase, these rules will be used by IWOA to select the optimal rules subset from the full rules set based on the used training dataset. Then, the selected rules will be applied on the testing dataset and result in candidate aspects list. This candidate will go to the last phase which is the pruning phase. In the final phase, which is the pruning phase, the correct aspects will be saved while the wrong aspects will be removed.

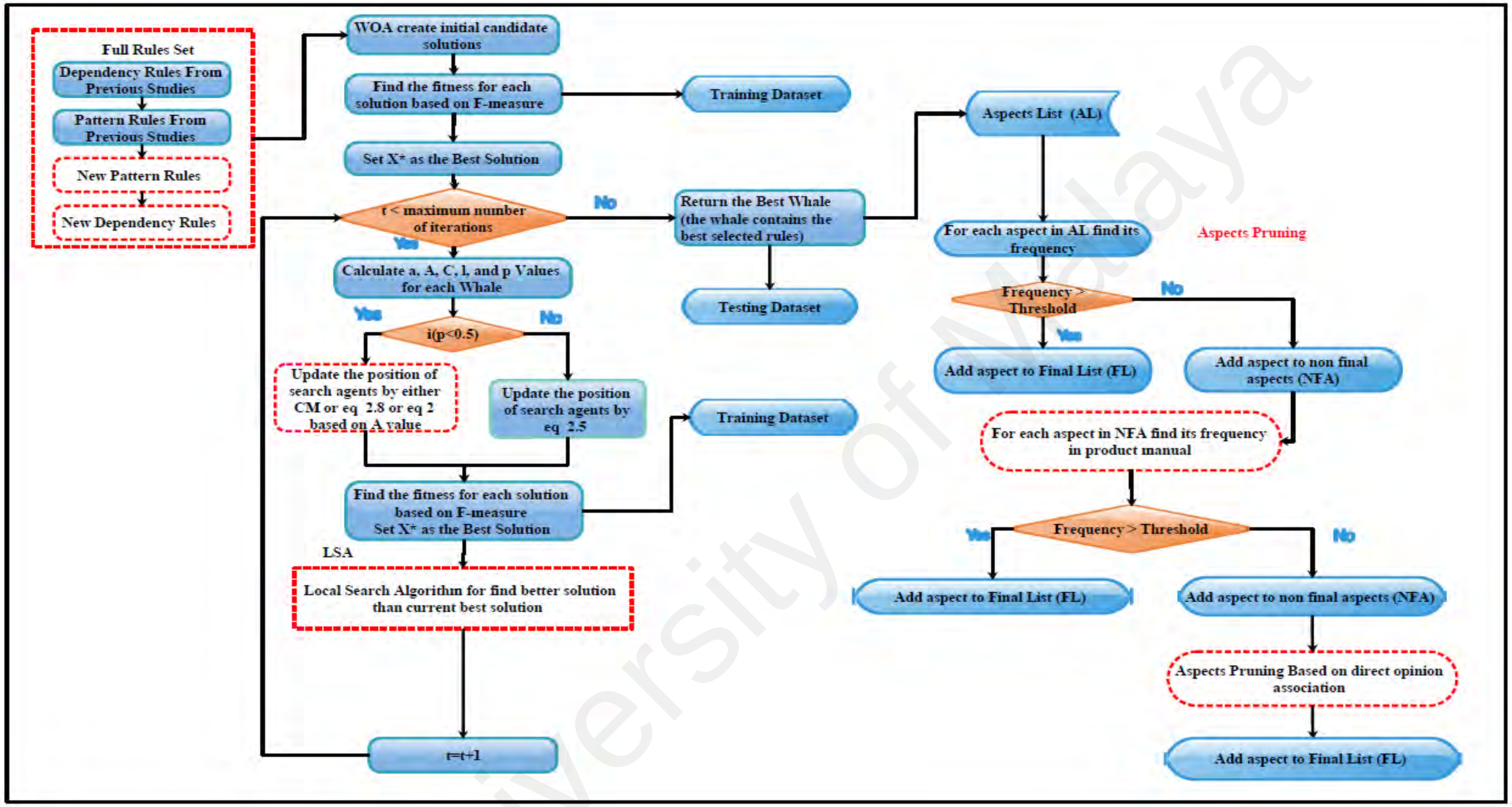

Figure 4.6 Proposed aspect extraction algorithm

4.4 Summary

This chapter introduced IWOA algorithm for rule selection with the improvements on the standard WOA which include the use of CM and LSA algorithm. IWOA was applied on the 126 rules set for selecting the optimal rules combination and discard low quality rules. In addition, to remove the non-aspect extracted words and keep the correct aspects, three-phase pruning algorithm was developed. The pruning algorithm was applied on aspects extracted from selected rules by IWOA. The proposed IWOA+PA algorithm capable of selecting optimal rules combination and discard low quality rules. In addition, it can further improve the performance by using pruning algorithm as discussed in Chapter 5.

CHAPTER 5: EXPERIMENTS RESULTS AND DISCUSSION

5.1 Introduction

This chapter describes the details about the used datasets, experiments, baseline, and achieved results. First, it gives description about the datasets in section 5.2, where these datasets are about five electronic products and were used by majority of research in aspects extraction. Then, section 5.3 gives some details about the performance metrices and baseline works, and section 5.4 presents the conducted experiments with the achieved results. There are four experiments: the first experiment was conducted by applying the full rules set on the datasets, while in the second experiment, IWOA was used for rules selection and aspects extraction. The results were compared with the state-of-the-art optimization algorithms. In the third experiment, PA algorithm was applied on the aspects obtained from IWOA application. The final experiment was conducted through comparison of IWOA+PA with other state of the art baseline works. Finally, section 5.5 concludes this chapter.

5.2 Evaluation Datasets

To test and evaluate the performance and effectiveness of the proposed aspect extraction algorithm, customer review datasets by Hu and Liu (2004a) was used. This dataset is a benchmark dataset for aspect extraction which was used by majority of researchers in aspect extraction (Rana & Cheah, 2016). It was used recently by many previous studies such as (Maharani et al., 2015; Liu, Gao, et al., 2016; Liu, Liu, et al., 2016; Poria et al., 2016; Asghar et al., 2017; Kang & Zhou, 2017; Rana & Cheah, 2017; Marcacini et al., 2018; Rana & Cheah, 2018a, 2018b; Saqia et al., 2018; Vo et al., 2018; Zhang et al., 2018).These datasets contain customer reviews on five electronic products which include reviews about two digital cameras types (Canon D1, Nikon D2), mobile phone (Nokia D3), MP3 player (Creative D4), and DVD player (Apex D5). Each dataset was annotated manually where each sentence was labeled with the mentioned aspect (Hu & Liu, 2004a). The detailed description and statistics of each included dataset is presented in Table 5.1. Some samples of the used datasets are presented in Appendix C.

Table 5.1: Detailed statistics of each dataset

Data	Product	#Reviews	#Sentences
D1	Canon digital camera	45	597
D2	Nikon digital camera	34	346
D3	Nokia Cellphone	41	546
D4	Creative Mp3 player	95	1716
D5	Apex DVD player	99	740

5.3 Evaluation Metrics and Comparison Baselines

The metrics used for algorithms evaluation includes precision, recall, and F-measure. Precision is defined as the ratio of the number of correctly extracted aspects based on gold standard to the total number of extracted aspects. Recall is the ratio of the number of correct extracted aspects based on gold standard to the total number of aspects in gold standard. F-measure is the harmonic mean of recall and precision.

The baseline methods used for the comparison include the state-of-the-art and the most recent aspect extraction methods. The following are the baseline works which were used in the comparison:

1. DP (Qiu et al., 2011), represents one of the famous algorithm used for aspect extraction. DP is based on eight dependency relations rules for extraction.
2. Htay (Htay & Lynn, 2013), is based on patterns rules for extraction as there are some pattern rules used for extraction in this thesis.
3. RubE (Kang & Zhou, 2017), represents one of the recent works which used DP rules and improved it with new rules.

4. Two-fold rule-based model (TF-RBM) (Rana & Cheah, 2017), represents one of the recent work which are based on sequential rules patterns for aspect extraction. It is selected as it performs some supervised learning which is similar to our approach.
5. Rule Selection using a Local Search algorithm (RSLS) (Liu, Gao, et al., 2016), represents one of the recent aspect extraction algorithms that is based on extending the DP rules only. In addition, they applied Simulated Annealing for rules selection. It was selected as it applied supervised learning which is also similar to our approach.
6. Convolutional Neural Networks with Linguistic Patterns (CNN + LP) (Poria et al., 2016), represents one of the recent supervised works for aspect extraction.

   All these baseline works used the same datasets (Hu & Liu, 2004a) in their experiments.

5.4 Experimental results

To evaluate the performance of the proposed work, four experiments were conducted. In the first experiment, all 126 aspect extraction rules were applied on the datasets. In the second experiment, IWOA was applied on the 126 rules for selecting best rules combination. In this experiment, one of the datasets was selected for IWOA training and the other dataset was used for testing. In addition, in this experiment the results of IWOA was compared with other state of the art optimization algorithms. The third experiment was conducted by applying the pruning algorithm on the extracted aspects, resulted from the second experiment. The fourth experiment represents comparison of IWOA+PA with other aspect extraction baseline works. In these experiments, Stanford Parser was used for dependency parsing and Stanford tagger for part-of-speech tagging. In addition, for all optimization algorithms, the population is set to 10 and iterations to 30. The parameter settings of the used optimization algorithms are shown in Table 5.2. In experiment two, f-measure is the fitness value which used for testing all optimization algorithms performance.

Table 5.2: Parameters setting of the used algorithms

Algorithm	Parameter
FFA	Alpha=0.5 Beta_min=0.20 Gamma=1
SCA	a = 2
SSA	c_2 and c_3 random numbers over [0,1]
PSO	Acceleration constants (C1=2 , C2=2) Inertia Weight (W1=0.2, W2=0.9)
IWOA	a = [2,0] b = 1 LSA iteration =10
WOA	a = [2,0] b = 1
MFO	b = 1
MVO	Maximum wormhole existence probability =1 Minimum wormhole existence probability = 0.2
GWO	a = [2,0]

5.4.1 Experiment 1: Results obtained from applying full set of rules on datasets

In this experiment, all rules are applied on each dataset. Table 5.3 displays the results of the performance in terms of precision, recall, and F-measure. As shown in the table, the recall is very high in comparison with precision in each dataset. Thus, from these results, it is confirmed that inclusion of all rules increased the recall. However, the precision is still low since there are some rules in this combination extract many incorrect aspect words. Thus, proper selection of the rules combination which used for aspects extraction is very important. Due to that, IWOA algorithm will be applied on 126 rules to select the optimal combination of these rules and discard the irrelevant and low-quality rules.

Table 5.3: Results obtained from applying full rules set on each dataset

Data	Precision	Recall	F-measure
D1	0.73	0.96	0.83
D2	0.74	0.97	0.84
D3	0.76	0.98	0.86
D4	0.74	0.97	0.84
D5	0.76	0.95	0.84
Avg	**0.75**	**0.97**	**0.84**

From the results, not all rules must be considered at the same time for extracting aspects, as there are some good rules for extraction that can extract aspects at high recall and at the same time with less non-aspects words.

5.4.2 Experiment 2: Results obtained from applying IWOA for rules selection

Experiment 2 was conducted to show the effectiveness of IWOA on aspects extraction. After applying IWOA on full rules set to select the optimal rules for aspect extraction, the results are shown in Table 5.4. In every session, one dataset was used for training and another dataset was used for testing as shown in Table 5.4.

Table 5.4: Results obtained from applying IWOA for rules selection

Testing Data	Precision	Recall	F-measure	Training Data
D1	0.86	0.94	0.90	D5
D2	0.84	0.94	0.89	D3
D3	0.89	0.94	0.91	D5
D4	0.85	0.96	0.90	D1
D5	0.86	0.92	0.89	D3
Avg	**0.86**	**0.94**	**0.90**	

From the table, there is a clear improvement on the precision with 11% and a little decrement in recall, which is about 3%. Furthermore, there is also an improvement on F-measure with 6%. However, there are some non-aspect words which decrease the precision. A pruning is required to remove non-aspect words and retain the correct aspects. Thus, in the next experiment, PA algorithm will be applied to remove the non-aspect terms.

Besides than that, a comparison between standard WOA and the proposed IWOA over all datasets was also conducted and the results are shown in Table 5.5. Based on the results from all datasets as shown in the table, it is obviously noticed that IWOA outperformed WOA in all metrices including precision, recall, and F-measure over all datasets. In addition, these results proved the ability of IWOA to escape from local optima based on the used LSA algorithm. Furthermore, it confirmed the ability of IWOA to generate diverse solutions based on using CM.

As shown in Figure 5.1, the improvements on the standard WOA which is called IWOA, present interesting results and improve the performance over all datasets. In addition, as shown in Table 5.5, IWOA outperforms original WOA with 2% in precision, 4% in recall, and 3% in term of F-measure respectively.

Table 5.5: Comparison between original WOA and IWOA

Testing Data	WOA			IWOA			Training Data
	P	R	F	P	R	F	
D1	0.85	0.91	0.88	**0.86**	**0.94**	**0.90**	**D5**
D2	0.83	0.90	0.86	**0.84**	**0.94**	**0.89**	**D3**
D3	0.82	0.92	0.87	**0.89**	**0.94**	**0.91**	**D5**
D4	0.84	0.92	0.88	**0.85**	**0.96**	**0.90**	**D1**
D5	0.85	0.87	0.86	**0.86**	**0.92**	**0.89**	**D3**
Avg	0.84	0.90	0.87	**0.86**	**0.94**	**0.90**	

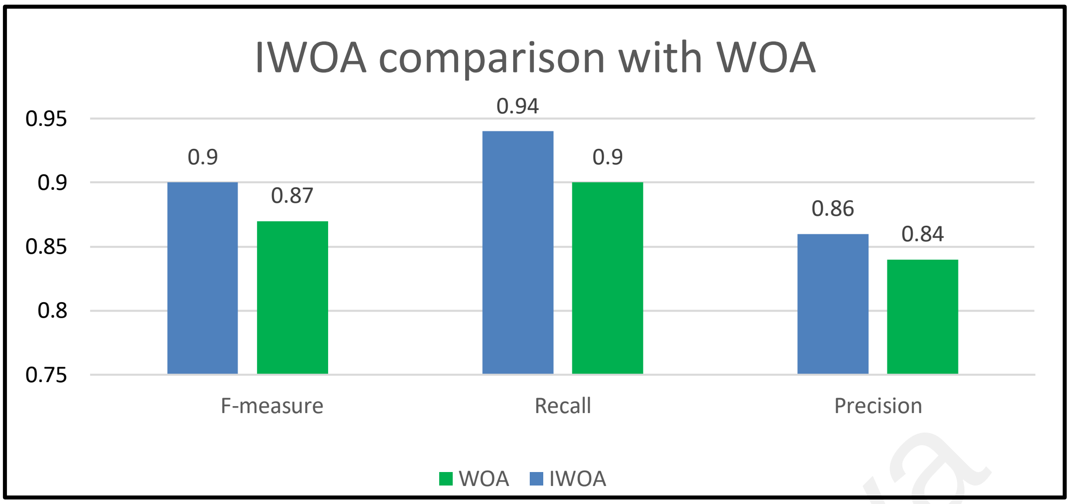

Figure 5.1 IWOA Comparison with WOA

To further evaluate and test the efficiency of IWOA algorithm, it was compared with 7 well-known optimization algorithms including FFA (Yang, 2009), SCA (Mirjalili, 2016), SSA (Mirjalili et al., 2017), PSO (Eberhart & Kennedy, 1995), MFO (Mirjalili, 2015), MVO (Mirjalili et al., 2016), and GWO (Mirjalili et al., 2014). The results of the comparisons are shown in Table 5.6. These selected baseline algorithms are selected as they represent the most recent optimization algorithms, well-known optimization algorithms, and some represents swarm algorithms types as WOA. In addition, they are selected as they were frequently used by the previous researches as a baseline to compare with such as in works (El Aziz et al., 2017; Elaziz et al., 2017; Singh & Singh, 2017; Abdel-Basset et al., 2018; Algabalawy et al., 2018; Aljarah, Ala'M, et al., 2018; Hegazy et al., 2018; Jadhav & Gomathi, 2018; Li et al., 2018; Luo et al., 2018; Sayed et al., 2018; Singh & Hachimi, 2018; Elaziz & Mirjalili, 2019; Nazari-Heris et al., 2019). As displayed in Table 5.6, it is clearly noticed that IWOA algorithm outperforms all other algorithms over all datasets as indicated by bold fonts in both tables. These results by IWOA algorithm was obtained because of its ability to balance between exploitation and exploration and the ability of IWOA algorithm to avoid being stuck into local optima.

Based on the results from IWOA comparison with other state-of-the-art algorithms, it confirmed the superiority of IWOA over all other algorithms. IWOA outperforms GWO with 1% in precision, 6% in recall, and 4% in term of F-measure respectively. The algorithm also, outperforms MVO with 2% in precision, 4% in recall, and 3% in term of F-measure respectively. Furthermore, IWOA outperforms MFO with 3% in precision, 3% in recall, and 3% in term of F-measure respectively. In addition, IWOA outperforms PSO with 5% in precision, 2% in recall, and 4% in term of F-measure respectively. The following rules which are numbered based on rules numbers represent a sample of frequently optimal selected rules by IWOA:

(4, 7, 8, 9, 16, 19, 22, 23, 26, 27, 30, 34, 36, 37, 40, 42, 43, 44, 48, 49, 53, 57, 58, 60, 61, 62, 63, 64, 65, 66, 71, 72, 75, 77, 81, 85, 89, 90, 92, 95, 96, 97, 98, 99, 104, 105, 108, 111, 113, 114, 115, 116, 117, 121, 122, 123, 124).

Table 5.6: Comparison between IWOA and other optimization algorithms

Test Data	FFA			SCA			SSA			PSO			MFO			MVO			GWO			IWOA			Train Data
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	
D1	0.85	0.93	0.89	0.85	0.91	0.88	0.81	0.89	0.85	0.81	0.93	0.87	0.81	0.93	0.87	0.83	0.93	0.88	0.83	0.90	0.86	**0.86**	**0.94**	**0.90**	**D5**
D2	0.81	0.92	0.86	0.83	0.93	0.88	0.83	0.92	0.87	0.80	0.91	0.85	0.81	0.91	0.86	0.83	0.91	0.87	0.83	0.87	0.85	**0.84**	**0.94**	**0.89**	**D3**
D3	0.86	0.89	0.87	0.83	0.88	0.85	0.82	0.91	0.86	0.84	0.92	0.88	0.88	0.88	0.88	0.86	0.88	0.87	0.87	0.85	0.86	**0.89**	**0.94**	**0.91**	**D5**
D4	0.82	0.93	0.87	0.82	0.92	0.87	0.83	0.93	0.88	0.81	0.93	0.87	0.81	0.93	0.87	0.84	0.90	0.87	0.84	0.91	0.87	**0.85**	**0.96**	**0.90**	**D1**
D5	0.82	0.87	0.84	0.82	0.91	0.86	0.84	0.87	0.85	0.81	0.90	0.85	0.85	0.90	0.87	0.85	0.86	0.85	0.84	0.86	0.85	**0.86**	**0.92**	**0.89**	**D3**
Avg	0.83	0.91	0.87	0.83	0.91	0.87	0.83	0.90	0.86	0.81	0.92	0.86	0.83	0.91	0.87	0.84	0.90	0.87	0.85	0.88	0.86	**0.86**	**0.94**	**0.90**	

When compare IWOA with SSA, IWOA outperforms SSA with 3% in precision, 4% in recall, and 4% in term of F-measure respectively. IWOA outperforms SCA with 3% in precision, 3% in recall, and 3% in term of F-measure respectively Finally, IWOA outperforms FFA approach with 3% in precision, 3% in recall, and 3% in term of F-measure respectively. Therefore, as indicated by the results, the proposed improvements make IWOA outperforms all other state-of-the-art optimization algorithms. In addition, as shown from Figure 5.2 it is clearly noticed the superiority of IWOA in comparison with other optimization algorithms. The outperformance of IWOA is due to its ability to escape from local optima and balance between exploration and exploitation.

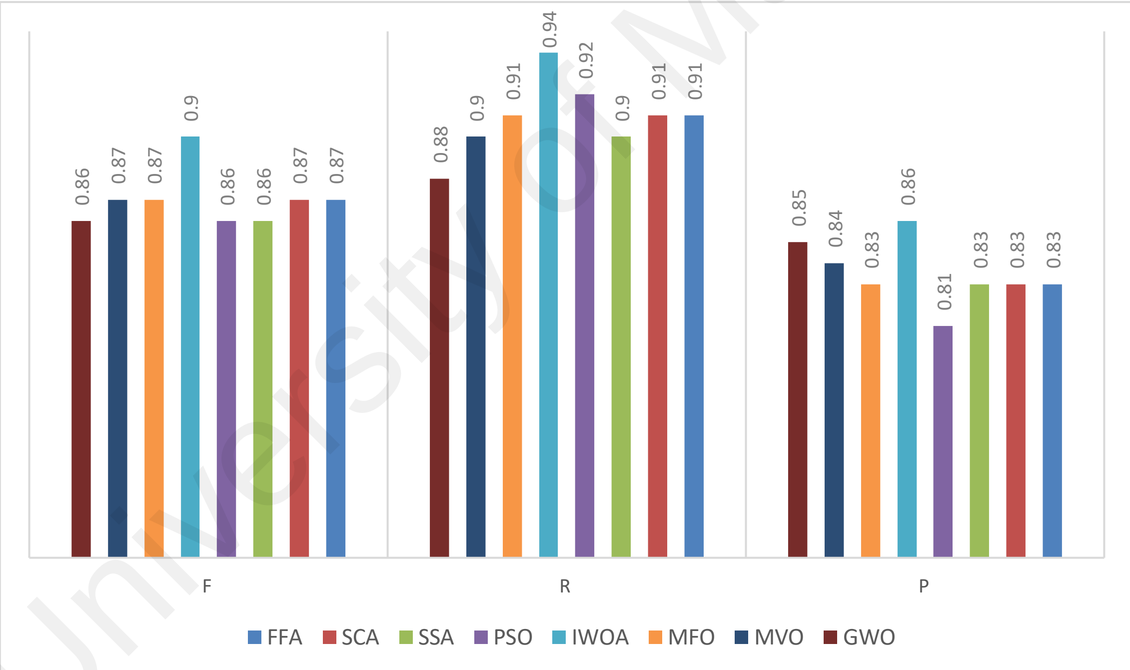

Figure 5.2 IWOA Comparison with other optimization algorithms

5.4.3 Experiment 3: Results obtained from applying pruning algorithm

In this experiment, the results are obtained after applying pruning algorithm on the aspects obtained from Experiment 2. The results are displayed in Table 5.7.

Table 5.7: Results obtained from applying pruning algorithm

Data	Precision	Recall	F-measure
D1	0.91	0.93	0.92
D2	0.91	0.92	0.91
D3	0.93	0.93	0.93
D4	0.93	0.95	0.94
D5	0.92	0.91	0.91
Avg	**0.92**	**0.93**	**0.92**

It is clearly noticed from Table 5.7 that there is a further improvement on precision value with 6% increase and only 1% drop in recall which considered as not too much effect on recall value. However, there is another further improvement on F-measure with 2% value. Figure. 5.3 shows the performance improvement resulted from each phase on aspect extraction algorithm. From the figure, it is clearly noticed that there is a noticeable improvement after the application of each phase.

Figure 5.3 Performance improvement at each phase

5.4.4 Experiment 4: Results comparison with baseline methods

As presented in Table 5.8, our proposed approach outperformed all others baseline works in all metrics including precision, recall, and F-measure as indicated by bold font. Thus, the proposed approach is performing more consistency as it balances between the recall and precision as shown from the results in bold. IWOA+PA outperforms DP with 4% in precision, 10% in recall, and 6% in term of F-measure respectively. Also, IWOA+PA outperforms RubE with 5% in precision, 5% in recall, and 5% in term of F-measure respectively. Furthermore, IWOA+PA outperforms TF-RBM with 5% in precision, 1% in recall, and 3% in term of F-measure respectively. In addition, IWOA+PA outperforms RSLS with 7% in precision, 2% in recall, and 4% in term of F-measure respectively. When compare IWOA+PA with Htay approach, IWOA+PA outperforms Htay work with 19% in precision, 7% in recall, and 13% in term of F-measure respectively. Finally, IWOA+PA outperforms CNN+LP approach with 2% in precision, 7% in recall, and 4% in term of F-measure respectively. Therefore, at overall IWOA+PA dominates all other state of the art and most recent aspect extraction works. Figure. 5.4 shows the efficiency of the proposed IWOA+PA in comparison to all other baseline works, where the results of each method was taken from their reported results as following DP (Qiu et al., 2011), RubE (Kang & Zhou, 2017; Rana & Cheah, 2017), TF-RBM (Rana & Cheah, 2017), RSLS (Liu, Gao, et al., 2016), Htay (Htay & Lynn, 2013), and CNN+LP (Poria et al., 2016).

As shown in Figure 5.4, IWOA+PA is outperforming all other baseline works in all performance metrices. This superiority of IWOA+PA resulted from first the use of different rules types. These rules overcome the weaknesses of previous studies and introduce new rules that are not explored by these studies. Also, these rules improve the recall but not all these rules relevant in extraction. To solve this issue, a proper selection of rules combination was conducted using IWOA. IWOA improve the performance because of its ability to balance between exploitation and exploration and escape from local optima. Finally, the use of PA after IWOA application has further improved the performance of the algorithm.

Table 5.8: Comparisons to other approach (Precision (P), Recall (R), and F-measure)

Data	DP			RubE			TF-RBM			RSLS			Htay			CNN + LP			IWOA+PA		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
D1	0.87	0.81	0.84	0.87	0.86	0.86	0.80	0.89	0.84	0.85	0.91	0.88	0.74	0.92	0.82	0.93	0.85	0.89	0.91	0.93	0.92
D2	0.90	0.81	0.85	0.90	0.86	0.88	0.87	0.93	0.90	0.89	0.94	0.91	0.71	0.81	0.76	0.83	0.87	0.85	0.91	0.92	0.91
D3	0.90	0.86	0.88	0.90	0.91	0.90	0.92	0.93	0.92	0.83	0.90	0.86	0.74	0.82	0.78	0.93	0.88	0.90	0.93	0.93	0.93
D4	0.81	0.84	0.82	0.87	0.90	0.88	0.86	0.93	0.90	0.82	0.91	0.86	0.70	0.76	0.73	0.93	0.86	0.89	0.93	0.95	0.94
D5	0.92	0.86	0.89	0.90	0.85	0.87	0.88	0.90	0.89	0.86	0.90	0.88	0.78	0.97	0.87	0.90	0.84	0.87	0.92	0.91	0.91
Avg	0.88	0.83	0.86	0.87	0.88	0.87	0.87	0.92	0.89	0.85	0.91	0.88	0.73	0.86	0.79	0.90	0.86	0.88	**0.92**	**0.93**	**0.92**

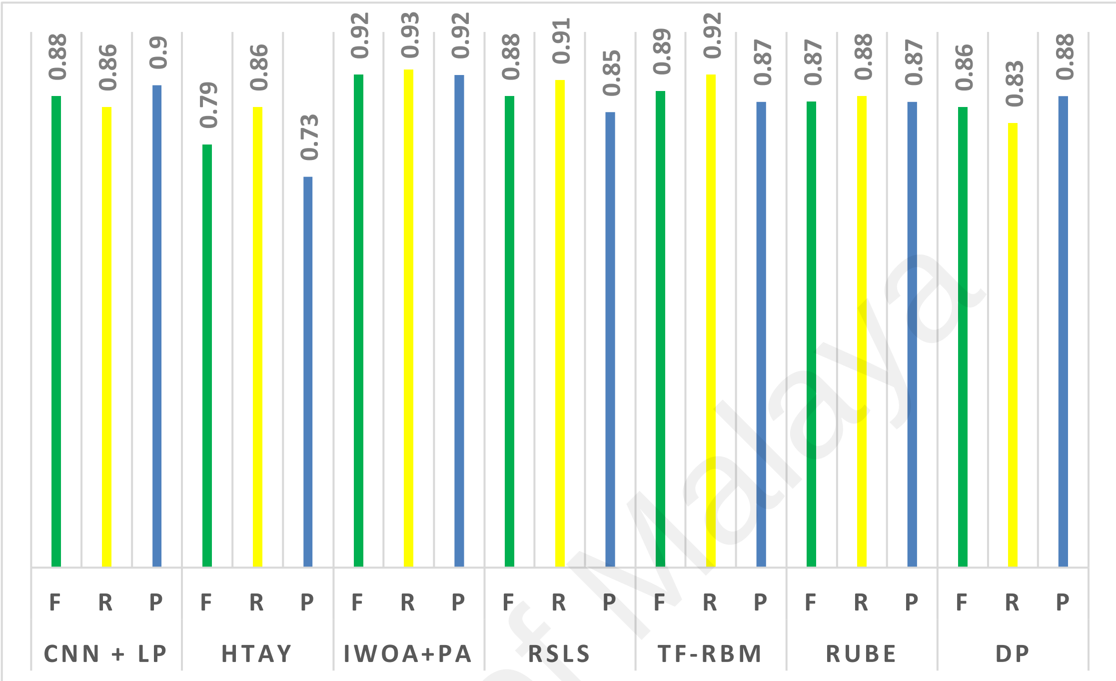

Figure 5.4 Comparison with baseline works

To further evaluate the ranks of each aspect algorithm and determine if there is a statistical difference between the proposed IWOA+PA and other algorithms, the Friedman Aligned-Rank test is applied on the obtained F-measure results from applying each algorithm over the 5 different datasets to find the rank for each algorithm (Hodges & Lehmann, 1962; Demšar, 2006). Then, to find the significance difference between these aspect extraction algorithms the paired *t*-test is applied, similar to the statistic test used in the baseline CNN+LP (Poria et al., 2016) . In the adopted paired *t*-test, the null hypothesis states that there is no difference between the performance of two aspect extraction algorithms. Therefore, the null hypothesis will be rejected if the significant test value of the aspect extraction algorithm is less than or equal to the value of significant level. The rejection of the null hypothesis means that there is a significant difference between the two algorithms. In this experiment, to find the significant difference between these algorithms, the value of the used significance level is 0.05. The obtained results from applying Friedman test is shown in Table 5.9.

Table 5.9: Average ranks by each method using Friedman Aligned rank test.

Algorithm	Ranking
IWOA+PA	**4.1**
TF-RBM	13.6
CNN + LP	16.8
RubE	17.5
RSLS	17.5
DP	24.9
Htay	31.6

Overall, the proposed IWOA+PA has the best average ranking in comparison to other state-of-the-art aspect extraction algorithms which were adopted in the experiments as shown in Table 5.9. To further test the significant difference between IWOA+PA and other algorithms, the results of applying paired *t*-test is shown in Table 5.10.

Table 5.10: Significance test values by paired t-test.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	IWOA_PA - DP	.06600	.03715	.01661	.01987	.11213	3.973	4	.017
Pair 2	IWOA_PA - RubE	.04400	.01517	.00678	.02517	.06283	6.487	4	.003
Pair 3	IWOA_PA - TF_RBM	.03200	.02950	.01319	-.00462-	.06862	2.426	4	.072
Pair 4	IWOA_PA - RSLS	.04400	.03209	.01435	.00415	.08385	3.066	4	.037
Pair 5	IWOA_PA - Htay	.13000	.06364	.02846	.05098	.20902	4.568	4	.010
Pair 6	IWOA_PA - CNN_LP	.04200	.01304	.00583	.02581	.05819	7.203	4	.002

Based on the obtained results from Table 5.10, the proposed IWOA+PA is significantly different from DP, RubE, RSLS, Htay, and CNN+LP, whereas IWOA+PA is not significantly different from TF-RBM. However, IWOA+PA obtained better precision, recall, F-measure, and ranking results than TF-RBM method.

5.5 Summary

This chapter introduced the details of the datasets used in all experiments. It also presented the details about the performance metrices and baseline works used. In addition, it provided the details about the experiments conducted and discussions on the results.

The first experiment was conducted using all 126 rules combination. The results achieved 75% precision, 97% recall, and 84% F-measure. From these results, it is confirmed that not all rules are suitable for aspects extraction. Thus, a proper rules selection must be carried out to select the most suitable rules for aspects extraction. To accomplish this task, the second experiment used IWOA to select the optimal rules combination and discard irrelevant rules. The results achieved 86% precision, 94% recall, and 90% F-measure. Therefore, the application of IWOA improved the performance by selecting the most suitable extraction rules. Moreover, to prove the superiority of IWOA, it was compared to the standard WOA and other well-known optimization algorithms including FFA, PSO, SSA, SCA, MVO, GWO, and MFO. The obtained results confirmed the outperformance of IWOA in comparison with other optimization algorithms. The outperformance of IWOA is based on its ability to balance between exploitation and exploration based on the two improvements, the use of CM improved population diversity, while LSA improved the best solution and avoid it from being stuck at local optima. After applying IWOA, there is a clear improvement in the performance, but there is a possibility to further improve the performance.

In the third experiment, PA was applied on aspects resulted from the application of IWOA selected rules. From this experiment the achieved results were 92% precision, 93% recall, and 92% F-measure. Finally, to investigate and evaluate the performance of IWOA+PA, in the fourth experiment, several experiments were conducted for comparison of IWOA+PA with other baseline works including DP, RubE, TF-RBM, RSLS, Htay, and CNN + LP. Thus, IWOA+PA outperforms all other baseline works over all datasets in all performance metrics. In addition, IWOA obtained the best average ranking among these algorithms using Friedman test. Based on paired *t*-test results, IWOA+PA is significantly difference from DP, RubE, RSLS, Htay, and CNN+LP algorithms.

As a conclusion, the use of different rules types improves the aspect extraction and resolves the weaknesses of the previous studies. Not all of these rules have the same quality, but some of these rules are extracting too many incorrect aspects. The IWOA algorithm can select optimal rules combination from these rules while it also discards low quality rules. IWOA which was improved by LSA and CM improvements, proved its ability to select optimal rules combination. Although IWOA improved the performance, but there were some non-aspect words which decreased the precision. Thus, PA was used after IWOA to remove these aspects. The use of IWOA with PA proved the ability of IWOA+PA algorithm in making a balance between different performance metrics.

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Introduction

Since explicit aspect extraction is important in different types of applications, it has attracted many researchers to develop different extraction techniques. Most of the previous studies focus on using either dependency-based approach or pattern-based approach. However, the available online reviews are of mixed types including structured and unstructured reviews. In addition, available rules-based methods in literature use either dependency or pattern approaches for aspects extraction, but no study in literature combine both approaches together. The dependency-based approach is better for structured reviews, while pattern approach is better for unstructured reviews. In addition, many extraction rules are still unexplored in both types.

To tackle the current issues in aspect extraction approaches, a number of contributions are proposed. Firstly, a combination of rules from previous studies together with the new developed rules is proposed. However, using all these rules together for explicit aspects extraction can decrease precision while it improves recall. Therefore, a proper selection of rules combination subset from full rules set is also required. The rules selection algorithm can be used to select optimal rules combination and discard low quality rules. Thus, an improvement to standard WOA were proposed which include the using of LSA at the end of WOA to search for better solution and improve its local search ability, and the use of CM to balance between local and global search of WOA and improve solutions diversity. The improved algorithm called IWOA was used to select optimal rules from the full rules set. IWOA results showed a clear improvement in performance metrics including F-measure and precision, but there is a possibility to further improve the results. Proper improvements to the existing pruning algorithms are proposed to improve extraction accuracy. This proposed pruning consists of three pruning phases including frequency, product manual, and direct opinion association phases. These pruning phases manage to improve the extraction performance. Thus, the major aim of this thesis is to improve explicit aspect extraction algorithm for extracting opinionated explicit aspects.

6.2 Summary of Thesis

Aspect based sentiment analysis (ABSA) is considered as an attractive area of researches by many researchers. This attractiveness comes as the result of huge volume of reviews available online on different shopping websites such as Amazon, and many more. These reviews grab the interest of both customers and manufactures. However, to read all these reviews by either individual customer or manufacture is an impossible and time consuming. ABSA solved this problem by extracting different opiniated aspects of products and opinion words expressed toward each aspect. Then, the semantic orientation about each product aspect will be determined based on the opinion words related to that aspect. Aspect extraction is the most important task in ABSA. Thus, the main aim of this thesis is to improve explicit aspects extraction algorithm for extracting opinionated explicit aspects from products reviews. To achieve this aim, a number of objectives are presented as in chapter 1. In the first objective, a number of new aspects extraction rules were formulated and developed. To achieve this objective, these new rules were combined with a number of rules from the previous studies. In the second objective, an improvement is proposed for WOA algorithm to solve its problems including population diversity, balance between exploitation and exploration, and fall in local optima. To achieve this objective LSA and CM is combined with WOA and the new algorithm is called IWOA. In addition, IWOA is used for rules selection. The third objective proposed improvements to pruning algorithm. To achieve that, a three-phase pruning algorithm is proposed, and it is known as PA. PA is composed of three phases including frequency pruning, product manual pruning, and direct opinion association pruning. Finally, in the fourth objective, a comparison between IWOA and other optimization algorithms is conducted. In addition, a comparison between IWOA+PA and other aspect extraction baseline works in literature was also conducted. The same benchmark dataset was used in these experiments.

6.3 Summary of thesis contributions

This thesis come up with number of contributions that can be summarized as the following:

1. **Formulation of a set of new explicit aspects extraction rules**: these new rules were formulated and developed to overcome the weaknesses of some available extraction rules and to improve the extraction performance.
2. **Combinations of different rules types:** the extraction rules are combined from different types including dependency-based and pattern-based rules from previous studies, and the new developed rules. In total the full set of rules includes 126 rules.
3. **Improved whale optimization algorithm for rules selection:** WOA is improved to solve its problems by using LSA and CM. LSA is combined with WOA at the end of each WOA iteration to improve its exploitation ability and escape from local optima. In addition, CM is combined with WOA to improve population diversity and balance between exploitation (local search) and exploration (global search). Finally, the improved WOA (IWOA) was applied on 126 extraction rules to select the best rules combination and discard the low-quality rules. IWOA algorithm has

proved its capability to select best rules combinations based on the results achieved in chapter 5 from the comparison with other optimization algorithms.

4. **Improved pruning algorithm (PA):** to overcome the limitations in pruning methods for the approaches proposed in the literature, PA algorithm was improved. In this improved PA, two new pruning phases where proposed including pruning based on product manual and pruning based on direct opinion association.

The proposed explicit aspect extraction algorithm is considered as promising algorithm based on the achieved results in comparison with other state-of-the-art extraction methods and optimization algorithms.

6.4 Limitations

Although the proposed aspect extraction algorithm IWOA+PA shows promising results in aspects extraction, but still there are some limitations as the following:

- **Implicit aspects extraction**: IWOA+PA only able to extract explicit aspect and cannot extract implicit aspect. Implicit aspect such as in example " *the phone is too large*", in this example *large* opinion word is an indicator for implicit aspect ***size***.
- **Explicit aspects extraction for other languages**: IWOA+PA only able to extract explicit aspect in English. However, it cannot extract explicit aspects in other languages such as Arabic and Malay.

6.5 Future work

As the main aim of this thesis is to extract opinionated explicit aspects. Therefore, there are more opportunities for future works. One possible direction is to define extraction rules for other languages such as Arabic and Malay and apply IWOA+PA for extracting aspects in these languages. One more possible future work is to apply the optimal selected rules by IWOA on datasets from different domains such as medical, movie, hotels, and different types of products or services.

Another possible future work which consider the extraction of implicit aspects. In this case, the proposed work by Tubishat and Idris (2017), will be developed to extract implicit aspects. The implicit extraction will be performed using a hybrid approach by utilizing both corpus-based approach and dictionary-based approach.

REFERENCES

Abdel-Basset, M., Abdle-Fatah, L., & Sangaiah, A. K. (2018). An improved Lévy based whale optimization algorithm for bandwidth-efficient virtual machine placement in cloud computing environment. *Cluster Computing*, 1-16.

Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., & Hussain, A. (2015). Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. *Cognitive Computation, 7*(4), 487-499.

Ahmed, M. M., Houssein, E. H., Hassanien, A. E., Taha, A., & Hassanien, E. (2017). *Maximizing lifetime of wireless sensor networks based on whale optimization algorithm.* Paper presented at the International conference on advanced intelligent systems and informatics.

Akalamkam, K., & Mitra, J. K. (2018). Consumer Pre-purchase Search in Online Shopping: Role of Offline and Online Information Sources. *Business Perspectives and Research, 6*(1), 42-60.

Akhtar, M. S., Gupta, D., Ekbal, A., & Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. *Knowledge-Based Systems, 125*, 116-135.

Ala'M, A.-Z., Faris, H., & Hassonah, M. A. (2018). Evolving Support Vector Machines using Whale Optimization Algorithm for spam profiles detection on online social networks in different lingual contexts. *Knowledge-Based Systems, 153*, 91-104.

Algabalawy, M. A., Abdelaziz, A. Y., Mekhamer, S. F., & Aleem, S. H. A. (2018). Considerations on optimal design of hybrid power generation systems using whale and sine cosine optimization algorithms. *Journal of Electrical Systems and Information Technology, 5*(3), 312-325.

Ali, M., & Pant, M. (2011). Improving the performance of differential evolution algorithm using Cauchy mutation. *Soft Computing, 15*(5), 991-1007.

Aljarah, I., Ala'M, A.-Z., Faris, H., Hassonah, M. A., Mirjalili, S., & Saadeh, H. (2018). Simultaneous feature selection and support vector machine optimization using the grasshopper optimization algorithm. *Cognitive Computation*, 1-18.

Aljarah, I., Faris, H., & Mirjalili, S. (2018). Optimizing connection weights in neural networks using the whale optimization algorithm. *Soft Computing, 22*(1), 1-15.

Almatrafi, O., Parack, S., & Chavan, B. (2015). *Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014.* Paper presented at the Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication.

Alzaqebah, A., Masadeh, R., & Hudaib, A. (2018). *Whale optimization algorithm for requirements prioritization.* Paper presented at the 2018 9th International Conference on Information and Communication Systems (ICICS).

Arulmurugan, R., Sabarmathi, K., & Anandakumar, H. (2017). Classification of sentence level sentiment analysis using cloud machine learning techniques. *Cluster Computing*, 1-11.

Asghar, M. Z., Khan, A., Zahra, S. R., Ahmad, S., & Kundi, F. M. (2017). Aspect-based opinion mining framework using heuristic patterns. *Cluster Computing*, 1-19.

Bagheri, A., Saraee, M., & de Jong, F. (2013). *An unsupervised aspect detection model for sentiment analysis of reviews.* Paper presented at the International Conference on Application of Natural Language to Information Systems.

Barham, R., & Aljarah, I. (2017). *Link prediction based on whale optimization algorithm.* Paper presented at the 2017 International Conference on New Trends in Computing Sciences (ICTCS).

Barrón-Estrada, M. L., Zatarain-Cabada, R., Oramas-Bustillos, R., & González-Hernández, F. (2017). *Sentiment Analysis in an Affective Intelligent Tutoring System.* Paper presented at the Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on.

Bauman, K., Liu, B., & Tuzhilin, A. (2017). *Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews.* Paper presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Bentouati, B., Chaib, L., & Chettih, S. (2016). *A hybrid whale algorithm and pattern search technique for optimal power flow problem.* Paper presented at the Modelling, Identification and Control (ICMIC), 2016 8th International Conference on.

Bhesdadiya, R., Jangir, P., Jangir, N., Trivedi, I. N., & Ladumor, D. (2016). Training multi-layer perceptron in neural network using whale optimization algorithm. *Indian J. Sci. Technol, 9*(19), 28-36.

Blazevic, V., Hammedi, W., Garnefeld, I., Rust, R. T., Keiningham, T., Andreassen, T. W., . . . Carl, W. (2013). Beyond traditional word-of-mouth: an expanded model of customer-driven influence. *Journal of Service Management, 24*(3), 294-313.

Brody, S., & Elhadad, N. (2010). *An unsupervised aspect-sentiment model for online reviews.* Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Canayaz, M., & Demir, M. (2017). *Feature selection with the whale optimization algorithm and artificial neural network.* Paper presented at the 2017 International Artificial Intelligence and Data Processing Symposium (IDAP).

Cao, J., Zeng, K., Wang, H., Cheng, J., Qiao, F., Wen, D., & Gao, Y. (2014). Web-based traffic sentiment analysis: Methods and applications. *IEEE transactions on Intelligent Transportation systems, 15*(2), 844-853.

Caputo, A., Basile, P., de Gemmis, M., Lops, P., Semeraro, G., & Rossiello, G. (2017). SABRE: A Sentiment Aspect-Based Retrieval Engine *Information Filtering and Retrieval* (pp. 63-78): Springer.

Ceyhan, M., Orhan, Z., & Domnori, E. (2017). Health service quality measurement from patient reviews in Turkish by opinion mining *CMBEBIH 2017* (pp. 649-653): Springer.

Chen, Mukherjee, A., & Liu, B. (2014). *Aspect extraction with automated prior knowledge learning.* Paper presented at the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Chen, Qi, L., & Wang, F. (2012). Comparison of feature-level learning methods for mining online consumer reviews. *Expert Systems with Applications, 39*(10), 9588-9601.

Chen, Y., Vepa, R., & Shaheed, M. H. (2018). Enhanced and speedy energy extraction from a scaled-up pressure retarded osmosis process with a whale optimization based maximum power point tracking. *Energy, 153*, 618-627.

Chinsha, T., & Joseph, S. (2015). *A syntactic approach for aspect based opinion mining.* Paper presented at the 2015 IEEE International Conference on Semantic Computing (ICSC).

Choi, Y., & Cardie, C. (2010). *Hierarchical sequential learning for extracting opinions and their attributes.* Paper presented at the Proceedings of the ACL 2010 conference short papers.

Colorni, A., Dorigo, M., & Maniezzo, V. (1992). *Distributed optimization by ant colonies.* Paper presented at the Proceedings of the first European conference on artificial life.

Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J., & Vallejo, C. G. (2013). 'Long autonomy or long delay?'The importance of domain in opinion mining. *Expert Systems with Applications, 40*(8), 3174-3184.

Dao, T.-K., Pan, T.-S., & Pan, J.-S. (2016). *A multi-objective optimal mobile robot path planning based on whale optimization algorithm.* Paper presented at the 2016 IEEE 13th International Conference on Signal Processing (ICSP).

Das, T. K., Acharjya, D., & Patra, M. (2014). *Opinion mining about a product by analyzing public tweets in Twitter.* Paper presented at the Computer Communication and Informatics (ICCCI), 2014 International Conference on.

De Marneffe, M.-C., & Manning, C. D. (2008). *Stanford typed dependencies manual.* Retrieved from

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of machine learning research, 7*(Jan), 1-30.

Ding, X., Liu, B., & Yu, P. S. (2008). *A holistic lexicon-based approach to opinion mining.* Paper presented at the Proceedings of the 2008 international conference on web search and data mining.

Du, R. Y., Hu, Y., & Damangir, S. (2015). Leveraging trends in online searches for product features in market response modeling. *Journal of marketing, 79*(1), 29-43.

Eberhart, R., & Kennedy, J. (1995). *A new optimizer using particle swarm theory.* Paper presented at the MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science.

Eid, H. F. (2018). Binary whale optimisation: an effective swarm algorithm for feature selection. *International Journal of Metaheuristics, 7*(1), 67-79.

Eirinaki, M., Pisal, S., & Singh, J. (2012). Feature-based opinion mining and ranking. *Journal of Computer and System Sciences, 78*(4), 1175-1184.

El Aziz, M. A., Ewees, A. A., & Hassanien, A. E. (2017). Whale optimization algorithm and moth-flame optimization for multilevel thresholding image segmentation. *Expert Systems with Applications, 83*, 242-256.

El Aziz, M. A., Ewees, A. A., & Hassanien, A. E. (2018). Multi-objective whale optimization algorithm for content-based image retrieval. *Multimedia Tools and Applications, 77*(19), 26135-26172.

Elaziz, M. A., & Mirjalili, S. (2019). A hyper-heuristic for improving the initial population of whale optimization algorithm. *Knowledge-Based Systems, 172*, 42-63.

Elaziz, M. A., Oliva, D., & Xiong, S. (2017). An improved opposition-based sine cosine algorithm for global optimization. *Expert Systems with Applications, 90*, 484-500.

Esuli, A., & Sebastiani, F. (2005). *Determining the semantic orientation of terms through gloss classification.* Paper presented at the Proceedings of the 14th ACM international conference on Information and knowledge management.

Formato, R. A. (2007). Central force optimization. *Prog Electromagn Res, 77*, 425-491.

Fu, X., Liu, W., Xu, Y., & Cui, L. (2017). Combine HowNet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis. *Neurocomputing, 241*, 18-27.

Gamboa, F., Malamud, M. A., Maguitman, A. G., & Chesñevar, C. I. (2016). *Citymis optree: Intelligent citizen management using sentiment analysis and opinion trees.* Paper presented at the Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance.

Ghahremani-Nahr, J., Kian, R., & Sabet, E. (2019). A robust fuzzy mathematical programming model for the closed-loop supply chain network design and a whale optimization solution algorithm. *Expert Systems with Applications, 116*, 454-471.

Gopalakrishnan, V., & Ramaswamy, C. (2017). Patient opinion mining to analyze drugs satisfaction using supervised learning. *Journal of Applied Research and Technology, 15*(4), 311-319.

Hai, Z., Chang, K., & Cong, G. (2012). *One seed to find them all: mining opinion features via association.* Paper presented at the Proceedings of the 21st ACM international conference on Information and knowledge management.

Hai, Z., Chang, K., Kim, J.-J., & Yang, C. C. (2014). Identifying features in opinion mining via intrinsic and extrinsic domain relevance. *IEEE Transactions on Knowledge and Data Engineering, 26*(3), 623-634.

Hassan, G., & Hassanien, A. E. (2018). Retinal fundus vasculature multilevel segmentation using whale optimization algorithm. *Signal, Image and Video Processing, 12*(2), 263-270.

Hatzivassiloglou, V., & 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.

Hatzivassiloglou, V., & Wiebe, J. M. (2000). *Effects of adjective orientation and gradability on sentence subjectivity.* Paper presented at the Proceedings of the 18th conference on Computational linguistics-Volume 1.

Hegazy, A. E., Makhlouf, M., & El-Tawel, G. S. (2018). Dimensionality reduction using an improved whale optimization algorithm for data classification. *International Journal of Modern Education and Computer Science, 10*(7), 37.

Hodges, J., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics, 33*(2), 482-497.

Holland, J. H. (1992). Genetic algorithms. *Scientific american, 267*(1), 66-73.

Htay, S. S., & Lynn, K. T. (2013). Extracting product features and opinion words using pattern knowledge in customer reviews. *The Scientific World Journal, 2013*.

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

Hu, M., & Liu, B. (2004b). *Mining opinion features in customer reviews.* Paper presented at the AAAI.

Huang, S., Liu, X., Peng, X., & Niu, Z. (2012). *Fine-grained product features extraction and categorization in reviews opinion mining.* Paper presented at the Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on.

Hussien, A. G., Hassanien, A. E., Houssein, E. H., Bhattacharyya, S., & Amin, M. (2019). S-shaped binary whale optimization algorithm for feature selection *Recent trends in signal and image processing* (pp. 79-87): Springer.

Jadhav, A. N., & Gomathi, N. (2018). WGC: Hybridization of exponential grey wolf optimizer with whale optimization for data clustering. *Alexandria Engineering Journal, 57*(3), 1569-1584.

Jadhav, A. R., & Shankar, T. (2017). Whale optimization based energy-efficient cluster head selection algorithm for wireless sensor networks. *arXiv preprint arXiv:1711.09389*.

Jakob, N., & Gurevych, I. (2010). *Extracting opinion targets in a single-and cross-domain setting with conditional random fields.* Paper presented at the Proceedings of the 2010 conference on empirical methods in natural language processing.

Jiang, P., Zhang, C., Fu, H., Niu, Z., & Yang, Q. (2010). *An approach based on tree kernels for opinion mining of online product reviews.* Paper presented at the Data Mining (ICDM), 2010 IEEE 10th International Conference on.

Jin, W., Ho, H. H., & Srihari, R. K. (2009). *A novel lexicalized HMM-based learning framework for web opinion mining.* Paper presented at the Proceedings of the 26th annual international conference on machine learning.

Jindal, N., & Liu, B. (2006a). *Identifying comparative sentences in text documents.* Paper presented at the Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.

Jindal, N., & Liu, B. (2006b). *Mining comparative sentences and relations.* Paper presented at the AAAI.

Jindal, N., & Liu, B. (2008). *Opinion spam and analysis.* Paper presented at the Proceedings of the 2008 international conference on web search and data mining.

Kang, Y., & Zhou, L. (2017). RubE: Rule-based methods for extracting product features from online consumer reviews. *Information & Management, 54*(2), 166-176.

Karlekar, N. P., & Gomathi, N. (2018). OW-SVM: Ontology and whale optimization-based support vector machine for privacy-preserved medical data classification in cloud. *International Journal of Communication Systems, 31*(12), e3700.

Kaveh, A., & Khayatazad, M. (2012). A new meta-heuristic method: ray optimization. *Computers & Structures, 112*, 283-294.

Kaveh, A., & Talatahari, S. (2010). A novel heuristic optimization method: charged system search. *Acta Mechanica, 213*(3-4), 267-289.

Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. *Journal of statistical physics, 34*(5-6), 975-986.

Kobayashi, N., Inui, K., & Matsumoto, Y. (2007). *Extracting aspect-evaluation and aspect-of relations in opinion mining.* Paper presented at the Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

Kumar, A., Bhalla, V., Kumar, P., Bhardwaj, T., & Jangir, N. (2018). Whale Optimization Algorithm for Constrained Economic Load Dispatch Problems—A Cost Optimization *Ambient Communications and Computer Systems* (pp. 353-366): Springer.

Kumar, V. R., & Raghuveer, K. (2013). Dependency driven semantic approach to product features extraction and summarization using customer reviews *Advances in computing and information technology* (pp. 225-238): Springer.

Law, D., Gruss, R., & Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications, 67*, 84-94.

Li, Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S., & Yu, H. (2010). *Structure-aware review mining and summarization.* Paper presented at the Proceedings of the 23rd international conference on computational linguistics.

Li, Qin, Z., Xu, W., & Guo, J. (2015). A holistic model of mining product aspects and associated sentiments from online reviews. *Multimedia Tools and Applications, 74*(23), 10177-10194.

Li, Zhang, N., Lai, X., Zhou, J., & Xu, Y. (2017). Design of a fractional-order PID controller for a pumped storage unit using a gravitational search algorithm based on the Cauchy and Gaussian mutation. *Information Sciences, 396*, 162-181.

Li, S., Fang, H., & Liu, X. (2018). Parameter optimization of support vector regression based on sine cosine algorithm. *Expert Systems with Applications, 91*, 63-77.

Li, S., Wang, R., & Zhou, G. (2012). *Opinion target extraction using a shallow semantic parsing framework.* Paper presented at the Twenty-sixth AAAI conference on artificial intelligence.

Li, S., Zhou, L., & Li, Y. (2015). Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures. *Information Processing & Management, 51*(1), 58-67.

Li, Z., Zhang, M., Ma, S., Zhou, B., & Sun, Y. (2009). *Automatic extraction for product feature words from comments on the web.* Paper presented at the Asia Information Retrieval Symposium.

Liu, Gao, Z., Liu, B., & Zhang, Y. (2015). *Automated Rule Selection for Aspect Extraction in Opinion Mining.* Paper presented at the IJCAI.

Liu, Gao, Z., Liu, B., & Zhang, Y. (2016). Automated rule selection for opinion target extraction. *Knowledge-Based Systems, 104*, 74-88.

Liu, Liu, B., Zhang, Y., Kim, D. S., & Gao, Z. (2016). *Improving Opinion Aspect Extraction Using Semantic Similarity and Aspect Associations.* Paper presented at the AAAI.

Liu, Xu, L., & Zhao, J. (2012). *Opinion target extraction using word-based translation model.* Paper presented at the Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.

Liu, Xu, L., & Zhao, J. (2015). Co-extracting opinion targets and opinion words from online reviews based on the word alignment model. *IEEE Transactions on Knowledge and Data Engineering, 27*(3), 636-650.

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

Liu, B., Mobasher, B., & Nasraoui, O. (2011). Web Usage Mining Web Data Mining: Springer, Berlin.

Liu, K., Xu, H. L., Liu, Y., & Zhao, J. (2013). *Opinion Target Extraction Using Partially-Supervised Word Alignment Model.* Paper presented at the IJCAI.

Liu, K., Xu, L., & Zhao, J. (2013). *Syntactic patterns versus word alignment: Extracting opinion targets from online reviews.* Paper presented at the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Liu, Y., Bi, J.-W., & Fan, Z.-P. (2017). Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. *Information Fusion, 36*, 149-161.

Luo, J., Chen, H., Xu, Y., Huang, H., & Zhao, X. (2018). An improved grasshopper optimization algorithm with application to financial stress prediction. *Applied Mathematical Modelling, 64*, 654-668.

Ma, B., Zhang, D., Yan, Z., & Kim, T. (2013). An LDA and synonym lexicon based approach to product feature extraction from online consumer product reviews. *Journal of Electronic Commerce Research, 14*(4), 304.

Mafarja, M., & Mirjalili, S. (2018). Whale optimization approaches for wrapper feature selection. *Applied Soft Computing, 62*, 441-453.

Maharani, W., Widyantoro, D. H., & Khodra, M. L. (2015). Aspect extraction in customer reviews using syntactic pattern. *Procedia Computer Science, 59*, 244-253.

Marcacini, R. M., Rossi, R. G., Matsuno, I. P., & Rezende, S. O. (2018). Cross-domain aspect extraction for sentiment analysis: A transductive learning approach. *Decision Support Systems, 114*, 70-80.

Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications, 41*(17), 7764-7775.

Maslowska, E., Malthouse, E. C., & Viswanathan, V. (2017). Do customer reviews drive purchase decisions? The moderating roles of review exposure and price. *Decision Support Systems, 98*, 1-9.

Miao, Y., Zhao, M., Makis, V., & Lin, J. (2019). Optimal swarm decomposition with whale optimization algorithm for weak feature extraction from multicomponent modulation signal. *Mechanical Systems and Signal Processing, 122*, 673-691.

Mirjalili, S. (2015). Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. *Knowledge-Based Systems, 89*, 228-249.

Mirjalili, S. (2016). SCA: a sine cosine algorithm for solving optimization problems. *Knowledge-Based Systems, 96*, 120-133.

Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., & Mirjalili, S. M. (2017). Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. *Advances in Engineering Software, 114*, 163-191.

Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in Engineering Software, 95*, 51-67.

Mirjalili, S., Mirjalili, S. M., & Hatamlou, A. (2016). Multi-verse optimizer: a nature-inspired algorithm for global optimization. *Neural Computing and Applications, 27*(2), 495-513.

Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software, 69*, 46-61.

Moghaddam, S. (2015). *Beyond sentiment analysis: mining defects and improvements from customer feedback.* Paper presented at the European Conference on Information Retrieval.

Moghaddam, S., & Ester, M. (2010). *Opinion digger: an unsupervised opinion miner from unstructured product reviews.* Paper presented at the Proceedings of the 19th ACM international conference on Information and knowledge management.

Moghaddam, S., & Ester, M. (2011). *ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews.* Paper presented at the Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.

Moodi, Y., Mousavi, S. R., Ghavidel, A., Sohrabi, M. R., & Rashki, M. (2018). Using Response Surface Methodology and providing a modified model using whale algorithm for estimating the compressive strength of columns confined with FRP sheets. *Construction and Building Materials, 183*, 163-170.

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

Mostafa, A., Hassanien, A. E., Houseni, M., & Hefny, H. (2017). Liver segmentation in MRI images based on whale optimization algorithm. *Multimedia Tools and Applications, 76*(23), 24931-24954.

Mukherjee, A., Chakraborty, N., & Das, B. K. (2017). *Whale optimization algorithm: An implementation to design low-pass FIR filter.* Paper presented at the 2017 Innovations in Power and Advanced Computing Technologies (i-PACT).

Musto, C., de Gemmis, M., Semeraro, G., & Lops, P. (2017). *A Multi-criteria Recommender System Exploiting Aspect-based Sentiment Analysis of Users' Reviews.* Paper presented at the Proceedings of the Eleventh ACM Conference on Recommender Systems.

Nasiri, J., & Khiyabani, F. M. (2018). A whale optimization algorithm (WOA) approach for clustering. *Cogent Mathematics & Statistics, 5*(1), 1483565.

Nazari-Heris, M., Mehdinejad, M., Mohammadi-Ivatloo, B., & Babamalek-Gharehpetian, G. (2017). Combined heat and power economic dispatch problem solution by implementation of whale optimization method. *Neural Computing and Applications*, 1-16.

Nazari-Heris, M., Mehdinejad, M., Mohammadi-Ivatloo, B., & Babamalek-Gharehpetian, G. (2019). Combined heat and power economic dispatch problem solution by implementation of whale optimization method. *Neural Computing and Applications, 31*(2), 421-436.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603-9611.

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

Osama, S., Darwish, A., Houssein, E. H., Hassanien, A. E., Fahmy, A. A., & Mahrous, A. (2017). *Long-term wind speed prediction based on optimized support vector regression.* Paper presented at the 2017 eighth international conference on intelligent computing and information systems (ICICIS).

Pan, W.-T. (2012). A new fruit fly optimization algorithm: taking the financial distress model as an example. *Knowledge-Based Systems, 26*, 69-74.

Pappula, L., & Ghosh, D. (2017). Synthesis of linear aperiodic array using Cauchy mutated cat swarm optimization. *AEU-International Journal of Electronics and Communications, 72*, 52-64.

Parambanchary, D., & Rao, V. M. (2018). WOA-NN: a decision algorithm for vertical handover in heterogeneous networks. *Wireless Networks*, 1-16.

Popescu, A.-M., & Etzioni, O. (2007). Extracting product features and opinions from reviews *Natural language processing and text mining* (pp. 9-28): Springer.

Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems, 108*, 42-49.

Poria, S., Cambria, E., Ku, L.-W., Gui, C., & Gelbukh, A. (2014). *A rule-based approach to aspect extraction from product reviews.* Paper presented at the Proceedings of the second workshop on natural language processing for social media (SocialNLP).

Prakash, D., & Lakshminarayana, C. (2017). Optimal siting of capacitors in radial distribution network using whale optimization algorithm. *Alexandria Engineering Journal, 56*(4), 499-509.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2009). *Expanding domain sentiment lexicon through double propagation.* Paper presented at the IJCAI.

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

Rana, T. A., & Cheah, Y.-N. (2016). Aspect extraction in sentiment analysis: comparative analysis and survey. *Artificial Intelligence Review, 46*(4), 459-483.

Rana, T. A., & Cheah, Y.-N. (2017). A two-fold rule-based model for aspect extraction. *Expert Systems with Applications, 89*, 273-285.

Rana, T. A., & Cheah, Y.-N. (2018a). Sequential patterns-based rules for aspect-based sentiment analysis. *Advanced Science Letters, 24*(2), 1370-1374.

Rana, T. A., & Cheah, Y.-N. (2018b). Sequential patterns rule-based approach for opinion target extraction from customer reviews. *Journal of Information Science*, 0165551518808195.

Rashedi, E. (2009). Nezamabadi pour, H., Saryazdi, S.:‘. *GSA: a gravitational search algorithm’, Inf. Sci, 179*, 13.

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

Saha, A., & Saikia, L. C. (2018). Performance analysis of combination of ultra-capacitor and superconducting magnetic energy storage in a thermal-gas AGC system with utilization of whale optimization algorithm optimized cascade controller. *Journal of Renewable and Sustainable Energy, 10*(1), 014103.

Sahu, P. R., Hota, P. K., & Panda, S. (2018). Power system stability enhancement by fractional order multi input SSSC based controller employing whale optimization algorithm. *Journal of Electrical Systems and Information Technology, 5*(3), 326-336.

Sai, L., & Huajing, F. (2017). *A WOA-based algorithm for parameter optimization of support vector regression and its application to condition prognostics.* Paper presented at the 2017 36th Chinese Control Conference (CCC).

Samha, A. K. (2016). Aspect-based opinion mining using dependency relations. *Int. J. Comput. Sci. Trends Technol.(IJCST), 4*.

Samha, A. K., Li, Y., & Zhang, J. (2014). Aspect-based opinion extraction from customer reviews. *arXiv preprint arXiv:1404.1982*.

Saqia, B., Khan, K., Khan, A., Khan, W., Subhan, F., & Abid, M. (2018). Impact of Anaphora Resolution on Opinion Target Identification. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 9*(6), 230-236.

Sayed, G. I., Darwish, A., & Hassanien, A. E. (2018). A new chaotic whale optimization algorithm for features selection. *Journal of Classification, 35*(2), 300-344.

Sayed, G. I., Darwish, A., Hassanien, A. E., & Pan, J.-S. (2016). *Breast cancer diagnosis approach based on meta-heuristic optimization algorithm inspired by the bubble-net hunting strategy of whales.* Paper presented at the International Conference on Genetic and Evolutionary Computing.

Sharawi, M., Zawbaa, H. M., & Emary, E. (2017). *Feature selection approach based on whale optimization algorithm.* Paper presented at the 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI).

Simon, D. (2008). Biogeography-based optimization. *IEEE transactions on evolutionary computation, 12*(6), 702-713.

Singh, N., & Hachimi, H. (2018). A new hybrid whale optimizer algorithm with mean strategy of grey wolf optimizer for global optimization. *Mathematical and Computational Applications, 23*(1), 14.

Singh, N., & Singh, S. (2017). A novel hybrid GWO-SCA approach for optimization problems. *Engineering Science and Technology, an International Journal, 20*(6), 1586-1601.

Sreenu, K., & Sreelatha, M. (2017). W-Scheduler: whale optimization for task scheduling in cloud computing. *Cluster Computing*, 1-12.

Storn, R., & Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization, 11*(4), 341-359.

Tharwat, A., Moemen, Y. S., & Hassanien, A. E. (2017). Classification of toxicity effects of biotransformed hepatic drugs using whale optimized support vector machines. *Journal of biomedical informatics, 68*, 132-149.

Tripathy, A., Anand, A., & Rath, S. K. (2017). Document-level sentiment classification using hybrid machine learning approach. *Knowledge and Information Systems, 53*(3), 805-831.

Tubishat, M., Abushariah, M. A., Idris, N., & Aljarah, I. (2018). Improved whale optimization algorithm for feature selection in Arabic sentiment analysis. *Applied Intelligence*, 1-20.

Tubishat, M., & Idris, N. (2017). *Explicit And Implicit Aspect Extraction Using Whale Optimization Algorithm And Hybrid Approach*. Paper presented at the International Conference on Industrial Enterprise and System Engineering (ICOIESE) Bandung, Indonesia.

Tubishat, M., Idris, N., & Abushariah, M. A. (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges. *Information Processing & Management, 54*(4), 545-563.

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.

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

Vo, A.-D., Nguyen, Q.-P., & Ock, C.-Y. (2018). Opinion–Aspect Relations in Cognizing Customer Feelings via Reviews. *IEEE Access, 6*, 5415-5426.

Wang, Deb, S., Gandomi, A. H., & Alavi, A. H. (2016). Opposition-based krill herd algorithm with Cauchy mutation and position clamping. *Neurocomputing, 177*, 147-157.

Wang, & Wang, H. (2008). *Bootstrapping both product features and opinion words from chinese customer reviews with cross-inducing.* Paper presented at the Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I.

Wang, Wang, W., Sun, H., & Rahnamayan, S. (2016). Firefly algorithm with random attraction. *International Journal of Bio-Inspired Computation, 8*(1), 33-41.

Wei, C.-P., Chen, Y.-M., Yang, C.-S., & Yang, C. C. (2010). Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and E-Business Management, 8*(2), 149-167.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation, 1*(1), 67-82.

Wu, & Law, R. (2011). Cauchy mutation based on objective variable of Gaussian particle swarm optimization for parameters selection of SVM. *Expert Systems with Applications, 38*(6), 6405-6411.

Wu, Wu, F., Wu, S., Yuan, Z., & Huang, Y. (2018). A hybrid unsupervised method for aspect term and opinion target extraction. *Knowledge-Based Systems, 148*, 66-73.

Wu, Zhang, Q., Huang, X., & Wu, L. (2009). *Phrase dependency parsing for opinion mining.* Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3.

Wu, F., Zhang, J., Yuan, Z., Wu, S., Huang, Y., & Yan, J. (2017). *Sentence-level Sentiment Classification with Weak Supervision.* Paper presented at the Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2018). Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. *arXiv preprint arXiv:1805.04601*.

Xu, L., Liu, K., Lai, S., Chen, Y., & Zhao, J. (2013). *Mining opinion words and opinion targets in a two-stage framework.* Paper presented at the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Xue, W., Li, T., & Rishe, N. (2017). Aspect identification and ratings inference for hotel reviews. *World Wide Web, 20*(1), 23-37.

Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR), 50*(2), 25.

Yan, Z., Xing, M., Zhang, D., & Ma, B. (2015). EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management, 52*(7), 850-858.

Yang, X.-S. (2009). *Firefly algorithms for multimodal optimization.* Paper presented at the International symposium on stochastic algorithms.

Yang, X.-S. (2010). A new metaheuristic bat-inspired algorithm *Nature inspired cooperative strategies for optimization (NICSO 2010)* (pp. 65-74): Springer.

Yu, J., Zha, Z.-J., Wang, M., & Chua, T.-S. (2011). *Aspect ranking: identifying important product aspects from online consumer reviews.* Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.

Yuan, P., Guo, C., Ding, J., & Qu, Y. (2017). *Synthesis of nonuniform sparse linear array antenna using whale optimization algorithm.* Paper presented at the 2017 Sixth Asia-Pacific Conference on Antennas and Propagation (APCAP).

Zamani, H., & Nadimi-Shahraki, M.-H. (2016). Feature selection based on whale optimization algorithm for diseases diagnosis. *International Journal of Computer Science and Information Security, 14*(9), 1243.

Zhang, Li, H., & Yu, W. (2017). *Fine-grained opinion mining: An application of online review analysis in the express industry.* Paper presented at the Computer and Communications (ICCC), 2017 3rd IEEE International Conference on.

Zhang, Liu, B., Lim, S. H., & O'Brien-Strain, E. (2010). *Extracting and ranking product features in opinion documents.* Paper presented at the Proceedings of the 23rd international conference on computational linguistics: Posters.

Zhang, Xu, H., & Wan, W. (2012). Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications, 39*(11), 10283-10291.

Zhang, Q., Chen, H., Heidari, A. A., Zhao, X., Xu, Y., Wang, P., . . . Li, C. (2019). Chaos-Induced and Mutation-Driven Schemes Boosting Salp Chains-Inspired Optimizers. *IEEE Access, 7*, 31243-31261.

Zhang, Y., Miao, D., Wang, J., & Zhang, Z. (2018). A cost-sensitive three-way combination technique for ensemble learning in sentiment classification. *International Journal of Approximate Reasoning*.

Zhao, H., Guo, S., & Zhao, H. (2017). Energy-related CO2 emissions forecasting using an improved LSSVM model optimized by whale optimization algorithm. *Energies, 10*(7), 874.

Zhao, K., Stylianou, A. C., & Zheng, Y. (2018). Sources and impacts of social influence from online anonymous user reviews. *Information & Management, 55*(1), 16-30.

Zhou, Q., & Zhang, C. (2017). Detecting dietary preference of social media users in China via sentiment analysis. *Proceedings of the Association for Information Science and Technology, 54*(1), 523-527.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing, 74*(2), 133-148.

Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). *Movie review mining and summarization.* Paper presented at the Proceedings of the 15th ACM international conference on Information and knowledge management.

Zou, Y., Liu, P. X., Yang, C., Li, C., & Cheng, Q. (2017). Collision detection for virtual environment using particle swarm optimization with adaptive cauchy mutation. *Cluster Computing, 20*(2), 1765-1774.

Zucco, C., Calabrese, B., & Cannataro, M. (2017). *Sentiment analysis and affective computing for depression monitoring.* Paper presented at the Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

1. Tubishat, M., Idris, N., & Abushariah, M. A. (2018). Implicit aspect extraction in sentiment analysis: Review, taxonomy, opportunities, and open challenges. *Information Processing & Management*, *54*(4), 545-563. (**Published**)

2. Tubishat, M., Abushariah, M. A., Idris, N., & Aljarah, I. (2018). Improved whale optimization algorithm for feature selection in Arabic sentiment analysis. *Applied Intelligence*, 1-20. (**Published**)

3. Tubishat, M., & Idris, N. (2017). Explicit And Implicit Aspect Extraction Using Whale Optimization Algorithm And Hybrid Approach. *International Conference on Industrial Enterprise and System Engineering (ICOIESE) 2017, November 2017.* (**Best Paper in the Conference**)