Sudan University of Science and Technology
College of Graduate Studies

Declaration

I, the signing here-under, declare that I'm the sole author of the Ph.D. thesis entitled Constructing an Arabic Opinion Mining Model with Special Reference to telecommunication Companies and Hotel Reviews

which is an original intellectual work. Willingly, I assign the copy-right of this work to the College of Graduate Studies (CGS), Sudan University of Science & Technology (SUST). Accordingly, SUST has all the rights to publish this work for scientific purposes.

Candidate's name: Limia Hassan Rahmatalla

Candidate's signature: .... Date: 30-4-2015

إقرار

أنا الموقع أدناه أقر بأننى المؤلف الوحيد لرسالة الدكتوراه المعنونة ....
بناء نموذج لتحليل الآراء العربيه : تطبيق على آراء
شركات الاتصالات والفنادق

وهى منتج فكري أصيل . وباختياري أعطى حقوق طبع ونشر هذا العمل لكلية الدراسات العليا جامعه السودان للعلوم والتكنولوجيا ،عليه يحق للجامعه نشر هذا العمل للأغراض العلمية .

اسم الدارس : لمياء حسن رحمة الله

توقيع الدارس : .... التاريخ : 30-4-2015

بسم الله الرحمن الرحيم

Sudan University of Science & Technology

جامعة السودان للعلوم والتكنولوجيا

College of Graduate Studies

كلية الدراسات العليا

SUST

كلية الدراسات العليا

Approval Page

Name of Candidate: Limia Hassan Rahamtalla Mohammed

Thesis title: Constructing an Arabic Opinion Mining Models: With Special Reference to Telecommunication Companies and Hotels Reviews

بناء نموذج لتحليل الاراء العربية: تطبيق على
آراء عملاء شركات الاتصالات والفنادق

Approved by:

1. External Examiner

Name: AbuBaker Elsadig

Signature: Abaker .Date: 24/12/2014

2. Internal Examiner

Name: Mohamed Elhafiz

Signature: .Date: 24/12/2014

3. Supervisor

Name: Altayeb Abudyamen

Signature: .Date: 24/12/2014

Sudan University of Science & Technology

College of Graduate Studies

Constructing an Arabic Opinion Mining Model: With Special Reference to Telecommunication Companies and Hotel Reviews

بناء نموذج لتحليل الآراء العربية :تطبيق على اراء شركات الاتصالات والفنادق

Submitted for the Degree of Doctor of Philosophy in Computer Science

By

Limia Hassan Rahamatalla

B.Sc. in computer science, University of Khartoum

M.Sc. in computer science, University of Khartoum

Supervisor: Prof .Eltayeb Salih Aubelyaman

December 2014

Dedication

To my lovely Parents, Husband, Kids, Brothers and Sisters for their care, continuous support and love.

Acknowledgement

I wish to express my sincere thanks to my supervisor Dr. Prof .Eltayeb Salih Aubelyaman for valuable encouragement, and his friendly attitude and much valued unlimited help.

Thanks are extended to the generous staff of SUST, particularly Dr. Mohammed El hafiz Mustafa , Ph.D. Program Coordinator .I would like to express my sincere gratitude to Professor Izzeldin Mohammed Osman, for his support and unlimited help

I am indebted, in various ways to the Drs at the Department of Arabic language, particularly Dr. Asia Wadatalla and Dr. Osman Ibrahim for their support.

My special thanks are due to Mujtaba Imad for crucial help with programming skills.

I thank my parents, my husband and my brothers ,sister for giving me endless encouragement and support. My kids, have been rooting for me like nobody else, despite the intrusion that this work has been. I thank them for their tolerance, and their smiles.

I am grateful to many friends and colleagues who have listened to my questions, often provided answers and have been patient with my idiosyncrasies.

Last but not least special acknowledgement and great appreciation are due to all those who helped me in one way or another.

Abstract

Due to the recent significant growth of e-commerce applications, most of the widely used products are marketed online. This triggered online assessment of products. As such, the success or failure of companies is partially measured by their ability to take assessments of their products seriously. Analysis of these assessments is necessary for ensuring continuous customer satisfaction and further improvements of current and future products. Naturally, understanding the preferences of customers is crucial for product manufacturer as it helps them in product development, marketing and consumer relationship management. On the other hand, customers use of reviews by other's online assessments influence their decision as to whether or not they purchase a product. Expectedly, assessment are given in unstructured texts of a natural language. Thus, their processing requires appropriate knowledge in different domains that include, but are not limited to: database, information retrieval, information extraction, machine learning, and natural language processing. However, it becomes difficult for product manufacturers or dealers to keep track of large number of assessments, hence forth will be called opinions and/or sentiments. In the past few years, researchers looked at different ways of taking further advantage of opinions in what is now known as opinion mining or sentiment analysis. The scope of opinion and sentiment includes characteristic, functionality and features of product. This thesis is about novel methods that addresses challenges of opinion mining of Arabic texts. To that end, a set of Arabic language corpora from hotel and telecommunication companies has been collected. The set was developed for evaluating the proposed sentiment analysis methodologies. As well, Arabic Sentiment Classifier (ASC) has been implemented at the document-level. This research focuses on improvement of the effectiveness of feature selection using Information Gain . It then proposes a generic framework on for feature-based level analysis. The Arabic Sentiment Analyzer (ASA) framework consists of two main modules: a language resource construction and an opinion miner. For the language resource construction module, the first phase proposes constructing an opinion lexicon for Arabic opinion word. It is based on a bootstrapping process over an online dictionary. A few seed sentiment words have been used for bootstrapping based on the synonym and antonym structures of the dictionary. This method is simple and efficient as it

gives reasonable results. During the second phase, features of objects are extracted based on frequent nouns, noun phrases, association rule mining and Natural Language Processing (NLP) techniques. This phase takes advantage of syntactic patterns to improve the accuracy of frequency- based techniques. Product features are stored in feature sets.

After a language resource is constructed, the opinion mining module uses a novel information summarizing and visualization approach. The approach is based on NLP techniques for defining sentiment sentences, identifying orientations of features and summarizing results. The visualization module is aimed at providing users an effective way of browsing the set of feature according to the polarity expressed by each assessments. In piratical results reflect efficiency of the proposed system.

المستخلص

هنالك العديد من الآراء المدونة في الإنترنت في مختلف المواقع مثل مواقع التواصل الاجتماعي و المدونات الشخصية ومواقع التقييم. معلوم أن هناك جهات سياسية وتجارية وبحثية و أمنية تحتاج لمعرفة هذه الآراء أيما حاجة، استجابة لهذه الحاجة ظهر علم تنقيب الآراء (Opinion Mining) أو تحليل الميول (Sentiment Analysis)،والذي يعتبر من روافد علم تنقيب البيانات. الهدف الأساسي من تنقيب الآراء هو عمل نظام حاسوبي قادر على التعرف على الآراء والمشاعر العامة الممثلة في النصوص الإلكترونية .ويمكن للرأي ان يكون مباشر على موضوع معين مثل رأي شخص ما في هاتفه المحمول   أو رأى مقارن بين موضوعين مثل مقارنة خصائص هاتفه المحمول مع خصائص الهاتف المحمول لصديقه. كما أنه يمكن أن يكون مفيداً في التسويق حيث أنه يساعد في الحكم على نجاح حملة إعلانية أو إطلاق منتج جديد، وتحديد إصدارات من المنتج أو تحديد الخدمة التي تحظى بقبول اكثر.

تهدف هذه الأطروحة لتطوير نماذج تصنيف الآراء ذات الكفاءة العالية والملائمة للغة العربية باستخدام تقنيات التنقيب عن البيانات. وعليه ،تم إنشاء مجموعتين من مستندات الآراء من موقع لحجز الفنادق البيانات و شركات الاتصالات السودانية .كما تم انشاء موديل لتصنيف اراء على مستوى مستند الراي ككل(Arabic Sentiment classifier) باستخدام ثلاثة تقنيات تنقيب بيانات للتصنيف وهي Support Vector Machine وNaive Bayes و K-Nearest Neighbor . ثم استخدمت تقنية   لاختيار الخصائص وهي   Information gain وذلك لتحسين مودل التصنيف . هذا بالإضافة تطبيق أثنين من طرق التحقيق وهي التحقيق المنقسم (split validation) والتحقيق المتبادل   (10-cross-validation)   لتقييم هذه النماذج المقترحة.

ثم بعد ذلك فحص مستند الراي على مستوى الخواص وتم تصميم اطار عام (Arabic Sentiment Analyzer) من وحدتين رئيسيتين:وحدة بناء الموارد اللغوية و مودل(نموذج) التعدين الرأي.
في وحدة بناء الموارد اللغوية، تم بناء معجم كلمات الرأي العربية (Arabic Opinion Lexicon) باستخدام مجموعة من الكلمات(صفات) المستخلصة من مستندات الراي و التي قسمت لمجموعتين كلمات موجبة وكلمات سالبة ثم استخدمت المرادفات لتوسيع المجموعتين من القاموس وقيم المعجم بواسطة خبراء في اللغة. اما في المرحلة الثانية، تم استخراج خواص المنتجات على أساس الأسماء المتكررة، والعبارات الاسمية، باستخدام تنقيب البيانات وتقنيات معالجة اللغات الطبيعية من خلال الأنماط النحوية لتحسين دقة الاستخلاص , وتم تخزينها في مجموعة الخواص . نتج من الجهد معجم لكلمات الراي ومجموعتان لكلمات الخواص المستخلصة من مستندات الراي الخاص بالفنادق والخاصة بشركات الاتصالات وقد استخدمت المجموعتين في بناء نموذج تعدين الراي الذي يستند على تقنيات معالجة اللغات الطبيعية لتحديد جمل الرأي و اتجاهاتها الموجب منها و السالب بناءً على معجم كلمات الراي ومن ثم تم تلخيص النتائج على مستوى مجموعة الخواص في صورة فعالة للمستخدمين. وفقا لهذه التجارب تبين أن النهج المقترح هو أكثر فعالية للغة العربية.

List of Publications

- A paper entitled “Constructing Opinion Mining model of Sudanese Telecom Products.” Submitted for publication.
- A paper entitled “تعدين الآراء التحديات والتطبيقات فى اللغة العربية” accepted for publication.
- Eltayeb Abuelyaman, Limia Rahmatullah, Wafaa Mukhtar, Muna Al-Ajabani, paper entitled “Machine Translation of Arabic Language: Challenges and Keys” published in 2014 5th International Conference on Intelligent Systems, Modelling and Simulation (ISMS2014). Langkawi, Malaysia
- A paper under preparation entitled “Creation of Arabic Opinion Lexicon”
- A paper under preparation entitled “Arabic sentiment classifier ”.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AOL	Arabic Opinion Lexicon
ASA	Arabic Sentiment Analyzer
ASC	Arabic Sentiment Classifier
FP-Growth	Frequent Pattern Tree Algorithm
Fpos	False Positive
FSM	Feature Segmentation Model
IG	Information Gain
IR	Information Retrieval
KNN	K-Nearest Neighbor
ML	Machine Learning
NB	Naive Bayes Classifier
NLP	Natural Language Processing
n-seed	Negative Seed
OM	Opinion Mining
PMI	Point Wise Mutual Information
POS	Parts-Of-Speech Tagging
p-seed	Positive Seed
SA	Sentiment Analysis
SMS	Short-Messaging-System
SO	Semantic Orientation
SVM	Support Vector Machine

TF-IDF	**Term Frequency–Inverse Document Frequency**
Tneg	**True Negative**
Tpos	**True Positive**

CHAPTER ONE

1 Introduction

1.1 Overview

Due to the huge growth of e-commerce applications, products are mostly bought and advertised online. As such, there are large amounts of opinion texts on the Internet regarding products and/or movie reviews. Opinions are central to almost all human activities and have a relevant impact on the human life. They convey how reality is perceived by people. Opinions are used to express points of views, beliefs as well as perceptions of reality and choices people make. Opinions are, to a considerable degree, conditioned upon how others see and evaluate the world. The success or failure of any company nowadays is measured by its ability to evaluate its customer reviews. Analysis of these reviews should be done in order to enhance customer satisfaction and help the design and planning of future products [1]. On the one hand, understanding preferences of customers is crucial from the product manufacturers' perspectives as they help in the improvement of product development, marketing and consumer relationship management. On the other hand, customers use reviews by others to support their decisions on whether to purchase products or not. Analysis and evaluation of opinions for the stated reasons is a newly emerging field the formal name for it is *opinion mining* and *sentiment analysis*.

Opinion mining can be defined as a sub-discipline of computational linguistics and information retrieval that concentrations on extracting people’s opinions, evaluations, attitudes, and expressed-emotions from huge data on the web. The recent expansion of the web encourages users to contribute and express their opinions using blogs, videos, social networking websites, and so on. These platforms provide a large amounts of valuable information that are worthy of analyzing.

Opinions are expressed on anything such as a product, a topic, an individual, etc. In opinion mining tasks, the orientation of an opinion on an object is identified by a set of components or attributes. Opinion Mining and Sentiment Analysis identify the new field of research devoted to designing and evaluating tools for automatic opinion analysis. It started in 2001[2], with contributions from researchers in the domains of machine learning, computational linguistic and information retrieval. Most of the research efforts are aimed at investigating Sentiment Analysis and Opinion Mining for English language text. Only a small portion of the available works deals with Sentiment Analysis for other languages[3].

1.2 Opinion Mining Terminologies

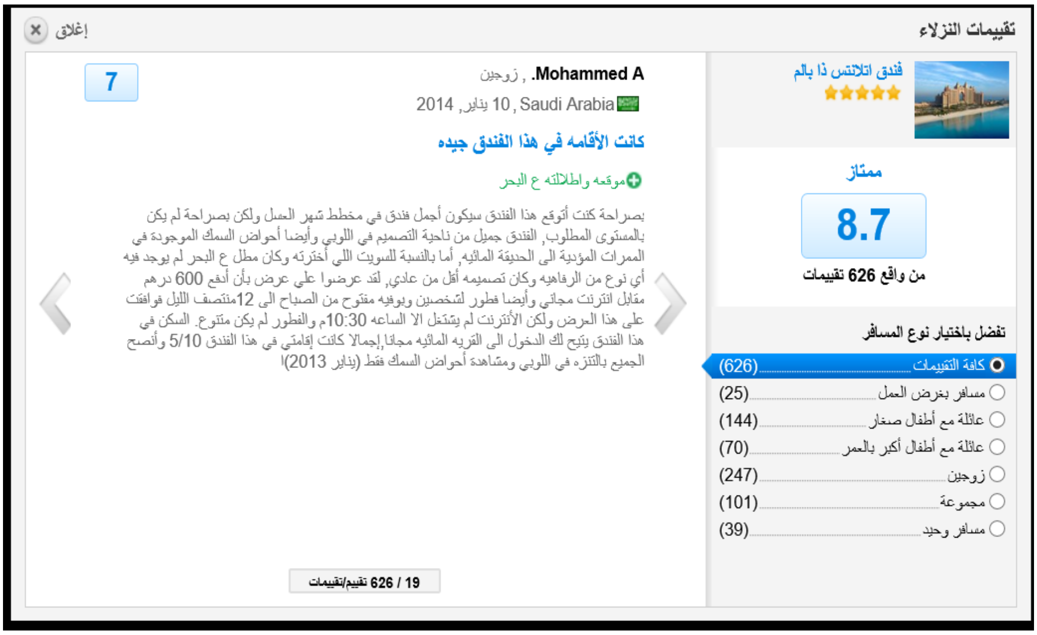
In this section, the basic terminology of opinion mining are reviewed.

- **Fact**: A fact is something that has really occurred or is actually the case.
- **Opinion**: An opinion is a belief about matters commonly considered to be subjective, and is the result of emotion or interpretation of facts.
- **Subjective/Opinionated Text**: A text is subjective or opinionated if it expresses personal feelings or beliefs, e.g. opinions.

- **Objective Text**: An objective text expresses some factual information about the world.
- **object**: An object is an entity which can be product, service, person, event, organization[4]. An object can be represented as a hierarchy of components, sub-components, etc. Each component has its own set of sub-components and attributes. In this hierarchy or tree, the root is the object itself. Each non-root node is a component or subcomponent of the object. Each link is a part-of relationship. Each node is associated with a set of attributes.
- **Review**: A review is a subjective text containing a sequence of words describing opinions of reviewer regarding a specific item. Review text may contain complete sentences, short comments, or both as in Figure(1.1).
- **Feature/Aspect:** An aspect (also called feature) is an attribute or component of the item that has been reviewed. If an aspect appears in a review, it is called explicit aspect; otherwise it is called implicit[5]. Current works mainly focus on extracting explicit aspects and only a few simple methods are proposed for identifying implicit aspects.
- **Explicit Feature**: Feature that are explicitly mentioned as nouns or noun phrases in a sentence, e.g., ‘picture quality’ in the sentence “The picture quality of this phone is great”.
- **Implicit Feature**: Feature that are not explicitly mentioned in a sentence but are implied, e.g., ‘price’ in the sentence “This car is so expensive.”, or ‘size’ in the sentence “This phone will not easily fit in a pocket”.
- **Sentiment**: Sentiment is a linguistic term which refers to the direction in which a concept or opinion is interpreted[5] .

Sentiment is used in a more specific sense as an opinion about an aspect. For example, ‘great’ is a sentiment for the aspect ‘picture quality’ in the sentence “It has great picture quality”.

- **Opinion Phrase**: An opinion phrase < h, m > is a pair of a head term **h** and a modifier **m [6]**. Usually, a head term is a candidate or a feature aspect, and a modifier is a sentiment that expresses some opinion towards the aspect, e.g. < الشبكة , قوية >.
- **opinion orientation or semantic orientation of an opinion:** The *semantic orientation* of an opinion on a feature ***f*** states whether the opinion is positive, negative or neutral.
- **Polarity**: Polarity is a two-level orientation scale. In this scale, a sentiment is either positive or negative.



تقييمات النزلاء

إغلاق ×

فندق اتلانتس ذا بالم

ممتاز

8.7

من واقع 626 تقييمات

تفضل باختيار نوع المسافر

● كافة التقييمات (626)
○ مسافر بغرض العمل (25)
○ عائلة مع أطفال صغار (144)
○ عائلة مع أطفال أكبر بالعمر (70)
○ زوجين (247)
○ مجموعة (101)
○ مسافر وحيد (39)

Mohammed A. , زوجين

Saudi Arabia, 10 يناير, 2014

7

كانت الأقامه في هذا الفندق جيده

موقعه واطلالته ع البحر

بصراحة كنت أتوقع هذا الفندق سيكون أجمل فندق في مخطط شهر العسل ولكن بصراحة لم يكن بالمستوى المطلوب, الفندق جميل من ناحية التصميم في اللوبي وأيضا أحواض السمك الموجودة في الممرات المؤدية الى الحديقة المائيه, أما بالنسبة للسويت اللي أخترته وكان مطل ع البحر لم يوجد فيه أي نوع من الرفاهيه وكان تصميمه أقل من عادي, لقد عرضوا علي عرض بأن أدفع 600 درهم مقابل انترنت مجاني وأيضا فطور لشخصين وبوفيه مفتوح من الصباح الى 12منتصف الليل فوافقت على هذا العرض ولكن الأنترنت لم يشتغل الا الساعه 10:30م والفطور لم يكن متنوع. السكن في هذا الفندق يتيح لك الدخول الى القريه المائيه مجانا,إجمالا كانت إقامتي في هذا الفندق 5/10 وأنصح الجميع بالتنزه في اللوبي ومشاهدة أحواض السمك فقط (يناير 2013)ا

19 / 626 تقييم/تقييمات

Figure 1.1A sample review from agoda.com

1.3 Levels of Analysis

Depending on the level of interest, there are three types of opinion mining. The analysis of opinions may be document-based, where consider a whole document is handled as a single entity and summarized as positive, negative or neutral [7, 8]. Opinions can be sentence-based, where individual sentences bearing sentiments are classified. This level of analysis is closely related to *subjectivity classification which firstly classifies sentences as objective or subjective. It then classifies subjective sentences as* positive or negative. Analyses at both levels do not lead to what exactly people liked and what they did not. *Feature level* (*feature-based opinion mining and summarization*) performs finer-grained analysis [9]. At this level, analysis is focused at the opinion itself. The idea is based on an opinion consists of a *sentiment* (positive or negative) and a *target* (of opinion). Realizing the importance of opinion targets also helps in understanding sentiment analysis problem better.

Based on feature-based level of analysis, a structured summary of opinions about entities and their aspects can be produced. The summary is a turning of an unstructured text to a structured data that can be used for all kinds of qualitative and quantitative analyses.

Automated Opinion Mining systems are advantageous over the traditional polling or focus groups. That is because they are consistent over time as companies using manual scoring will realize changes in results due to personnel turnover. Additionally, these systems operate in near real time. They assimilate vast amounts of information from the Web; a feature that makes them relatively inexpensive.

1.4 Aims and significance

There are thousands of reviews of customers, so it is difficult for the company and customers to have an idea about the service from these large reviews. Thus, this work emphasizes the need of developing an opinion mining system that can analyze opinions expressed in Arabic online opinion resources .Automatic extraction of customer opinions can benefit both companies and customers

1.5 Contribution

The contributions of this dissertation to sentiment analysis of opinions that are expressed in Arabic texts are as follows:

- a set of corpora that is automatically constituted by hotel and telecommunication companies reviews written in Arabic language has been collected. Each review is a short text. Hotel reviews have an overall polarity rating indicator that is aimed at representing the expressed polarity within the review. Corpora which have been developed in order to perform evaluation of the proposed methodologies for sentiment analysis, could be used in the future by other researchers as nothing like it is available for the Arabic language.
- Domain-dependent classifier or Arabic Sentiment Classifier (ASC) has been implement. Feature selection has been investigated to improve its effectiveness.
- A generic framework aimed at defining automatic tools dedicated to feature based classification has been implemented. The Arabic

Sentiment Analyzer (ASA) framework consists of two main modules: a language resource construction and an opinion mining.

- The language resource construction module first constructs an opinion lexicon for Arabic opinion word. It obtained through a bootstrapping process using online dictionary. A few seed sentiment words have been used for bootstrapping based on the synonym and antonym structure of the dictionary. This method is simple and efficient as it gives reasonable results. Features of objects are extracted based on frequent nouns, noun phrases, the association rule mining and NLP techniques. The features are stored in feature set.
- The opinion mining module uses a novel information summarizing and visualization approach. The approach is based on NLP techniques for defining sentiment sentences, identifying orientations of features and summarizing the results. The visualization module is aimed at providing users an effective way to browse the set of feature according to the polarity expressed by each review.

1.6 Dissertation Outline

This section describes the organization of the remaining chapters as follows:

Chapter 2 is the Literature Review: in this chapter, we will present an overview of opinion mining studies. It covers the basics method of opinion lexicon creation , the types of feature based summarization method, and methods that been used in Arabic opinion mining areas.

Chapter 3, Research Methodology: this chapter presents the methodology used in this research. A methodology is generally a guideline for solving a research

problem. It contains the generic framework of the research and the steps required to carry out the research systematically.

Chapter 4, Arabic Sentiment Classifier(ASC) : this chapter provides an overall opinion on an entity, topic or event by using opinion mining classification at document level. Constructing Arabic sentiment classifier(ASC) is relying on manually annotated corpus. This chapter propose a supervised machine learning technique :naïve Bayesian, and support vector machines (SVM) and KNN classifier .

Chapter 5 Creation of opinion lexicon: the main goal of this chapter is to develop Arabic opinion mining lexicon which used in ASA later . The method is able to quickly acquire a large opinion lexicon by bootstrapping from a extracted adjective seeds. Experiment results from two domains demonstrate that the lexicon generated with our approach reach an excellent precision and could get many sentiment words in a special domain.

Chapter 6, Arabic Sentiment Analyzer: this chapter aims to introduce Arabic Sentiment Analyzer to mine and summarize opinions from customer reviews. By taking advantage of both frequency- and relation-based approaches to identify opinion sentiment. ASA first mines a set of feature from frequent noun phrases in the review texts , also uses a novel technique to group synonymous feature . In addition, it determine whether an opinion is positive or negative and generate a summary

Chapter 7, Conclusion and Future Work: this chapter provides the overall

CHAPTER TWO

2 literature Review

2.1 Introduction

As discussed in the previous chapter, there are two main tasks in the problem of aspect-based opinion mining: aspect extraction, and aspect sentiment classification. Liu, [4] classified aspect extraction techniques into four categories: finding frequent nouns and noun phrases,using opinion and target relations, using supervised learning, and using topic models. For aspect sentiment classification there are two main approaches, the supervised learning approach and the lexicon-based approach. Supervised learning is dependent on the training data, a model or classifier trained from labeled data in one domain often performs poorly in another domain. The current methods are also mainly used for document level sentiment classification as documents are long and contain more features for classification than individual sentences or clauses. Thus, supervised learning has difficulty to scale up to a large number of application domains. To avoid the difficulty of supervised method, the lexicon-based approach has been shown to perform quite well in a large number of domains. Such methods are typically unsupervised. They use a sentiment lexicon (which contains a list of sentiment words, phrases, and idioms), composite expressions. In this chapter the related work of identify opinion word extraction in section 2 and feature (or topic) extraction method in section 3 and relate work of opinion mining in Arabic text in section4 will be defined.

2.2 Identify opinion word synonyms

In the literature survey, *sentiment words* are also called *opinion words*, *polar words*, or *opinion-bearing words these words* that convey positive or negative polarities. They are critical for opinion mining. Positive sentiment words are used to express some desired states or qualities while negative sentiment words are used to express some undesired states or qualities. Examples of Arabic positive sentiment words are ,جميل ممتاز, and مدهش. Examples of Arabic negative sentiment words are سيئ,ضعيف , and زهيد. Sentiment words, can be founded in the sentence as adjective e.g. ممتاز or verb e.g. يحب or sentiment phrases and idioms. When collected, they are called *sentiment lexicon* (or *opinion lexicon*).

Sentiment words can be divided into two types: *base type* and *comparative type*. All the words exampled in the previous paragraph are of the base type. Sentiment words of the comparative type (which include the superlative type) are used to express comparative and superlative opinions. Examples of such words are احسن, افضل,جدا which are comparative and superlative forms of their base adjectives or adverbs, e.g., جيد and سيئ. Unlike sentiment words of the base type, sentiment words of the comparative type do not express a regular opinion on an entity but a comparative opinion on more than one entity, e.g., "*سعر الرسائل في الشبكة(أ) ارخص من سعر الرسائل في الشبكة(ب)*." This sentence does not express an opinion saying that any of the two networks is good or bad ,it just compares prices of short message sentence (SMS).

The key difficulty in finding opinion words is that opinions expressed by many of them are domain or context dependent. Several researchers that have studied the problem of finding opinion words have proposed many approaches to compile sentiment words. Three main approaches are: *manual approach*, *dictionary-based approach*, and *corpus-based approach*. The manual approach is labor intensive and time consuming, and is thus not usually used alone but combined with automated approaches as the final check, because automated methods make mistakes.

2.2.1 Dictionary -based Approach

Using a dictionary to compile sentiment words is an obvious approach because most dictionaries (e.g., WordNet) list synonyms and antonyms for each word. Thus, a simple technique in this approach is to use a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. The main algorithm of this technique as following.

1. A small set of sentiment words (seeds) with known positive or negative orientations is first collected manually, which is very easy.
2. The algorithm then grows this set by searching in the WordNet or another online dictionary for their synonyms and antonyms.
3. The newly found words are added to the seed list.
4. The next iteration begins.
5. The iterative process ends when no more new words can be found.

This approach was used by Hu and Liu,Valitutti et al., [9] , [10]. They used a manual inspection step to clean up the list. Kim and Hovy,[11] also used a similar method, adding a probabilistic method to clean up the resulting words (to remove errors) and to assign a sentiment strength to each word. Mohammad et al [12] came up with a new method to increase the coverage by exploiting many antonym generating affix patterns like *X* and dis *X* (e.g., honest-dishonest).

Kamps et al., [13], proposed a sophisticated approach using a WordNet distance-based method to determine the sentiment orientation of a given adjective. The distance d ($t1$, $t2$) between terms $t1$ and $t2$ is the length of the shortest path that connects $t1$ and $t2$ in WordNet. The orientation of an adjective term t is determined by its relative distance from two reference (or seed) terms *good* and *bad*, i.e., $SO(t) = (d(t, \text{bad}) - d(t, \text{good}))/d(\text{good, bad})$. t is positive if $SO(t) > 0$, and is negative otherwise. The absolute value of $SO(t)$ gives the strength of the sentiment.

Williams and Anand, 2009 [14]studied the problem of assigning sentiment strength to each word by building an adjective graph using WordNet to measure semantic distance between words seed words and the target word.

Previous work in this area was extended by using a small training data set to learn an optimal predictor of polarity strength and to reduce polarity assigned to non-polar adjectives.

In Blair-Goldensohn et al [15], Different bootstrapping method that used three different seed sets (positive, negative and neutral). A directed, weighted semantic graph used in this approach; where neighboring nodes are synonyms or antonyms of words in WordNet and are not part of the seed neutral set. The neutral set is used to stop the propagation of sentiments. Pre assigned the edge weights based on a scaling parameter for different types of edges, i.e., synonym or antonym edges.

In Zhu and Ghahramani, [16]. A modified version of the label propagation algorithm was used to assigning a sentiment value to each word. At the beginning, (+1 is given to positive seed word, -1 for negative seed, and all other words are given 0). These initial value are revised. after a number of iterations the propagation stops, the final scores after a logarithmic scaling are assigned to words as their degrees of being positive or negative.

Rao and Ravichandran,[17], tried to separate positive and negative words using three graph-based semi-supervised learning methods; given a positive seed set, a negative seed set, and a synseed sets of positive, negative, and neutral words. Then their synonyms was found in WordNet. However, expanded sets have many errors. Each word closeness to each category (positive, negative, and neutral) is compute using Bayesian formula to determine the most probable.

Hassan, [18] used WordNet synonyms and hypernyms to present a Markov random walk model for building a word relatedness graph to create a sentiment estimate for a given word. they defined, *mean hitting time* $h(i| S)$ measure, i *refer to* a node and S a set of nodes (words), which is the average number of steps that a random walker, starting in state $i \notin S$, will take to enter a state $k \in S$ for the first time. Given a set of positive seed words S^{+} and a set of negative seed words S^{-}, to estimate the sentiment orientation of a given word w, it computes the hitting times $h(w |S^{+})$ *and* $h(w| S^{-})$.If $h(w |S^{+})$ is greater than $h(w|S^{-})$, the word is classified as negative, otherwise positive.

Hassan et al [19], defined multilingual method which is finding sentiment orientations of foreign words. They build a word graph for both English words and foreign words. Using meanings in dictionaries for different languages words are connected.

Turney and Littman,[20], measured the association strength using PMI to compute the sentiment orientation of a given word. Specifically, it computes the orientation of the word from the strength of its association with a set of positive words (good, *nice, excellent, positive, fortunate,* *correct*, and *superior*), minus the strength of its association with a set of negative words (*bad*, *nasty*, *poor*, *negative*, *unfortunate*, *wrong*, and *inferior*).

Esuli and Sebastiani, [21]starting with a two sets of seed words P of positive seed words and N of negative seed words ; built a supervised learning classifier to classify words into positive and negative classes. Expanding the two seed sets using synonym and antonym relations in an online dictionary (e.g., WordNet) to build the expanded sets P' and N', which form the training set. The algorithm then uses all the glosses in the dictionary for each term in P' $^{U}N'$ to generate a feature vector. The classifier can be constructed and an updated by running process iteratively, added to the training set the newly identified positive and negative terms and their synonyms and antonyms.

Esuli and Sebastiani, [22] the objective seed set was expanded using hyponyms, in addition to synonyms and antonyms. They then did the three-class classification trying different strategies. It utilized these classifiers to construct the SentiWordNet, a lexical resource in which each synset of WordNet is associated with three numerical scores Obj(s), Pos(s), and Neg (s), describing Objective synset, Positive synset, and Negative synset

The method of Kim and Hovy, [23] also started with three seed sets of positive, negative, and neutral words. It then finds their synonyms in WordNet. The expanded sets, however, have many errors. The method then uses a Bayesian formula to compute the closeness of each word to each category (positive, negative, and neutral) to determine the most probable class for the word.

Andreevskaia and Bergler,[24] proposed a more sophisticated bootstrapping method with several techniques to expand the initial positive and negative seed sets and to clean up the expanded sets (removing non-adjectives and words in both positive and negative sets). In addition, their algorithm also performs multiple runs of the bootstrapping process using non-overlapping seed subsets.

Each run typically finds a slightly different set of sentiment words. A net overlapping score for each word is then computed based on how many times the word is discovered in the runs as a positive word and as a negative word. The score is then normalized to [0, 1] based on the fuzzy set theory.

In Kaji and Kitsuregawa, [25, 26] The dataset was collected from HTML documents based on Web page layout structures which have a column clearly indicate positive or negative orientations. Many heuristics were used to build a sentiment lexicon from this dataset. Adjective phrases are then extracted from these sentences and assigned sentiment orientations based on different statistics of their occurrences in the positive and negative sentence sets, respectively.

Velikovich et al.,[27] also proposed a method to construct a sentient lexicon using Web pages. It was based on a graph propagation algorithm over a phrase similarity graph. It again assumed as input a set of positive seed phrases and a set of negative seed phrases. The nodes in the phrase graph were the candidate phrases selected from all n-grams up to length 10 extracted from 4 billion Web pages. Only 20 million candidate phrases were selected using several heuristics, e.g. Frequency and mutual information of word boundaries. A context vector for each candidate phrase was then constructed based on a word window of size six aggregated over all mentions of the phrase in the 4 billion documents. The edge set was constructed through cosine similarity computation of the context vectors of the candidate phrases. All edges (*vi*, *vj*) were discarded if they were not one of the 25 highest weighted edges adjacent to either node *vi* or *vj*. The edge weight was set to the corresponding cosine similarity value. A graph-propagation method was used to calculate the sentiment of each phrase as the aggregate of all the best paths to the seed words.

Another; but very different bootstrapping method ;was proposed by Dragut et al., [28] using WordNet. Given a set of seed words, instead of simply following the dictionary, the authors proposed a set of sophisticated inference rules to determine other words' sentiment orientations through a deductive process i.e. the algorithm takes words with known sentiment orientations (the seeds) as input and produces synsets (sets of synonyms) with orientations. The synsets with the deduced orientations can then be used to further deduce the polarities of other words.

Peng and Park, [29] presented a sentiment lexicon generation method using constrained symmetric nonnegative matrix factorization (CSNMF). The method first uses bootstrapping to find a set of candidate sentiment words in a dictionary and then uses a large corpus to assign polarity (or sentiment) scores to each word. This method thus uses both dictionary and corpus. Xu et al., [30]presented several integrated methods as well using dictionaries and corpora to find emotion words. Their method is based on label propagation in a similarity graph(Zhu and Ghahramani), [16]

In summary, the advantage of using a dictionary-based approach is that one can easily and quickly find a large number of sentiment words with their orientations. Although the resulting list can have many errors, a manual checking can be performed to clean it up, which is time consuming (not as bad as people thought, only a few days for a native speaker) but it is only a one-time effort. The main disadvantage is that the sentiment orientations of words collected this way are general or domain and context independent. In other words, it is hard to use the dictionary-based approach to find domain or context dependent orientations of sentiment words.

Many sentiment words have context dependent orientations, for example for a speaker phone; if it is quiet; it is usually negative however, for a car, if it is quiet, it is positive. The sentiment orientation of quiet is domain or context dependent. The corpus-based approach can help deal with this problem.

2.2.2 Corpus -based Approach

The corpus-based approach has been applied to two main scenarios:

(1) given a seed list of known(often general-purpose) sentiment words, discover other sentiment words and their orientations from a domain corpus;

(2) adapt a general-purpose sentiment lexicon to a new one using a domain corpus for sentiment analysis applications in the domain.

However, the issue is more complicated than just building a domain specific sentiment lexicon because the meaning of the word is context depend e.g."سعر المكالمات مرتفع" negative in the context but "قوة الشبكة مرتفعة" positive in another.

Hatzivassiloglou and McKeown,[31]are the first to deal with opinion classification. They focus on adjectives and studied phrases where adjectives are connected with conjunction words such as “and” or “but”. They construct a log-linear regression model so as to clarify whether two adjectives have the same orientation thereafter they perform clustering to separate the adjectives into two classes, and assumed the cluster with the highest frequency to be the positive orientation cluster

Kanayama and Nasukawa,[32]used domain dependent corpus to find sentiment words and their orientations in Japanese text by introducing the concepts of intra-sentential (within a sentence), and inter-sentential (between neighboring sentences) sentiment consistency, which they call *coherency*. The intra-sentential consistency is similar to the idea above.

Inter-sentential consistency simply applies the idea to neighboring sentences. That is, the same sentiment orientation is usually expressed in consecutive sentences. Sentiment changes are indicated by adversative expressions such as "*but"* and *however*. Some criteria were also proposed to determine whether to add a word to the positive or negative lexicon.

Moreover, finding domain-specific sentiment words and their polarity are useful, however this is inadequate in practice. Ding et al., [33] showed that many words in the same domain can have different orientations in different contexts these often occur with quantifiers like(*long, short, large, small)*. e.g., in the camera domain, the word " long " clearly expresses opposite opinions in the following two sentences: "*The battery life is long* " (positive) and "*It takes a long time to focus* " (negative) whereas in a car review, the sentence "*This car is very quiet* " is positive, but the sentence "*The audio system in the car is very quiet* " is negative. Thus, finding domain-dependent sentiment words and their orientations is insufficient. The authors found it important to extract both the aspect and the sentiment expressing words. then proposed to use the pair (aspect, sentiment word) as an opinion context, e.g., (" battery life", " long "). To determine sentiment words and their orientations whether a pair is positive or negative, the above intra-sentential and inter-sentential sentiment consistency rules about connectives are still applied.

Wu and Wen,[34], Their method is based on syntactic patterns as in[8], and also use the Web search hit counts to solve the problem in Chinese language. However, they only focused on pairs in which the adjectives are quantifiers such as big, small, low and high.

Lu et al., [35]used the same context definition as well. Ding et al., [33] assumed that the set of aspects was given. assigning each pair the positive or negative sentiment is considered as an optimization problem with a number of constraints. The objective function and constraints were designed based on clues such as a general-purpose sentiment lexicon, for rating sentiment of each review, they used synonyms and antonyms, as well as conjunction “and” rules,“ but ” rules, and “negation” rules.

Takamura et al., 2007, Turney, 2002[8, 36]method can also be considered as an implicit method for finding context-specific opinions, but they did not use the sentiment consistency idea. Instead, they used the Web to find their orientations.

However, it should be noted that all these context definitions are still not sufficient for all cases, e.g., consuming a large amount of resources.

Wilson et al., 2005[37] at the phrase or expression level the contextual subjectivities and sentiments was studied. Contextual sentiment means that although a word or phrase in a lexicon is marked positive or negative, but in the context of the sentence expression it may have no sentiment or have the opposite sentiment. they first labeled the subjective expressions in the corpus which contain subjective words or phrases in a given subjectivity lexicon. Note that a subjectivity lexicon is slightly different from a sentiment lexicon as subjectivity lexicon may contains words that indicate only subjectivity but no sentiment, e.g., *feel*, and *think*. The main aim of the study was to classify the contextual sentiment of the given expressions in the subjectivity lexicon. a supervised learning approach (algorithm BoosTexter AdaBoost) was applied with two steps. firstly, it determines whether the expression is subjective or objective. In the second step, it determines whether the subjective expression is positive, negative, both (means there are both positive and negative sentiments), or neutral. Neutral is still included because the first step can make mistakes and left some neutral expressions unidentified.

For subjectivity classification, a large and rich set of features was used, which included *word features*, *modification features* (dependency features), *structure features* (dependency tree based patterns), *sentence features*, and *document features*. For the second step of sentiment classification, it used features

such as *word tokens*, *word prior sentiments*, *negations*, *modified by polarity*, *conj polarity*, etc.

Choi and Cardie,[38]studied the problem of adapting a general lexicon to a new one for domain specific expression level sentiment classification. In their technique they utilize the expression-level polarities in the domain to generate a new lexicon, the adapted word-level polarities were used to improve the expression-level polarities. the problem was solved using integer linear programming and modeled polarity relationships between the word-level and the expression-level as a set of constraints. This work assumed that there was a given general-purpose polarity lexicon *L*, and a polarity classification algorithm *f* (*el*, *L*) that can determine the polarity of the opinion expression el based on the words in el and L. Jijkoun et al., [39]proposed a related method to adapt a general sentiment lexicon to a topic specific one as well.

Du et al., [40] proposed algorithm for adapting the sentiment lexicon from one domain (not a general-purpose lexicon) to another domain. the algorithm takes as input, a set of labeled documents from in-domain sentiment, a set of sentiment words from these in-domain documents, and a set of out-of-domain documents. The task was to make the in-domain sentiment lexicon adapted for the out-of-domain documents. they used two ideas were first, a document should be labeled as positive (or negative) if it contains many positive (or negative) words, and a word should be positive (or negative) if it appears in many positive (or negative) documents. These are mutual reinforcement relationships. Second, even though the two domains may be under different distributions, it is possible to identify a common part between them (e.g. the same word has the same orientation). The sentiment lexicon adaption was solved using the information bottleneck framework. The same problem was also solved by Du and Tan, [41].

Wiebe and Mihalcea,[42] investigated the possibility of assigning subjectivity labels to word senses based on a corpus ,the method was based on distributional similarity. Lin, [43] conducted two different studies. The first study investigated the agreement between annotators who manually assigned labels *subjective, objective*, or *both* to WordNet senses and evaluated a method for automatic assignment of subjectivity labels / scores to word senses in the second study. Subjectivity is a property that can be associated with word senses this is one of observer result study, and word sense disambiguation can directly benefit from subjectivity annotations. A subsequent work was reported by Akkaya et al.,[44]. Su and Markert,[45] also studied the problem and performed a case study for subjectivity recognition. In 2010 the same authors investigated this problem and applied it in a cross-lingual environment.

Brody and Diakopoulos,[46] presented an automatic way to leverage this association to detect domain sentiment and emotion words. By studied the lengthening of words (e.g., *slooooow*) in microblogs. The authors showed that lengthening is strongly associated with subjectivity and sentiment.

Feng et al.,[47] proposed a graph-based method based on mutual reinforcement to solve the problem of producing a connotation lexicon. A connotation lexicon differs from a sentiment lexicon in that the latter concerns words that express sentiment either explicitly or implicitly, while the former concerns words that are often associated with a specific polarity of sentiment, e.g., *award* and *promotion* have positive connotation and *cancer* and *war* have negative connotation.

For building a general-purpose sentiment lexicon the dictionary-based approach is usually more effective as a dictionary has all words however, the corpus-based approach may also be used to if a very large and very diverse corpus is available.

Dictionary-based approaches are generally not suitable for finding domain specific opinion words as dictionaries contain little domain specific information, however, domain and context-dependent sentiments remain to be highly challenging even with so much research.

the key difficulties of constructing lexicon are: (1) how to compact with context or domain dependent opinion words without any prior knowledge from the user, (2) how to deal with many important language constructs which can change the semantic orientations of opinion word such as negation word.

2.2.3 Summary

Table(2.1) illustrates a summary of the advantages and disadvantages of the approaches of creating lexicon

Table 2.1Summary of what are reviewed in the literature

Approach	Advantage	Disadvantage
Corpus based approach	possibility to identify multi-word opinion-bearing expressions find domain dependent orientations lexicon	These methods require a great amount of data to be processed.
Dictionary based approach	can easily and quickly find a large number of sentiment words with their orientations is the possibility to explore well-defined, formally coded and validated semantic relations between the words and a vast lexical base	Although the resulting list can have many errors Time consuming to clean the errors it is hard to use the dictionary-based approach to find domain or context dependent orientations of sentiment words , slang and social attributed connotations not contemplated in the thesaurus or dictionary are not accessible.
Translate based approach	since in some languages, linguistic resources are not available	the great challenge of translating a word or expression to another language maintaining its original sense
Manual based approach	possibility to identify multi-word opinion-bearing expressions find domain dependent orientations lexicon	It is labor intensive and time consuming

2.3 Feature-based opinion mining

Feature-level (aspect level) performs finer-grained analysis which attempts to discover a *target* (or feature entities) from sentences and identify opinion words (positive or negative) as associated with each entity. Instead of looking at language constructs (documents, paragraphs, sentences, clauses, or phrases), directly looks at the opinion itself.

In many applications, opinion targets are described by entities and/or their different aspects. Thus, the goal of this level of analysis is to discover sentiments on entities and/or their aspects. For example, the sentence

" *سعر المكالمات رخيص لكن الشبكة ضعيفة.*" evaluates two aspects: *سعر المكالمات* and *الشبكة*, of *network service* (entity). The sentiment on سعر المكالمات is positive, but the sentiment on its *الشبكة* is negative. The *سعر المكالمات* and *الشبكة* are the opinion targets.

As a result of analysis in this level, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses. Both the document-level and sentence-level classifications are already highly challenging. The feature-level is even more difficult, Various methods applied on feature-based opinion mining with two tasks (Fig. 1):

1. **Feature extraction**: extracts aspects that have been evaluated.
2. **Feature sentiment classification**: determines whether the opinions on different aspects are positive, negative, or neutral.

Observations: The important feature are identified according to: (a) the feature of a product that are usually commented by a large number of reviews; and (b) customer’ opinions on the important feature greatly affect their overall opinions on the product.

2.3.1 Feature extraction

Various methods for feature extraction and refinement have been applied on feature-based opinion mining .There are four main approaches:

- Extraction based on frequent nouns and noun phrases
- Extraction by exploiting opinion and target relations
- Extraction using supervised learning
- Extraction using topic modeling

This method focuses on two approaches(extraction based on frequent nouns and noun phrases extraction by exploiting opinion and target relations) which more compatible with Arabic corpus review due to the limitation like availability of label data set, since these are supervised techniques, they need manually labeled data for training. That is, one needs to manually annotate aspects and non-aspects in a corpus. Where topic modeling is an unsupervised learning method that assumes each document consists of a mixture of topics and each topic is a probability distribution over words. A topic model is basically a document generative model which specifies a probabilistic procedure by which documents can be generated. The output of topic modeling is a set of word clusters. Each cluster forms a topic and is a probability distribution over words in the document collection, so it depend on size of corpus.

2.3.1.1 Finding *Frequent* Nouns and Noun Phrases

This method finds explicit aspect expressions that are nouns and noun phrases from a large number of reviews in a given domain. The advantage of this method is that it is a simple empirical method that gives good results particularly for product reviews. Its disadvantage is that no normalization of features and it may need different heuristics given a different domain.

Hu and Liu,[9], established feature-based opinion summarization, they used association rule mining algorithm, to extract frequent item sets as explicit product features only in the form of noun phrases identified by a part-of-speech (POS) tagger. Apriori algorithm was used for finding frequent words; however, their method does not consider the position of the words in a sentence.

Popescu and Etzioni,[48]removed noun phrases which does not contain any features rather than on determining sentence or review polarity, by computing a point wise mutual information (PMI) score between the phrase and some meronymy discriminators associated with the entity class.

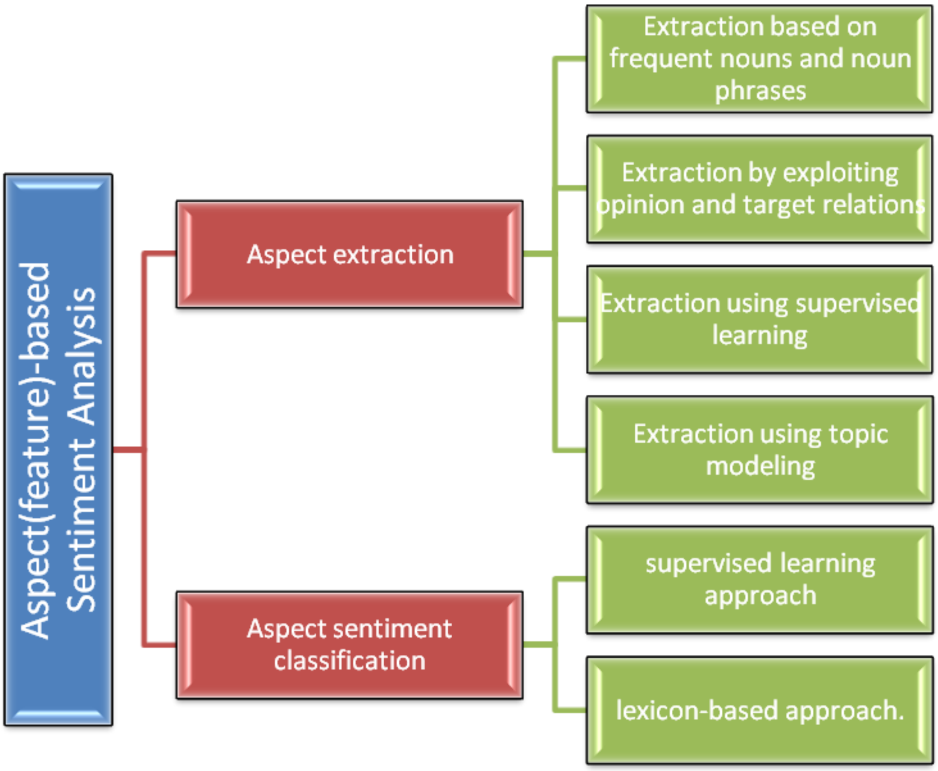

Aspect(feature)-based Sentiment Analysis
Aspect extraction
Aspect sentiment classification
Extraction based on frequent nouns and noun phrases
Extraction by exploiting opinion and target relations
Extraction using supervised learning
Extraction using topic modeling
supervised learning approach
lexicon-based approach.

Figure 2.1 Aspect(feature)-based Sentiment Analysis

Blair-Goldensohn et al.,[15], considered mainly those noun phrases that are in sentiment-bearing sentences or in some syntactic patterns which indicate sentiments. Ku et al., [49]made use of the TF-IDF scheme considering terms at the document level and paragraph level. Moghaddam and Ester, [50], augmented the frequency-based approach with an additional pattern-based filter to remove some non-aspect terms. Their work also predicted aspect ratings . Scaffidi et al.,[51] compared the frequency of extracted frequent nouns and noun phrases in a review corpus with their occurrence rates in a generic English corpus to identify true aspects. Long et al., [52]extracted feature (nouns) based on frequency and information distance whereas Jeong et al., [53] proposed an enhanced feature extraction and refinement method that effectively extracts correct features from review data by exploiting both grammatical properties and semantic characteristics of feature words and refines the features by recognizing and merging similar ones[4].

2.3.2 Using Opinion and Target Relations

Since an opinion unit is defined as a triple consisting of a product feature (targets), an expression of opinion, and an emotional attitude(positive or negative), they are obviously related. Their relationships can be exploited to extract aspects which are opinion targets because sentiment words are often known. Hu and Liu,[9] used this method to extracting infrequent feature. They considered that, the same sentiment word can be used to describe or modify different features. If a sentence does not have a frequent aspect but has some sentiment words, they extracting the nearest noun or noun phrase to each sentiment word and assign as infrequent feature. Since no parser was used by [9], the “nearest” function approximates the dependency relation between sentiment word and noun or noun phrase that it modifies, which usually works quite well. For example, in the following sentence ,“The software is amazing.” If we know that “amazing” is a sentiment word, then “software” is extracted as an aspect. This idea is more useful even to find all feature.

Blair-Goldensohn et al., [15] proposed sentiment patterns method used a similar idea. Furthermore, they used this method to discover important or key aspects (or topics) in opinion documents and be useful method because an aspect or topic is suspect to be important if there is no opinion or sentiment expressing about it.

Zhuang et al. and Somasundaran et al., [54, 55] used a dependency parser to identify such dependency relations for aspect extraction employed. whereas [56] form candidate aspects as noun or verb phrases .A phrase dependency parser was used for extracting noun phrases and verb phrases rather than a normal dependency parser thereafter they filtered out unlikely aspects by employing a language model.

All previous work in a normal dependency parser identifies dependency of individual words only, but a phrase dependency parser identifies dependency of phrases, which can be more appropriate for aspect extraction.

The dependency idea was further generalized into the double-propagation method for simultaneously extracting both sentiment words and aspects by[57, 58].

2.3.3 Feature sentiment classification

The feature extraction and sentiment determination process are tightly coupled together. Determining the orientation of sentiment expressed on each aspect in a sentence had been studied by two main approaches, these are the supervised learning approach and the lexicon-based approach

In the present study , the second approach is studied first, i.e., determining the orientation of sentiment expressed on each aspect in a sentence.

2.3.3.1 supervised learning approach

Wei and Gulla, [59] proposed a hierarchical classification model. However, they mentioned crucial question is how to determine the scope of each sentiment expression, i.e., whether it covers the aspect of interest in the sentence.

Jiang et al.,[60]proposed a dependency parser, a set of aspect dependent features is generating for classification task.

Boiy and Moens,[61] used a related approach which weighs each feature based on the location of the feature relative to the target feature in the parse tree.

Since a classifier trained from labeled data to build model, the model for one domain often performs poorly in another domain. The current methods are mainly used for document level sentiment classification as documents are long and contain more features for classification than individual sentences or clauses. Thus, supervised learning has difficulty to scale up to a large number of application domains.

2.3.3.2 The lexicon-based approach

Ding et al., Hu and Liu, [9, 33], applied the lexicon-based approach to avoid some of the issues, and seen that it will get good result in a large number of domains. Such methods use a sentiment lexicon (which contains a list of sentiment words, phrases, and idioms), composite expressions, rules of opinions, and (possibly) the sentence parse tree to determine the sentiment orientation on each feature in a sentence and may also consider sentiment shifters, but-clauses and many other constructs which may affect sentiments.

Ding et al.,[33] introduced one simple lexicon-based method which has four steps

1. Mark sentiment words and phrases
2. Apply sentiment shifters
3. Handle but-clauses
4. Aggregate opinions

This simple algorithm performs quite well in many cases. Hu and Liu [9] counted the sentiment scores of all sentiment words in a sentence or sentence segment whereas[11, 62] used multiplication of sentiment scores of words.

Blair-Goldensohn et al.,[15] proposed a method that integrated the lexicon-based method with supervised learning to enhance the above method .

To make this method even more effective, parsing is needed to find the dependency between the words to determine the scope of each individual sentiment word. and then discover automatically the sentiment orientation of context dependent words such as “long” above.

Table 2.2Summary of Feature extraction method

Method	Technique	Strength	Limitations
frequency-based methods[9, 48]	apply a set of constraints on high-frequency noun phrases to identify aspects.	very simple quite effective	produce too many non-aspects miss low-frequency aspects require the manual tuning of various parameters (thresholds) makes them hard to port to another

			database
Relation-based Methods[63, 64]	sentiment expresses an opinion on an aspect and sentiments are often known or easy-to-find their relationship can be used for identifying new aspects (and sentiments).	find low- frequency aspects.	produce many non-aspects matching with the relation patterns
Supervised Learning Techniques[65, 66]	The current state-of-the-art sequential learning methods are HMM and CRF	• supervised learning approaches overcome the limitations of frequency- and relation- based • methods by learning the model parameters from the data	need manually labeled data for training
Topic Modeling Techniques[67, 68]	Topic modeling is an unsupervised learning method that assumes each document consists of a mixture of topics and each topic is a probability distribution over words. The output of topic modeling is a set of word clusters	no need for manually labeled data. perform both aspect extraction and grouping at the same time in an unsupervised manner	need a large volume of (unlabeled) data to be trained accurately

2.4 Arabic opinion mining

Opinion Mining in Arabic text is not popular among researches due to a number of limitations[69] :

A. Structure and morphology

The language is complex in terms of both structure and morphology, since many different parts of speech are possible. Furthermore, it is highly inflectional and derivational language with many word forms and special labels called diacritics used instead of vowels [69] , e.g. the sentence “علم احمد” can be tagged as either noun phrase when the word "علم" is taken as “flag” or as a verb phrase if taken as “knew”. The same three-letter root can give rise to different words with different meanings. When using stemming , the same word can have several different forms with different diacritics.

Also Arabic have different types of sentence structures: verbal, where the sentence starts with a verb phrase, and nominal, where the sentence starts with a noun phrase.

B. Standard Arabic Forms

The lack of opinions written in classical or Modem Standard Arabic forms. Such text is hard to find in domains such as movie and product reviews, which are the standard domains addressed. However, the forms used in forums and blogs are mostly dialects. This complicates the use of semantic approaches for mining opinions. It is important to emphasize that the opinion is limited to a specific locality, e.g. “الشبكة بلشت تقطع” can be viewed as either negative or positive depending on the viewer. That is, a Sudanese would view the statement as a positive opinion whereas a Lebanese would view it as negative.

C. Lack of Labeling

Most of the approaches proposed in previous work in Arabic Opinion Mining used Supervised Learning Algorithms , while few existing approaches use Unsupervised Learning Algorithms .This raised the challenge of having to build annotated text corpora for the purpose of evaluating these proposed approaches.

There is unlabeled-classical-Arabic-text, which is required for input into a Supervised Learning Algorithm.

D. Opinion Lexicon

The absence of an opinion Lexicon for the Arabic language, which thwarts polarity measurement of extracted subjective text.

These issues pose a challenge for sentiment mining, which generally requires both semantic analysis of words and grammatical analysis of text , now, we discuss some of the existing works that tried to deal with Arabic language.

Ahmad et al.,[70, 71] used Local Grammar to identified domain-specific key words by looking for frequent words that exist in a corpus of financial news but infrequently in a general corpus, build a local grammar to extract sentiment-bearing phrases by using the context around these words. Their approach are applied to Arabic, English and Chinese. As result their system achieved accuracy rates between 60-75% for extracting the sentiment bearing phrases and evaluated the system manually and. Note that the proposed system language in depend.

Ahmad [70] applied their work to documents from the financial news domain. They identified domain-specific key words by looking for words that occurred often in a corpus of financial news but relatively infrequently in a general corpus. Using the context around these words they built a local grammar to extract sentiment-bearing phrases. They applied their approach to Arabic, English and Chinese. They evaluated the system manually and achieved accuracy rates between 60-75% for extracting the sentiment bearing phrases. Importantly, the proposed system could be used to extract the sentiment phrases in financial domain for any language.

Abbasi et al.,[72]worked on document-level by using syntactic and stylistic features and a feature selection algorithm that they developed and named Entropy Weighted Genetic Algorithms (EWGA)) combine genetic algorithms with information gain (IG) to perform the feature selection for both Arabic and English. In specific, IG is used to select the initial set of features for the initial stage of the GA, and is also applied during the cross-over and mutation stages. EWGA is applied to select features for sentiment analysis in a corpus of Web forum data containing multiple languages. They avoided semantic features because they are language dependent and need lexicon resources, while the limitation of their data prevent the use of linking features. The paper evaluates the proposed system on a benchmark tested consisting of 1000 positive and 1000 negative movie reviews. Using this system, they achieved an accuracy rate of 91% while other systems achieved accuracy rates between 87-90% on the movie review data set. They were also able to achieve 92% accuracy on Middle Eastern forums and 90% on US forums using the EWGA feature selection method.

Elhawary and Elfeky,[73] used Arabic financial reviews to build a system for sentiment analysis ,with specific objective of building a web search engine that would automatically annotate returned pages with sentiment scores. The system has several components. The first component classifies whether an Internet page is a review or not. The task of classifier is to assign a tag from the set (review, forum, blog, news, shopping store) to Arabic document. They collected 2000 URLs and more than 40% of them were found to be reviews through manual labeling to build an Arabic review classifier data set, by searching the web using keywords that usually exist in reviews (such as "the camera is very bad"). they translated the lists of keywords collected and add to them a list of Arabic keywords that usually appear in opinionated Arabic text. The final list contained 1500 features and was used to build an AdaBoost classifier, using 80% of the data for training and the rest for testing. After a document is classified as belonging to the Arabic review class or not, a second component of the system analyzes the document for its sentiment. They build an Arabic lexicon based on a similarity graph for use with the sentiment component. The final component of the system is designed to provide the search engine with an estimate of the sentiment score assigned to a document during the search.

Farra et al.,[74]proposed two sentence-level sentiment analysis approaches, one of them relies on grammatical features of the Arabic language. Which is based on Arabic grammatical structure and combines the verbal and nominal sentence structures in one general form based on the idea of actor/action. In this approach, the subjects in verbal and nominal sentences are actors and verbs are actions. Manual POS tagging of words was applied and used as features for vectors Their feature vector constitutes the following dimensions: sentence type (verbal or nominal), actor, action, object, adjective, type of pronoun and noun, transition (the type of word linking the current sentence with the previous sentence), word polarity (positive, negative, neutral) and sentence class. Using SVM classifier was reported accuracy in the 80%

The second approach proposed by Farra et al., [74] was jointed syntactic and semantic features like frequency of negation, opinionated words (positive, negative, and neutral words) and special emphasis words(e.g., "really" and "especially"). For extracting the semantics of the words, a semantic interactive learning dictionary which stores the semantic polarity of word roots extracted by stemmer was developed. The system asks the user for the polarity of a word if it has not yet been learned. For evaluation of the grammatical approach, only 29 sentences are annotated manually with part-of-speech tags. They report 89.3% accuracy using an SVM classifier with 10-fold cross validation. Classification accuracy ranged from 60% to 80%. Manual results than the interactive dictionary because many words of different polarity have the same stem and were incorrectly tagged by the dictionary

Sentences from 44 random documents are used for evaluating the semantic and syntactic approach using a J48 decision tree classifier. They report 80% accuracy when the semantic orientation of the words extracted and assigned manually is used, and 62% when the dictionary is used. They also classified the documents by using all sentence features and chunking the document into different parts, reporting 87% accuracy with an SVM classifier when documents divided into 4 chunks and neutral class excluded.

The work of Rushdi-Saleh and Martín-Valdivia, [75] used supervised learning to build classifiers using both the OCA and EVOCA corpora from movie reviews. using both Support Vector Machines (SVMs) and

Naive Bayes (NB) classifiers, reporting 90% F-measure on OCA and 86.9% on EVOCA using SVMs. They show that SVMs outperform the NB classifier, which is common in text classification tasks. Our result showed that there is no difference between using term frequency (TF) ,and term frequency-inverse document frequency (TF-IDF) as weighting schemes. Experiments also show three no need for stemming words before feature extraction and classification because it degrade the results.

El-Halees,[76] proposed a combined classification approach for document level sentiment classification.by applied different classifiers in a consecutive way. A lexicon-based classifier is first used to estimate the sentiment of a document based on an aggregation of all the opinion words and phrases in the document. However, lacking of enough opinion words in some documents, he used lexicon-based classifier, in phase two used a maximum entropy classifier. The input from first classifier , classified documents are used as the training set for second classifier, which is then used to compute the probability that a given document belongs to a certain sentiment class. In particular, if the probability is greater than a threshold of 0.75, then the document is assigned a class, and otherwise the document is passed to the next stage. The final phase is a k-nearest neighbors (KNN) classifier that finds the nearest neighbors for the unannotated document using the training set coming from the previous two classifiers. The corpus used for evaluation consisted of 1134 documents collected from different domains (e.g., education, politics, and sports), with 635 positive documents (with 4375 positive sentences) and 508 negative documents (with 4118 negative sentences). For preprocessing phase, first remove HTML tags and non-textual contents. Corrected misspelled words

,alphabets are normalized. Then tokenized, removing stop words, stemming the words using Arabic light stemmer, and TF-IDF is used for term weighting. Their result is f-measure of 81.70% averaged over all domains for positive documents and 78.09% F-measure for negative documents. The best F-measure is obtained in the education domain (85.57% for the positive class and 82.86% for the negative class)

In the field of knowledge-based techniques, a study conducted by Al-Subaihin et al., 2011 [77]proposed the implementation of a new tool that can be used for Arabic sentiment analysis which accept input with informal Arabic language. Two techniques are combines in their system, natural language processing and human computation. Their system contains two parts: game-based lexicon and sentiment analyzer. constructing the lexicon based on human computation used online computer game in the first part . The game presents many phrases and words extracted from Qaym.com to the player and (s) he has to decide whether they are positive, negative or neutral. They constructing the lexicon automatic to avoid the problems of manual construction. Sentences patterns is another output for this game which contains positive, negative, natural and negation tags ,then they stored theses tags with their polarities into a database. The second part of this tool is sentiment analyzer that takes each review and performs sentences segmentation. After that, for each sentence the words will be tagged to be POS, NEG, ENT and NO to represent positive, negative, neutral and negation words according to the game-based lexicon. After tagging, the sentence polarities can be detected by matching the resulted patterns with the ones that are stored in the database. Then the polarity for the review will be determined according to the maximum polarities

Opinion corpus for Arabic (OCA) is a corpus of text from movie review sites by Rushdi-Saleh et al.,[78] and includes a parallel English version called EVOCA. The corpus consists of 500 reviews, half negative and half positive. The raw reviews contained a number of challenges which the authors attempted to fix manually, including filtering out spurious and unrelated comments, Romanization of Arabic, multi-language reviews, differing spellings of proper names, and movie reviews that were more opinions of the cultural and political themes of a movie than the film itself. (As an example of the latter issue, the movie “Antichrist” has a rating of 6.7 in IMDB but a rating of 1 in the reviews on the Arabic blog.) OCA and EVOCA performed standard pre- processing on the corpus, including correcting spelling mistakes and deleting special characters, and also have made available unigram, bigram, and trigrams for the dataset MPQA subjective lexicon & Arabic opinion holder corpus: Another corpus for Arabic

Abdul-Mageed, Korayem, and Diab,[79-81] built subjectivity and sentiment analysis systems exploiting them based on sentence-level annotated Arabic corpora. In their systems used various types of features, including language independent features, Arabic-specific morphological features, and genre-specific features.

Abdul-Mageed and Korayem, [79, 81] extended the previous work by classifying MSA news data at the sentence level for both subjectivity and sentiment. They use a two-stage SVM classifier, where a subjectivity classifier is first used to separate subjective from objective sentence. In a second stage, classified subjective sentences into positive and negative. They make use of both language-independent and Arabic-specific features reported 95.52% accuracy of Classification their results showed that the adjective feature is very important, as it improved the accuracy by more than 20% and the unique and domain features are also helpful.

Abdul-Mageed et al.,[82]presented SAMAR, an SVM-based system for Subjectivity and Sentiment Analysis (SSA) for Arabic social media genres which tackles the problem of sentiment analysis in social media from a mostly linguistic perspective, including how to best represent lexical information, whether standard features are useful, how to treat Arabic dialects, and whether genre specific features have a measurable impact on performance.

Their system is based on support vector machine (SVM) classifiers and carries out SO determination in two steps. In the first step, distinguishing between subjective (opinionated) and objective case. In the second step, another classifier is used to determine the polarity (positive or negative) of subjective input. They are not use neutral and mixed cases in their system. Some of the features used by the classifiers include morphological features, part of speech (POS) tags, and matches made with entries in an adjective polarity lexicon which simply classifies adjectives as either positive or negative. The dialectical performance of the system was evaluated using the Tagged dataset which consists of 3015 Arabic divided into 1466 written in MSA 1549 tweets written in different dialects. 80% of each of the datasets were used for training, 10% for developments, and 10% for testing.

The highest accuracy reported through the dialect-specific sentiment experiments was 71.15% with an F-score of 29.4%for positive cases and an F-score of 81.8% for negative ones.

The fact that the dialect specific dataset consists mostly of negative tweets, has balanced out the low positive F-score when assessing the overall accuracy.

Mountassir et al.,[83]investigated sentiment classification in an Arabic context. they used two Arabic corpora with sizes different(ACOM ,OCA). ACOM is a corpus that have been developed (from Aljazeera's site) and annotated manually with two main categories: POSITIVE and NEGATIVE. It consists of two data sets; DS1 which is a collection of 368 comments about a series reviewing and DS2 which is a collection of 1000 comments from sport domain. OCA is movie-reviews collected by Rushdi-Saleh et al., [78] who aimed to investigate some settings like stemming type, term frequency threshold, term weighting, and n-gram words model , that yield the best results. Common three classifier are used, Naïve Bayes, Support Vector Machines and k-Nearest Neighbor. The authors compared between these three classifiers and the effectiveness of an Arabic context. Their results showed that the best setting for almost all classifiers on all data sets was the application of light-stemming, the elimination of hapaxes (as threshold), the combination of unigrams and bigrams, and the use of a presence-based weighting.

Elarnaoty et al.,[84] proposed an opinion holder and subjectivity lexicon. Created an Arabic news corpus by crawling 150 MB of Arabic news and do manually annotating by three different people for 1 MB of the corpus for opinion holder. Using majority voting to remove any conflict emerging. For preprocessing the corpus, using for handle the morphological analysis of Arabic sentences and assign parts of speech (POS) tags, the Research and Development International (RDI) tool (http://www.rdi- eg.com). Finally, semantic analysis of the words were done. Arabic Named Entity Recognition (ANER) [3] was used for extracting names from documents

Misbah and Imam, [85]presented an optimized approach for mining opinions in Arabic Religious Decrees using an improved "Semantic Orientation using Point wise Mutual Information" Algorithm. Their approach executed a number of steps to classify a religious decree into either halal (Allowed) or “haram” (Prohibited) which including Data Collection, Simple Text Preprocessing, Manual Data labeling, Advanced Text Preprocessing, Weight Calculation and experimentation using Supervised and Unsupervised Learning Algorithms. Results obtained by original approach gave an accuracy rate of 73.08%. The new approach utilizes an improved SO-PMI Algorithm that executes a series of advanced steps to improve the calculation of the weights. The improved algorithm increased the accuracy rate of the Unsupervised Learning Algorithm up to 2000 but produced poor results for the Supervised Learning Algorithm. It is recommended that Subjectivity Classification be executed before Advanced Text Preprocessing. In this step, a classifier would be used to classify sentences into Objective and Subjective. Subjective sentences would be checked with their relevancy against the asked question and only those correlated with the question should be used. After this process is done, Advanced Text Preprocessing and Weight Calculation in the proposed approach should be executed against extracted sentences.

Using this improvement would guarantee that tokens extracted would be opinion-oriented tokens. The tokens will also be closely correlated with the topic of the decree. It is expected that this would increase the accuracy rate of Sentiment Classification of the decrees

Itani et al.,[86]presented the application of two different approaches to classify Arabic Facebook posts. The first one using common patterns used in different Arabic dialects to express opinions depended on syntactic features. These patterns achieved high accuracy in determining the polarity of a sentiment even when tested against new corpus. This approach acts on informal Arabic text, which has not been addressed before. Different setups were tried and the highest coverage and accuracy achieved were 49.5% and 83.4 % respectively. The second approach is used Naïve Bayes classifier an ordinary probabilistic model, which assumed the independence of features in determining the class the highest coverage achieved in this approach was 60.5% in the first setup and 91.2% when Naïve search was used as a binary classifier to classify the posts as objective or subjective.

CHAPTER THREE

3 Research Methodology

3.1 Introduction

This chapter presents the different phases of this research work and discusses the methodology used for development of the proposed Arabic Sentiment Classifier model and Arabic Sentiment Analyzer model to achieve the objectives of this research. . In this research The Arabic Sentiment Classifier(ASC) is built on document level, and a feature-based opinion mining method is proposed to build the Arabic Sentiment Analyzer model(ASA). The domain problem is the Arabic customer review in a telecommunication company in Sudan and hotel review. Since Arabic customer review data set does not exist one of the contribution of this thesis is the creation of two Arabic corpuses and second contribution is the constructed opinion lexicon. The two corpuses and opinion lexicon are content of language resource. And the third contribution is extracted feature of the product and build opinion mining model , and fourth contribution is constructed Arabic sentiment classifier.

As a result of the literature survey, dictionary based approach has been used to create opinion lexicon , Frequency- and Relation-based Approaches have been used to build opining mining model. Manual annotation corpus is also employed as a benchmarking dataset for evaluation. Fig (3.1) explains the proposed phase.

3.2 Phase 1: Problem Domain Identification

The research activities described in this thesis aim at investigating and proposing different techniques for Sentiment Analysis applied to customer reviews written in the Arabic Language. User-generated contents are written in natural language with unstructured-free-texts scheme. Manually scanning through large amounts of user- generated contents is time consuming and sometimes impossible. In this case, opinion mining is a better alternative .It has a wide range of applications such as: product reviews, advertising systems, market research, public relations, financial modelling and many others .Most research efforts in the area of opinion mining deal with English texts.

Some new research works have deal with other languages, but in Arabic, which is a language for Millions of people, there is a little work.

The need of automatic tools for Sentiment Analysis is justified by the huge amount of opinionated contents available on the Web (e.g.: review sites, blogs, forums) and their continuous growth rate**.** Most reviews are in unstructured text format. Some reviews are long and contain only a few sentences expressing opinions on the product. Therefore, it is not an easy task for a potential user (either a customer or a company) to locate, read, understand and analyze each review that may be relevant to his or her decision making. So it becomes essential , to develop an Arabic Sentiment Analyzer that can analyze and summarize opinions ,expressed in Arabic opinion resources to provide useful information for potential users and better understanding of customer’s satisfaction**.**

This phase contains two stages. The first one is the study of customer review related to particular service. The second stage is how to have an idea about the service from these large review , Customers may comments on two types of format.[4]

Format 1 - Pros, Cons, and the detailed review: The reviewer first describes some brief pros and cons separately and then writes a detailed/full review. An example of such a review is given in Fig(3.2).

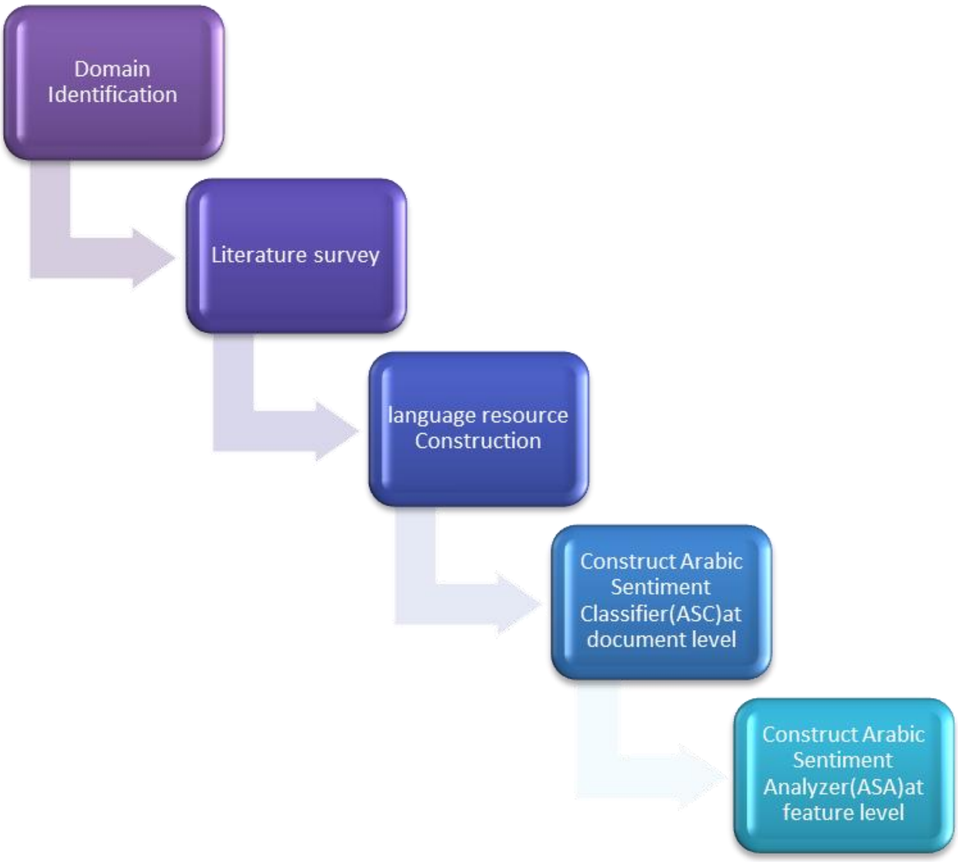
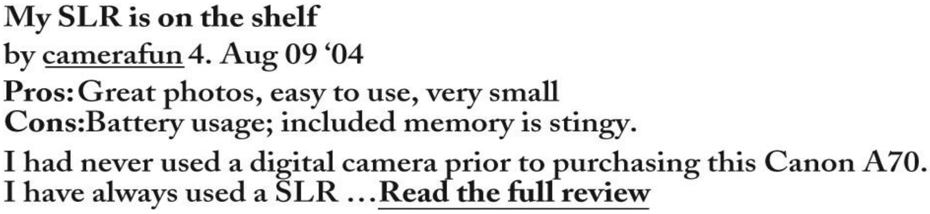

Domain Identification
Literature survey
language resource Construction
Construct Arabic Sentiment Classifier(ASC)at document level
Construct Arabic Sentiment Analyzer(ASA)at feature level

Figure 3.1 The proposed phase

Format 2 - Free format: The reviewer writes freely, i.e., no brief pros and cons. An example of such a review is given in Figure (3.3). In this research the review format 2 (free format) ,in this research we used format2 written in Arabic language

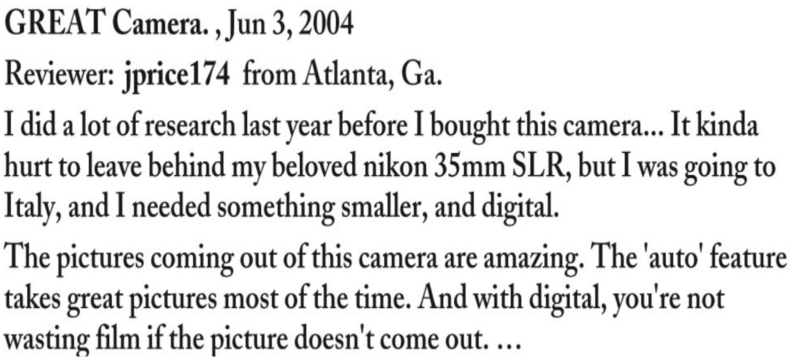
3.3 Phase Two: Literature Survey

In this phase we read and analyze a lot of scientific papers to give a solid background of the opinion mining task. This phase contains three stages, the first one is concerned with methods of creating opinion lexicon which a play central role in aspect based techniques. The second stage focuses on aspects based method which is used to develop opinion mining models and the third stage is concerned with Arabic opinion mining. As a result of this phase mostly aspect based techniques have been identified and most Arabic research in opinion mining is listed .



My SLR is on the shelf
by camerafun 4. Aug 09 '04
Pros: Great photos, easy to use, very small
Cons: Battery usage; included memory is stingy.
I had never used a digital camera prior to purchasing this Canon A70. I have always used a SLR … Read the full review

Figure 3.2 An example of a review of format 1.



GREAT Camera., Jun 3, 2004
Reviewer: **jprice174** from Atlanta, Ga.
I did a lot of research last year before I bought this camera... It kinda hurt to leave behind my beloved nikon 35mm SLR, but I was going to Italy, and I needed something smaller, and digital.
The pictures coming out of this camera are amazing. The 'auto' feature takes great pictures most of the time. And with digital, you're not wasting film if the picture doesn't come out. …

Figure 3.3 An example of a review of format 2

3.4 Phase Three: Language Resource Construction:

In the language resource construction module has different steps ,firstly collects the objects (e.g., customer reviews) from company sites. Then it performs a Simple text preprocessing to the collected data so, to prepare it ,Save it in text file .

The main objective of this stage is to prepare two opinion corpuses for the selected telecommunication Company in Sudan and hotel review site

(agoda .com) are created.

3.4.1 Telecommunication corpus:

The following are the activities conducted to prepare opinion corpus :

- Interviewing Contact Center Manager to identify how they get feedback from customer .
- The company call center takes the customers feedback by asking them different questions about the provided service or free talk
- Those conversions are converted to a text and later saved by using excel sheets
- About 600 documents(2 data set) were Selected randomly
- Each row in excel represent a single review
- Separate the documents into a single review and save them into a single file.
- For Arabic scripts some repeated letters have been cancelled
- Some wrongly spelt words are corrected.

In this is corpus the sentence is short and it looks like an answer to a question that is asked by a call center For example : «سعر المكالمات داخل الشبكة مناسب و خارج الشبكة مناسب، خدمة الانترنت ممتازة و الشبكة ممتازة، سعر الرسائل جيد و الخدمات جيدة»

3.4.2 Hotel corpus

The second corpus gathers data on hotel reviews using information available publicly on the Agoda Website (http://www.agoda.co.th), an online hotel reservation service. We have selected Agoda because they have a greater number of customer reviews in the Arabic language unlike other Web sites that translate to Arabic . Our data covered 50 hotels in UEA with 1200 customer reviews. The entry for every hotel listed on Agoda contains the general information about the hotel; such as size, location, pictures and reviews from their previous customers, the review format for a hotel corpus has used format2 and may some Arabic Dialects. After collecting the data :

- A single review is extracted after stripping out the HTML tags and non-textual contents.
- Save into a single file(text file).
- For Arabic scripts some repeated letters have been cancelled (that happens in discussion when the user wants to insist on some words).
- some of the wrongly spelt words are corrected.

Assign label according to rating system (positive for ratting >=seven)and negative for ratting <seven, as result about 1090 customer reviews(751 positive ,339negative)

In this is corpus the sentences are too long and may have some Arabic Dialects

«موقع ممتاز الموقع جيد جدا و الخدمة جيدة و السعر مناسب حمام السباحة لا يعمل و
في الصيانة قنوات التليفزيون تحتاج لتحسين و زيادة في عدد القنوات يفضل جعل الانترنت

بدون مصروفات للأسف لم تكن الغرفة جيدة فكل ما فيها غير نظيف مواقف السيارة - بهو
الفندق جميل- استقبال الموظفين رائع»

3.5 Phase four : Construct Arabic Sentiment Classifier(ASC) at document level

This phase to Construct Arabic Sentiment Classifier(ASC) which aims to classify an Arabic opinion review as expressing either a positive or negative opinion or sentiment. The task is also commonly known as the *document-level sentiment classification* because it considers the whole document as a basic information unit.

3.6 Phase five : Construct Arabic Sentiment Analyzer model(ASA)

A generic framework aimed at defining automatic tools dedicated to feature based classification which has been designed and implemented. The ASA framework consists of two main module opinion lexicon construction and opinion mining modules which consist of two phases the first phase is used for identifying features and their orientation while the second one is for generating summary. these following steps are adopted in phase five

3.6.1 Preprocessing step

- The sentences are tokenized.
- Stop words removed.

- Obtained vector representations for the terms from their textual representations by performing (Term occurrences).
- Association rule mining are used to discover correlations among a set of items in database.

3.6.2 Extract frequent word

This step is intended to find frequent words that are most popular in a text (nouns and adjectives) . In order to do this, we use association rule mining (Agrawal and Srikant 1994) to find all frequent itemsets.

The generated frequent itemsets, which are also called candidate frequent words, are stored as two sets(frequent adjective and frequent noun or noun phrase) for further processing .

3.6.3 Creation of Arabic opinion lexicon

We now identify opinion words. These are words that are primarily used to express subjective opinions or words that convey positive or negative sentiments, are instrumental for sentiment analysis[87]. Previous work on subjectivity [88]has established a positive statistically significant correlation with the presence of adjectives. In this thesis we use adjectives as opinion words to construct opinion lexicon using dictionary based approach . A set of frequent adjectives are expanded to create Arabic opining lexicon

3.6.4 Design of opinion mining modules

3.6.4.1 Extract feature of the product

Using the extract frequent noun as a product feature of the objects which represent the entity of the product [9, 58, 89] . The feature is stored in a feature set .

3.6.4.2 Opinion Summarizing

a novel information summarizing and visualization approach, based on NLP technique:

- Define sentiment sentence. Note that these opinion sentences must contain one or more product features identified above.
- Identify the orientation whether it is positive or negative of each feature in this sentence depending on opinion lexicon
- Aggregate on each feature
- Summarize the results.

The visualization module is aimed at providing users with an effective way to browse the set of feature according to the polarity expressed by each review.

3.7 Evaluation

The main objective of this phase is to identify evaluation criteria for the proposed ASA models and validate them to choose the suitable model for each corpus.

Evaluating the accuracy and precision of ASC model as well as evaluating ASA from this perspectives:

1. the accuracy of opinion lexicon

2. The accuracy of feature extraction.

If the necessary ground truth is available, the performance of a method for aspect-based opinion mining can be evaluated by measures such as accuracy, precision and recall. However, in real-life data sets such ground truth is typically not available[90]. In some of the works some human judges have been asked to read a set of reviews and manually create a set of opinion word , a set of “true” aspects and their ratings for the reviewed item as “gold standard”. Precision and recall of aspect extraction are then computed versus this gold standard. So the of this work depends on two experts in Arabic language .

3.8 Summary

This chapter presented the research phases, how each phase was conducted, and how these phases are related. A general overview of this research methodology is summarized in Table(3. 1)

Table 3.1 summary of the Research Methodology

Phase	Activities	Objective(s)	Outputs
Problem Domain Identification	The first stage is the studying of customer review related to one services. The second stage is how to have idea about the service from these large review	justified The need of automatic tools for Sentiment Analysis is: • The huge amount of opinionated contents available on the Web (e.g.: review sites, blogs, forums) and their continuous growth rate**.** • Most reviews are in unstructured text format. Some reviews are long and contain only a few sentences expressing opinions on the product. • It is not an easy task for a potential user (either a customer or a company) to locate, read,	Investigating and proposing different techniques for Sentiment Analysis applied to customer reviews written in the Arabic Language

		understand and analyze each review that maybe	
Literature Survey	• Reading and analysis scientific papers	• To build Back ground of opinion mining aspect based • used method of creating opinion lexicon • To identify the mostly research in Arabic opinion mining	• The mostly used aspect based techniques • The mostly used construct opinion lexicon techniques • The pros and cons of these techniques • most Arabic research in opinion mining are listed
language resource Construction			
Creation of corpuses	• Interviewing Contact Center Manager to identify how the get feedback from customer	• To Prepare the two corpus • Assigns label	• Hotel review corpus • Tele review corpus
Design of opinion mining modules			
Preprocessing of corpus	• Manually cleaned to remove spelling mistakes. • Tokenizing • Removing of stop words, • Part-of-Speech Tagging (POS) • Stemming . • Create word vector .	• To Prepare high quality corpus	
Extract frequent word	• Applying association rule mining and NLP technique to tow corpora • use adjectives and noun or noun phrase	• Adjective set • Noun set	
Creation of	• use adjectives set as opinion words	• construct opinion lexicon	• Opinion lexicon.

opinion lexicon	• apply dictionary based approach to construct opinion lexicon		
Extract feature of the product	• Use the noun or noun phrase set	• Extract feature of the objects which represent the entity of the product	• Two feature set
Opinion Summarizing	• Define sentiment sentence. • Identify the orientation • Aggregate on each feature • Summarizing the results	• aimed at providing users an effective way to browse the set of feature according with the polarity	• visualization module

CHAPTER FOUR

4 Construct Arabic Sentiment Classifier(ASC) at document level

4.1 Introduction

This chapter aims to classify an Arabic opinion review as expressing either a positive or negative opinion or sentiment. The task is also commonly known as the *document-level sentiment classification* because it considers the whole document as a basic information unit. The majority of research papers on this topic classifies online reviews. Thus the definition of the problem in the review context, but the definition is also applicable to other similar contexts

Problem definition: Given an opinion document 'D' evaluating an entity, determine the overall sentiment s of the opinion holder about the entity, i.e., determine s expressed on aspect GENERAL in the quintuple

(_, GENERA L, s, _, _),

Where the entity e, opinion holder h, and time of opinion t are assumed known or irrelevant (do not care). It's defined as a classification problem if it formulations based on the type of categorical values that is takes, e.g., positive and negative .

To ensure that the task is meaningful in practice, existing research makes the following implicit assumption [3]

Assumption: Sentiment classification or regression assumes that the opinion document d (e.g., a product review) expresses opinions on a single entity e and contains opinions from a single opinion holder h.

In practice, if an opinion document evaluates more than one entity, then the sentiments on the entities can be different. For example, the opinion holder may be positive about some entities and negative about others. Thus, it does not make practical sense to assign one sentiment orientation to the entire document in this case. It also does not make much sense if multiple opinion holders express opinions in a single document because their opinions can be different too.

This assumption holds for reviews of products and services because each review usually focuses on evaluating a single product or service and is written by a single reviewer. However, the assumption may not hold for a forum and blog post because in such a post the author of the review may express opinions on multiple entities and compare them using comparative sentences.

In this chapter the aim is constructing Arabic Sentiment Classifier(ASC) at document level using hotel corpus that created in chapter 3. The rest of this chapter is organized as follows. Section 2, gives a brief description of algorithm .Section 3, shows experimental Setup. Section 4 describes the valid and feature selection methods .Section 5 presents evaluation results. And section 6 gives the conclusion

4.2 Classification Method

In this section we review fundamental aspects of three popular supervised classifiers: Naıve Bayes, Support Vector Machines and KNN.

4.2.1 Naıve Bayes

naïve Bayes is a probabilistic learning method that assumes terms occur independently. Given a collection of N documents $\{d_j\}_{j=1}^{N}$, where each document is represented as a sequence of T terms d_j = {t_1, t_2, . . . , t_T}, the probability of a document d_j occurring in class c_k is given as

$$p(c_k|d_j) = p(c_k)\prod_{i=1}^{T} p(t_i|c_k) \quad (1)$$

where $p(t_i|c_k)$ is the conditional probability of term $t_i$occurring in a document of class c_k and $p(c_k)$ is the prior probability of a document occurring in class c_k.) . $p(t_i|c_k)$ and $p(c_k)$ are estimated from the training data. [91]

Naïve Bayes Classification Method has two Advantages the first one it is easy to interpret, the second is more efficient computation. However the disadvantage of Naïve Bayes is the assumptions of attributes being independent, which may not be necessarily valid.

4.2.2 Support Vector Machines

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples fig (4.1).

SVM is a linear learning method that finds an optimal hyper plane to separate two classes. As a supervised classification approach, SVM seeks to maximize the distance to the closest training point from either class in order to achieve better generalization/classification performance on test data[92]. The solution is based only on those training data points which are at the margin of the decision boundary.

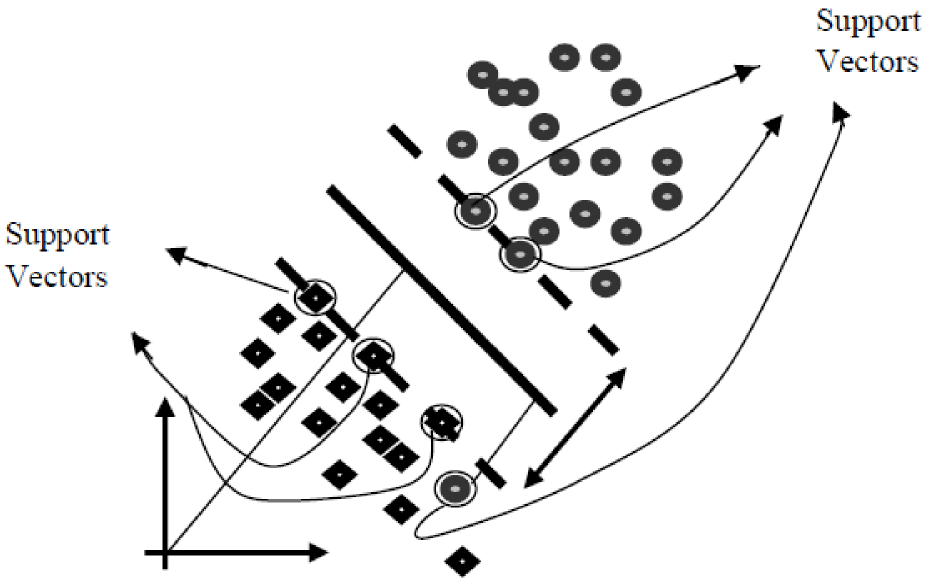


Figure 4.1 Support Vector Machines

The advantages of the Support Vector Machine Method are: it's very good performance on experimental results. And have a low dependency on data set dimensionality. And one disadvantage of SVM is i.e. in case of categorical or missing value it needs pre-processed and the difficult interpretation of the resulting model.

4.2.3 K Nearest Neighbor

KNN is a simple machine learning algorithm. In this algorithm, the objects are classified based on the majority of its neighbor. The class assigned to the object is most common among its k nearest neighbors. The KNN classification algorithm classifies the instances or objects based on their similarities to instances in the training data . In KNN, selection is based on majority voting or distance weighted voting.

KNN is unsupervised text classification algorithm and its work efficiently when the training set is large. Consider the vector A and set of M labeled instances {ai, bi}1M. The classifier predicts the class label of A on the predefined N classes. The KNN classification algorithm finds the k nearest neighbors of A and determines the class label of A using majority vote [91] . KNN classifier applies Euclidean distances as the distance metric.

Dist (x,y)=$\sqrt{\sum_{I=1}^{D}(X_i - Y_i)}$ (2)

4.3 Experimental Setup

To evaluate our approach, a set of experiments was designed and conducted. In this section we describe the experiments design including the hotel corpus, the preprocessing stage, the feature selection used methods and evaluation metrics. Fig (4.2) show the experiment steps :

4.3.1 Preprocessing

After getting the data associated with hotel domains (751 positive ,339negative) obtaining vector representations for the terms from their textual representations by performing TFIDF (Term Frequency–Inverse Document Frequency) weight which is a well-known weight presentation of terms often used in text mining[93], the sentences are tokenized, stop words removed and Arabic light stemmer applied .

4.3.2 Sampling technique

Shuffled sampling: The Shuffled sampling builds random subsets of the dataset. Examples are chosen randomly for making subsets.

Stratified sampling: The Stratified sampling builds random subsets and ensures that the class distribution in the subsets is the same as in the whole data Set.

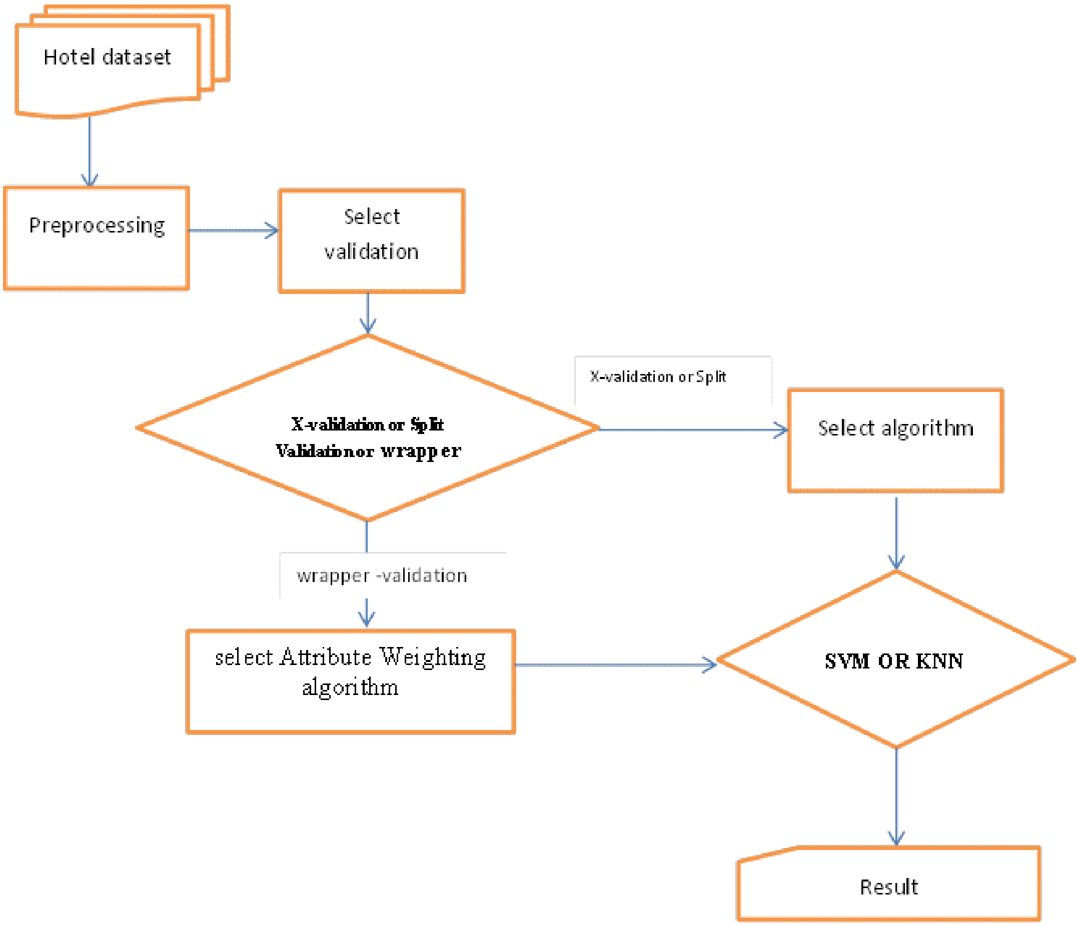


Figure 4.2 Method of ASC

4.3.3 Validation method and feature selection

4.3.3.1 X-Validation

The X-Validation is a nested operator. It has two sub processes: a training sub process and a testing sub process. The training sub process is used for training a model. The trained model is then applied in the testing sub process. The performance of the model is also measured during the testing phase.

The data is partitioned into *k* subsets of equal size. Of the *k* subsets, a single subset is retained as the testing data set (i.e. input of the testing sub process), and the remaining $k - 1$ subsets are used as training data set (i.e. input of the training sub process). The cross-validation process is then repeated *k* times, with each of the *k* subsets used exactly once as the testing data. The *k* results from the *k* iterations then can be averaged (or otherwise combined) to produce a single estimation. The value *k* can be adjusted using the *number of validations* parameter.

4.3.3.2 Split Validation

The Split Validation is a nested operator. It has two sub processes: a training sub process and a testing sub process. The training sub process is used for learning or building a model. The trained model is then applied in the testing sub process. The performance of the model is also measured during the testing phase.

The data set is partitioned into two subsets. One subset is used as the training set and the other one is used as the test set. The size of two subsets can be adjusted through different parameters. The model is learned on the training set and is then applied on the test set. This is done in a single iteration, as compared to the X-Validation operator that iterates a number of times using different subsets for testing and training purposes.

4.3.3.3 Wrappers validation

Wrappers provide better results as regards final predictive learning algorithm accuracy than filters as feature selection is optimized for a particular learning algorithm. But as a learning algorithm evaluates every feature set considered, wrappers are very costly to run, and are intractable for large databases having many features Further, as feature selection is combined with a learning algorithm, wrappers are not as common as filters. They should also be re-run when moving from one learning algorithm to another.

4.3.4 Information Gain

The Information gain procedure calculates an instance’s probability because it is a segment border and compares it to a segment border probability where a feature has a specific value [72]. The higher the probability change, the more useful is the feature. This simple ranking process is regularly used in text categorization applications where voluminous data prevents the use sophisticated attribute selection techniques. Decreasing class entropy reveals additional class information provided by the attribute and is called information gain

4.4 Evaluation of sentiment classification

In general, the performance of sentiment classification is evaluated by using four indexes. They are Accuracy, Precision . The common way for computing these indexes is based on the confusion matrix as shown in table(4.1) where:

#	Predicted positive	Predicted negative
Actual positive instances	Number of true positive instances(TP)	Number of false negative instances(FN)
Actual negative instances	Number of false positive instances(FP)	Number of true negative instances(TN)

Table 4.1 confusion matrix

True class P (TP) - correctly classified into class P

False class P (FP) - incorrectly classified into class P

True class N (TN) - correctly classified into class N

False class N (FN) - incorrectly classified into class N

precision = TP / (TP + FP) (3)

recall = TP / (TP + FN) (4)

accuracy = (TP + TN) / (TP + TN + FP + FN) (5)

Accuracy is the portion of all true predicted instances against all predicted instances. An accuracy of 100% means that the predicted instances are exactly the same as the actual instances. Precision is the portion of true positive predicted instances against all positive predicted instances. Recall is the portion of true positive predicted instances against all actual positive instances. F1 is a harmonic average of precision and recall.

4.5 Experimental Results

This section presents the results of experiments using two different sampling techniques . Evaluation of opinion classification relies on a comparison of results on the hotel corpus .

First, we have evaluated the accuracy of the data sets using two validation methods x-validation with 10 folds and spilt validation with (70:30) ,it performs stemming and it runs without stemming . Table (1,2) gives the accuracy and precision with methods which are usually used in Arabic opinion mining which are: K-nearest neighbor (kNN), Naïve Bayses (NB), and support vector machine (SVM). The (SVM) gives the higher accuracy (77.13%) with x-validation method and stratified sampling technique.

Second evaluations are calculated with using wrapper validation. This wrapper improve the accuracy of (NB) form 68.22% to 72.23% with x-validation and stratified sampling but it decreases the accuracy of SVM form 77.13 to 73.19%. and the KNN algorithm their accuracy increases with shuffled sampling form (69.46 to 69.55) with x-validation.

The first observation ,the use of stemming does not increase the accuracy expect when it used with split validation in two sampling techniques.

The other observation that use of x-validation is more effect than spilt validation expect with wrapper and with stemming.

Fig (4.3) shows that the SVM gives a good accuracy with stratified sampling and fig (4.4) show that the KNN gives a good accuracy with stratified sampling and shuffled sampling ,fig (4.5) show that the NB gives a good accuracy with wrapper.

As result the observation is that the use of stemming doesn't improve the accuracy of model.

Table 4.2 The accuracy with stratified sampling

Algorithm	X-Validation		X-Validation with stemming		Split Validation		Split Validation with stemming	
	Accuracy	precision	Accuracy	precision	Accuracy	precision	Accuracy	precision
SVM	77.13%	80.04%	76.92%	76.89%	76.21%	73.91%	76.90%	72.15%
KNN	69.46%	64.91%	69.97%	69.94%	64.14%	48.96%	64.48%	49.49%
NB	68.22%	70.50%	68.84%	64.23%	67.24%	68.42%	71.03%	78.12%

Table 4.3 The accuracy with shuffled_ sampling

Algorithm	X-Validation		X-Validation with stemming		Split Validation		Split Validation with stemming	
	Accuracy	precision	Accuracy	precision	Accuracy	precision	Accuracy	precision
SVM	76.38%	79.42%	75.66%	72.86%	75.17%	92.68%	73.79%	79.25%
KNN	69.55%	68.95%	68.83%	66.35%	63.45%	50.63%	62.41%	48.72%
NB	66.85%	63.59%	69.02%	65.20%	63.79%	56.25%	66.21%	62.86%

4.6 Conclusion

Sentiment classification at the document level provides an overall opinion on an entity, topic or event. This level of classification has some shortcomings for applications. In many applications, the user needs to know additional details, e.g., what aspects of entities are liked and disliked by customers ,It does not perform such fine-grained tasks, which require in-depth natural language processing , so document sentiment classification failed to extract such details for more details ,then go to feature based level which will discussed in the next chapter.

Table 4.4 The stratified sampling

Algorithm	Wrapper X-Validation		Wrapper X-Validation with stemming		Wrapper Split Validation		Wrapper Split Validation with stemming	
	Accuracy	precision	Accuracy	precision	Accuracy	precision	Accuracy	precision
SVM	73.19%	73.52%	75.57%	81.82%	74.48%	74.14%	75.86%	77.59%
KNN	68.33%	55.19%	68.22%	55.63%	68.62%	65.71%	68.97%	60.71%
NB	72.23%	75.59%	70.99%	67.92%	65.86%	100.00%	66.55%	100.00%

Table 4.5 shuffled sampling

Algorithm	Wrapper X-Validation		Wrapper X-Validation with stemming		Wrapper Split Validation		Wrapper Split Validation with stemming	
	Accuracy	precision	Accuracy	precision	Accuracy	precision	Accuracy	precision
SVM	72.86%	72.59%	74.21%	75.22%	72.41%	77.55%	74.48%	82.35%
KNN	66.44%	52.64%	68.83%	56.34%	66.55%	52.64%	68.83%	56.34%
NB	71.16%	73.94%	72.00%	70.45%	63.79%	100.00%	64.83%	100.00%

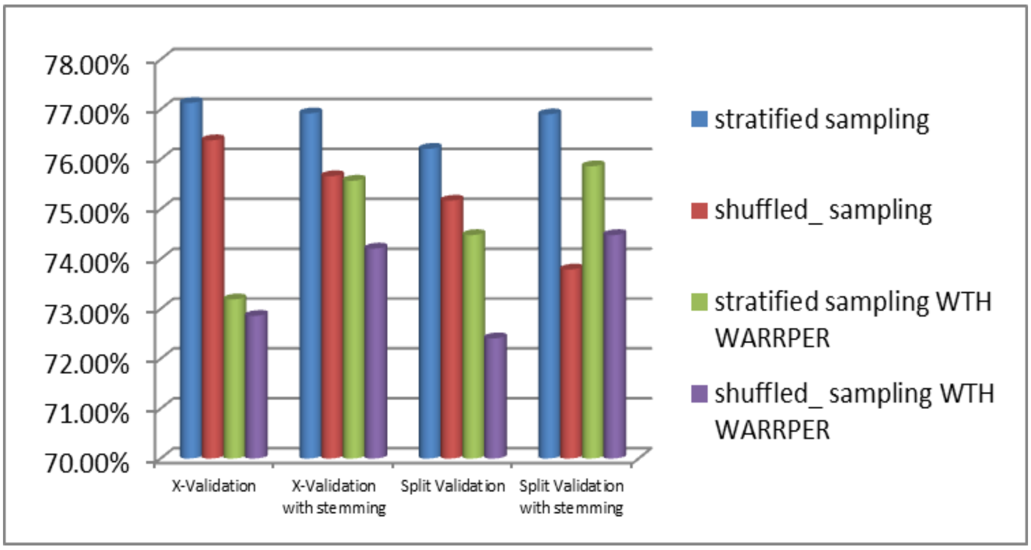


Figure 4.3 Summary of SVM accuracy with different sampling method

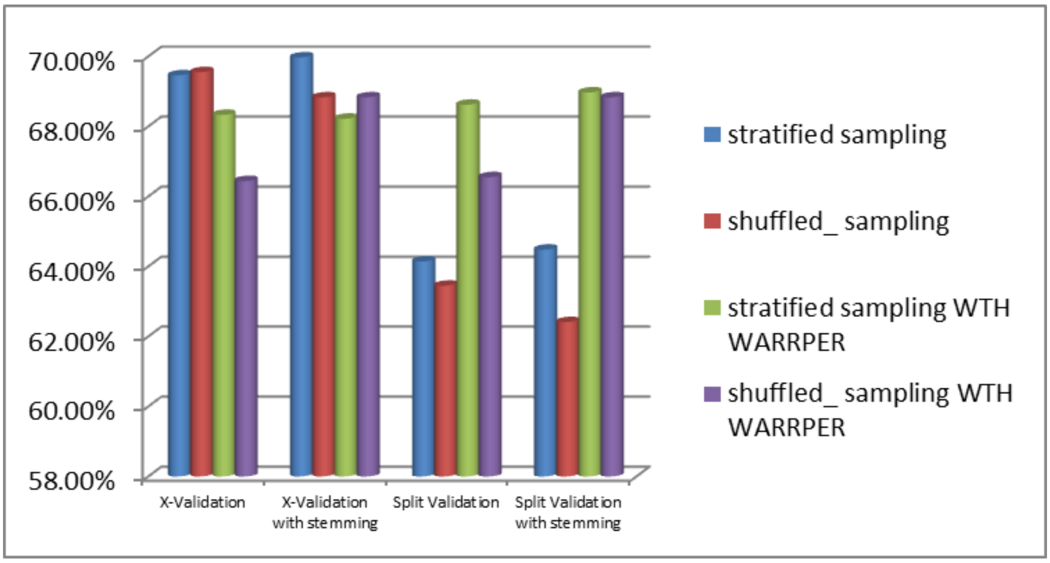


Figure 4.4 Summary of KNN accuracy with different sampling method

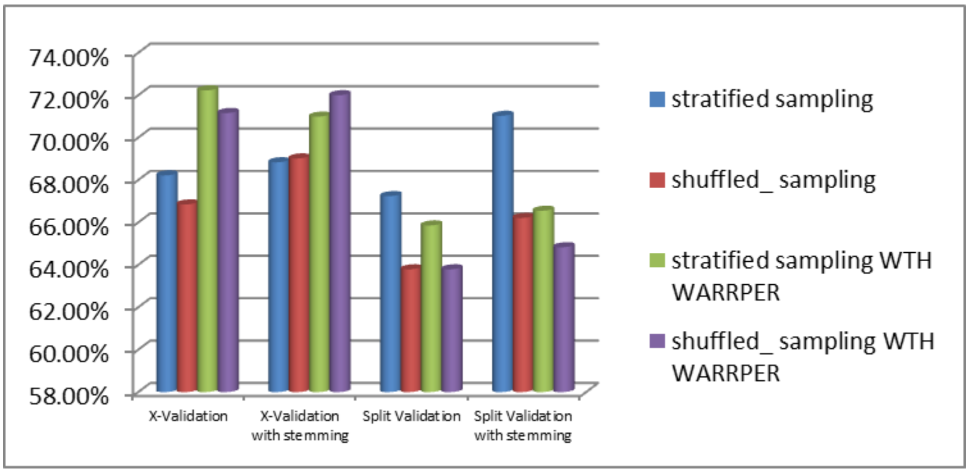


Figure 4.5 Summary of NB accuracy with different sampling method

CHAPTER FIVE

5 Creation of Opinion Lexicon

5.1 Introduction

The Arabic Sentiment Analyzer (ASA) framework consists of two main modules: language resource constructions and feature based model. Language resource construction consists of: opinion corpus and a set opinion word that indicate positive or negative.

This chapter deals with a set of opinion words which are called "opinion lexicon". The opinion lexicon plays a central role in feature based model to generate a summarization of the opinion applications [7-9, 33, 48, 49, 68, 94-97]. On the other hand, it is well known that there is no universally optimal opinion lexicon since the polarity of words is sensitive to the topic domain.

The step of generating lexicon is concerned with Arabic grammar of sentence see appendix (1). There are many different approaches used for generating lexicon. These are : manual approach, dictionary-based approach, corpus-based approach, and multilingual/translation approach.

Semantic orientation for each opinion word needs to be identified to be used to predict the semantic orientation of each feature in an opinion sentence. The semantic orientation of a word indicates the direction that the word deviates from the norm for its semantic group. Words that encode a desirable state (e.g. جميل) have a positive orientation, while words that represent undesirable states have a negative orientation (e.g., قبيح). While orientations apply to many adjectives, there are also those adjectives that have no orientation (e.g. أصفر)[98]. In this chapter, only positive and negative orientations will be discussed.

Semantic orientation information for each word is not contained in WordNet or dictionaries. Previous work on detecting semantic orientation depended on using a supervised learning algorithm to gather the semantic orientation of adjectives from constraints on conjunctions [31]. However, this approach relies on statistical information of large corpus, and needs a large amount of manually tagged training data.

Turney,[8], calculated semantic orientation of each phrase using mutual information between phrase and given word “excellent” minus the mutual information between the given phrase and the word “poor”, however, they do not define the semantic orientations of individual words/phrases in their results. Moreover, this technique also relies on statistical information from a rather big corpus.

In this thesis a dictionary–based approach will be used. A simple and yet an effective method by utilizing the adjective synonym set and antonym set in online dictionary to create an opinion lexicon and predict the semantic orientations of word (adjective) will be proposed.

The rest of this chapter is organized as follow: section two defines the features of opinion mining, whereas section three gives a brief description of different approaches of creating lexicon, , section four describes the proposed method section five presents evaluation results, and lastly section six gives the conclusion.

5.2 Features of Opinion Mining.

Feature engineering is an extremely basic and essential task for Opinion Mining. Converting a piece of text to a feature vector which is the basic step in any data driven approach to Opinion.[2]. The feature extraction phase deals with feature types (which identifies the type of features used for opinion mining).

Types of features used for opinion mining could be:

A. Term Presence vs. Frequency:

Term frequency has always been considered essential in traditional Information Retrieval and Text Classification tasks. But it is found that term presence is more important to Sentiment Analysis than term frequency. Pang [7] achieved better performance by using presence rather than frequency. That is, binary- valued feature vectors in which the entries merely indicate whether a term occurs (value 1) or not (value 0). This finding may be indicative of an interesting difference between typical topic-based text categorization and polarity classification: While a topic is more likely to be emphasized by frequent occurrences of certain keywords, overall sentiment may not usually be highlighted through repeated use of the same terms.

B. Parts of Speech:

In opinion mining the most commonly used is Part-of-speech (POS) information. . One of the most significant reasons of using POS is that they can be responsible for a crude form of word sense disambiguation. Hatzivassiloglou and Wiebe[98] revealed a high correlation between the presence of adjectives and sentence subjectivity. Adjectives have been employed as features by a number of researchers [99]. This finding has often been taken as evidence that adjectives are good indicators of sentiment, and sometimes has been used to guide feature selection for sentiment lexicon in that a number of approaches focus on the presence or polarity of adjectives when trying to create an opinion lexicon [13, 100] .

C. Term Position:

Words appearing in certain positions in the text carry more sentiment or weight than words appearing elsewhere. This is similar to IR where words appearing in topic Titles, Subtitles or Abstracts etc. are given more weight than those appearing in the body[101]. In many examples, although the text contains positive words throughout, the presence of a negative sentiment at the end sentence plays the deciding role in determining the sentiment.Hu[9] used the word position to find in frequent feature.

D. Negation:

Handling negation can be an important concern in sentiment analysis. When treating negation, one must be able to correctly determine what part of the meaning expressed is modified by the presence of the negation. Most of the times, its expression is far from being simple, and does not only contain obvious negation words, such as (ليس, ما). Researches in the field have shown that there are many other words that invert the polarity of an opinion expressed, such as diminishes / valence shifters e.g., "وجدت الغرفة ليست نظيفة", connectives "بالرغم من ان الغرف واسعة لكنها متسخة". As can be seen from these examples, modeling negation is a difficult yet an important aspect of sentiment analysis.

E. Syntactic Dependency Tree Patterns:

A syntax dependency tree is a syntax tree structure that is constructed by the syntax relation between a word (a head) and its dependents. Dependency structures identify useful semantic relationships. In syntactic dependency trees structures, each word or phrase is one leaf node, and two nodes are connected by one edge. [102] The relations among nodes are based on dependency grammars. The parent word is known as the head in the structure, and its children are known as modifiers. Many researchers have focused on this field to get efficient and accurate parsing tree patterns for sentiment analysis. Works such as (Collins ,Lin ,sha et al., Sang et al., Blache et al. ,Nakagawa et al.) [103] [43] [104] [105] [106] [107] have applied the syntactic dependency trees to sentiment analysis and obtained higher performance than using Bag-of-Word features. Words, phrases or patterns are usually given certain thresholds to be treated as features for machine learning models, the thresholds that measure effective frequency of occurrence. Syntactic dependency tree patterns are structured patterns, so they could occur very few times in a corpus, especially the longer syntactic patterns.

5.3 Approach Of Creating Lexicon

Identification of a term orientation usually falls in one of three approaches: a corpus-based approach, a multilingual/translation approach, and a lexicon/dictionary-based approach.

The first uses the relations encountered in large-corpora between words and expressions to determinate their polarity. Then it finds co-occurrence patterns of words to determine the sentiments of words or phrases, Works as [8, 31, 108] fall in this category. Their advantage is the possibility to identify multi-word opinion-bearing expressions but it requires a great amount of data to be processed (labeled training data) to achieve high accuracy. However, Arabic linguistic resources are not as available as other language, therefore, this method will be ignored.

The second approach is the multi-lingual and translation-based methods which explores available resources some languages, as in English, to be used in different language. This is the main advantages of these methods, since in some languages, linguistic resources are not available. Simple translation, however, using standard dictionaries or using machine translation, are not very efficient as most words have many different possible translations, depending on: the context, part-of-speech, etc[109]. They must deal, nevertheless, with the huge challenges of translating a word or expression to another language maintaining its original sense and this is much harder when it comes to Arabic due to the complexity of Arabic morphological.

The challenges are quite clear when one compares the gradually emerging Arabic to English Machine Translation with the rapidly developing Many-to-English Machine Translations available today. The lagging of the former maybe attributed to two reasons. The first is the frequent need for diacritical marks necessary for disambiguation of Arabic words. The second is the incompatibility of existing machine-translation techniques with the Arabic language.

Finally, the third approach explores the semantic relations annotated in resources such as thesauri and dictionaries using synonyms and antonyms in WordNet to determine word sentiments based on a set of seed opinion words. Representatives of such methods are the work of [13] who had made use of the WordNet relation of synonymy to determine polarity; or [22] who had used an online dictionary and the WordNet relations.

The advantage of using a dictionary-based approach is that one can easily and quickly find a large number of sentiment words with their orientations. Although the resulting list can have many errors, a manual check can be performed to clean it up. This cleanup will consume time (not as bad as people think, only a few days for a native speaker), more over it is a one-time only process.

The main disadvantage is that the sentiment orientations of words collected in this way are general or domain and context independent. In other words, it is hard to use the dictionary-based approach to find domain or context dependent orientations of sentiment words. Table (5.1) summarizes list of opinion lexicon approaches in different language

Table 5.1 Different method of creating opinion lexicon

Opinion Lexicon	**Language**	**Types of Words**	**Approach**	**Description**
OpinionFinder [110]	English	Adjective +noun+ verb	Manual approach	The lexicon was compiled from manually developed resources augmented with entries learned from corpora and it contains 6,856 unique entries that are also associated with a polarity label, indicating whether the corresponding word or phrase is positive, negative, or neutral.
SentiWordNet ([22]	English	Adjective +noun+ verb	Dictionary based approach	The lexicon contains 100,000 words. It was built on top of WordNet.
[111]	German	Adjective +noun+ verb	Manual approach	Extracts a list of 8,000 nouns, verbs, and adjectives in German annotated for polarity and strength.
([17]	English, French, and Hind	adjective	Dictionary based approach	Was built based on induction method which uses the WordNet graph and the relationships it entails to

				extend polarity classification using graph based semi-supervised learning algorithms
[23]	German	Verb+ adjective	Translation approach	Had built a lexicon for German starting with a lexicon in English, this time focusing on polarity rather than subjectivity.
[112]	French	noun, +adjective + verb	bootstrapping technique	Had used SVM classifier trained on a feature space produced from Latent Semantic Analysis over a large corpus in the new Language.
([25])	Japanese		Corpus based	Starting with one billion HTML documents, about 500,000 polar sentences were collected, with 220,000 being positive and the rest negative. Manual verification of 500 sentences, carried out by two human judges, indicated an average precision of 92%, which showed that reasonable quality can be achieved using this corpus construction method.
[113].	Romanian		translation	Generate a subjectivity lexicon for Romanian by starting with the English subjectivity lexicon (6,856 entries) from OpinionFinder and translating it using an English-Romanian bilingual dictionary
[32]	Japanese,		corpus-based method	Constructing polarity lexicons focusing on domain-specific

				proposition.
[39] .	Dutch		Translation approach and dictionary-based approach	Applying an online automatic translation system and the WordNet to improve the results.
[13]	English	Adjective	Semantic relations	Depended on the hypothesis that synonyms share the same semantic orientation. Had used an initial set of polar words - a seed set - that was expanded through the exploration of synonymy relations.
[100]	Arabic	Adjective	Manual approach	Had used a manually compiled list of approximately 4,000 Arabic adjectives from the newswire domain annotated for polarity.

5.4 The proposed method

5.4.1 Bootstrapping

The proposed method is able to quickly acquire a large opinion lexicon by bootstrapping from a selected seed. At each iteration, the seed set is expanded with related words found in an online dictionary, which are filtered by using a measure of word similarity. The bootstrapping process is illustrated in Fig (5.1).

5.4.2 Seed Set

The first step of the a proposal algorithm is to extract adjectives using association rule mine to find most frequent adjective of one item set with support of 0.1 and save this set as seed set (Past work has demonstrated that significant correlation with the presence of adjectives and subjectivity [98, 114]). The seeds are selected from two resources: the most frequent adjective in hotel corpus and telecommunications corpus.

The extracted adjectives seeds were merged and then manually classified into Negative and Positive seeds. Each type of seeds was saved in a separate file.

Table 2 shows a sample of the entries in the initial seed set of 101 that have been extracted. Although an isolated adjective may indicate subjectivity, there may be an insufficient context to determine semantic orientation. That is why the second step was to split these seeds set into two sets positive seed set (P_i-seeds) and negative seed set (N_i-seeds), containing words that indicate positive opinion and indicate negative opinion respectively, where I= {1,2} identifying the corpus. This step will be repeated for each seed.

Then, P_i-seeds will be merged to have one positive seed P-seed, as well as N_i-seeds to have one N-seed. After merging repeated words in the same seed will be eliminated and words that may appear in both P-seed and N-seed due to context depend meaning and domain depend, will be excluded. This step is done by two experts in the Arabic language. At the end of this stage the initial seed will be created.

A part-of-speech tagger was applied to the review beforehand.

This is step is done by tow expert in the Arabic language. First a part-of-speech tagger was applied to the review .

5.4.3 Bootstrapping Iterations

In general, adjectives share the same orientation as their synonyms and opposite orientations as their antonyms. Starting with the seed set, new related words are added based on the entries found in a dictionary[39].This idea is used to predict the orientation of an adjectives. To do this, the synset of the given adjective and the antonym set are searched. If a synonym or antonym has known orientation, then the orientation of the given adjectives could be set similarly.

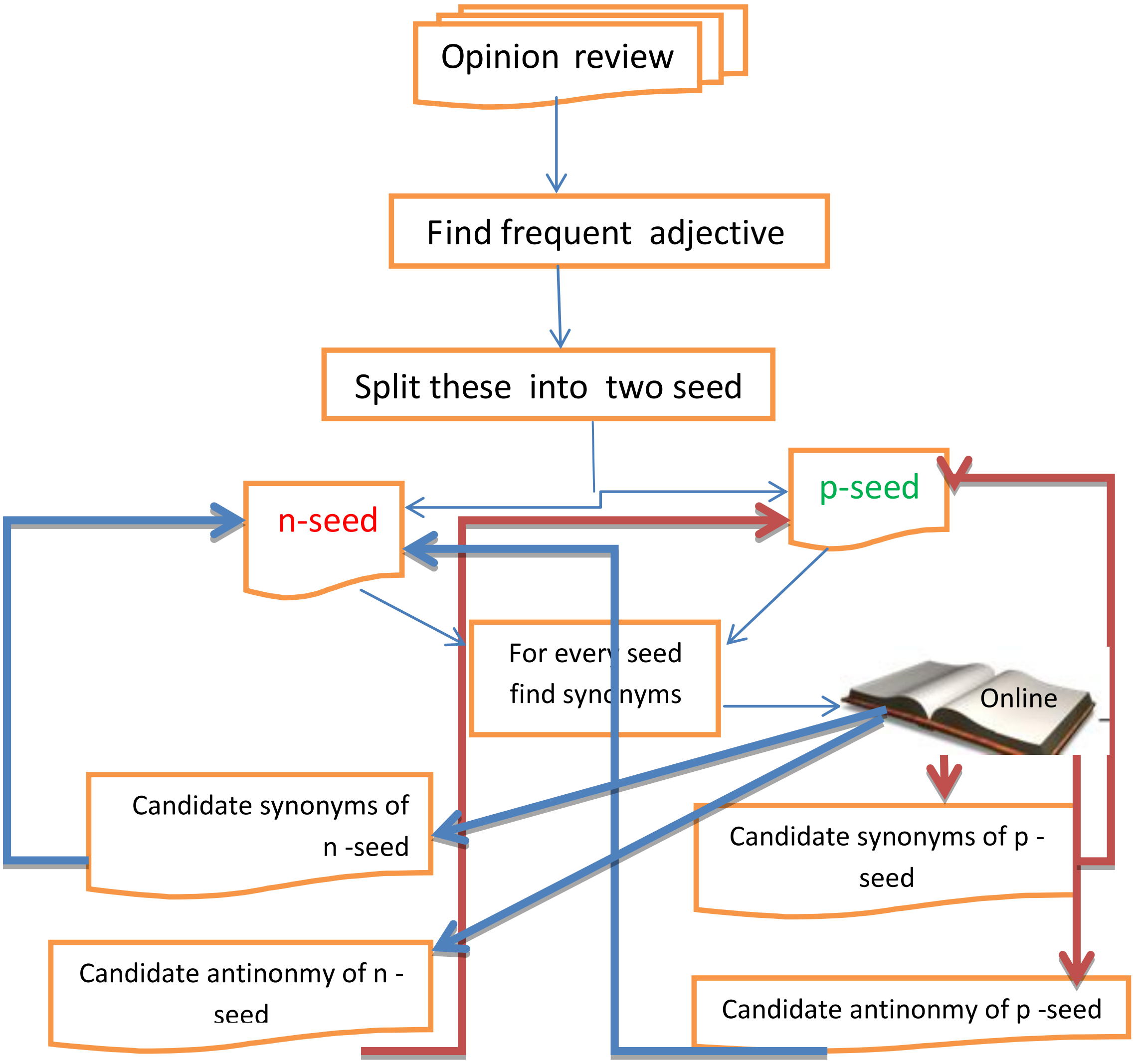

Opinion review
Find frequent adjective
Split these into two seed
n-seed
p-seed
For every seed find synonyms
Online
Candidate synonyms of n -seed
Candidate synonyms of p -seed
Candidate antinonmy of n -seed
Candidate antinonmy of p -seed

Figure 5.1 Method of creating lexicon

Table 5.2 Initial seed set

Positive word	Negative word
وحيد وجود واقع واسع هادئ نظيف نابض موجود مهم منفصل منظم مناسب مميز ممكن ممتع ممتاز ملائم مكتمل مقنع مقبول مغر معقول معتدل مشهور مشمول مزدوج مريح مرون مرغوب مرحب مرتب مرتاح مذهل مدهش محترم محبب متوقع متوفر متوسط متواصل متواجد متنوع متميز متكامل متعدد متعاون متاح مباشر لطيف لبق كويس كثير كبير كامل قيم قو قريب فسيح فخم ضخم صحيح سهل سعيد سريع زهيد رخيص رائع دقيق حديث جيد جميل جديد باهر انيق	منخفض ممل مليء مكلف معيب مزعج مزدحم مرتفع محدود متواضع متكرر متقلب متعب متدن قديم غال عاد ضعيف صغير صخب سيئ سخيف ردئ بعيد بطئ بسيط باهظ

Given enough seed adjectives with known orientations, the orientations of almost all the adjective words in the review collection, can be predicted.

Thus, our plan is to use previously extracted sets of seed adjectives (P-seed and N-seed) and then expand this set by searching in the an online dictionary. Abound finding a new adjective, the adjective’s orientation is predicted, and it will be added to the seed set. Next a new iteration begins. The iterative process ends when no more new adjectives can be found.

1. Procedure expand the lexicon (P_seed, N- seed, online dictionary)

2. Begin

```
3. For each adjective $W_I$ in p-seed
4. Begin
5. Find the synset of the way  in online dictionary
4   If ($W_I$ has synonyms in p-seed)
5   $w_i$'s orientation= s’s orientation;
6   Add $W_I$ with orientation top-seed;
7   Find all antonym of the way  in online dictionary
8   If  ($W_I$ has antonym a in n-seed)
9   $w_i$’s orientation = opposite orientation of a’s orientation;
10  add $w_i$ with orientation to n-seed; }
11  end for;
12  repeat this step {3-11}for n-seed
13  end
```

5.4.4 Filtering

In order to remove noise from the lexicon, we implemented a filtering step which is performed by calculating a measure of similarity between the original seeds and each of the possible candidates. We experimented with two corpuses based measures of similarity, namely the Point wise Mutual Information [115] and cosine similarity. After each iteration, only candidates with an PMI score higher than 0.5(deduced empirically) between the original seed set and the candidates are considered to be expanded in the next iteration

5.4.5 Semantic Orientation From Association

The Point wise Mutual Information (PMI) between two words, word1 and word2, is defined as follows[116]:

PMI(word1, word2) = log2p(word1 & word2) p(word1) p(word2)(1)

Here, p(word1 & word2) is the probability that word1 and word2 co-occur. If the words are statistically independent, then the probability that they co-occur is given by the product p(word1) p(word2). Thus a measure of the degree of statistical dependence between the words is the ratio between p(word1 & word2) and p(word1) p(word2). The log of this ratio is the amount of information that we obtain about the presence of one of the words when we observe the other

The semantic orientation of a given word is calculated from the strength of its association with a set of positive word, minus the strength of its association with a set of negative word we consider the strong positive and negative word is:

Positive word = "جيد"

Negative word = "سيئ"

SO(word) = PMI(word, “جيد”) - PMI(word, “سيئ”)(1)

The reference words “جيد” and “سيئ” were chosen because, in the five star review rating system, it is common to define one star as “سيئ”

After calculating semantic orientation of word in expand seeds with threshold of 0.5, if SO >0.5 the word is positive otherwise is negative. Then it is evaluated by measures such as accuracy, precision and recall where True positive (TP) - correctly classified into positive seed, False positive (FP) - incorrectly classified into positive seed, True negative (TN) - correctly classified into negative seed, False negative (FN) - incorrectly classified into negative seed. Table (3) show the accuracy of PMI

Table 5.3 The Accuracy of PMI

PRECITION	RECALL	ACCUARY
0.951773	0.965468	0.952922

5.5 Evaluation Result

For the evaluations which play an important role while suggesting a new method , we use a subjectivity lexicon obtained through several iterations of bootstrapping, with below mentioned strategies.

5.5.1 Human Judgment

This method is usually used for languages with limited resource. In this method, some manual annotators are appointed whose task is to tag the generated lexicon into positive and negative and compare the generated lexicon word , and the annotators usually is an expert in the Arabic language .

In this method of evaluation, we have appointed two manual annotators who are language experts in Arabic. We asked each annotator to tag the words generated by our system on the scale of 2 (negative:-1, positive:1). After getting the list annotated by all the annotators, we had two votes for each word and we took the majority call. Table 5.4 are reports accordance with Arabic lexicon generated using our system with manual annotation. Calculate the accuracy of each annotators.

among the annotators is that many words in Arabic show ambiguous nature. Their polarity depends on the sense in which they are used for e.g., "رخيص"is positive "ثمين"which indicates positive opinion but "رخيص"indicates positive opinion it also depends on context and domain.

Table 5.4 Agreement of our lexicon with the annotators

Annotator	accuracy	Seed set
Annotator 1	89.75%	for negative word
Annotator 1	94.32%	for positive word
Annotator 1	92.03%	for lexicon
Annotator 2	85.5%	for negative word
Annotator 2	94.20%	for positive word
Annotator 2	94.20%	for lexicon
Overall Agreement of our lexicon with the annotators		90.94%

5.5.2 Cosine Similarity

Cosine similarity is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval applications, that measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them, when comparing documents in text mining this measure is used . Given two vectors of attributes, A and B, the cosine similarity, θ, is represented using a dot product and magnitude as [117, 118]

When documents are represented as term vectors, the similarity of two documents correspond to the correlation between the vectors, are usually the TF vectors of the documents.

$$\text{similarity} = \cos\theta = \frac{A.B}{\|A\|\|B\|} \quad (2)$$

For this evaluation strategy, we perform Cosine Similarity on four seed set described in Section. On these seeds, we perform TD-IDF vector before determining the TD-IDF. It uses tokenize to split document into words and finds the stem of each word using the light Stemmer algorithm. It eliminates all stop words ,and only keeps nouns with non-stop word stems.

then calculates cosine similarity fig(5.2) shows the similarity graph . Table 3 reports the results of cosine similarity which shows that big similarity between initial positive seed and expand positive seed and less similarity between two initial seeds(p-seed and n-seed)

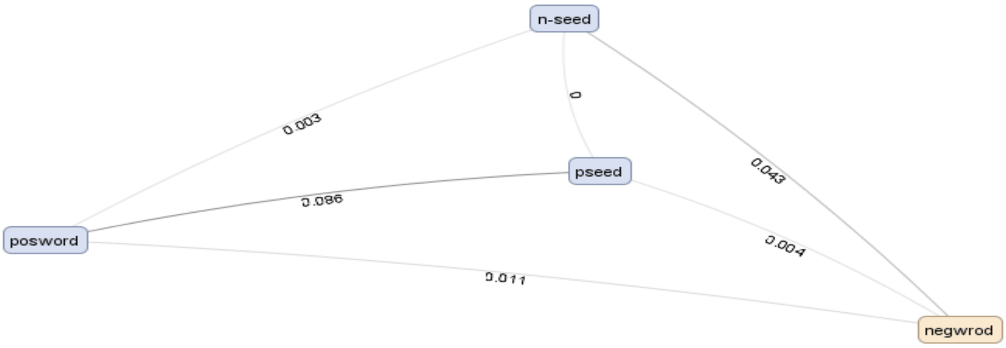


Figure 5.2 Similarity graph

Table 5.5 Cosine similarity between p-seed and n-seed with expand seeds

Seed set	Seed set	Similarity
negwrod	positive	4%
negwrod	negative	47%
negwrod	posword	11%
positive	negative	0.0%
positive	posword	86%
negative	posword	3%

5.6 Conclusion

We have presented our semi-automatic approach to construct Arabic sentiment lexicon using most frequent adjectives as initial seed and online dictionary to find synonyms and antinomies to expand the lexicon . In particular, we calculate PMI values between each expanded word positive (negative) seed words to filter these seeds . as result a lexicon of 1174 words (671 positive words,503 negative words)was created. Experiment results from two domains demonstrate that the lexicon generated with our approach has reached an excellent precision and could get many sentiment words in a special domain.

CHAPTER SIX

6 Construct Arabic Sentiment Analyzer(ASA)

6.1 Introduction

Identifying the feature of product review is crucial in opinion mining, which includes the extraction of product entities from product review. Potential customer finds it difficult to read the large database of these reviews in order to make a decision on whether to buy the product or not , as we have seen in the chapter 4 representation at document level does not give a good representation on what a customer like or dislike in granule level . Recent works observed that feature (aspect) word depends on the noun or noun phrase.

As discussed in the chapter 2, most of the early works on feature-based opinion mining are frequency based approaches ,which provide a good set of candidate aspects that needs to be filtered to get actual ones. Relation-based approaches use the feature-sentiment relationships to identify features and sentiments. One of the relationships that is mainly used is the syntactic relation between aspects and sentiments. In this work, we take advantage of both approaches and propose a method, called an Arabic Sentiment Analyzer , for identifying features and defining semantic orientations using opinion lexicon that created in the previous stage . A simple way to merge these approaches is to use a set of predefined syntactic patterns for filtering. However, syntactic patterns can only be used for the language and the type of text (full sentences, sentence segments, phrases, etc.) they are defined for. In other words, each language or text type has its own grammatical structure(see appendix) and therefore syntactic patterns.

In this research, we propose a method, called an Arabic Sentiment Analyzer, to mine and summarize opinions from the customer reviews , sentiment analyzer takes review texts as input, and outputs a set of aspects with their polarity.

Arabic sentiment analyzer first segment the review in sentence segment ,then use this segment as transaction to find frequent nouns or noun phrases ,and for filtering frequent noun phrases we mine a set of opinion patterns from the given text (review) .In addition, Arabic sentiment analyzer determines the polarity of each feature depending on Arabic Opinion lexicon (AOL) which was created in the chapter 5.

This chapter is organized as follows: In the next section, we describe Challenges of feature based opinion mining. Sections 6.3 presents feature based extraction using frequent pattern mining and 6.4 presents Arabic sentiment analyzer (ASA) . In Section 6.5 we report the results of our experimental evaluation on a dataset from two corpuses. Finally, Section 6.6 concludes this chapter with a summary and discussion.

6.2 Challenge of Feature Based Opinion Mining

The feature -based opinion mining seems to be facing different challenges : the first challenge in identifying aspects is that different reviewers may use different words or phrases to express the same aspect, e.g.,

- خدمة الاستقبال في الفندق مضيافة.
- موظف الاستقبال يقوم بمهام على بصورة سريعة.

Likewise, different reviewers tend to use different sentiments for expressing the same rating, e.g.

- فندق رايق وحلو وسعرها ممتاز , ويمتاز بالبساطة والقرب من كل شيء صراحه حلو وممتاز للي يستخدم المترو
- غير مقبول مواقف و وسيع الفندق سي جدا استفلالي غير نظيف الانترنت غير مجاني وبطي جدا الفطور بارد ومحدود الاختيارات

Another challenge is noisy information. Full text reviews normally include a large amount of irrelevant information, e.g., opinion about the manufacturer of the product and information about the reviewer.

- فندق محبب موقع ممتاز موقع جميل في شارع الشيخ زايد حيث يقرب منك الكثير من الاماكن تقع بجانبك خدمتين تموينيه الجمعية ومغسلة الملابس ودبي مول ومحطة المترو محطة الخليج التجاري

While explicit aspect/sentiment extraction has been studied extensively, limited research has been done on extracting implicit ones. However, there are many aspects/sentiments in reviews which are implicit

- غير صحى وليس مناسبا للأسر الإقامة غاية في السوء و لم تعجبني نهائيا و لن أكررها ثانية في هذا الفندق وكل شيء به غير صالح للاستعمال و محتويات الغرفة رديئة جدا ولا يوجد وحدة تحكم بالمكيف و يظل بارد جدا جدا جدا

The last challenge that we want to discuss here is Identifying opinions in comparative sentences is also very challenging. A comparative opinion expresses a relation between two or more items and/or a preference of the reviewer based on some shared aspects of the items, e.g.

- افاد ان الشبكة تضعف فى كثير من الاحيان فى المنطقة لزا يستخدم زين بصورة اكبر

6.3 Feature based extraction using frequent pattern mining

Feature-based opinion extraction system takes as input a set of user reviews for a specific product or service and produces a set of relevant feature. In general, the opinions can be expressed on anything, e.g., a product, an individual, an organization, an event or a topic. The general term “object” will be adopted to denote the entity as recommended in [3]. An object has a set of components and attributes or properties. For example, a network service is an object which has a set of components, i.e. a short-messaging-system (SMS), voice calls and internet. As well, an SMS has a set of attributes, e.g., SMS quality, price and reliability. The voice also has its set of attributes, e.g., local voice call price, international voice call price. In describing or criticizing a product, users do not usually use objects, even though they describe components and attributes. In the context of this paper, a feature is to represent both components and attributes.

If a feature F appears explicitly in an evaluative text T, it is called an explicit feature in T. For example, if the T is “الشبكة ضعيفة”“poor network connection”, then the explicit feature F is “network connection”. If F does not explicitly appear in T but lesser frequent than explicit ones. Consequently, only explicit features will be discussed beyond this point.

6.4 Arabic sentiment analyzer (ASA)

Most of primary works on feature-based opinion mining are frequency based approaches. They provide a set of candidate words, since some words could represent features and some are not, thus they still need filtering to get actual features. Relation-based approaches use the feature-sentiment relationships to identify features and sentiments. One of these relationships is mainly used as a syntactic relation between features and sentiments. Arabic Sentiment Analyzer (ASA) is proposed by taking the advantages of both approaches, for identifying feature and define semantic orientations using the Arabic Opinion Lexicon(AOL).

A simple way to merge these approaches is to use a set of predefined syntactic patterns for filtering, however syntactic patterns can only be used for the language and the type of text (full sentences, sentence segments, phrases, etc.) in which they are defined. In other words, each language or text type has its own grammatical structure and moreover its syntactic patterns.

fig (6.1) describes Arabic sentiment analyzer (ASA) that can mine and summarize opinions from customer reviews, it takes review texts as input, and outputs a set of features with their polarity. It first segment the review into segments ,then use these segments as transactions to find frequent nouns or noun phrases. Filtering frequent noun phrases is done by syntactic relation to group synonyms features.

6.4.1 Preprocessing step

The sentences are tokenized. Stop words removed. Obtained vector representations for the terms from their textual representations by performing (Term occurrences).

6.4.2 Parts-of-Speech Tagging (POS)

The Stanford POS tagger applied to produce tags for each word and identify simple noun and noun groups. For instance, e.g. <ROOT <SQ <S <S <NP <NN سعر> <NP <DTNNS المكالمات>>> <NP <NN داخل> <NP <DTNN الشبكة>>> <ADJP <JJ ممتاز>>> > < NN’> indicates a noun and <NP> indicates a noun phrase. The POS tagged information of each word is then saved in the transaction file.

6.4.3 Feature -based sentence segmentation

Customer reviews of products might be in incorrect syntactic form, sentence fragments, short phrases, or missing punctuations. The presence of adjectives in a sentence usually means that the sentence is subjective and contains opinions[5].

For a review sentence (opinionate sentence) that contains multiple feature (aspects), one of the key issues for feature-based opinion is to split such a multi-feature sentence into multiple single-feature units as the basis for feature-based opinion. To tackle this problem, we propose a feature segmentation model(FSM)

that takes a review sentence as input and produces single feature segments. For example, the review sentence:

"غير راضي عن خدمة الرسائل ,سعر الاتصال خارج الشبكة غالية . الشبكة متذبذب "

can be segmented into two single feature units, as :

"سعر الاتصال خارج الشبكة غالية " و " الشبكة متذبذب ".

Our first intuition is to treat single-feature segmentation by taking dependency relations for nominal sentence ((adjective (صفة) and noun (موصوف)).

Table(2) shows the relation between an adjective and the noun it describes, as well as the dependencies that relate pairs of nominal (predicate, apposition and specification). The compound relation is used to form numbers from single digit words.

6.4.4 N-gram model

The N-gram word model is a method that finds a series of consecutive word of length n. The most commonly used ones are unigram, bigram and trigram models. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram"; size 3 is a "trigram". Larger sizes are sometimes referred to by the value of n, e.g., "four-gram", "five-gram", and so on. For e.g. " سرعة المكالمات داخل الشبكة" its bigram will be as follows:

" سرعة المكالمات", "المكالمات داخل", "داخل الشبكة"

Table 6.1 1Dependency relations for nominal sentences

Relation	Arabic Name	Dependency	Dependent → Head
adj	صفة	Adjective	adjective → noun
poss	مضاف إليه	Possessive construction	second noun → first noun
pred	مبتدأ وخبر	Predicate of a subject	predicate → subject
app	بدل	Apposition	second noun → first noun
spec	تمييز	Specification	second noun → first noun
cpnd	مركب	Compound	second number → first number

6.4.5 Frequent Features Generation

Up to this step, features of more interest to customers will be extracted. To that end, a tool that discovers frequent patterns is used. In our context, an item set is a set of words or a phrase that occurs together. Association rules are used to discover correlations among a set of items in database. These relationships are based on co-occurrence of the data items rather than the inherent properties of the data themselves (as with functional dependencies). Association rule and frequent item set mining is becoming an extensive researched area resulting in development of faster algorithms. This thesis will use the fast and scalable algorithm, Frequent Pattern tree algorithm, FP-Growth[119]. Association rule mining takes this sentence as transaction but it is not suitable for this task because association rule mining is unable to consider the sequence of words, which is very important in natural language texts. Thus the pre-processing methods are used in order to find patterns to extract features (n- gram).The FP-Growth module was then applied to generate the frequent itemsets, the nouns or noun phrases that appear in more than 1% (minimum support). Table (6.2),(6.3) represent the candidate frequent features, which stored to the feature set for further processing. Figure 6.2 explains this method.

6.4.6 FP-Growth Algorithm

Let I ={a1, a2, . . . , am} be a set of items, and a transaction database DB=_T1, T2, . . . , Tn_, where Ti (i ∈ [1 . . . n]) is a transaction which contains a set of items in I . The support of a pattern A, where A is a set of items, is the number of transactions containing A in the database. A pattern A is frequent if A's support is no less than a predefined minimum support threshold, minsup . Given a transaction database DB and a minimum support threshold, the problem of finding the complete set of frequent patterns is called the frequent-pattern mining problem.

The FP-Growth algorithm allows frequent itemsets discovery without candidate generation and works in two steps. In the first step, it builds a compact data structure known as FP tree for itemsets from a set of transactions that satisfy a user-specified minimum support. In the second step, it extracts frequent itemsets directly from the FP tree. In addition, only frequent itemsets with maximum of four words will be considered since a product feature contains no more than three words.

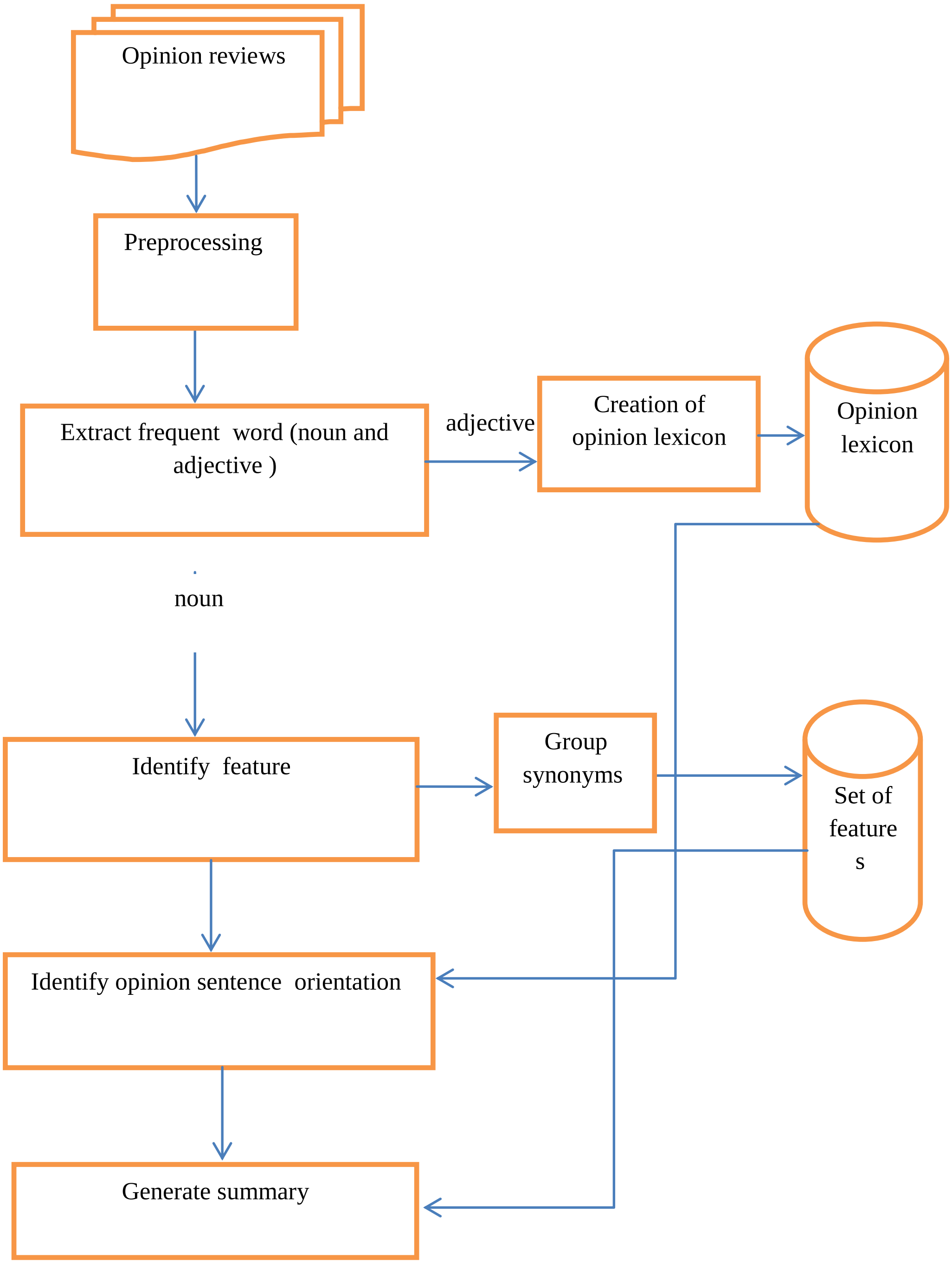


Figure 6.1 Arabic Sentiment Analyzer (ASA Method)

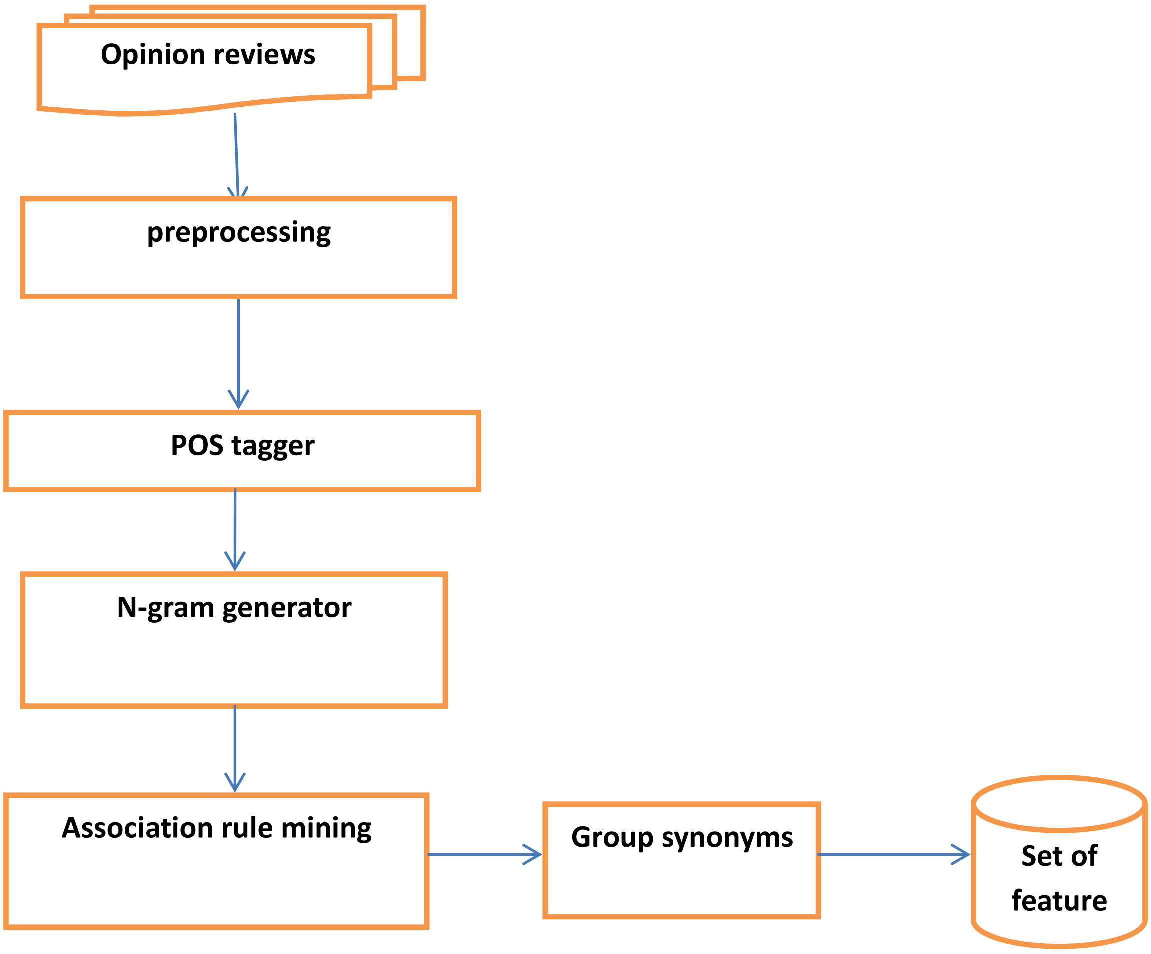


Figure 6.2 Method of extract feature

Table 6.2Extracted feature from Tel corpus

support	item
0.450	سعر_المكالمات
0.162	سرعة_الانترنت
0.155	سعر_المكالمات_خارج_الشبكة
0.150	سعر_المكالمات_داخل_الشبكة
0.144	سعر_المكالمات_العالمية
0.141	الخدمات_الاخرى
0.124	سعر_الرسائل

6.4.7 Grouping Candidate Features

Since different people use different words or phrases to express the same feature, grouping synonyms helps reducing the size of the extracted feature set. Although most of the previous methods do not consider feature grouping at all. More or less they use synonyms grouping, however, some synonyms might give many errors in the result of feature set generation. Using the syntactic role to group word. Using traditional Arabic grammar of iʿrāb (إعراب) which assigns a syntactic role to each word in a sentence. Pairs of syntactic units are related through directed binary dependencies table (6.1) shows the syntactic relation between noun and other words that define the first word e.g "سعر المكالمات" is Possessive construction "مضاف ومضاف اليه"

Table 6.3 Extracted feature form hotel corpus

support	item
0.562	الفندق
0.051	موقع_الفندق
0.085	خدمات_الفندق
0.048	الانترنت
0.044	طاقم_العمل

6.4.8 Summary Generation

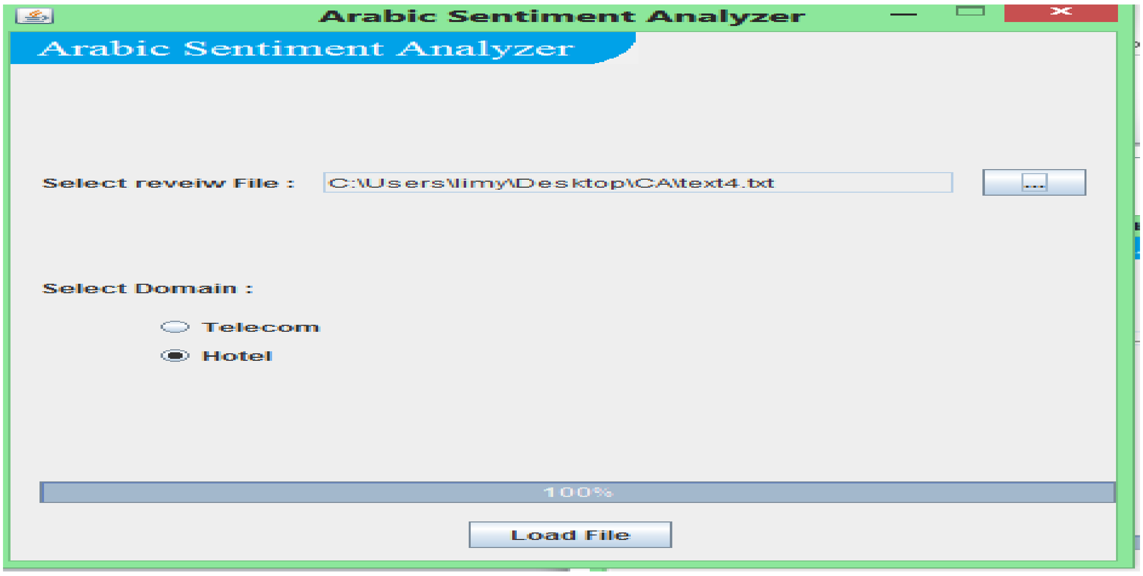
After all the previous steps, generating the A novel information summarizing based on and NLP technique, which is straightforward and consists of the following steps:

1. Fetch review from the directory (figure 6.3).
2. POS tagger are applied
3. Fragment the reviews in the sentences. Depending on adjective
4. Fetch a feature from a feature list(figure 6.4)

5. Assign weight to the each feature in sentences based on the opinion lexicon . (lexicon contains positive words, negative words)
6. If negation word is found it usually reverse the opinion expressed in a sentence. (Negation words include traditional words such as “no”, (eg.ليس, غير وما))
7. Sum up the weight(positive\negative) of the each feature to get weight and displayed
8. Sum up the positive\negative sentence of the each feature to get overall text summary.

Table 6.4 Association rules without n-gram

9. size	support	item1	item2	item3	item4
4	0.155	سعر	المكالمات	الشبكة	خارج



Select reveiw File : C:\Users\limy\Desktop\CA\text4.txt

Select Domain :

○ Telecom

● Hotel

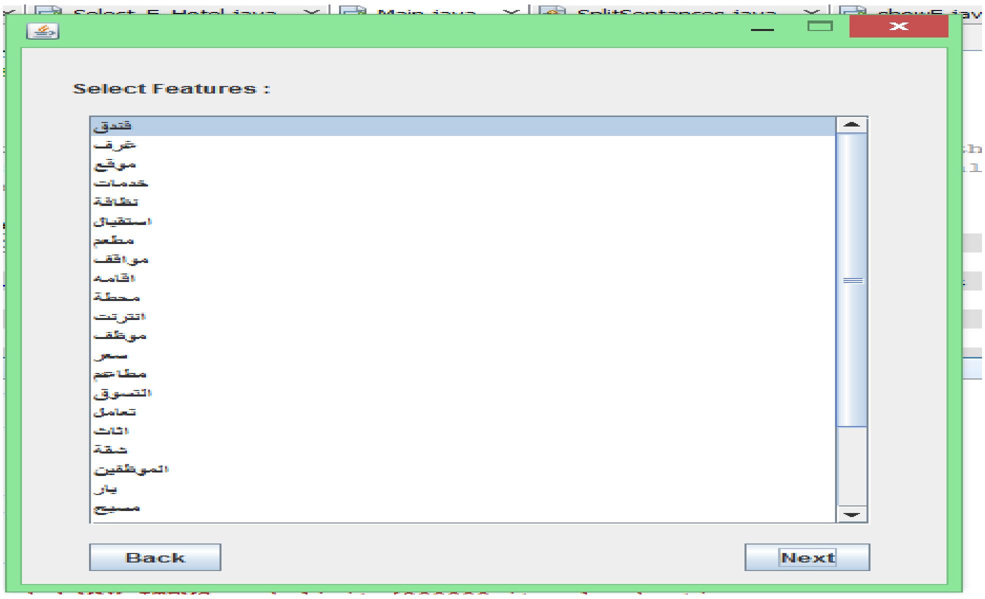
Figure 6.3 Select domain

Table 6.5 Association rules with n-gram

size	support	Item
1	0.155	سعر_المكالمات_خارج_الشبكة

6.5 Discussion of results

The proposed technique was evaluated to see how effective it is in identifying product features from a set of corpora that is automatically constituted by hotel and telecommunication companies reviews written in Arabic language which have collected. Each review is a short text .For Arabic scripts, some alphabets have been normalized (e.g. the letters which have more than one form) and some repeated letters have been cancelled, some of the wrongly spelt words are corrected.



فندق
غرف
موقع
خدمات
نظافة
استقبال
مطعم
مواقف
اقامه
محطة
انترنت
موظف
سعر
مطاعم
التسوق
تعامل
اثاث
شقة
الموظفين
بار
مسبح

Figure 6.4 Select feature

Table 6.6 Set of feature with same meaning

Feature
سعر المكالمات
سعرها
السعر
المكالمات سعرها

Table 6.7 Evaluation of extracted feature

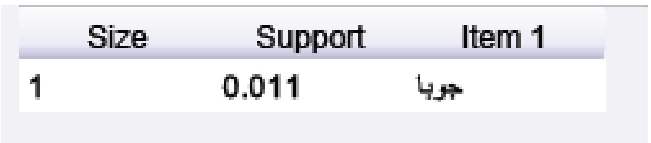
Data set	Extract feature	Actual feature	accuracy
Tel2	71	48	67.6%
Tel 1	21	18	85.7 %
All(tel1+tel2)	92	66	71.7%
Hotel	65	56	86.1%

Table (6.4) shows the execrated feature when using association rule without n-gram , the solution of this problem is shown in table (6.5). Table (6.6) represents a set of features with the same meaning to minimize the size of set of candidate feature this group of words will be " سعر المكالمات". Table (6.7) shows the number of frequent features generated for each company, column 1 lists the company name while column 2 lists the number of extracted features, column 3 gives the number of actual features and columns four gives the accuracy of frequent feature generated for each product. It can be observed that the accuracy of Tel2 is relatively low because the reviews were taken randomly and contain implicit features. While the accuracy of Tel1 is high because they concentrate their reviews on specific services, but the number of extracted features are small. However, there is a problem that some frequent nouns and noun phrases such as town names may not be real product features as seen in Fig 6.5). The accuracy of hotel corpus is relatively high compared with the telecom domain because the web side mentions the main service .

Fig (6.6) shows an example of summary for the feature “سرعة الانترنت”. Fig(6.7) shows that the proposed technique has handled the negation word .

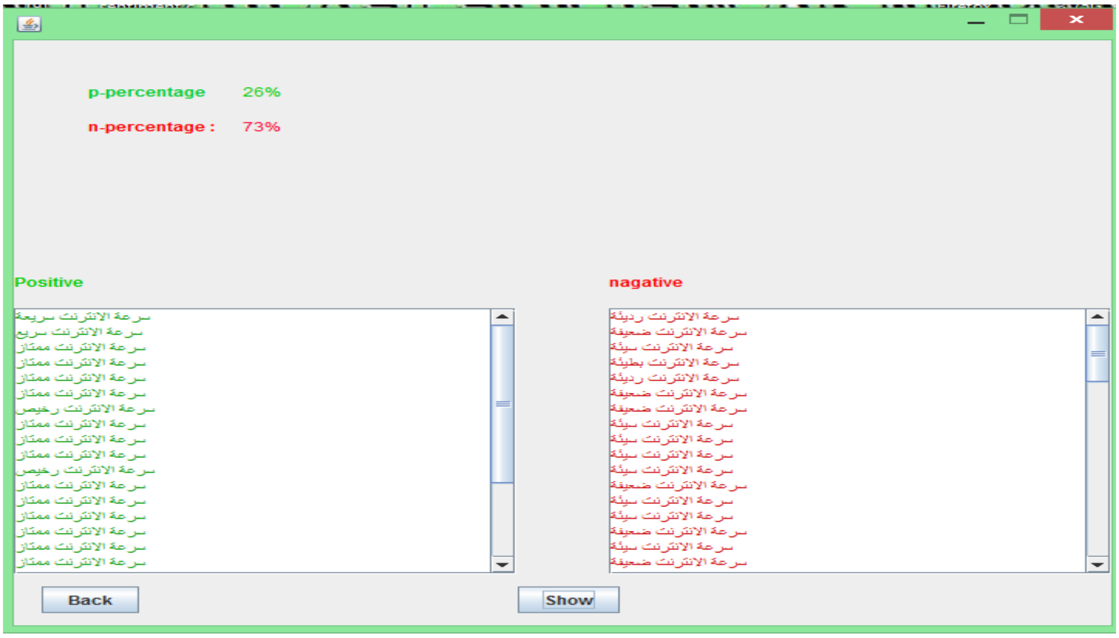
6.6 Conclusion

This chapter proposed a set of techniques for mining and summarizing product reviews based on data mining and natural language processing methods. The main objective is to construct ASA to provide a feature-based summary of a large number of Arabic customer reviews. The experimental results indicate that the proposed techniques are very talented in performing their tasks. And it believe that this problem will become increasingly important as more people are buying and expressing their opinions on the Web. Summarizing the reviews is not only useful to common shoppers, but also crucial to product manufacturers.



Size	Support	Item 1
1	0.011	جوربا

Figure 6.5 Town frequency



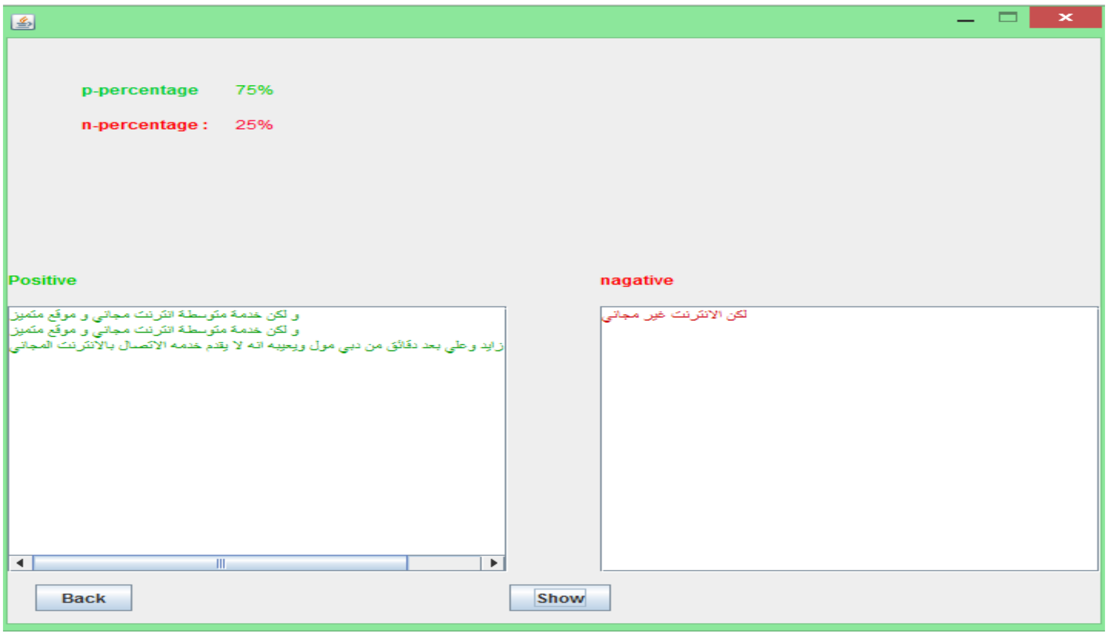
p-percentage 26%

n-percentage : 73%

```
سرعة الانترنت سريعة
سرعة الانترنت سريع
سرعة الانترنت ممتاز
سرعة الانترنت ممتاز
سرعة الانترنت ممتاز
سرعة الانترنت ممتاز
سرعة الانترنت رخيص
سرعة الانترنت ممتاز
سرعة الانترنت ممتاز
سرعة الانترنت ممتاز
سرعة الانترنت رخيص
سرعة الانترنت ممتاز
سرعة الانترنت ممتاز
سرعة الانترنت ممتاز
سرعة الانترنت ممتاز
سرعة الانترنت ممتاز
سرعة الانترنت ممتاز
```

```
سرعة الانترنت رديئة
سرعة الانترنت ضعيفة
سرعة الانترنت سيئة
سرعة الانترنت بطيئة
سرعة الانترنت رديئة
سرعة الانترنت ضعيفة
سرعة الانترنت ضعيفة
سرعة الانترنت سيئة
سرعة الانترنت سيئة
سرعة الانترنت سيئة
سرعة الانترنت سيئة
سرعة الانترنت ضعيفة
سرعة الانترنت سيئة
سرعة الانترنت سيئة
سرعة الانترنت ضعيفة
سرعة الانترنت سيئة
سرعة الانترنت ضعيفة
```

Figure 6.6 Summary Generation



p-percentage 75%

n-percentage : 25%

```
و لكن خدمة متوسطة انترنت مجاني و موقع متميز
و لكن خدمة متوسطة انترنت مجاني و موقع متميز
زايد وعلي بعد دقائق من دبي مول ويعيبه انه لا يقدم خدمه الاتصال بالانترنت المجاني
```

```
لكن الانترنت غير مجاني
```

Figure 6.7 Handel the negation word

CHAPTER SEVEN

7 Conclusion

Opinion mining has turned into a captivating investigation range because of the accessibility of an enormous volume of client produced ,substance e.g., reviewing web sites, forums, and blogs. This thesis has examined the issues of characterizing and assessing novel techniques for Sentiment Analysis that concentrate on the Arabic language.

The work represents the first endeavor to tackle this specific sort of issues for the Arabic language and the current results, can't be contrasted and whatever other exploration work done on the other language.

The feature-based opinion mining, which aims to extract item feature and Classify the opinion of customer into positive and negative classes and find the size of each class for each feature, is a relatively new sub-area that has attracted a great deal of attention recently. This thesis focuses on this problem because of its key role in the area of opinion mining. The extracted aspects not only ease the process of decision making for customers, but also can be utilized in other opinion mining systems.

Chapter 1,2,3 defines this problem formally and reviews the state-of-the-art approaches presented in the literature , creating an Arabic opinion Corpora . Chapter 4 has proposed a supervised algorithm, based on several different learning methods including Naive Bayesian ,KNN and SVM classifiers, with a view to determining the Overall Opinion

Polarity of a product review. In particular, we have identified a set of document representation features, partly borrowed from the literature of the English language, aimed at properly representing the ASC of a given document. Information Gain has been exploited as feature selection criteria in order to improve the accuracy of the classification activity. The result described in Chapter 4 show how the proposed approach works well on reports representing a single domain, with an accuracy, in the best case, of 77.13% ,however the Information Gain improves the accuracy of Naive Bayesian classifiers from 68.22% to 72.23% .

In Chapter 5, introduces Arabic Opinion Lexicon , for the considered problem . in chapter 6 ASA takes advantages of both frequency- and relation-based approaches to identify aspects and classifying the opinion . ASA finds the aspect-sentiment relations by mining a set of opinion patterns from reviews. Then, it uses the mined pattern to filter out non-aspects from frequent noun phrases. It also uses a novel technique for grouping synonymous aspects., ASA precisely determines the strength of positiveness or negativeness of an opinion feature by classifying them according to AOL and then generate a summary. Evaluation of results showed that combining the idea of frequency and relation-based approaches can effectively improve the accuracy of aspect extraction.

The work of this thesis is based on represents a first methodological approach to Sentiment Analysis, and more specifically for Arabic language. The two opinion corpuses and ASA framework, thanks to its modularity and flexibility, could be used in the future to investigate new methodologies and resources for Sentiment Analysis.

7.1 Future Research Directions

This thesis suggests many promising directions for future research in the field of Arabic opinion mining. In this section, we briefly discuss such directions

7.1.1 Directions in the area of natural language processing:

- **Identify Implicit aspects** : Most of the current works extract only explicit aspects. However, there are usually many types of implicit aspect expressions in a review. Adjectives and adverbs are perhaps the most common types because most adjectives describe some specific attributes or properties of entities, e.g., "غالى"describe "السعر " .Implicit aspects can be verbs too. In general, implicit aspect expressions can be very complex, e.g., "هذا الفندق في منتهى الضيافة" " منتهى الضيافة"indicates the feature of "طاقم الفندق". Although there have been some works considering extraction of implicit aspects, further research is still needed.

- **Expand opinion lexicon with more complex sentiments word:** Most sentiments are expressed through adjectives and adverbs. However, nouns (e.g., الجمال, القبح) and verbs (e.g., يكرهand يحب) can also be used to express sentiments.

- **Other types of opinionated documents** (such as a forum): other forms of opinion text such as forum discussions and commentaries are much harder to deal with because they are mixed with all kinds of non-opinion contents and often talk about multiple items and involve user interactions

- **Extraction of opinion phrases**: bag-of-opinion phrases models can outperform bag-of-words topic models and using the semantic relationship between words pays off for extracting opinion phrases. More sophisticated methods for extracting opinion phrases are needed to further improve the accuracy of aspect-based opinion mining.

REFERENCE

[1] Somprasertsri, G. Mining feature-opinion in online customer reviews for opinion summarization. *Journal of Universal …*, 162010), 938-955.

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

[3] Liu, B. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,*2010), 1-38.

[4] Liu, B. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5, 1 2012), 1-167.

[5] Liu, B. *Web data mining*. Springer, 2007.

[6] Lu, Y., Zhai, C. and Sundaresan, N. *Rated aspect summarization of short comments*. ACM, City, 2009.

[7] Pang, B., Lee, L. and Vaithyanathan, S. *Thumbs up?: sentiment classification using machine learning techniques*. Association for Computational Linguistics, City, 2002.

[8] Turney, P. D. *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews*. Association for Computational Linguistics, City, 2002.

[9] Hu, M. and Liu, B. Mining and Summarizing Customer Reviews. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004).*2004).

[10] Valitutti, A., Strapparava, C. and Stock, O. Developing Affective Lexical Resources. *PsychNology Journal*, 2, 1 2004), 61-83.

[11] Kim, S.-M. and Hovy, E. *Determining the sentiment of opinions*. Association for Computational Linguistics, City, 2004.

[12] Mohammad, S., Dunne, C. and Dorr, B. *Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus*. Association for Computational Linguistics, City, 2009.

[13] Kamps, J., Marx, M., Mokken, R. and Rijke, M. D. *Using wordnet to measure semantic orientations of adjectives*. City, 2004.

[14] Williams, G. K. and Anand, S. S. *Predicting the Polarity Strength of Adjectives Using WordNet*. City, 2009.

[15] Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A. and Reynar, J. *Building a sentiment summarizer for local service reviews*. City, 2008.

[16] Zhu, X. and Ghahramani, Z. *Learning from labeled and unlabeled data with label propagation*. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[17] Rao, D. and Ravichandran, D. *Semi-supervised polarity lexicon induction*. Association for Computational Linguistics, City, 2009.

[18] Hassan, A. Identifying Text Polarity Using Random Walks. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,*2010), 395-403.

[19] Hassan, A., Abu-Jbara, A., Jha, R. and Radev, D. R. *Identifying the Semantic Orientation of Foreign Words*. City, 2011.

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

[21] Esuli, A. and Sebastiani, F. *Determining the semantic orientation of terms through gloss classification*. ACM, City, 2005.

[22] Esuli, A. and Sebastiani, F. Determining term subjectivity and term orientation for opinion mining. *Proceedings of EACL*, 22006), 193-200.

[23] Kim, S.-M. and Hovy, E. *Identifying and analyzing judgment opinions*. Association for Computational Linguistics, City, 2006.

[24] Andreevskaia, A. and Bergler, S. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. *Proceedings of EACL*2006), 209-216.

[25] Kaji, N. and Kitsuregawa, M. *Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents*. City, 2007.

[26] Kaji, N. and Kitsuregawa, M. *Automatic construction of polarity-tagged corpus from HTML documents*. Association for Computational Linguistics, City, 2006.

[27] Velikovich, L., Blair-Goldensohn, S., Hannan, K. and McDonald, R. *The viability of web-derived polarity lexicons*. Association for Computational Linguistics, City, 2010.

[28] Dragut, E. C., Yu, C., Sistla, P. and Meng, W. Construction of a sentimental word dictionary. *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*2010), 1761.

[29] Peng, W. and Park, D. H. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. *Urbana*, 512004), 61801.

[30] Xu, G., Meng, X. and Wang, H. *Build Chinese emotion lexicons using a graph-based algorithm and multiple resources*. Association for Computational Linguistics, City, 2010.

[31] Hatzivassiloglou, V. and McKeown, K. R. *Predicting the semantic orientation of adjectives*. Association for Computational Linguistics, City, 1997.

[32] Kanayama, H. and Nasukawa, T. *Fully automatic lexicon expansion for domain-oriented sentiment analysis*. Association for Computational Linguistics, City, 2006.

[33] Ding, X., Liu, B. and Yu, P. S. *A holistic lexicon-based approach to opinion mining*. ACM, City, 2008.

[34] Wu, Y. and Wen, M. *Disambiguating dynamic sentiment ambiguous adjectives*. Association for Computational Linguistics, City, 2010.

[35] Lu, Y., Castellanos, M., Dayal, U. and Zhai, C. *Automatic construction of a context-aware sentiment lexicon: an optimization approach*. ACM, City, 2011.

[36] Takamura, H., Inui, T. and Okumura, M. *Extracting Semantic Orientations of Phrases from Dictionary*. City, 2007.

[37] Wilson, T., Wiebe, J. and Hoffmann, P. *Recognizing contextual polarity in phrase-level sentiment analysis*. Association for Computational Linguistics, City, 2005.

[38] Choi, Y. and Cardie, C. *Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification*. Association for Computational Linguistics, City, 2009.

[39] Jijkoun, V., de Rijke, M. and Weerkamp, W. *Generating focused topic-specific sentiment lexicons*. Association for Computational Linguistics, City, 2010.

[40] Du, W., Tan, S., Cheng, X. and Yun, X. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*2010), 111.

[41] Du, W. and Tan, S. Building domain-oriented sentiment lexicon by improved information bottleneck. *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*2009), 1749.

[42] Wiebe, J. and Mihalcea, R. *Word sense and subjectivity*. Association for Computational Linguistics, City, 2006.

[43] Lin, D. *Automatic retrieval and clustering of similar words*. Association for Computational Linguistics, City, 1998.

[44] Akkaya, C., Wiebe, J. and Mihalcea, R. *Subjectivity word sense disambiguation*. Association for Computational Linguistics, City, 2009.

[45] Su, F. and Markert, K. *From words to senses: a case study of subjectivity recognition*. Association for Computational Linguistics, City, 2008.

[46] Brody, S. and Diakopoulos, N. *Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs*. Association for Computational Linguistics, City, 2011.

[47] Feng, S., Bose, R. and Choi, Y. *Learning general connotation of words using graph-based algorithms*. Association for Computational Linguistics, City, 2011.

[48] Popescu, A.-M. and Etzioni, O. *Extracting product features and opinions from reviews*. Springer, City, 2007.

[49] Ku, L.-W., Liang, Y.-T. and Chen, H.-H. *Opinion extraction, summarization and tracking in news and blog corpora*. City, 2006.

[50] Moghaddam, S. and Ester, M. *Opinion digger: an unsupervised opinion miner from unstructured product reviews*. ACM, City, 2010.

[51] Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H. and Jin, C. *Red Opal: product-feature scoring from reviews*. ACM, City, 2007.

[52] Long, C., Zhang, J. and Zhut, X. *A review selection approach for accurate feature rating estimation*. Association for Computational Linguistics, City, 2010.

[53] Jeong, H., Shin, D. and Choi, J. FEROM: Feature Extraction and Refinement for Opinion Mining. *ETRI Journal*, 33, 5 2011), 720-730.

[54] Zhuang, L., Jing, F. and Zhu, X.-Y. Movie review mining and summarization. *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06*2006), 43.

[55] Somasundaran, S., Namata, G., Getoor, L. and Wiebe, J. *Opinion graphs for polarity and discourse classification*. Association for Computational Linguistics, City, 2009.

[56] Wu, Y., Zhang, Q., Huang, X. and Wu, L. *Phrase dependency parsing for opinion mining*. Association for Computational Linguistics, City, 2009.

[57] Qiu, L., Zhang, W., Hu, C. and Zhao, K. *Selc: a self-supervised model for sentiment classification*. ACM, City, 2009.

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

[59] Wei, W. and Gulla, J. A. *Sentiment learning on product reviews via sentiment ontology tree*. Association for Computational Linguistics, City, 2010.

[60] Jiang, L., Yu, M., Zhou, M., Liu, X. and Zhao, T. *Target-dependent twitter sentiment classification*. Association for Computational Linguistics, City, 2011.

[61] Boiy, E. and Moens, M. A machine learning approach to sentiment analysis in multilingual Web texts. *Information retrieval*2009), 1-30.

[62] Zhu, J., Wang, H., Tsou, B. K. and Zhu, M. Multi-aspect opinion polling from textual reviews. *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*2009), 1799.

[63] Liu, B., Hu, M. and Cheng, J. *Opinion observer: analyzing and comparing opinions on the web*. ACM, City, 2005.

[64] Baccianella, S., Esuli, A. and Sebastiani, F. *Multi-facet rating of product reviews*. Springer, City, 2009.

[65] Jin, W., Ho, H. H. and Srihari, R. K. *OpinionMiner: a novel machine learning system for web opinion mining and extraction*. ACM, City, 2009.

[66] Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zhang, S. and Yu, H. *Structure-aware review mining and summarization*. Association for Computational Linguistics, City, 2010.

[67] Mei, Q., Ling, X., Wondra, M., Su, H. and Zhai, C. *Topic sentiment mixture: modeling facets and opinions in weblogs*. ACM, City, 2007.

[68] Titov, I. and McDonald, R. A joint model of text and aspect ratings for sentiment summarization. *Urbana*, 512008), 61801.

[69] Korayem, M., Crandall, D. and Abdul-Mageed, M. *Subjectivity and sentiment analysis of arabic: A survey*. Springer, City, 2012.

[70] Ahmad, K., Cheng, D. and Almas, Y. *Multi-lingual sentiment analysis of financial news streams*. City, 2006.

[71] Almas, Y. and Ahmad, K. *A note on extracting 'sentiments' in financial news in English, Arabic & Urdu*. City, 2007.

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

[73] Elhawary, M. and Elfeky, M. Mining Arabic Business Reviews. *2010 IEEE International Conference on Data Mining Workshops*2010), 1108-1113.

[74] Farra, N., Challita, E., Assi, R. A. and Hajj, H. Sentence-Level and Document-Level Sentiment Mining for Arabic Texts. *2010 IEEE International Conference on Data Mining Workshops*2010), 1114-1119.

[75] Rushdi-Saleh, M. and Martín-Valdivia, M. Bilingual Experiments with an Arabic-English Corpus for Opinion Mining. *Proceedings of Recent Advances in Natural Language Processing*2011), 740-745.

[76] El-Halees, A. *ARABIC OPINION MINING USING COMBINED CLASSIFICATION APPROACH*. City, 2011.

[77] Al-Subaihin, A. A., Al-Khalifa, H. S. and Al-Salman, A. S. *A proposed sentiment analysis tool for modern arabic using human-based computing*. ACM, City, 2011.

[78] Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A. and Perea-Ortega, J. M. OCA: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology*, 62, 10 2011), 2045-2054.

[79] Abdul-Mageed, M. and Korayem, M. *Automatic identification of subjectivity in morphologically rich languages: the case of Arabic*. City, 2010.

[80] Abdul-Mageed, M. and Diab, M. T. *AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis*. City, 2012.

[81] Abdul-Mageed, M. Subjectivity and sentiment analysis of Modern Standard Arabic. *Proceedings of the 49th …*2011), 587-591.

[82] Abdul-Mageed, M., Kübler, S. and Diab, M. *Samar: A system for subjectivity and sentiment analysis of arabic social media*. Association for Computational Linguistics, City, 2012.

[83] Mountassir, A., Benbrahim, H. and Berrada, I. An empirical study to address the problem of Unbalanced Data Sets in sentiment classification. *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*2012), 3298-3303.

[84] Elarnaoty, M., AbdelRahman, S. and Fahmy, A. A Machine Learning Approach For Opinion Holder Extraction In Arabic Language. *arXiv preprint arXiv:1206.1011*2012).

[85] Misbah, A. M. and Imam, I. *Mining opinions in Arabic text using an improved "Semantic Orientation using Pointwise Mutual Information" Algorithm*. IEEE, City, 2012.

[86] Itani, M. M., Hamandi, L., Zantout, R. N. and Elkabani, I. *Classifying sentiment in arabic social networks: Naïve search versus Naïve bayes*. IEEE, City, 2012.

[87] Wiebe, J. *Learning subjective adjectives from corpora*. City, 2000.

[88] Bruce, R. F. and Wiebe, J. M. Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, 5, 2 1999), 187-205.

[89] Hu, M. and Liu, B. Opinion extraction and summarization on the web. *Proceedings of the national conference on artificial …*2006), 1621-1624.

[90] Abbasi Moghaddam, S. *Aspect-based opinion mining in online reviews*. Applied Sciences: School of Computing Science, 2013.

[91] Russell, S. Artificial Intelligence: A Modern Approach Author: Stuart Russell, Peter Norvig, Publisher: Prentice Hall Pa2009).

[92] Friedman, J., Hastie, T. and Tibshirani, R. *The elements of statistical learning*. Springer Series in Statistics New York, 2001.

[93] Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24, 5 1988), 513-523.

[94] Wilson, T., Wiebe, J. and Hwa, R. *Just how mad are you? Finding strong and weak opinion clauses*. City, 2004.

[95] Kobayashi, N., Inui, K. and Matsumoto, Y. Opinion mining from web documents: Extraction and structurization. *Information and Media Technologies*2007).

[96] Pang, B. and Lee, L. Opinion mining and sentiment analysis. *CSI Communications* /2008), 22-23.

[97] Gamon, M., Aue, A., Corston-Oliver, S. and Ringger, E. *Pulse: Mining customer opinions from free text*. Springer, City, 2005.

[98] Hatzivassiloglou, V. and Wiebe, J. M. Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of the 18th conference on Computational linguistics -*, 12000), 299-305.

[99] Whitelaw, C., Garg, N. and Argamon, S. *Using appraisal groups for sentiment analysis*. ACM, City, 2005.

[100] Abdul-Mageed, M., Diab, M. T. and Korayem, M. *Subjectivity and Sentiment Analysis of Modern Standard Arabic*. City, 2011.

[101] Shelke, N. M., Deshpande, S. and Thakre, V. Survey of Techniques for Opinion Mining. *International Journal of Computer Applications*, 572012).

[102] Meng, Y. *Sentiment Analysis: A Study on Product Features*. 2012.

[103] Collins, M. *Three generative, lexicalised models for statistical parsing*. Association for Computational Linguistics, City, 1997.

[104] Sha, F. and Pereira, F. *Shallow parsing with conditional random fields*. Association for Computational Linguistics, City, 2003.

[105] Sang, E. F. Memory-based shallow parsing. *The Journal of Machine Learning Research*, 22002), 559-594.

[106] Blache, P. and Balfourier, J.-M. *Property Grammars: a Flexible Constraint-Based Approach to Parsing*. City, 2001.

[107] Nakagawa, T., Inui, K. and Kurohashi, S. *Dependency tree-based sentiment classification using CRFs with hidden variables*. Association for Computational Linguistics, City, 2010.

[108] Riloff, E. and Shepherd, J. *A corpus-based approach for building semantic lexicons*. City, 1997.

[109] Izwaini, S. Problems of Arabic machine translation: evaluation of three systems. *The British Computer Society (BSC), London*2006), 118-148.

[110] Wiebe, J. and Riloff, E. *Creating subjective and objective sentence classifiers from unannotated texts*. Springer, City, 2005.

[111] Clematide, S. and Klenner, M. *Evaluation and extension of a polarity lexicon for German*. City, 2010.

[112] Pitel, G. and Grefenstette, G. *Semi-automatic Building Method for a Multidimensional Affect Dictionary for a New Language*. City, 2008.

[113] Mihalcea, R., Banea, C. and Wiebe, J. *Learning multilingual subjective language via cross-lingual projections*. City, 2007.

[114] Wiebe, J. Learning subjective adjectives from corpora. *Proceedings of the National Conference on Artificial …*2000).

[115] Turney, P. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the Twelfth European Conference on Machine Learning (pp491-502)Berlin: Springer-Verlag.*2001), 491-502.

[116] Church, K. W. and Hanks, P. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16, 1 1990), 22-29.

[117] Huang, A. *Similarity measures for text document clustering*. City, 2008.

[118] Guntur, A. CLUSTERING BASED ON COSINE SIMILARITY MEASURE.

[119] Agrawal, R. and Srikant, R. *Fast algorithms for mining association rules*. City, 1994.

Appendix

A. Arabic language grammar

Introduction

Traditional Arabic grammar defines a detailed part-of-speech hierarchy which applies to both words and morphological segments. Fundamentally, a word may be classified as a nominal ism (اسم), verb fiʿil (فعل) or a particle ḥarf (حرف). The set of nominal include nouns, pronouns, adjectives and adverbs. The particles include prepositions, conjunctions and interrogatives, as well as many others. Morphological annotation in the Arabic divides words into multiple segments. Each segment is assigned a part-of-speech tag. These tags are detailed in the following sections. In addition to part-of-speech tags, each segment is annotated using a set of multiple morphological features

there are two basic types of sentence, based on what the sentence's first word . the verbal sentence, where the sentence's first word is a verb and subject follows ,which the subject is the subject of a verb. The verbal sentence is composed of verb "الفعل" which followed by "subject" "الفاعل".

Where the nominal sentence, where the sentence's first word is a noun or subject which represents the entity of (person, place, animal, etc.) about which the sentence is talking .and the others follow, which the subject is the topic .The nominal sentence is composed of "starting" " المبتدأ" which is followed by "information" "الخبر". Information is the part of the phrase to complete the information about starting.

Nominal

The first of the three basic parts-of-speech are the nominal ism اسم (literally "names" in Arabic). The tags for nominal in the Arabic are shown in table A.1:

	Tag	Arabic Name	Description
Nouns	N	اسم	Noun
	PN	اسم علم	Proper noun
Derived nominals	ADJ	صفة	Adjective
	IMPN	اسم فعل أمر	Imperative verbal noun
Pronouns	PRON	ضمير	Personal pronoun
	DEM	اسم اشارة	Demonstrative pronoun
	REL	اسم موصول	Relative pronoun
Adverbs	T	ظرف زمان	Time adverb
	LOC	ظرف مكان	Location adverb

Table A.1Part-of-speech tag set for nominal

Proper Nouns

Proper nouns are annotated using the PN tag in the Arabic languge. In Arabic orthography, there is no distinction between a proper noun and a noun, whereas in English these are written with the first letter capitalized. Proper nouns in Arabic are known by convention and through the fact that they have the grammatical property of being definite even though they do not carry the *al* ال determiner prefix. The set of proper nouns includes personal names such as "the prophet *ibrāhīm*". In Arabic, proper nouns as known as اسم علم.

Pronouns

Three types of pronoun are identified in the using the tags PRON, DEM and REL. The personal pronouns (PRON) are those which are found in English ("I", "we", "you", "them", "us") together with pronouns found only in Arabic language, such as those inflected for the dual or feminine (for example *antumā*, انتما "you two"). When segmenting words for morphological annotation, the PRON tag is also used to identify attached pronoun segments, which are suffixes that appear at the end of words. In the case of nouns these are possessive pronouns. For example "his book" is fused into a single Arabic word-form (*kitābuhu,*كتابه). Suffixed pronouns attached to verbs will be either subject pronouns or object pronouns.

The DEM tag is used to identify demonstrative pronouns ("this", "that", "these", "those"). In Quranic Arabic, these are termed *ism ishāra أسم إشارة*(literally, "the names of pointing"). The REL tag is used to identify relative pronouns which connect a relative clause to its main clause (for example "the book <u>that</u> you bought"). In Arabic grammar, relative pronouns are known as *ism mawṣūl* أسم موصول("the names of connection").

Verbs

The second of the three basic parts-of-speech is the verb. All verbs in the Arabic language are tagged using the V (verb) tag. Each verb is also annotated using multiple morphological features to specify conjugation. In Arabic, verbs can be conjugated according to three different grammatical *aspects* (perfect, imperfect and imperative) as well as moods of the imperfect (indicative, subjunctive and jussive). Nouns derived from verbs – such as active and passive participles – are tagged as N (noun) and are annotated using the "derivation" feature.

	Tag	Arabic Name	Description
Verbs	V	فعل	Verb

Table A.2. Verb part-of-speech tag

Adjectives

Arabic Adjectives(**صفة**) are words that describe or modify another person or thing in the sentence. And are closely related to nouns in Arabic language , and it is sometimes not straightforward to distinguish between the two as both carry the same morphological features.

There are two things we must remember about adjectives in Arabic:

Firstly, they come *after* the nouns that they describe, unlike in English, where they occur before the nouns. So, whereas in English we would say 'a narrow sword', in Arabic we say 'a sword narrow':

Secondly, the adjective must *agree* with the noun it describes in three ways:

1. Definiteness: If the noun is definite, its adjective must also be definite; if the noun is indefinite, its adjective must also be indefinite eg. السيارة الكبيرة

2. Gender eg. ولد صغير

3. Number eg. الوالدان الذكيان

A nominal tagged as an adjective will directly follow the noun that it describes.

Particles

The third of the three basic parts-of-speech is the particle. Particles include prepositions, *lām* لام prefixes, conjunctions and others.

Interrogative particles are tagged using INTG, which includes the independent particle *hal* and the prefixed interrogative *alif*. Negative particles in the Quranic Arabic corpus are tagged as NEG. Certain negative particles may place a following imperfect verb into the subjunctive or jussive mood. The VOC tag is used to identify vocative particles and prefixes such as in *yā-rabbi* يا ربي.In English this would be roughly translated using the archaic vocative particle "O", as in "O my Lord".

Negation

Negation in English is achieved by using the word "not" (be not, do not). In Arabic, there are many words that are used to form negative statements, each one having its specific uses and conditions. However, there are four principle negative words that are commonly used in modern standard Arabic. In standard Arabic, you simply insert ليس (laysa), conjugated to match the noun

Negative Words Commonly Used in Modern Formal Arabic		
Word		**Usage**
لَيْسَ	**lays(a)**	**Before the predicate in present tense be-sentences** **(sentences without verbs)**
	is not	
مَاْ	**maa**	

	not	
لَمْ	lam	Before verbs (past tense)
	≡ did not	
لَنْ	lan	Before verbs (future tense)
	≡ will not	
لا	laa	Before verbs (present tense & imperative)
	not	

Table A.3 Negative Words

The word *r̠ayru* غَيْرُ= "other than" is often used in a similar way; however, that word forms a genitive construction with the noun following it and will not be prefixed to it غَيْرُ مَسْؤُوْل.

Although it is seldom used in formal Arabic, negative *maa* ما is the most commonly used negative particle in the modern spoken dialects of Arabic.

Syntactic Relations

The traditional Arabic grammar of *iʿrāb* (إعراب) assigns a syntactic role to each word in a sentence. Pairs of syntactic units are related through directed binary dependencies. In the Arabic language these relations are represented as directed edges on dependency graphs. The following tables list dependencies which are used to relate morphological segments, words, phrases and clauses.

Nominal Dependencies

Relations between nominal are shown in tableA.4. These include the relation between an adjective and the noun it describes, as well dependencies that relate pairs of nominal (*predicate*, *apposition* and *specification*). The *compound* relation is used to form numbers from single digit words.

Relation	Arabic Name	Dependency	Dependent → Head
adj	صفة	Adjective	adjective → noun
poss	مضاف إليه	Possessive construction	second noun → first noun
pred	مبتدأ وخبر	Predicate of a subject	predicate → subject
app	بدل	Apposition	second noun → first noun
spec	تمييز	Specification	second noun → first noun
cpnd	مركب	Compound	second number → first number

Table A.4. Dependency relations for nominal

Verbal Dependencies

Verbs are related to their arguments through subject and object dependencies are shown in tableA.5, with certain special verbs taking a subject and predicate as arguments. Imperfect verbs (فعل مضارع) may form part of an imperative expression through the *imperative* and *prohibition* relations.

Relation	**Arabic Name**	**Dependency**	**Dependent → Head**
subj	فاعل	Subject of a verb	subject → verb
pass	نائب فاعل	Passive verb subject representative	subject representative → verb
obj	مفعول به	Object of a verb	object → verb
subjx	اسم كان	Subject of a special verb or particle	subject → verb or particle
predx	خبر كان	Predicate of a special verb or particle	predicate → verb or particle
impv	أمر	Imperative	imperfect verb → imperative particle
imrs	جواب أمر	Imperative result	result → imperative verb
pro	نهي	Prohibition	imperfect verb → prohibitive particle

Table A.5Dependency relations for verbs.

Phrases and Clauses

A preposition phrase is formed from a preposition and its genitive noun. Preposition phrase attachment is annotated through the *link* dependency. Conjunction particles relate two clauses as either a coordinating conjunction, or through a subordinating conjunction which introduces a subordinate clause. Another common pair of dependencies which relates clauses are the condition and result relations as shown in table A.6

Relation	Arabic Name	Dependency	Dependent → Head
gen	جار ومجرور	Preposition phrase	preposition → noun
link	متعلق	PP attachment	PP phrase → verb or noun
conj	معطوف	Coordinating conjunction	second phrase → first phrase
sub	صلة	Subordinate clause	subordinate clause → particle
cond	شرط	Condition	condition → conditional particle
rslt	جواب شرط	Result	result → conditional particle

Table A.6Dependency relations for phrases and clauses.

Several relations link a noun to its verb to form an adverbial expression Table A.7 shows this relations. In each of these constructions, the noun will always be found in the accusative case *manṣūb* (منصوب). These include accusatives of circumstance and purpose, the cognate accusative and the commutative object.

Relation	Arabic Name	Dependency	Dependent → Head
circ	حال	Circumstantial accusative	accusative → verb
cog	مفعول مطلق	Cognate accusative	accusative → verb
prp	المفعول لأجله	Accusative of purpose	accusative → verb
com	المفعول معه	Comitative object	accusative → verb

Table A.7Dependency relations for adverbial expressions

Particle Dependencies

Certain types of particle occur frequently in Arabic language, and due to their individual nature they are each assigned unique syntactic relations. For example, the vocative حروف نداء particles each affect the case ending of nouns that they modify according to different grammar rules. A list of dependencies for particles is shown in table below:

Relation	Arabic Name	Dependency	Dependent → Head
emph	توكيد	Emphasis	verb → emphatic particle
intg	استفهام	Interrogation	verb → interrogative particle
neg	نفي	Negation	imperfect verb → negative particle
fut	استقبال	Future clause	imperfect verb → future particle
voc	منادي	Vocative	noun → vocative particle
exp	مستثني	Exceptive	noun → exceptive particle
res	حصر	Restriction	noun → restriction particle
avr	ردع	Aversion	dependent → aversion particle
cert	تحقيق	Certainty	dependent → particle of certainty
ret	اضراب	Retraction	dependent → retraction particle
prev	كاف	Preventive	preventive particle → accusative particle
ans	جواب	Answer	dependent → answer particle
inc	ابتداء	Inceptive	dependent → inceptive particle
sur	فجاءة	Surprise	dependent → surprise particle
sup	زائد	Supplemental	dependent → supplemental particle
exh	تحضيض	Exhortation	dependent → exhortation particle

exl	تفصيل	Explanation	dependent → explanation particle
eq	تسوية	Equalization	verb → equalization particle
caus	سببية	Cause	imperfect verb → particle of cause
amd	استدراك	Amendment	dependent → amendment particle
int	تفسير	Intepretation	dependent → particle of intepretation

Table A.8 Dependency relations for particles

Dependency Graphs

The syntax of traditional Arabic grammar is represented in the Quranic Arabic corpus using dependency graphs. Graphs are mathematical structures which consist of *nodes* and *edges* which link nodes together. In linguistic terms, a dependency graph is a way to visualize the structure of a sentence by showing how different words relate to each other using directed links called *dependencies*. In most variations of dependency grammar the nodes of a graph consist of words. That is, only links between words are allowed. However in traditional Arabic grammar the basic syntactic unit is not always a word. In most cases the syntactic unit is a morphological segment and the grammar explains how various segments are related across words. A syntactic unit may also be a complete word (with all its morphological segments) or a continuous sequence of words (such as a phrase or clause). This flexible approach to dependencies allows relations to be described between word segments, entire words or between phrases. Figure is a simple dependency graph which describes the syntax relation . The graph shows a dependency relation between the words in the verse, with the link pointing from the left dependent node to the right head node

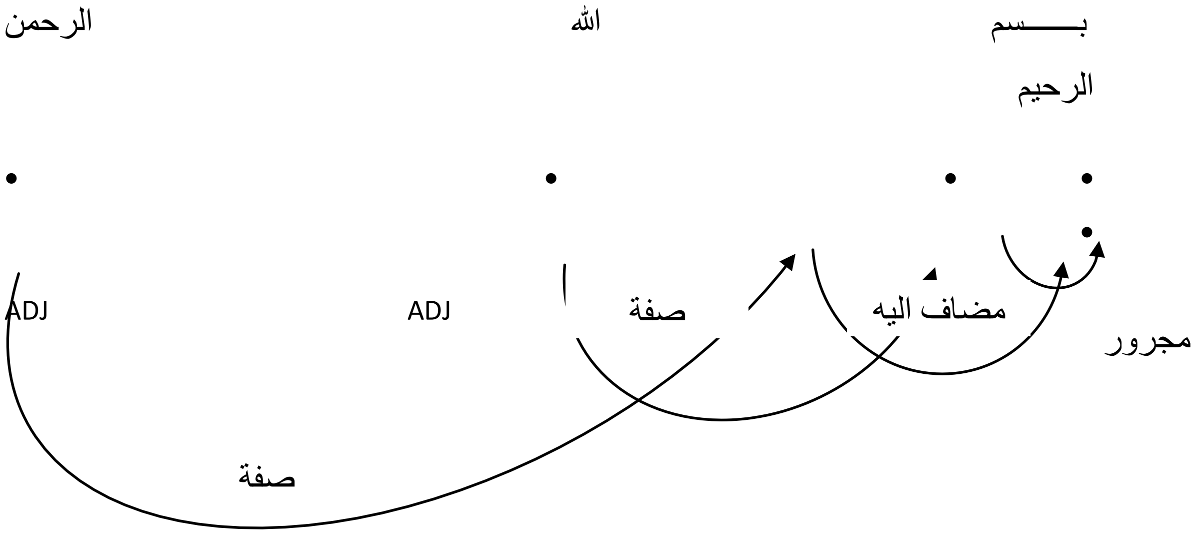


Figure1.Dependency graph