Benha University
Faculty of Engineering (Shoubra)
Electrical Engineering Department

Opinion Extraction for Arabic Reviews

A Thesis Submitted to

Electrical Engineering Department

Faculty of Engineering "Shoubra"- Benha University

In Partial Fulfillment of the Requirements

For the M.Sc. Degree in Computer Systems Engineering

Prepared By

Shimaa Ismail Mohamed Mostafa

B.Sc. in Computer Systems Engineering

Supervised By

Prof. Dr. Tarek El-Shishtawy

Faculty of Computers and Informatics

Benha University

Egypt

Assoc. Prof Abdulwahab Alsammak

Computer Engineering Departement

Benha University

Shoubra-Cairo, Egypt.

(2015)

Acknowledgments

Firstly, I would like to express my sincere gratitude to ALLAH - the most gracious, the most merciful for achieving this work and helping me to bear the difficulties and the pressures. All praise to ALLAH.

Many people supported me to finish this work in that form. I am grateful for the continuous support and advising from my supervisors: Prof. Tarek El-Shishtawy, and Prof. Abdulwahab Alsammak for their guidance and insight throughout the research. They improved and enhanced my work to be extracted in this accuracy. Without their assistance, this study would not have been successful.

I would like to thank my committee members for their comments and constructive feedbacks and all who believes me in their prayers. Also I would thank all my friends.

I am very grateful to my husband who convinced in my ability to realize this study in proper manner. He encouraged, supported me by his open mind and understanding spirit during the many hours I dedicated to achieve this work. To my son who fill my life happiness and indescribable sense of motherhood. To my parents who always encourage and pray for me all times. I would to thank all my family, sisters and brother who love and trust me a lot.

Finally, I hope this thesis would be a useful in the sentiment orientation field for Arabic language and for Arabic natural language processing.

Abstract

The web and social media contains millions of pages whose text review objects or events. It will be very helpful if one benefits of other's published opinions and experiences before taking decisions concerning these entities. Also, for opinions to be comprehensive, analysis should provide the attitude for the entity as well as its basic aspects or features. In this work, we propose a domain independent approach that extracts both of the entity aspects and their attitudes for Arabic reviews. The proposed approach does not exploit predefined sets of features, nor domain ontology hierarchy. Instead we add sentiment tags on the pattern and root levels of Arabic lexicon and used these tags to extract the opinion carrying words and their polarities.

The proposed approach relies on dividing the opinion mining task into three dependent subtasks at word, sentence, and document levels. The word level concerns with extracting the opinion carrying, negation, and intensifier words. The sentence level concerns with extracting the candidate aspects using syntactic patterns for Arabic sentences and based on the opinion-carrying words. The document level aggregates the lemma forms of the extracted aspects to summarize the entity orientation. The nondeterministic nature of some roots used in different ways in different domains affects the degree of sentiment role certainty. A certainty factor is proposed to express the percentage of orientation certainty of each aspect and declaring its effect on the system accuracy.

The proposed system is evaluated on the entity-level using a dataset of 500 movie reviews with accuracy 96%. Then the system is evaluated on the aspect-level using 200 Arabic reviews in different domains (Novels, Products, Movies, Football game events and Hotels). It extracted aspects, at 89% recall and 85% precision with respect to the aspects defined by domain experts. This proves that the proposed system can be used for generic domains beyond the limited coverage of existing ontologies.

Table of Contents

Chapter 4: Aspect-Based Opinion Mining

Chapter 5: Experiments and Results

Chapter 6: Conclusions and Future Work

References

Appendixes

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
ARBL	Arabic Root Based Lemmatizer
ASG	Arabic Similarity Graph
CF	Certainty Factor
EDU	Elementary Discourse Unit
EOU	Elementary Opinion Unit
IR	Information Retrieval
ME	Maximum Entropy
MSA	Modern Standard Arabic
NB	Naive Bayes
NE	Named Entities
Neg	Negative
Neut	Neutral
OM	Opinion Mining
PMI	Pointwise Mutual Information
POS	Part Of Speech
Pos	Positive
SA	Sentiment Analysis
SO	Sentiment Orientation
SSA	Subjectivity Sentiment Analysis
SVMs	Support Vector Machines

Publications

"Domain Independent Aspect-Based Sentiment Analysis for Arabic Reviews", Abdulwahab Alsammak, Shimaa Ismail, Tarek Elsheshtawy, submitted to Journal of Intelligent Systems (JISYS), July 2015.

Chapter 1

Chapter 1
Introduction

1.1. What Is Opinion Mining?

Opinion Mining (OM) or Sentiment Analysis (SA) is one of the most recent topics of research in the information extraction area. Opinion mining refers to a broad area of Natural Language Processing, Computational Linguistics Processing which is concerned with the opinions expressed in the documents and Text Mining that extracts information from the reviews in the web (Esuli and Sebastiani, 2006; Stavrianou and Chauchat, 2007; Harb et al. 2008). The basic concept is that people can benefit from the opinions and experiences of others through the growing availability of opinion resources such as online review sites and personal blogs (Pang and Lee, 2008). By extracting useful information from reliable amounts of feedback data in automatic or semi-automatic ways and presenting the information by the most effective way to serve the chosen objectives. This process is known as Opinion Mining.

Opinion Definition:

Opinion is a subjective statement or an attitude about an entity. Entity can be a product, service, person, event, organization, or topic. The attitude may be a judgment or an evaluation, their affective state (the emotional state of the author when writing) or the intended emotional communication (the emotional effect the author wishes to have on the reader).

Opinion Mining Tasks*:*

OM field is concerned with multiple tasks. 1) Determining the aspect-based opinion summarization which extract the aspects/features of the entity/object and arrange them according to its frequency in the reviews. 2) Determining the sentiment orientation (SO) or the polarity of the document into one of these classes positive, negative or neutral. 3) Determining the subjectivity of the document into two classes subjective or objective. 4) Determining the strength of document orientation which is strongly, mildly, weakly document. 5) Sentiment analysis of comparative sentences which compare the object with some other similar objects. 6) Identifying the opinion spam (Liu, 2010).

***Opinion Mining Utility*:**

OM is more suitable to various types of intelligence applications (e.g. businesses and organizations). Its appearance strongly associated with Web search or information retrieval. It is used for many decision making tasks such as obtaining an accurate opinion about a particular topic, improving the performance of a product or a service that presented by organizations, satisfying the customers, and others (Hu and Liu, 2006; Ding et al. 2009).

1.2. Motivation

When an individual needs to take a decision about something as purchasing any product, he/she typically asks for opinions from friends and families. Also, when an organization needs to find opinions of the general public about its products or services, it conducts surveys and focused groups to obtain a decision.

Now there are some opinionated documents on the World Wide Web (called user generated content) like blogs, forums, social media and social network sites. For a popular product, the number of reviews can be in hundreds or even thousands. This makes it difficult for a potential customer to read them to make a decision whether to buy the product or not. It also makes it difficult for the manufacturer of the product to keep track and manage customer opinions. Since for each manufacturer, there are many merchant sites may sell the product and he normally produces many kinds of products.

This leads to the importance of automatically mining the web content to summarize the opinions of users from a wide range of reviews, blogs, and tweets. For opinions to be comprehensive it is not sufficient to have opinion analysis only at the entity level. In many real-life applications, in order to make product improvements, one needs to know what components and/or aspects of the entity are liked and disliked by consumers. For instance, in a product review sentence, it identifies product aspects that have been commented on by the reviewer and determines whether the comments are positive or negative. For example, in the sentence, “The battery life of this camera is too short,” the comment is on “battery life” of the camera object and the SO is negative.

1.3. Problem Definition

Nowadays, most of the research works deal with the sentiment analysis of the documents but little work has been done for extracting the aspects of the object/entity especially for Arabic text. Many difficulties are faced when dealing with the aspect-based opinion summarization problems such as:

1. Expression: Since the opinion could express people emotions about something, it is hard to be expressed with keywords. Opinion may be written in unstructured-free-texts scheme and some written in the vernacular.

2. Domain Considerations: Most of the research concerns a specific domain and exploit a pre-built dictionary containing most of the opinion words concerned with this domain. As some opinion carrying words have different sentiment orientation in different domains. For example, the word "big, كبير" has a positive orientation in hotel domain and a negative orientation in technology domain. The objective is to propose a generic sentiment analyzer and without relying on a previously built opinion word lists.

3. Using Fixed List (Lack of context): Most of the aspect-based research used aspects list for a specific domain in extracting the aspects of the entity. This process is very tedious in case of domain independent sentiment analysis.

4. Negation: It represents the opposite emotion about something. As the negation words can change the meaning of the sentence as the word "NOT", it can lead to faulty orientation decisions as in the sentence "the product is not excellent" which does not mean that it is bad.

5. Intensification: It represents the degree of expressiveness of the opinion carrying words in the text. It takes many forms such as the use of adverbs (e.g. good) and/or some amplifiers (e.g., very). The overall sentiment result may be misleading due to the excessive use of intensifiers for some reviewers.

6. Resources Availability: Usually sentiment analysis for Arabic text suffers from the lack of available resources. Little resources are available for Arabic data sets, Lexicons, Stemmers, and Sentiment Analyzers.

1.4. Contribution

In this research, we presented a generic approach for automatically extracting the entity aspects and their attitudes. As we intended to analyze domain independent aspect level sentiments, the proposed approach does not exploit a predefined set of features, nor domain ontology hierarchy. Opinion tags are added to an existing accurate Arabic Root Based Lemmatizer ARBL lexicon (El-Shishtawy and El-Ghannam, 2012), at the root and pattern levels. This eliminates the need of opinion word lists and allows analysis for generic domains and entity types. Also, the proposed algorithm relies on a new task decomposition technique, based on the concept that each opinion has a target aspect or entity. Therefore, when an opinion carrying word is recognized, the algorithm scans the sentence to extract the intended target. The mining tasks are decomposed into the following subtasks:

1- Detecting the opinion-carrying words at word and sentence levels. This includes intensification and negation.
2- Exploiting the detected opinion-carrying words to extract the target noun phrases as candidate aspects or the general entity.
3- Extracting the entity aspects according to the syntactic patterns used for sentiment expressions.
4- Estimating the overall sentiment score and attitude by aggregating the orientations of the lemma-form candidate aspects.

1.5. Thesis Organization

The remaining content of the thesis is organized as follows:

Chapter 2: Introduces a survey for the related work and an overview for the main techniques used in sentiment analysis and aspect extraction.

Chapter 3: Presents the preprocessing methods and tools used to analyze the Arabic Review text on the word level and produce the Part Of Speech POS tags of each word. It includes Arabic lemmatizer and polarity lexicon.

Chapter 4: Presents the proposed generic approach for automatically extracting the entity aspects and their attitudes.

Chapter 5: Presents the Data Sets used to experiment the proposed approach. It illustrates the results of experiments compared to similar work and human experts with discussion.

Chapter 6: Presents the conclusions and also the problems that not solved yet. Besides to the suggestions for future work.

Chapter 2

Chapter 2
Background and Related Work

This chapter reviews the recently published works on Aspect-based opinion summarization. The sentiment analysis methods and identification of the sentiment strength will be discussed. Also, the techniques used for automatically extracting the entity aspects are introduced.

There are two types of textual information on the web; facts and opinions. Facts are objective statements about entities. Opinions are *emotional statements or thought that reflect people's attitudes about entities and objects*. Most of search engines dealt with facts that matched with topic keywords but little dealt with opinions.

Definition (Entity): *An entity can be a product, service, person, event, organization, or topic* (Zhang, 2012). For example: a reviewer can express an opinion on the entity itself as "Sony phones" in the expression "I do not like Sony phones.", or expressed on one of its aspects, e.g. "picture quality" in the expression "The picture quality of Sony phone is not good".

Definition (Aspect): *The aspects of an entity are the attributes, features or components of the entity* (Zhang, 2012).

Aspects are usually expressed as nouns and noun phrases, and can also be expressed as adjectives, adverbs, and verb phrases (Zhang, 2012). The aspects that expressed in nouns and noun phrases are called *Explicit Aspects*. For example, "sound" is explicitly appeared in "The sound of this phone is clear". But other expressions presented *Implicit Aspects*. For example, the adverb "heavy" in "The phone is too heavy." expressed on the implicit aspect "weight". There are many implicit aspect expressions for adjectives and adverbs, e.g., expensive for the (price) aspect, slow for the aspect (speed) ...etc. An example for the verb aspect is "lasts" in the sentence "The phone lasts all day" expressed on the battery of the phone. In this thesis, we focus on

extracting explicit aspect expressions, since most of aspect expressions in opinion documents are explicitly expressed.

The opinion has positive or negative attitude about an entity or its aspect. Positive, negative or neutral are called *Sentiment Orientation* SO. Other names for SO are opinion orientation, sentiment classification, semantic orientation, or polarity.

To determine the SO for the entities and their attributes, we need to define the review formats on the web. There are three different review formats which need different techniques to be handled (Liu et al. 2005).

Format (1) - pros and cons: The reviewer is asked to describe pros and cons separately as in C|net.com.

Format (2) - pros, cons, and detailed review: The reviewer is asked to describe pros and cons separately and also write a detailed review as in Epinions.com.

Format (3) - free format: The reviewer can write freely, that is, no separation of pros and cons as in Amazon.com.

For the review formats (1) and (2), opinion/semantic orientations (positive or negative) of the aspects are known because pros and cons are separated as in (Liu et al. 2005). The entity aspects only need to be determined. In our thesis, we concentrate on review format (3). Since we need to identify and extract both entity aspects and sentiment orientations.

This task goes to the sentence level to discover, what aspects of an object that people liked or disliked. For instance, in a product review sentence, it identifies product features that have been commented on by the reviewer and determines whether the comments are positive or negative. For example, in the sentence, "*The breakfast buffet in the hotel is variant.*" the aspect is "breakfast buffet" for the entity "hotel" and the SO is positive.

2.1. Sentiment Orientation

Earlier research works on opinion mining started by extracting the attitude of the whole object/subject as positive, negative or neutral. This task was commonly known as the *document-level sentiment classification* because it considered the whole document as the basic information unit, which assumed that the document was known to be opinionated as in (Pang et al. 2002; Turney, 2002). A positive document did not mean that the author had positive opinions on every aspect of the entity. Also, a negative document did not mean that the author disliked everything about the object. For example, in a product review, the reviewer usually wrote both positive and negative aspects of the product, although the general sentiment on the product could be positive or negative according to the maximum percentage of the polarity.

Likewise, the sentiment classification could be applied to individual sentences. However, each sentence couldn't be assumed to be opinionated in this case. It needed first to be classified as opinionated or not opinionated, which was called subjectivity classification. The resulting opinionated sentences were classified as expressing positive or negative opinions. It was called the *sentence-level sentiment classification* (Riloff et al. 2003; Wilson et al. 2004; Wilson et al. 2005). Wilson et al. (2004) classify each sentence in the document, and count the number of sentences that have positive or negative orientation. According to this number the whole text will be assigned as positive or negative. The concept was improved to find orientation at the phrase level for sentences that have multiple attitudes (Wilson et al. 2005).

2.1.1. Sentiment Orientation Classification

To determine the sentiment orientation for documents, several researchers have studied the problem of defining the opinion words (Liu, 2010) then classifying them to its category as positive, negative or neutral.

Definition (Opinion Word): An opinion word is a term used to refer to the word that usually qualifies an object or an attribute of this object. They are usually adjectives and adverbs, but they can also be nouns and verbs e.g., (beautiful, magnificent, nice, smooth, love, liked) for positive SO, (bad, terrible, damage, poor, hate) for negative SO.

Example: The adjective "clear" in the sentence "The sound of this phone is clear" is considered as an opinion word refers to the aspect "sound" with positive SO. Also the verb "hate" in the sentence "I hate this camera" is an opinion word refers to the entity "camera" with negative SO.

Most of the researchers were concerned with the automatic identification of opinion words. This is because the manual technique required a lot of human efforts and it was costly. Other approaches could be grouped into corpus-based and dictionary-based approaches for defining the opinion words or sentiment classification.

Corpus-based approaches considered syntactic and statistical properties such as word co-occurrence (Hatzivassiloglou and McKeown, 1997; Qiu et al. 2011; Khan et al. 2014). But it faced the problem of the domain dependent because of using seed list method. This method had a small seed list of adjectives as a start and expanded every time a new opinion word was found.

Hatzivassiloglou and McKeown, (1997) were the first researchers that tried to solve the problem of defining the opinion words. They used the corpus-based approach and classified these words by analyzing the pairs of adjectives conjoined by some constraints such as ('and', 'but', 'either-or' or 'neither-nor') that was extracted from a large unlabelled documents. Using these constraints, one adjective could infer opinion polarities of unknown adjective based on the known ones.

Double propagation method is proposed by Qiu et al. (2011). This method used dependency relations to extract both opinion words and product aspects. It used an initial set of opinion word seeds as the input then tried to find the relation between them and the target aspects. Also the opinion word seeds had relations among themselves too. This propagation or bootstrapping process ends when no more opinion words or aspects can be found. Also (Khan et al. 2014) used a list of subjective adjectives to define and classify the opinion words.

The opinion words that defined by Pang et al. (2002), were extracted manually. But they used several completely prior-knowledge-free supervised machine learning (ML) techniques (Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVMs)) for the sentiment classification. Although ML classifiers perform well, their

performance drops on topics or texts that are different from those that they were trained on (Gamon and Aue, 2005).

On the other hand of the supervised algorithm, (Turney, 2002) presented a simple unsupervised learning algorithm that evaluated each two consecutive words from the review if their tags conformed to predefined patterns. He categorized the seed list using point wise mutual information (PMI) to detect document sentiment based on selected phrases. These phrases are chosen via a number of pre-specified part-of-speech (POS) patterns from POS tagger of (Brill, 1994), that including adjectives and adverbs. PMI also measured the strength of semantic association between the two words by comparing its similarity to a positive reference word ("excellent") and its similarity to a negative reference word ("poor").

Dictionary-based approaches used hierarchies such as WordNet (*an online lexical reference system that organize words in synonym sets, called synsets*) in order to identify the sentiment orientation for the opinion words (Hu and Liu 2004; Esuli and Sebastiani, 2006; Kim and Hovy, 2006; Popescu and Etzioni, 2007). Sentiment dictionaries have a great role in determining the accuracy of sentiment analysis systems. But it faced the problem of lack of context information in these hierarchies (Hu and Liu, 2006a).

Hu and Liu, (2004) used the WordNet for extracting the opinion words. The adjectives were organized as bipolar clusters as the word with its synonym set and antonym set. They used NLProcessor linguistic parser to generate the POS tags of each word. Also (Liu et al. 2005) used this parser for tagging, but the opinion words were already classified, since they used pros and cons of review format 2. This opinion observer enabled the analysts to correct the errors using a convenient user interface which called semi-automatic tagging. (Popescu and Etzioni, 2007) used a mixture of WordNet information (e.g., antonyms are placed in the same cluster) and lexical pattern information (e.g., “clean, almost spotless” suggests that “clean” and “spotless” are likely to refer to the same property). They used a novel relaxation-labeling technique to determine the semantic orientation of potential opinion words as a tuple of (word, feature, and sentence) with a set of SO labels.

Somprasertsri and Lalitrojwong (2010) used existing dictionaries such as the General Inquirer (Stone, 1966). Dictionaries were built in different ways: manually, making use of existing resources, or automatically. In manual approach, a corpus of opinion-bearing words is built and manually tagged. For example, in the work of Taboada et al. (2011), a corpus of 400- review text was used to extract 2,252 adjective entries, 1,142 nouns, 903 verbs, and 745 adverbs. Terms were ranked in a single scale combining sentiment polarity and strengths, ranging from −5 for extremely negative to +5 for extremely positive. Some researchers (Hu and Liu, 2004a; Kim and Hovy, 2006) make use of WordNet as lexical resource. In SentiWordNet (Baccianella et al. 2010), all WordNet synsets are automatically annotated according to their degrees of polarity. Each term is annotated with three numerals: positive, negative, and neutral. The score for each word is calculated by its proximity with respect to one or more seed words.

2.1.2. Negation

Negation words are used to reverse the polarity of the opinion words e.g., (not good). Examples of the negation words are "not, none, nobody, never, lack, and nothing".

Many researcher handled the negation using different techniques as in (Kennedy and Inkpen, 2006; Taboada et al. 2011; Choi and Cardie, 2008). Always the negation words are found by searching backward from the opinion word till finding a punctuation marker. Then the orientation score will be reversed as if the adjective "good" has orientation value (+1), so "not good" will has a orientation value (-1).

2.1.3. Intensification

Intensification is the process of using special words besides the opinion words to increase or decrease the semantic orientation score. They usually are neighboring adverbs as in (Benamara et al. 2007). These intensifiers were classified into two categories; Amplifiers (e.g., very, too, extremely) and Downtoners (e.g., slightly)

(Quirk et al. 1985). Some researchers such as (Kennedy and Inkpen 2006; Polanyi and Zaenen 2006) have implemented the intensifiers using simple addition and subtraction. As the amplifiers are positive scores added to the positive orientation opinion words. But the downtoners are negative scores added to the negative orientation opinion words.

Taboada et al. (2011) handled the intensifiers by using different technique. They ranked each adverb (intensifier) by special modifiers. This modifier is added or subtracted from 100% according to the sign of this value, then multiplied the result by the opinion word score. For example the combination of "somewhat sleazy" will be computed as the SO of "sleazy" is -3, the modifier of "somewhat" is -30%, so the total weight is [- 3 * (100 % - 30 %) = -2.1].

2.1.4. Arabic Sentiment Orientation

Arabic is the official language of 23 countries and is spoken by more than 379 million people. Arabic is the fastest-growing language on the web (with growth rate of 5,296.6% of Internet users in 2014, compared to 2,721.8% for Russian, 1,910.3% for Chinese and 468.8 % for English)[1]. There are about 135.6 million Arabic users online, or about 4.8% of the global Internet population.

Arabic is a Semitic language and consists of many different regional dialects (Versteegh, 1997). While these dialects are true native language forms, they are typically used only in informal daily communication and are not standardized or taught in schools (Habash, 2010). There is one formal written standard that is commonly used in written media and education throughout the Arab world called Modern Standard Arabic (MSA). There is a large degree of differences between MSA and most Arabic dialects, and, interestingly, MSA is not actually the native language of any Arabic country or group. MSA is syntactically, morphologically, and phonologically based on Classical Arabic (CA) (Habash, 2010), which is the language of the Qur'an (Islam's Holy Book).

Arabic has a very rich inflectional system and is considered one of the richest languages in terms of morphology (Habash et al, 2009). Arabic sentential forms are divided into two types, nominal and verbal constructions (Farra et al. 2010). In the

[1] Internet world status (http://www.internetworldstats.com/stats7.htm.)

verbal domain, Arabic has two word order patterns (i.e., Subject-Verb-Object and Verb-Subject-Object). In the nominal domain, a normal pattern would consist of two consecutive words, a noun (i.e., subject) then an adjective (subject descriptor).

Recently, several efforts have been proposed for subjectivity and sentiment analysis for Arabic documents. Korayem et al. (2012) and Medhat et al. (2014) survey the different techniques used for subjectivity and sentiment analysis for Arabic. The sentiment analysis for Arabic texts was conducted by few researchers, may be due to the scarcity of the available resources. So we need to know the availability of annotated corpora and lexicons for training and testing to enable progress on sentiment recognition systems. Collecting these data (and particularly the annotations) can be very labor-intensive. Different corpora and lexicons used by many researchers will be reviewed.

Definition (Corpora): *is a collection of data sets (reviews) about specific topic such as movies, sports, and politics.* As the Penn Arabic Tree Bank (PATB) which is an existing collection of newswire stories in different domains (e.g. sports, politics, finance, etc.) and Opinion Corpus for Arabic (OCA) which is a corpus of text from movie review sites (Rushdi-Saleh et al. 2011).

Definition (lexicon): *is the vocabulary of language, or branch of knowledge* that assigns specific information to specific words or sentences as classifying the adjectives polarity into positive, negative or neutral also for classifying the sentences subjectivity into objective or subjective.

Abbasi et al. (2008) used a corpus of 1000 positive and 1000 negative movie reviews to test their approach. They used Entropy Weighted Genetic Algorithms to select language features for both Arabic and English. They used two types of features, stylistic features and lexical features and achieved an accuracy rate of 91%. Rushdi-Saleh et al. (2011) built an Opinion Corpus for Arabic (OCA) which contains 500 movie reviews, 250 of them considered as positive and other 250 as negative. They used both Support Vector Machines and Naive Bayes classifiers, reporting 90% F-measure on OCA using SVMs.

An Arabic Lexicon for Business Reviews was proposed by Elhawary and Elfeky (2010). They built corpora of dataset collected from 2000 URLs. They used about

1500 translated adjectives. As a starting of 600 positive words/phrases and more than 900 negative words/phrases collected by manually looking at the corpora of the reviews classifier training; and a seed list of almost 100 neutral words/phrases, collected from the top frequent Arabic words/phrases used over the web. The researchers performed the label propagation mechanism on an Arabic similarity graph ASG which was the core of the Arabic lexicon. ASG consisted of two columns where the first column was the word/phrase and the second column represented the score of the word which was the sum of the scores of all edges connected to this node. The positive and negative scores were normalized before adding them to compensate for the negative skew (bias) in the scores. Finally, the scores were filtered by eliminating the scores below some cutoff and the log is taken.

Lexicon-based opinion classifier proposed by El-Halees (2011) to define the sentiment orientation of the whole document. They initially used the word list that included in the SentiStrength software after translating it from English into Arabic to build their lexicon classifier. They improved the list to be more applicable to Arabic words and phrases by omitting some unrelated words, other common Arabic words and some synonym words for the words in the online dictionary. After using the lexicon classifier, they presented two other machine language approaches (Maximum entropy and k-nearest). As lexicon based method was used to classify as much documents as possible. The resultant classified documents were used as training set for maximum entropy method which subsequently classified some other documents. Finally, k-nearest method used the classified documents from lexicon based method and maximum entropy as training set and classified the rest of the documents.

Lazhar and Yamina (2012) also used a lexicon for the adjectives with their polarity (bag of sentiment words) as positive or negative to detect the semantic orientation of the overall content of a text. They presented a domain dependent analyzer that identified opinions for Arabic text using domain ontology. In their approach each concept and each property is associated to a set of labels that correspond to their semantics.

AWATIF is a multi-genre corpus for MSA presented by Abdul-Mageed and Diab (2012) which used a collection of data sets from three different resources:

PATB, Wikipedia user talk pages and conversation threads from web forums. They also used a lexicon for the sentiment classification that was created manually by defining the adjective's polarity.

Abdul-Mageed and Diab (2012) built their manually annotated corpus of Modern Standard Arabic together with a new polarity lexicon by using a machine translation procedure to translate the available English lexicons. Abdul-Mageed et al. (2014) produced a domain-dependent supervised machine learning system for Arabic Subjectivity Sentiment Analysis called SAMAR using the buckwalter Arabic transliteration scheme that convert the Arabic letters into English letters as the way in which a word is pronounced according to a database. They manually created a lexicon of 3982 adjectives labeled with one of the following tags {positive, negative or neutral}. Their results suggest that they need individualized solutions for each domain and task, but that lemmatization is a feature in all the best approaches

Mourad and Darwish (2013) introduced a new tweet corpus for Subjectivity and Sentiment Analysis SSA. They adopted a random graph walk approach to extend the Arabic SSA lexicon using Arabic/English phrase tables, leading to improvements for SSA on Arabic microblogs. They also used different features for subjectivity and sentiment classification including stemming, part-of-speech tagging, as well as tweet specific features.

2.2. Aspect-Based Opinion Summarization

The sentiment analysis on the document and sentence levels hides many important details about the object to be reviewed. To obtain more fine-grained sentiment analysis, we need to delve into the aspect level. This idea leads to the Aspect-based Opinion Mining whose basic task is to extract and summarize the reviewers opinions expressed on aspects of the entities.

For example, in the sentence "I bought a Nikon camera yesterday, and its picture quality is great," the aspect-based opinion analysis system should identify that the author expresses a positive opinion on the picture quality aspect for the entity Nikon camera.

Aspect-Based Opinion Summarization is the process of obtaining a summary of the entity ranked attributes/features with their sentiment orientations. The different approaches used to identify the entity aspects are discussed in the following subsections.

2.2.1. Aspect Categories

The entity aspects are classified into two main categories; explicit and implicit (Su et al. 2006). The explicit aspects are clearly appeared and can be extracted easily from the text. On the other hand, the implicit aspects are not clearly appeared and hard to be identified. The implicit aspects for a specific product are identified by assigning some adjectives in a lexicon to a set of pre-defined product aspects in a polarity lexicon then finding the relationship between those opinion words and the aspects. (Su et al. 2006) proposed an automatic identification for implicit product aspects expressed in the automobile reviews in the context of opinion question answering.

2.2.2. Aspect Extraction

There are three approaches used for extracting the explicit aspects. The first approach is gathering the aspects that belong to a particular domain in a database then matching the extracted aspects with the database (Liu, 2007; Lazhar and Yamina, 2012). The second approach relies on extracting the nouns/noun phrases that have more frequencies as a candidate features then pruning them (Hu and Liu, 2004; Liu et al. 2005; Popescu and Etzioni, 2007; Ghorashi et al. 2012).

Liu et al. (2005) presented an Opinion Observer system that used to identify product features from Pros and Cons in reviews of format (2). Also it gave a solution for the comparison of consumer opinions of multiple (competing) products. They used a supervised mining rule to generate language patterns to identify the features. They extracted the nouns/noun phrases produced by NLProcesser as candidates for the explicit features. For the implicit features, they used a list of candidate features with the actual aspects.

(Popescu and Etzioni, 2007) presented an unsupervised system called OPINE to mine the reviews. They built a model of important product features. When the entity is known, a list of its explicit features and adjectives can be utilized to extract feature-based opinions. They used WordNet's IS-A hierarchy to differentiate between the parts and properties of the entity.

Similar to Liu work, Ghorashi et al. (2012) collects frequent nouns and noun phrases as product features. However, they overcome different writing styles by analyzing extracted phrases to produce patterns using frequent pattern mining algorithm called H-Mine as shown in figure (2.1) then pruning these features using a minimum support value of 1% to remove any redundant features.

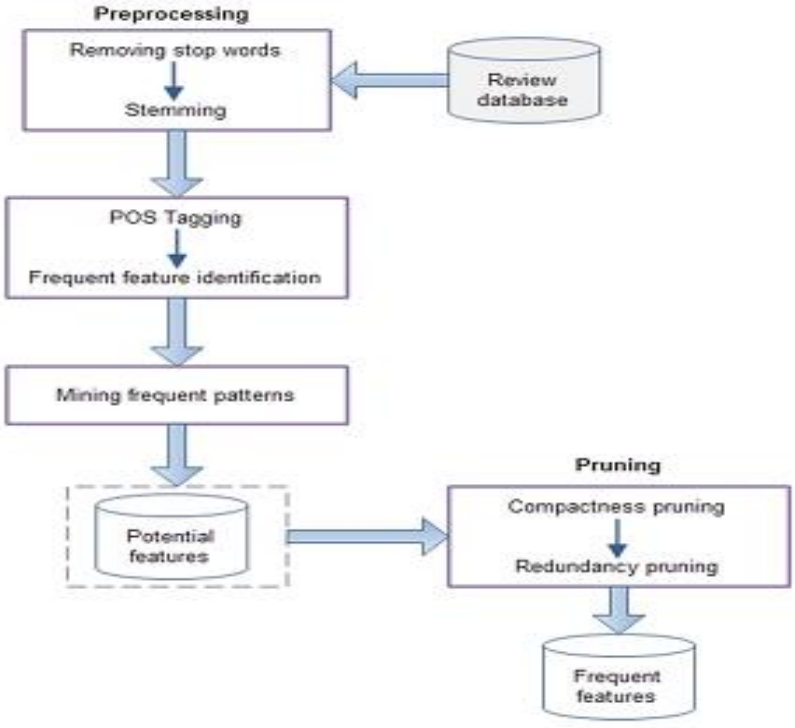

Figure 2.1: H-Mine framework (Ghorashi et al. 2012)

Extracting product features can be done also by utilizing patterns (hybrid dependency patterns) which are based on dependency relation between opinion terms presented by opinion lexicon and product features presented by noun (Khan et al. 2014). The hybrid pattern is a combination of four different patterns. The opinion lexicon is a list of subjective adjectives that have positive or negative polarities. The process of extracting the features is shown in figure (2.2).

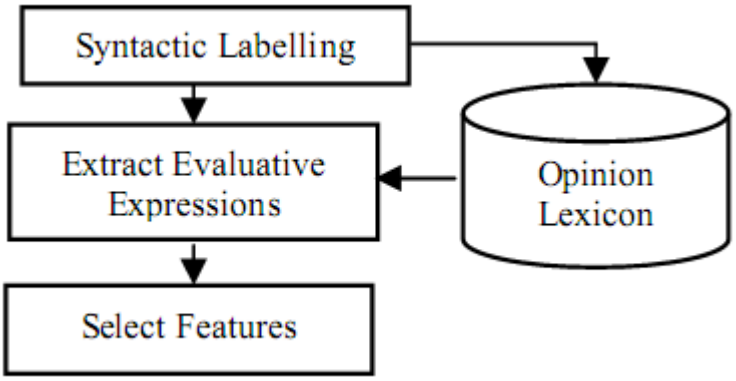

Figure 2.2: Product Feature Extraction (Khan et al. 2014)

The third approach is relying on the use of an opinion word list, where the nearest nouns/noun phrases are considered as the candidate aspects. Hu and Liu (2004) presented an opinion mining system based on this idea as shown in figure (2.3). They extracted all the nouns/noun phrases as frequent features then using these features to extract the nearest adjectives as opinion words to expand the opinion word list. They used a simple heuristic method. This method states that the sentences that had no frequent features but had one or more opinion word will be examined to extract the nearest nouns/noun phrases for the opinion words as infrequent features. This technique had a problem that some nouns/noun phrases were irrelevant to the given product. But the researchers neglected them since the infrequent features number was small comparable to the frequent features number and they was obtained for completeness.

Definition (Frequent Features) are the hot features that most of the customers expressed about in their opinions.

Definition (Infrequent Features) are the features that little people mentioned in their reviews. These features may be useful for many customers and for the manufacturer to develop its product.

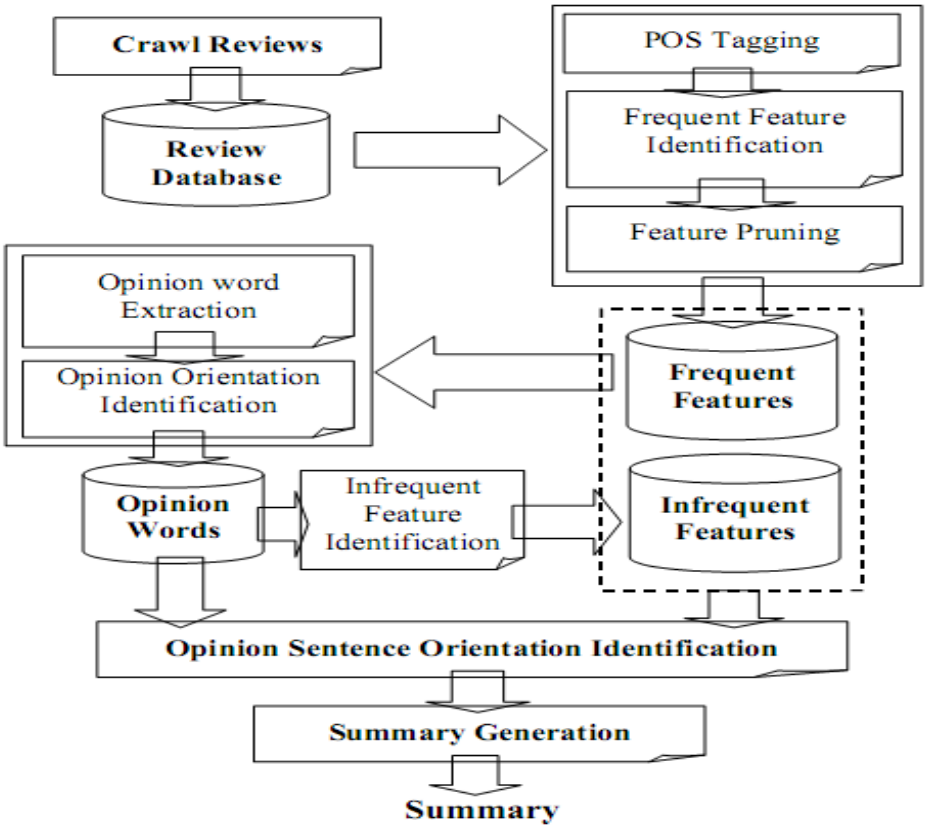

Figure 2.3: Feature-based opinion summarization system (Hu and Liu, 2004)

The idea of using the modifying relationship of opinion words and aspects to extract aspects can be generalized to using dependency relation. Zhuang et al. (2006) employed the dependency relation to extract aspect-opinion pairs from movie reviews. After parsed by a dependency relation parser (e.g. MINIPAR1), words in a sentence were linked to each other by a certain dependency relation. Figure (2.4) shows the dependency grammar graph of an example sentence, “This movie is not a masterpiece.”, where “movie” and “masterpiece” had been labeled as aspect and opinion respectively, a dependency relation template could be found as the sequence “NN - nsubj - VB - dobj - NN”. Zhuang et al. (2006) first identified reliable dependency relation templates from training data, and then used them to identify valid aspect-opinion pairs in test data.

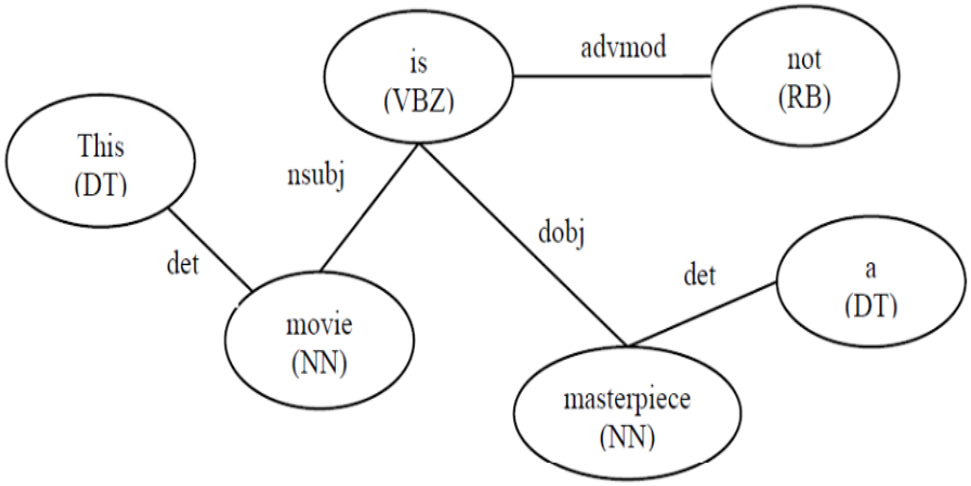

Figure 2.4: The dependency grammar graph for the sentence "This movie is not a masterpiece." (Zhuang et al. 2006)

Some works was mixed between extracting the implicit features and the explicit features (Liu et al. 2005; Popescu and Etzioni, 2007; Cruz et al. 2010; Ghorashi et al. 2012). Most of researches that published to define the implicit features used the technique of constructing a matrix that concludes the co-occurrence between each feature and words in the same sentence. Besides to a list contains all the implicit features of a specific object. (Schouten and Frasincar, 2014) used a threshold score which is defined first. After training analysis, they choose the most performing threshold value to be used in the evaluation part. But (Zhang and Zhu, 2013) add another matrix called the word modification matrix to define the implicit features.

2.2.3. Aspect Extraction for Arabic Text

There are little researches had been introduced in extracting the entities attributes/features for Arabic text. Most of these efforts are concerned with extracting the explicit features of the entities. One approach used knowledge representation models to discover the different characteristics of a product or an object. Only the expressions of opinions (adjectival and adverbial) were extracted, then a summary was produced to show for each characteristic, the positive and the negative opinions and the total number of these categories (Turney, 2002). The main limitation of this approach is the large number of extracted features and a lack of organization.

Another approach uses taxonomies to build a hierarchical organized list of features. The taxonomy is a list of terms organized hierarchically through a sort of "is a kind of". Ontologies aimed to organize the features using elaborated representation models. Unlike taxonomies, ontology is not restricted to a hierarchical relationship between concepts, but can describe other types of paradigmatic relations such as synonymy, or more complex relationships such as relations of composition or spatial relationships.

Figure (2.5) explains the architecture that presented by (Lazhar and Yamina, 2012). The system defined an elementary discourse unit (EDU) as a clause contains at least an elementary opinion unit (EOU) or a sequence of clauses that address a rhetorical relation to a segment expressing an opinion. EOU was an explicit opinion expression composed of an explicit noun, an adjective or a verb with its possible modifiers.

For each extracted EDU, the system:

- Extracted EOUs using an approach based on rules;
- Extracted the features that correspond to the process of terms extraction using the domain ontology;
- Associating or linking, for each feature within the EDU, the set of opinion expressions.

The authors proved that the use of ontologies improved the extraction of features and facilitated the association between opinions expressions and opinionated features of the object. Also, domain ontology is useful within its list of concepts which carry much semantic data in the system. The use of ontology concepts labels can recognize terms that refers to the same concepts and provides a hierarchy between these concepts. On the other hand, ontology is useful to its list of properties between concepts that can recognize the opinions expressed on the implicit features.

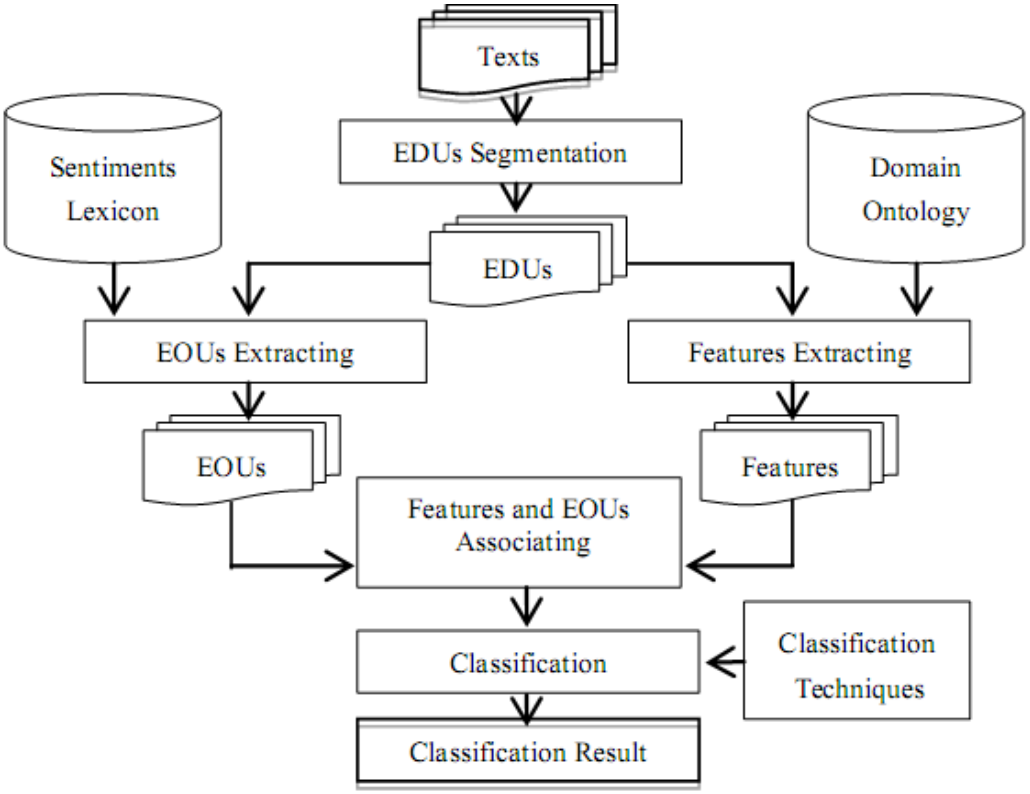

Figure 2.5: EDU architecture (Lazhar and Yamina, 2012)

Table (2.1) summarizes the distinguishing efforts dealing with aspect extraction and sentiment analysis according to the classification level, domain, learning, POS tagging, feature extraction method, and sentiment analysis approaches.

Table 2.1: An overview for the related works about aspect extraction and sentiment orientation approaches for English reviews.

Author	Turney 2002	Pang et al, 2002	Hu and Lui 2004		Lui et al, 2005	Popescu and Etzioni, 2007		Khan et al., 2010	Ghorashi et al, 2012	Khan et al, 2014
Features Type	Whole Document	Whole Document	Frequent Explicit	Infrequent Explicit	Explicit & Implicit	Explicit	implicit	Explicit & Implicit	Explicit	Explicit
Identifying opinion sentiment	Yes	Yes	Yes		Yes	Yes		N/A	N/A	Yes
Classification Level	Document Level	Document Level	Sentence Level		Sentence Level	Sentence Level		Sentence Level	Phrase Level	Phrase Level
Domain	Different Domains	Specific (Movie)	Independent		Specific	Specific (Scanners, Hotels)		Specific (Hotels)	Specific (Technology)	Specific
The Learning algorithm	unsupervised	supervised			supervised	unsupervised			N/A	Semi-Supervised
POS Tagger	part-of-speech tagger of (Brill,1994)	Oliver Mason's Qtag program	NLProcessor Linguistic Parser		NLProcessor Parser	MINIPAR Parser		WordNet	POS tagger developed in PHP language	Stanford Parser
Feature Extraction Approach	N/A	N/A	Extract nouns / noun phrases as candidate features	Extract nouns / noun phrases dependent on the opinion words list	Extract nouns / noun phrase according to Association rule mining	Extract nouns / noun phrases as a candidate features	Assign adjectives by predefined list of adjectives & features	Extract nouns based on features seed list	Extract nouns / noun phrases besed on Frequent pattern mining algorithm called H-Mine	Extract noun phrases based on hybrid dependency patterns
Feature Pruning Technique	N/A	N/A	P-support value	N/A	Frequent term strategy	Threshold value for high frequency	N/A	Nouns has 2 occurrence or more	P-support value	N/A
Opinion tags (opinion words)	Two-word phrases as pattern tags	List of adjectives extracted manually by human	The nearby adjectives (effective opinions)		Pros, Cons	Opinion Phrases according to expression rules	Adjectives list	Auxiliary verbs & opinion words list	N/A	List of adjectives
Linguistic Characteristics	Statistical					Morphological		Statistical	N/A	Syntactic
Sentiment Analysis Approach	(PMI-IR) Pointwise Mutual Information + Information retrieval	Machine Learning Techniques (NB - ME - SVM)	WordNet (synonym & antonym) as Seed list of adjectives		Already classified as Pros and cons	Relaxation Labeling as tuple (word, feature, sentence)	WordNet-Based rules and Web-Based rules	Sentences with auxiliary verbs and without	N/A	Opinion Lexicon
Orientation Strength	AltaVista Advanced Search engine	N/A	N/A		N/A	Web-derived constraints & adjective strength		N/A	N/A	N/A
Precision (Extraction)	N/A	N/A	0.79	0.72	0.791	0.79		N/A	N/A	0.789
Recall (Extraction)	N/A	N/A	0.67	0.8	0.824	0.76		N/A	0.33	0.718
Precision (Polarity)	N/A	N/A	0.64		N/A	0.86		N/A	N/A	N/A
Recall (Polarity)	N/A	N/A	0.7		N/A	0.89		N/A	N/A	N/A
Accuracy	0.74	0.744	N/A		N/A	N/A		N/A	N/A	N/A

Chapter 3

Chapter 3
Text Pre-Processing

3.1. Introduction

In Natural Language Processing (NLP), pre-processing aims to reduce the complexity of the vocabulary of documents, and relates words that have the same meaning. Pre-processing eliminates the punctuation, filters the function words and normalizes morphological variants. Arabic is a very rich language in categorizing words, and hence, numerous stemming techniques have been developed for morphological analysis and POS tagging. Morphological analysis is an important phase of the pre-processing of the input text. It affects the output performance because it allows a search term to focus more on the meaning of a term and closely related terms instead of specific character matches (El-Shishtawy and El-Ghannam, 2012).

Root, stem and lemma are three different word analysis levels that are used in information retrieval (IR) techniques. So we need to distinguish between these words;

Root: *is the foundation of the Arabic word. Each Arabic word originated from three-letters (trilateral) or four-letters (quadrilateral).* Various vowels, prefixes and suffixes are used with the root letters to create desired inflection of meaning. In this thesis, we will assign x, y, z for the original tri-root. For example (فكر - he thinks, xyz), and (مفكر - Thinker, M xyz).

Stem: *is the form after removing the affixes (prefixes and suffixes) from the word.* This process called stemming process which can result correct roots with some words. But it fails when using to return the word from past tense form to its present tense form as the word (یرى - see - Yxz : رأى - saw - xyz). Also when getting the singular noun form from the broken (irregular) plural nouns as the word (أفكار - ideas - AxyAz : فكرة - idea - xyzA).

Lemma*: is the canonical form, or citation form of a set of words. It refers to the set of all word forms that have the same meaning, and hence capture semantic similarities between words.* For nouns and adjectives, it is the singular indefinite form, and for verbs, it is the perfective third person masculine singular form. For examples (خدمة -

service - xyzA & خدمات - services - xyzAT) have the same lemma of (خدمة), also the two verbs (يأخذ - takes - Yxyz & أخذ - took - xyz) have the same root (أخذ).

The problem of using the root as a standard representation level in IR systems is the over-semantic classification. Many words have the same root, but don't have similar semantic interpretations. Stemming and lemmatization shares the common purpose of reducing words to an acceptable abstract form, suitable for NLP applications. But stemming process suffers from under-semantic classification. It can't detect the syntactic similarities between the singular noun and its broken plural form.

But lemmatization in NLP field refers to the process of relating a given textual item to the actual lexical or grammatical morpheme (Dichy, 2001). So it is the best choice to be the basic step in building our system of aspect-based opinion mining.

3.2. Arabic Lemmatizer

In our proposed system, we will make use of an existing accurate Arabic Root Based Lemmatizer ARBL (El-Shishtawy and El-Ghannam, 2012). The Lemmatizer exploited Arabic language knowledge in terms of roots, patterns, affixes, and a set of morpho-syntactic rules to generate accurate lemma form and its relevant morpho-syntactic features that support information retrieval purposes. Morpho-syntactic features are required also, to capture the important semantic senses of the language as expecting the correct word category and verified it.

The lemmatizer consists of two phases: POS tagging phase and Lemma generation phase that will be explained in the following subsections.

3.2.1. POS tagging

To generate more accurate word features including the POS tags, more information about the word are needed to be collected. The lemmatizer produced a set of POS tags for each word representing its class (noun, verb, adjective, preposition …), gender (male, female), count (single, plural), and tense (past, present). Also, it extracted the word root, stem and pattern.

ARBL based on the open source root-based stemmer of (khoja, 1999). This stemmer removed the possible infixes from a word, found corresponding matched pattern, and extracted the word root without POS tags. It uses a lexicon contains 3800 trilateral and 900 quad literal roots. Also, Khoja system recognized a list of 168 Arabic stop words. It achieved high accuracy compared with Buckwalter Analyzer and Tri-literal Root Extraction Algorithm (Sawalha and Atwell, 2008).

El-Ghanamm modified both the data and the basic algorithm flow that were necessary to add Arabic knowledge. As using different knowledge resources of Arabic language: prefixes, suffixes, patterns, and rules. Limited size auxiliary dictionaries were used to augment morphological and syntactic rules in recognizing words, and resolving their ambiguity. The dictionaries included only words that were expected to fail in tagging by rules. In most cases, the ambiguity was due to the absence of the short vowels in the electronic Arabic documents, or non templatic word stems (El-Shishtawy and El-Ghannam, 2012). The algorithm outline is shown in figure 3.1.

```
For each word (WO) Do
 Begin word_block
   Search a word in proper noun dictionary
   If exists POS = N, with features set, exit word_block.
   Check the existence in closed set word dictionary
   If found, POS = article with features set, exit word_block;..
   For each affix -longest first- Do
    Begin affix_block
      If affix cannot be removed from W then exit affix_block;
      Remove affix to extract the (W) form
      Check if (W) matches a pattern (P) with root R
      If (P) exists
       Begin POS_block
         Apply POS identification rules;
          If rules failed POS =N;
         Apply syntactic rules to detect Adjectives
       End POS_block;
     End  affix_block;
    End loop;
  End word_block;
      End loop
```

Figure 3.1: The outline of determining the POS tagging of ARBL

The first stage of this algorithm started with the analysis of checking a closed set of 346 Arabic words that are categorized into 16 groups (e.g., prepositions, conjunctions, adverbs, numerals, etc...). The basic flow of the algorithm starts by removing the longest suffix and the longest prefix in turn. After every elimination process, the algorithm checked a list of 69 patterns, if matched a pattern, the root is extracted and verified by checking the list of 3829 tri-roots. The output of this stage is the suffix, prefix, word pattern, and root.

The second stage was to tag POS of the word and to extract the corresponding features. The features for nouns and adjectives were definite case, count, and gender. POS tagging and word feature extraction were completed through many levels. The following subsections describe each level.

3.2.1.1. Nouns and Verbs Identification

In Arabic language, some verbs or inflected nouns can have the same orthographic form due to the absence of short vowels. For example (Verb: contributes أسهَم, Noun: stocks أسهُم). The lemmatizer tried to overcome this problem by the following rules:-

a) The word is categorized as a noun if it comes after one of the noun articles such as (أمام – إلى – فوق – بعد - ...). Also, the word is categorized as a verb if it comes after one of the verb articles such as (عندما – قد – لن - لم – كلما - كي ....).

b) Applying some syntax rules. For example, one rule stated that if the previous word was a verb, the current word couldn't be also a verb, since Arabic language did not permit two successive verbs to exist.

c) Applying some morphological rules during stemming. For example, affixes were categorized into three classes: affixes used by nouns only, affixes used by verbs only, and those that were used by either nouns or verbs.

d) The fourth level was the pattern-level that illustrated in next subsection.

3.2.1.2. Pattern Identification

The collected information about words included word pattern. Arabic stem-patterns have interesting semantic features that give rise to senses of words. For example, syntactic patterns recognized a given word as being the agent of an action, the instrument of that action, or the place at which the event occurs.

In the lemmatizer, patterns played essential role in recognizing lexical word category. Arabic patterns were classified into three classes:

- Verb Patterns: That used for verbs only.
- Noun Patterns: That used for nouns only.
- General Patterns: This might be used for verbs or nouns according to different vocalization and not-written diacritics.

3.2.1.3. Adjectives Identification

Traditionally, Arabic does not include adjective as one of its main POS. An adjective in Arabic is actually a noun that happens to describe something (الحملاوي، ١٨٩٤). Adjectives take almost all the morphological forms of nouns. The word is considered as an adjective, if it has the same count and gender with the previous word. Also it followed the previous word as definite or indefinite.

3.2.2. Lemma Generation

The second phase of the lemmatizer algorithm is generating the abstract lemma form of the word that will be used in the process in the aspect extraction.

3.2.2.1. Verb lemma

Verb lemma is the perfective, singular verb form. In most cases, lemma was extracted by removing prefixes and suffixes. For example, the word (يكتب, Yxyz) has the same form for both the root and lemma (كتب, xyz). But in other cases, the word (يستخرجون, YSTxyzON) has the root form (خرج, xyz) and the lemma form (استخرج, ESTxyz).

3.2.2.2. Noun lemma

Lemma form of noun (or adjective) is the singular indefinite form. In Arabic, there are two types of noun and adjective plural forms: regular plurals, and broken (irregular) plurals.

Regular Plural: The lemma form of the masculine plural was generated simply by removing suffixes "ون" or "ين" from the noun form. Lemma singular form of feminine plural nouns had two cases; feminine or masculine single form. Feminine singular form is generated by adding "ة" to indicate its feminine nature (e.g. جمعية / جمعيات).

Broken Plural nouns: Another issue with the lemma generation for nouns and adjectives is broken plurals. There are about 27 pattern forms for the broken plural (الحملاوي، ١٨٩٤). There exist many possibilities for the singular form for each pattern. For example the broken plural pattern (فعائل) of the broken plural words (رسائل ، صحائف) has two different patterns for the singular form (فعالة ، فعيلة) in which the single word forms (رسالة ، صحيفة) are generated.

The lemmatizer uses a dictionary to store only ambiguous cases, i.e., that had a lot of probabilities for the singular form.

3.3. Sentiment-Annotated Lexicon

To determine the sentiment orientation for the Arabic reviews, most researches used a classified lexicon (seed list) that contained the words/adjectives that are usually used to express opinions i.e. *opinion word*s. This technique suffers from the problems of lack of context and domain dependent. To overcome these problems, we proposed a new approach which exploits lexical features and a lemma based analyzer annotated with opinion tags at the root and pattern levels. This allows sentiment analysis to be performed for generic domains beyond the limited coverage of existing ontologies.

In Arabic language, actually in all Semitic languages, a single root with associated patterns can generate many lemma forms; with each has a different semantic meaning. For example, the different patterns for the Arabic root (xyz, write "كتب"), can generate many words that have different semantic senses, such as (MxyzH, "مكتبة ", "library"), (xAyz, "كاتب", "writer") and (xyAz, "كتاب ", "book"), originating from the same root. Also, the word pattern provides a mean to infer if the given word is the agent of an action, the instrument of the action, or the place at which the action occurs. Therefore, Arabic word generation is a process of applying one pattern forming rule to a specific root. Motivated by this computational behavior of Arabic language, the proposed approach depends on annotating both roots and patterns with opinion tags, to allow the system to extract sentiment carrying words, while keeping the dictionary in minimum size. With an analogy to English language, the infinitive form 'success' carries a positive orientation and so its derived words (successful, successfully, succeed, or succeeded). Similarly, fail, failure or failed have the negative orientation effect.

3.3.1. Pattern-level Tags

In all existing Arabic lexicons, patterns are classified according to POS (Khoja, 1999). We extended the classification to include sentiment tags at the pattern level, as added the forms of sentiment carrying patterns and comparator patterns in table (3.1). With the assistance of two Arabic language specialists 39 patterns are tagged as pattern-carrying opinion out of the available 69 patterns collected by the ARBL.

Table 3.1: Syntactic and Sentiment Patterns

	Sentiment Pattern Classifications	**Pattern Class**	**Pattern form Examples**	**Word Examples**
Syntactic Pattern Classification (69)	Neutral Patterns (30)	Verb Patterns	ENxyz "انفعل" ESTxyz "استفعل"	- انتبه "pay attention" - استقام " unbend "
		Noun Patterns	MxyOz "مفعول" ExTyAz "افتعال"	- مكتوب " written" - اكتساب " gain "
		General Patterns	xAyz "فاعل" TxAyz "تفاعل"	- شاعر " poet " - تقابل "meet"
	Sentiment Patterns (39)	Sentiment – Carrying Patterns (37)	xyEz "فعيل" MxyAz "مفعال" xyOz "فعول"	- جميل "beautiful" - ممتاز "excellent" - كسول "lazy"
		Comparator Patterns (2)	Axyz "أفعل" xyzA "فعلى"	- أفضل " best or better for boy" - فضلى " best or better for girl "

3.3.2. Root-level Tags

The root is the origin of all derivative words in the Arabic language. The number of the tri-roots in ARBL is about 3829 roots. The role of each root in sentiment is studied carefully and polarity information is added for these roots to improve the process of determining the sentiment orientation. If all the words derived from a root have a common orientation, then the root could be annotated as sentiment root (مختار عمر، ٢٠١٣).

One common problem for lexicon-based approach is the context-dependent sentiment word, i.e., the different sentiment orientation in different domains. For example, the word "big, كبير" has positive orientation in hotel domain, while has negative orientation in technology domain. Tagging at the root level adds a second source of uncertainty, because the same root can generate different orientation words with different patterns. For example, the root (xyx, "خلف") can be positive if it has a form (MxTyz, "different", "مختلف"), while it has a negative orientation, if it takes the form (MTxyz, "lagging", "متخلف"). In our work, we tagged these roots as uncertain roots either due to context dependent or different patterns sentiments.

In our work, 213 roots[2] are manually marked as positive, 260 roots as negative, and 107 as uncertain oriented roots, out of 3829 roots recognized by ARBL. Examples of these roots are shown in table (3.2).

Table 3.2: Examples for positive, negative and uncertain roots

Positive (213)	**Negative (260)**	**Uncertain (107)**
Example: kind "لطف" succeed "نجح" surprise " دهش" satisfy "رضي" ...etc	Example: damage " تلف" harm "ساء" poison "سمم" poor "هزل" .....etc	Example: old "قدم" big "كبر" cold "برد" long "طول" ....etc

The uncertain roots are categorized into two types as the root that has most common positive orientation will be uncertain positive root and the root that has most common negative orientation will be uncertain negative root.

Table (3.3) summarizes the inter-annotator agreement during the lexicon annotation process. We have adopted only the common roots between the two experts and added the roots they differed to the uncertain roots.

Table 3.3: Inter-annotator agreement during the lexicon annotation process

	No of positive roots	**No of negative roots**	**No of uncertain roots**
I st Expert	260	303	124
2 nd Expert	225	265	130
Common Roots	**213**	**260**	**107**

Some words are written in Arabic language as introductory words such as (clearly: في الحقيقة ,briefly: بإختصار). These words should be excluded from the list of candidate aspects and so their roots. We tagged about 70 roots as Excluded roots[3].

[2] More positive, negative and uncertain roots are found in appendix (A).

[3] A list of Excluded roots can be found in appendix (A).

3.3.3. Negation

Negation is an important parameter that affects the sentiment meaning. Negation words like (no, not, none, nobody, never, without, and nothing). They represent the opposite indication of the extracted sentiment. For example the expression "Service is not good" has the opposite attitude of the expression "Service is good". In Arabic, about 12 words[4] are tagged as negation words.

3.3.4. Intensification

Intensifier has another important effect on the sentiment meaning of the text (Taboada et al. 2011). It can strengthen or weaken the sentiment meaning by using some neighboring adverbs like (جداً : very, مطلقاً : absolutely, قليلاً : slightly... etc) (Beamer et al. 2007). In Arabic, there are similar adverbs strengthen the sentiment as positive or negative. In Arabic, about 27 words[5] are used as sentiment intensifiers.

[4] Arabic negation words are found in appendix (A).

[5] Samples of the intensifiers are found in appendix (A).

Chapter 4

Chapter 4

Aspect-Based Opinion Mining

The proposed approach for the aspect-based opinion mining summarization topic is divided into three main tasks: 1) Identifying the opinion-carrying words using sentiment analysis, 2) Extracting the detailed entity aspects and their attitudes by making use of the detected opinion words, 3) Determining the orientation of the whole text by aggregating the orientation values for the opinion-carrying words. Figure (4.1) shows the overview for the proposed approach.

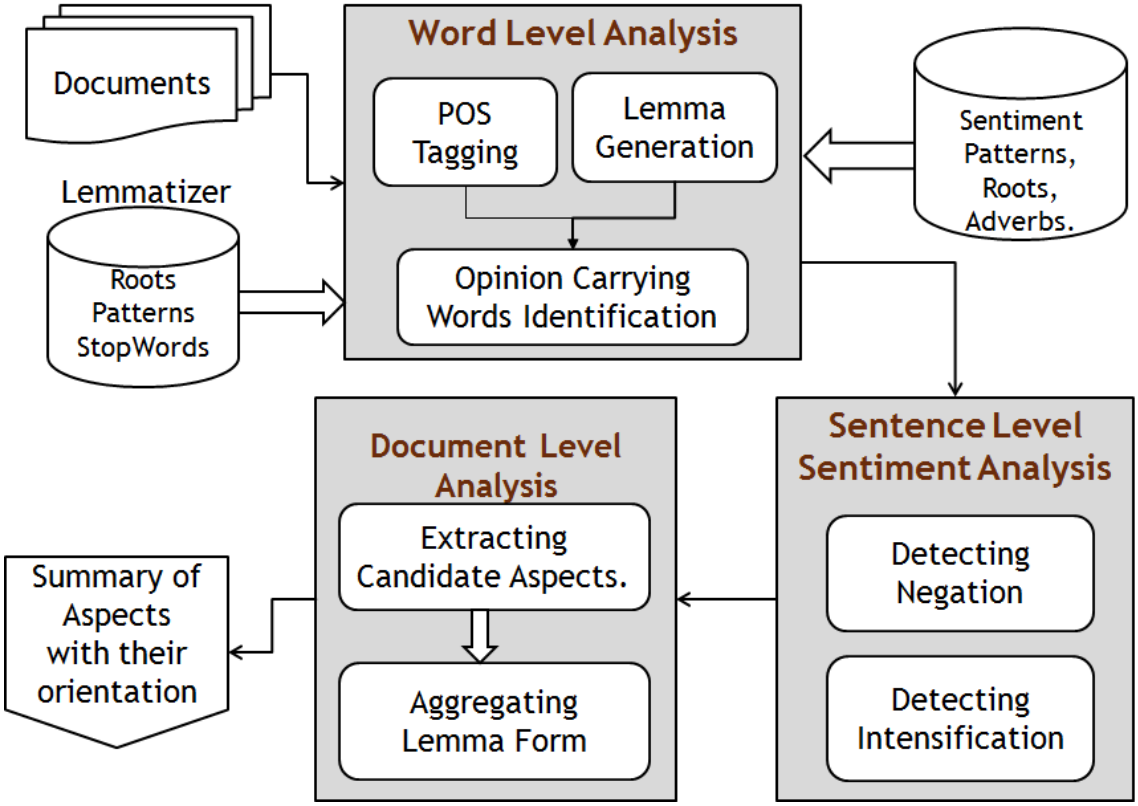

Figure 4.1: The Proposed System Overview

The achievement of these tasks requires the analysis on the three levels as follows:

1- Word-level analysis to extract syntactic, lexical and opinion tags.
2- Sentence-level sentiment analysis to detect negation and intensification.
3- Document-level analysis to extract the aspects with their SO.

4.1. Word-Level Analysis

In this phase, all words are analyzed to extract their basic features that were augmented by ARBL provides the following information:

1) Syntactic information: POS tags (noun, verb, adjective, preposition...), Gender (male, female) and Count (single, plural).

2) Lexical Information: The word root, pattern, and its lemma form.

Part-of-speech (POS) information is commonly exploited in sentiment analysis and opinion mining. It helps in defining the opinion-carrying words as well as its polarity. Also it is used for extracting the nouns that will be considered as the aspects of the entities. Using the lemma form will serve in capturing all semantic features of the word and preventing occurrence of data redundancy.

4.1.1. Opinion-Carrying Words

Instead of using opinion word lists to detect the sentiment orientation, we proposed the sentiment annotated lexicon that has sentiment tags (polarity and strength) at the root and the pattern levels. In this approach, the word is considered as opinion-carrying or opinion tag if it meets the following two conditions:

1) Its pattern matches one of the orientation patterns, and
2) Its root matches one of the positive, negative or uncertain roots.

The pseudo code of the algorithm that explains these conditions is shown in figure (4.2). The algorithm starts by examining the pattern of each word. If the pattern belongs to one of the orientation patterns, it will check its root. If the root is one of the sentiment roots (positive, negative, uncertain positive or uncertain negative), it will consider this word as an opinion-carrying word and will be used in the next phase of extracting the aspects of the entities. Also the algorithm determines the polarity value for each opinion-carrying word.

```
1. For each word do
2.      Check the orient-pattern (if_exist)
3.          Check certain positive roots (if_exist)
4.              Check the strength of pattern (sentiment / comparator)
5.                  Assign positive value w.r.t pattern strength.
6.                  Assign a certain value equal 1.
7.          Else check certain negative roots (if_exist)
8.              Check the strength of pattern (sentiment / comparator)
9.                  Assign negative value w.r.t pattern strength.
10.                 Assign a certain value equal 1.
11.         Else check uncertain positive roots (if_exist)
12.             Check the strength of pattern (sentiment / comparator)
13.                 Assign positive value w.r.t pattern strength.
14.                 Assign an uncertain value equal 0.5.
15.         Else check uncertain negative roots (if_exist)
16.             Check the strength of pattern (sentiment / comparator)
17.                 Assign negative value w.r.t pattern strength.
18.                 Assign an uncertain value equal 0.5.
19.         Else it is a neutral word.
20.     Else it is a neutral word.
```

Figure 4.2: The outline algorithm for detecting the opinion-carrying words.

4.1.2. Polarity Value

Several lexicon based approaches have expressed the semantic orientation as a numerical value range to express the word's strength, (Wiebe et al. 1999; Hu and Liu 2004; Kim and Hovy 2006; Taboada et al. 2011). In our work, we followed another approach, where all opinion-carrying words are handled as 'like' or 'dislike' binary opinions, whatever is the strength of vocabulary used in the review. This gives more importance to the number of reviewers who liked (or disliked) an entity (or aspect) rather than their use of strong synonym words. The assumption of equal opinion weights is proposed for the following reasons:

(1) In spite of previous efforts of building and ranking dictionary words - for example giving the sentiment word "love" a stronger weighting than the word "like". A criticism still raised that the dictionaries are unreliable, as they are either built automatically or hand-ranked by humans (Andreevskaia and Bergler 2008).

(2) The overall sentiment result may be misleading. As an example adapted from Taboada et al. (2011), the opinion of one reviewer who used the word 'masterpiece' (ranked +5), will dominate the opinions of four other reviewers used the word 'delay' (ranked -1).

(3) Reviewers were not have the chance to choose specific opinion word from a closed terms arranged by strength from highly positive to highly negative. Therefore, reviewers - in most cases - express opinions based on their culture background and mode.

Following the assumption of "prior polarity" of words, (Osgood et al. 1957), we assigned each root a context-independent semantic orientation. The orientation is manually tagged and expressed as a numerical value (+1, -1) for positive or negative orientation, respectively. Also, the comparator patterns (e.g., words 'smaller', 'smallest') takes the value to +2 or -2 according to the root polarity and will be used to amplify the polarity value of the corresponding sentiment carrying word. Thus, the result of the sentiment analysis at the word level is a list of opinion carrying words with polarity values range from +2 for strong positive to -2 for strong negative.

For example, the word (beautiful, جميل), has a positive orientation with value (+1) due to a positive root (جمل, xyz) and its sentiment-carrying pattern (فعيل, xyEz). Similarly, the word (Ugliest, أبشع) has a negative orientation with value (-2) due to the negative root (بشع, xyz) and the comparator pattern (أفعل, Axyz).

For adjacent opinion-carrying words in the same sentence, a special criterion is used for handling this point. As adding the orientation value for the current and the previous opinion-carrying word. This criterion is shown in the following algorithm in figure (4.3).

```
1. If the current word is an opinion-carrying word
2.     Check the previous orientation word
3.         If its sentiment orientation is (positive/negative)
4.             Restore the orientation value for the previous word.
5.             Add the current orientation value to the previous value.
```

Figure 4.3: The criterion for detecting adjacent opinion-carrying words.

For the uncertain roots that have double sentiment orientation according to different patterns or different domains, a certainty value is assigned for them. Each opinion-carrying word is assigned a certainty value (1 or 0.5) according to its root as shown in figure (4.2) in lines (6, 10, 14, and 18). Certainty value is set to 0.5, if the word root matches one of the uncertain roots; else it is set to 1. It is important to note that uncertain roots do not affect aspect extraction, as both positive and negative roots are used to locate aspects.

4.2. Sentence-Level Sentiment Analysis

The purpose of sentence level analysis is to detect the word intensification (e.g. very good) and negation (e.g., not good). In Arabic, both of intensification and negation are long-distance phenomenon, and therefore should be detected at the sentence level. In this work, we first compute the strength and orientation independently at the word level, and then applying the detecting algorithms to update the opinion value and/or polarity of the opinion carrying word.

4.2.1. Detecting Intensification

Intensifier parameter assesses the semantic of a word, using some neighboring adverbs like (very, extremely, absolutely, etc...) (Benamara et al. 2007). The effect of intensifier words is to increase the value by (1) in its polarity direction as shown in figure (4.4). When the program detects one of the intensifier words that mentioned in section (3.3.4), it checks the previous word. If this word is an opinion carrying word and has positive orientation, it adds (1) to the saved orientation value and vice versa. For example the total weight for the sentence "The efficiency of this phone is very good." is equal to (+2) due to the polarity value of positive root and pattern of the word "good" plus the value of the intensifier word "very".

```
1. If the current word belongs to the adverbs closet set
2.     Check the previous orientation word
3.         If its sentiment orientation is (positive)
4.             Increase the last orientation value by 1.
5.         If its sentiment orientation is (negative)
6.             Decrease the last orientation value by 1.
```

Figure 4.4: The criterion for handling the intensifier effect

4.2.2. Detecting Negation

Negation is an important parameter that affects the orientation for the detected opinion carrying words. In most cases, negation reverses the word orientation. Usually in MSA writing style, negation precedes opinion words. Therefore, starting from the opinion-carrying word, the system scans for the existence of negation word in backward direction within the sentence. Once a negation word is detected, both of the opinion tag orientation and value are reversed. For example the expression "Service is not good" has the opposite orientation of the expression "Service is good".

At the end of this phase, the orientation type, score, and certainty value of each sentence containing opinion-carrying words are determined. Table (4.1) shows an example of the output of a sentence " This hotel is very nice; هذا الفندق لطيف جداً" after this level of analysis.

Table 4.1: The sentiment orientation analysis for the sentence "هذا الفندق لطيف جداً "

Word	هذا	الفندق	لطيف	جداً
Lemma	هذا	فندق	لطيف	جداً
No_Char	3	6	4	3
Suffix				
Prefix		ال		
Pattern			فعيل	
Root			لطف	
Type	اسم اشارة	اسم	اسم	Intensifier (كلمة مبالغة)
Count		مفرد	مفرد	
Gender		مذكر	مذكر	
Orient_type			Pos	
Orient_value			1	
Certain_value			1	

4.3. Document-Level Analysis

Analysis on the word and sentence levels provides an overall opinion of the general discussed entity. The text can be given a single scale combining sentiment polarity and strength of all sentiment words. However, this does not provide the required comprehensive level at the aspect level. In a typical review text, people express their opinion about an entity or product by discussing both positive and negative aspects of the entity. This level of detailed analysis is quite useful for many real life applications that need feedbacks from consumers to improve their products. This leads to the importance of extracting the object features with their polarity.

Subsequent work on subjectivity detection revealed a high correlation between the presence of adjectives and sentence subjectivity (Hatzivassiloglou and Wiebe, 2000). This finding has been taken as evidence that certain adjectives are good indicators of sentiment, and can be used to guide the feature selection process for sentiment classification. As we intend to extract automatically domain independent aspects or features, the proposed approach does not exploit predefined set of features, nor domain ontology hierarchy. Instead, the identified opinion-carrying words are used for extracting entity level aspects and their orientations.

The presented system analyzes sentences to extract all target noun phrases as candidate aspects. The second step is to aggregate the candidates based on their lemma-form frequencies after removing entity names. The sentiment weight and attitude is then calculated for each aspect and the general entity. In this section, we mainly focus on the two mining tasks:

1- Extracting target noun phrases as candidate aspects.
2- Aggregating lemma-based candidate aspects.

4.3.1. Extracting candidate aspects

Each opinion-carrying word has a target aspect or entity, and the problem is how to locate these aspects. By analyzing sample reviews, we have identified repeated patterns of word categories representing aspects or features. These patterns are found in two directions; backward and forward. It is important to note that the search direction is

language dependent. In Arabic language, the search direction is forward when the category of the opinion-carrying word is verb; else it is backward for all other POS opinions.

4.3.1.1. Backward Direction

To extract the target noun phrases, we used a method based on a set of syntactic rules to determine the allowed sequence of words of n-gram terms according to their POS tags (El-Shishtawy and Al-Sammak, 2012). For example, candidate aspect can start only with some sort of nouns; the candidate aspect can end with noun phrase. Table (4.2) shows a sample of the allowed syntactic patterns to extract the target noun phrases. The abbreviation symbols of the following table are found in Appendix (B).

Table 4.2: Examples of syntactic patterns for detecting candidate aspects

Syntactic Pattern	Extracted Aspect	Example
NN+DTNN2+Prep+DTNN1+Particle	NN + DTNN2	**وضوح الشاشة** فى الهاتف كان.
Prep + DTNN2 + Particle + DTNN1	DTNN1	فى الشقة كانت **الغرف**.....
Prep + DTNN1 + Particle + NN	NN	فى الحقيقة كانت **خدمة**.....
Prep + DTNN2 + DTNN1	DTNN1	فى الواقع **المباراة** .....
Prep + DTNN2 + NN2 + NN1	NN2 + NN1	فى الفندق **منتجع صحى** .....
NN + DTNN2 + Prep + DTNN1	NN + DTNN2	**خدمة الغرف** فى الفندق .....
DTNN1 + Particle	DTNN1	**الإضاءة** كانت .....
DTNN2 + Prep + DTNN1	DTNN2	**السرد** فى القصة.....
DTNN1 + Prep + NN	NN	المطعم فى **موقع** .....
NN + DTNN1	NN + DTNN1	**بوفية إفطار** .....
DTNN1	DTNN1	**الرواية** .....
NN	NN	**هاتف** .....

When the algorithm detects an opinion-carrying word as described in section (4.1.1), it goes backward to extract the target noun as the candidate aspect for the entity according to the syntactic sentence patterns in table (4.2). When the sentence is located, its corresponding target candidate aspect is extracted - as shown in the second column of table (4.2). In Arabic example the bold words represents the extracted aspects in each pattern and the doted points represents the place of the opinion-carrying words.

The pseudo code for extracting the aspects according to the syntactic rules is displayed in figure (4.5).

```
1.  For each opinion carrying word  found
2.    Check the orientation type (positive/negative)
3.      Loop backward until word_no=0:
4.        Check the POS word
5.          If a negation word
6.              Opposite the orientation
7.              Go backward (word_no --)
8.          If the root belongs to the Excluded Roots
9.              Go backward (word_no --)
10.         If a preposition
11.             If(word_no!=0)
12.               Remove the feature value ( feature= " ")
13.               Go backward (word_no --)
14.             End if
15.         If any POS tag unless noun
16.             Go backward (word_no --)
17.         If a noun
18.             Update the feature score
19.             Go backward (word_no --)
20.         End if
21.     End loop
22.  Save the feature with its orientation type and score
23. End for
```

Figure 4.5: The outline of extracting the entity aspects for the backward direction

The system examines the POS type for the current word; if it is not a noun, it will ignore this word and go backward searching for the nearest noun. An exception for the intended nouns is the introductory words such as (clearly: في الحقيقة ,briefly: بإختصار). If the existing noun has a root belonging to one of the Excluded Roots, described in section (3.3.2), it will be ignored by the system.

The aspect sentiment orientation is determined according to the corresponding opinion-carrying word. If the opinion-carrying word has a positive sentiment orientation, the aspect will also have a positive sentiment orientation. Also, the score of the aspect will be determined by the opinion carrying word value plus the intensifier value according to the polarity. The output of this algorithm is a list of candidate aspects with their orientation type, score and certainty value. For example, the candidate aspect (room service 'خدمة غرفة', Pos, +1, 1).

4.3.1.2. Forward Direction

The fact that adjectives and nouns are good predictors of a sentence being subjective does not, however, imply that other POS do not contribute to express opinion

or sentiment. Verbs like “love, أحب” or "recommend, أوصي" can be used as strong indicators for the sentiment orientation (Riloff et al. 2003). So we tagged about 52 positive and negative roots[6] used as opinion-carrying words but in the forward direction. For example (loved "أحببت", recommend "أوصى") have positive orientation and (lacks "يفتقر", suffers "يعانى") have negative orientation. The algorithm for defining the aspects in forward direction is shown in figure (4.6).

In the backward direction, the algorithm searches for all nouns till the start of the sentence. But in this direction, the search will go forward within the next three words only. Since the most extracted aspects doesn't exceed three words. If the algorithm finds a negation word, the orientation will be reversed. Also the orientation score will be increased or decreased, if an intensifier is found. The nouns that have roots belonging to the Excluded roots will be eliminated. The extracted aspect is assigned by the same sentiment (type and score) and certainty value as its base opinion carrying word.

```
1.  For each opinion-carrying word found
2.      Check the type of the orientation (positive/negative)
3.      Check the POS of the next 3 words only
4.          If a negation word
5.              Opposite the orientation type
6.          If an intensifier
7.              Increase/decrease the orientation value w.r.t SO of root
8.          If the root belongs to the Excluded roots
9.              Eliminate it
10.         If a noun
11.             Update the feature
12.         End if
13.     Save the feature with the orientation type and score.
14. End for
```

Figure 4.6: The outline of extracting the entity aspects for the forward direction

4.3.2. Aggregating lemma-based candidate aspects

The purpose of this task is to group similar candidate aspects and compute their sentiment and certainty scores. Two main problems face the process of aggregating candidate aspects. The first problem is that the same aspect can be represented in different lexical forms in different reviews (e.g.' خدمة الغرفة', 'خدمات الغرف'). The second problem is the presence of the entity name inside some of the candidate aspects which

[6] Samples of the positive and negative roots for forward direction are shown in Appendix (A).

leads to the existence of extra different forms of the same aspects. For example, 'hotel team work' and 'team work' refers to the same aspect 'team work' in the hotel reviews.

To overcome the first problem, we represent the aspect words with their canonical lemma forms. For nouns and adjectives, lemma represents the abstract form of the words that have the same meaning, and hence capture semantic similarities between words. For verbs, it is perfective third person masculine singular form. The lemma form is proved to be the smallest form that captures all semantic features of the word. Lemmatization transforms the inflected word form to its dictionary lemma look-up form. For example, an aspect in the hotel domain can take different lexical forms as shown in table (4.3).

Table 4.3: Different lexical forms reduced to one lemma form

Different lexical Forms	One Lemma Form
الخدمات بالغرف	خدمة غرفة
خدمة الغرفة	
خدمات الغرف	
الخدمة بالغرفة	

To overcome the second problem, we adopted a simple assumption that the 'entity name' usually has the highest frequency in the review text. Therefore, all single and compound noun terms are counted, and the highest frequency term is removed from all extracted candidate aspects.

The sentiment score of the lemma-based aspect is represented by the sum of sentiment scores of all different lexical forms of the aspect. The certainty factor of each lemma-based aspect is the average of all certainty scores of the different lexical forms of the aspect. Thus, the aggregation process outputs the non-repeated aspects along with their total sentiment scores, and average certainty values (e.g. room service 'خدمة غرفة', +4, 0.75).

Chapter 5

Chapter 5
Experiments and Results

In this chapter, the testing datasets that used to experiment the proposed approach will be clarified along with the evaluation metrics used to measure the accuracy of the proposed algorithm. The experiments will be carried out in two levels; the first level is on the entity-level as determining the whole document orientation and the second one is on the aspect-level that is concerned with the aspect extraction process.

5.1. Dataset Description

One of the major limitations for Arabic research is the lack of adequate resources that could help in testing the system to get good evaluation for the system performance. The intended dataset is generic customer's reviews about any topic, product, object or company. Most of the existing datasets for reviewing objects are written in English Language and very little is written in Arabic.

We used two types of datasets to evaluate the performance of the proposed approach. The first dataset contains 500 movie reviews collected from different web pages and blogs in Arabic, 250 of them considered as positive reviews, and the other 250 as negative opinions. To our best knowledge, this is the only Arabic dataset available to the scientific community that can be used for sentiment analysis[7] and is called "Opinion Corpus for Arabic" OCA (Rushdi-Saleh et al. 2011).

In order to evaluate the performance of the proposed approach, Arabic reviews from different domains must be used. The second dataset contains 200 Arabic reviews in four different domains: hotels, novels, products, and football game events. The source of the dataset includes comments from different websites (e.g. tripadvisor.com.eg, goodreads.com, unlimit-tech.com, android4ar.com and Al-ahly.com). The dataset is made available for researchers in Arabic sentiment analysis[8]. Table (5.1) summarizes the testing datasets.

[7] http://sinai.ujaen.es/oca-corpus-en/

[8] https://www.scribd.com/eng.shismail

Table 5.1: The testing datasets

Dataset	Domain	No. of Entities	No. of Reviews	Avg. Token/Review	Avg. Sentence/Review
OCA	Movies	15	500	431	16
Our Dataset	Hotels	3	75	65	11
	Novels	3	46	96	13
	Products	3	58	47	7
	Events	3	21	104	12

5.2. Experiments

Two experiments were carried out to test the performance of the proposed system. The first experiment aimed to measure the efficiency of the proposed system at the entity level in different domains. This experiment is carried out using the two datasets. The results of applying our algorithm, using the first dataset, are compared with the results obtained by the author of the dataset.

The second experiment evaluates the efficiency of extracting the entities' aspects. In this experiment the results obtained by the system were compared to those extracted by human experts. Two domain-oriented human judges are asked to determine the proper aspects of each object reviews along with their polarities as shown in the form found in Appendix (C). The selected aspects and scores were automatically processed to ensure that there are no redundant aspects extracted for the same entity in different reviews. The processing includes lemma form generation, aggregating similar aspects, and computing sentiment scores for each aspect.

A sample output of the extracted aspects by one of the human judge with their orientation score is shown in table (5.2).

Table 5.2: Sample of the extracted aspects by a human judge

Aspects	Positive No	Negative No	Orientation score	Sentiment Orientation
خدمة	3	0	7	Positive
موقع	5	0	8	Positive
حمام سباحة	0	1	-2	Negative
شقة	2	0	3	Positive
غرف	5	0	6	Positive
أثاث	1	1	0	Neutral

5.3. Evaluation Measures

Many measures of retrieval effectiveness have been proposed. In this work, Precision (P), Recall (R), F-measure, and accuracy metrics are used to evaluate the performance of the proposed system for aspect extraction. Precision is an estimate of the probability that a given model identifies an aspect as relevant to a user's aspects (How many selected aspects are relevant?). Recall is an estimate of the probability that, if an aspect is relevant to a user's aspects, then a given model will classify it as relevant (How many relevant aspects are selected?). Both recall and precision take on values between 0 and 1.

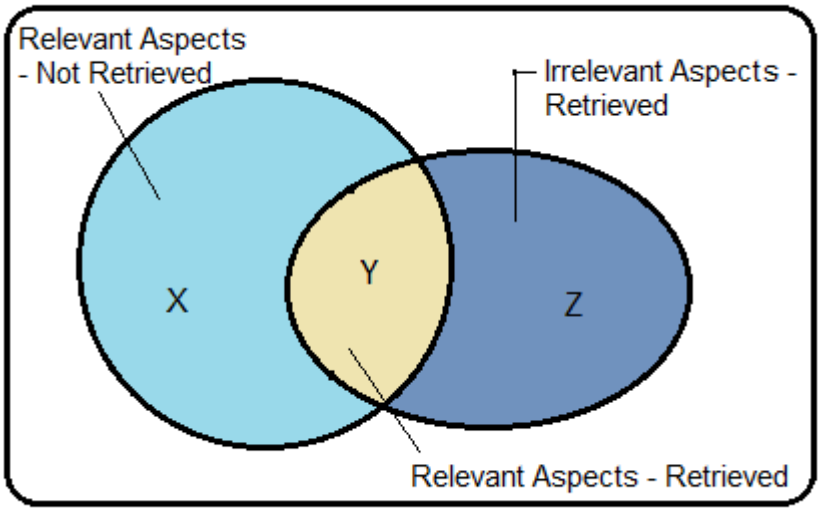

Figure 5.1: Precision and Recall Description

Using the description shown in figure 5.1, precision (P) and recall (R) are given by the following equations.

$$P = Y/Z \qquad (5.1)$$

$$R = Y/X \qquad (5.2)$$

Where: X: aspects (or opinions) identified by experts.

Y: Intersection of aspects identified by both of the system and expert.

Z: aspects (or opinions) identified by the system.

In many cases, it is important to evaluate precision and recall in conjunction, because it is easy to optimize either one separately. The two measures are integrated together in what is called F-Measure. The F-Measure consists of a weighted combination of precision and recall which is sometimes called harmonic mean (Mehlitz et al. 2007). The general form of F-Measure is given by equation (5.3).

$$F - measure = \frac{(\alpha^2 + 1) * P * R}{\alpha^2 * P + R} \qquad (5.3)$$

Where α is a weighting factor that determines the relative importance of precision and recall (Hripcsak and Rothschild, 2005). However, in most experiments, there is no particular reason to favor precision or recall, so most researchers use a balanced weighting measure between precision and recall with α=1 as shown in equation (5.4).

$$F - measure = \frac{2 * precision * recall}{precision + recall} \qquad (5.4)$$

Accuracy is a measure used to evaluate the percentage of agreement between the output of the proposed system and the output of the domain expert at the entity-level. It is computed by dividing the number of agreed documents by the total number of documents as shown in equation (5.5)

$$Accuracy = \frac{No\ of\ true\ orientation\ documents}{Total\ No\ of\ documents} \qquad (5.5)$$

5.4. Experiment 1: Evaluating Aspect Extraction

In this experiment the aspects extracted by the system are compared with aspects defined by two human experts. The common aspects between experts are taken as the reference aspects of the entity. Table (5.3) shows the extracted aspects by the system, experts and the common aspects between them for an event reviews.

Table 5.3: The extracted aspects by both the system and the expert for an event

Number	System Aspects	Expert Aspects	Common Aspects	Expert Orientation	System Orientation
1	إدارة	إدارة	إدارة	Pos	Pos
2	إمكانية فنية فريق	إمكانية فنية فريق	إمكانية فنية فريق	Neg	Neg
3	الزمالك	الزمالك	الزمالك	Pos	Pos
4	اهلى	اهلى	اهلى	Neut	Neut
5	ايمن حفنى لاعب	ايمن حفنى لاعب	ايمن حفنى لاعب	Pos	Pos
6	تحكيم	تحكيم	تحكيم	Neg	Neg
7	تسلل	تسلل	تسلل	Neg	Neg
8	تعادل	تعادل	تعادل	Pos	Pos
9	جمهور	جمهور	جمهور	Pos	Pos
10	جهاز فني	جهاز فني	جهاز فني	Pos	Pos
11	حكم	حكم	حكم	Neg	Neg
12	حماس مهارة	حماس مهارة	حماس مهارة	Pos	Pos
13	ضربة جزاء	ضربة جزاء	ضربة جزاء	Pos	Neut
14	فريق الزمالك	فريق الزمالك	فريق الزمالك	Neut	Neut
15	فوز	فوز	فوز	Pos	Pos
16	لاعب	لاعب	لاعب	Neg	Neg
17	لاعب دفاع	لاعب دفاع	لاعب دفاع	Pos	Pos
18	مباراة	مباراة	مباراة	Neg	Neg
19	محمد صلاح	محمد صلاح	محمد صلاح	Pos	Neg
20	مدرب	مدرب	مدرب	Neg	Neg
21	مستوى حسام عاشور	مستوى حسام عاشور	مستوى حسام عاشور	Pos	Pos
22	مستوى محمد رزق	مستوى محمد رزق	مستوى محمد رزق	Pos	Pos
23	مستوي	مستوي	مستوي	Neg	Neg
24	منافسة	منافسة	منافسة	Neg	Neg
25	مهارة وليد سليمان تريزيجيه	مهارة وليد سليمان تريزيجيه	مهارة وليد سليمان تريزيجيه	Pos	Pos
26	نادي الزمالك	نادي الزمالك	نادي الزمالك	Pos	Pos
27	هدف	هدف	هدف	Neg	Neg
28	ثبات	نقطة			
29	دوري	انفراد			
30	صلاح الدين	جهد بدن فن			
31	فريق				
32	قمة				
33	نجيب				

From table (5.3), it is clear that the common aspects are 27 aspects out of 30 aspects by experts and 33 aspects by system. The left and down highlight words are the aspects that are not common. This experiment is applied to the whole documents of the testing dataset (4 domains) and the total number of extracted aspects by the system, experts, and the common aspects between them are shown in table (5.4).

Table 5.4: Number of the extracted aspects by system, experts, and the common

Review Document	System (Z)	Experts (X)	Common Aspects (Y)
Hotel1	62	56	53
Hotel2	62	60	55
Hotel3	62	58	51
Novel1	58	58	51
Novel2	42	37	34
Novel3	47	46	41
Laptop	53	53	46
Phone	43	43	38
Windows 8	36	35	30
Event 1	34	35	30
Event 2	24	23	20
Event 3	33	30	27

Table (5.5) shows the precision, recall and f-measure of the extracted aspects by the system and the experts.

Table 5.5: Precision, Recall and F-measure of extracted aspects

Category	**Precision**	**Recall**	**F- Measure**
Hotel1	0.855	0.946	0.898
Hotel2	0.887	0.916	0.901
Hotel3	0.850	0.879	0.865
Average Hotel	**0.864**	**0.914**	**0.888**
Novel1	0.879	0.879	0.879
Novel2	0.810	0.919	0.861
Novel3	0.872	0.891	0.881
Average Novel	**0.854**	**0.896**	**0.874**
Laptop	0.868	0.868	0.868
Phone	0.884	0.884	0.884
Windows 8	0.833	0.857	0.845
Average Product	**0.862**	**0.870**	**0.866**
Event1	0.882	0.857	0.869
Event2	0.833	0.869	0.851
Event3	0.818	0.900	0.857
Average Event	**0.844**	**0.875**	**0.859**
Total Average	***0.856***	***0.889***	***0.872***

Table (5.5) shows the average recall of the system that ranges from 85% to 91% for the four domains (hotels, novels, products and events). The average precision of the system ranges from 81% to 88% for the same domains. The results show that the recall values are higher than the precision because the number of extracted aspects by the system is higher than that extracted by the experts. The precision and recall values for the product and event reviews (football games) are slightly lower than other domains due to the difficulty of extracting the event aspects. This difficulty comes from the fact that the event may contain many actors upon which the reviewer can comment about their aspects such as the players, referee, audience (fans), and the environment in the football game events. Also, the product reviews contain more uncertain opinion words like (كبير، صغير، خفيف، ثقيل). The results show comparable accuracy values for extracting entities' aspects from reviews in different domains with an average precision 85% and average recall 89%.

5.5. Experiment 2: Evaluating opinions

The objective of this experiment is to measure the efficiency of the proposed system at the entity level in different domains. The first part of the experiment concerns applying our algorithm on the first dataset (OCA) which includes 500 movie reviews[9] (250 positive and 250 negative). The precision, recall and accuracy are computed and compared to the corresponding values obtained by Pang et al. (2002) and Rushdi-Saleh et al. (2011) using the same dataset. Table (5.6) shows the results of this comparison.

Table 5.6: The testing results compared to Pang & OCA

	Precision	**Recall**	**Accuracy**
Pang	0.8619	0.8450	0.8535
OCA	0.8738	0.9520	0.9060
Our approach	**0.9528**	**0.9680**	**0.9600**

Although OCA system has been trained by using data in the same domain, its accuracy does not exceed 90%, while it reached 96% using our proposed system. This proves that the use of our sentiment annotated lexicon and aspect extraction algorithm

[9] A sample of positive and negative orientation of Movie Reviews is shown in Appendix (D).

outperforms the use of domain-oriented corpus and machine learning algorithms (SVM, NB).

Then we applied our algorithm on the second dataset, which is a collection of Arabic reviews about entities in different domains. In this experiment we compare the opinion orientations obtained by our proposed system with that obtained by two human experts. We compute the percentage of agreement between aspects extracted by the proposed system and aspects extracted by the two experts using equation (5.6).

$$Percentage\ of\ agreement = \frac{Number\ of\ matched\ aspects}{Number\ of\ common\ aspects\ bw\ judge\ \&\ system} \quad (5.6)$$

The entity aspects shown in table (5.3) declare that most of the extracted aspects have the same orientation except two aspects only (highlighted in the right side). So the accuracy of this entity is 25/27 (92.6%). Table (5.7) displays the percentages of opinion agreement between the aspects identified by the experts and the system for all entities with the average of each domain.

Table 5.7: Percentage of orientation agreement at entity level

Domain	**Entity**	**System vs. Expert1 (%)**	**System vs. Expert2 (%)**	**Average Agreement per Domain**
Hotel	Hotel1	96.2	93.6	93.7%
	Hotel2	95.4	92.4	
	Hotel3	94.1	90.5	
Novel	Novel1	90.2	88.4	89.35%
	Novel2	88.2	86.5	
	Novel3	92.7	90.1	
Product	Laptop	84.8	83.2	90%
	Phone	97.3	93.7	
	Windows 8	91.4	89.6	
Event	Event1	90.0	87.9	91.38%
	Event2	95.0	92.3	
	Event3	92.6	90.5	
Total Average				***91.10%***

The results show high degrees of agreement with experts ranges from 89% in Novel domain to 93% in hotel domain with an average agreement 91%. The high accuracy rates of the current work promotes the proposed methodology using the

sentiment annotated lexicon instead of exploiting predefined lists of opinion words or entity features which is domain dependent. Also, our adopted assumption of equal opinion weights (like or dislike) gives better results than using different weights for strong and weak opinion word synonyms. This assumption reduced the impact of some negation errors as ("the service is not excellent") which needs special treatment to avoid reversing the polarity. But, using the simple negation algorithm, the equal opinion weights rule returns -1 compared to -4 or -5 in different approaches using different weights for strong opinion words.

To investigate the effect of assigning sentiment tags at the root and pattern levels and hence ignoring context-dependent sentiment words, the certainty factor for each aspect is calculated. Certainty factor (CF) is used to measure the percentage of certainty of the extracted aspects orientations using the following equation.

$$Certainty\ Factor = \frac{Certain\ Score}{Total\ Aspcet\ Frequency} \qquad (5.7)$$

For the certain root, we are sure that it is positive or negative, so we give it the value 1. But for the uncertain root, we are not sure so we give it the value 0.5. Table (5.8) shows the certainty factor for each aspect separately and the total average for the entity.

For example: an aspect x appeared 3 times, 2 of them used certain roots, and one uncertain. So the certain score for this feature is 2.5 and the certainty factor is = 2.5/3 = 0.83%.

Table 5.8: Average Certainty Factor for an event entity

Number	Common Aspects bw System & Expert	Frequency	Certain Score	Certainty Factor
1	إدارة	2	2	1
2	إمكانية فنية فريق	1	1	1
3	الزمالك	2	1	0.5
4	اهلى	3	3	1
5	ايمن حفنى لاعب	1	1	1
6	تحكيم	2	2	1
7	تسلل	1	1	1
8	تعادل	1	1	1
9	جمهور	1	1	1
10	جهاز فني	1	1	1
11	حكم	1	0.5	0.5

12	حماس مهارة	1	1	1
13	ضربة جزاء	2	2	1
14	فريق الزمالك	2	2	1
15	لاعب	1	0.5	0.5
16	فوز	1	1	1
17	مباراة	4	4	1
18	محمد صلاح	1	0.5	0.5
19	مدرب	1	1	1
20	مستوى حسام عاشور	1	1	1
21	مستوى محمد رزق	1	1	1
22	مستوي	1	1	1
23	لاعب دفاع	1	1	1
24	منافسة	1	1	1
25	مهارة وليد سليمان تريزيجيه	1	0.5	0.5
26	نادي الزمالك	2	2	1
27	هدف	1	1	1
Average Certainty Factor for this Entity				0.907407

The average certainty factor for each entity reflects the percentage of assurance of the entity orientation. Table (5.8) shows that the percentage of assurance of an event aspects orientation is 90.7%. Table (5.9) shows the certainty factor of each entity and the average certainty of each domain.

Table 5.9: The average certainty factor for each entity and domain

Domain CF	**Document**	**CF (%)**	**Average CF per Domain**
Hotel	Hotel1	0.889	92.5%
	Hotel2	0.943	
	Hotel3	0.943	
Novel	Novel1	0.892	87.4%
	Novel2	0.853	
	Novel3	0.878	
Product	Laptop	0.853	88.4%
	Phone	0.904	
	Windows 8	0.896	
Event	Match1	0.921	91.3%
	Match2	0.910	
	Match3	0.907	
Total Average			***89.9%***

We can conclude from the results of table (5.7) and (5.9) that the certainty factor of aspect opinions affects the average agreement between the system and human experts. For example, the hotel domain that has the highest average CF of 92%, which leads to the highest average agreement of 93%. Also the lowest CF of 87% with novel domain leads to the lowest average agreement of 89%.

This proves that the proposed system is generic and able to extract the object aspects and orientation from Arabic reviews in different domains.

Chapter 6

Chapter 6
Conclusions and Future Work

In this chapter, the conclusion of the proposed approach is presented in section 6.1. Section 6.2 describes some problems and limitations facing the proposed approach. Section 6.3 presents some suggestions for future work in this area of research.

6.1. Conclusions

In this thesis, we presented a generic approach for extracting the aspects of objects and events of Arabic reviews as well as their orientation. The idea that the object aspects and their orientations are usually correlative is adopted in this research. The proposed approach does not exploit predefined set of features, nor domain ontology hierarchy. Instead we add sentiment tags on the pattern and root levels of Arabic lexicon and used these tags to extract the opinion carrying words and their relevant aspects. We can conclude the main contributions of the thesis as follows:

1) Building a sentiment-annotated lexicon by adding semantic tags for the roots and patterns of the lexicon to define its sentiment role. This makes the proposed approach more suitable for use in various domains beyond the product and service reviews.
2) Exploiting the sentiment-annotated lexicon to extract the opinion-carrying words taking into consideration the negation and intensification effects.
3) Extracting the entities aspects according to syntactic patterns for Arabic sentences and based on the opinion-carrying words. The lemma forms of the candidate aspects are used to overcome the problem of redundant aspects.
4) Summarizing these extracted aspects with their orientations and scores to determine the entity orientation.
5) Presenting a certainty factor to express the percentage of orientation certainty of each aspect and declaring its effect on the system accuracy.

The system is evaluated on the entity-level using 500 movie reviews with accuracy 96%. Then the system is tested on the aspect-level using 200 Arabic reviews in different

domains (Novels, Laptops, Mobile phones, Windows 8, Football game events and Hotels).

The results of extracted aspects from the system along with their orientation are compared with that defined by human experts. The system results are very close to that of the experts. On average, the proposed system achieves a recall 89%, a precision 85% and F-measure 87%. Thus, the proposed system proves its ability to rely upon in helping both customers and object authors by summarizing the existing object reviews.

6.2. Problems

The proposed system suffers from some limitations that have been discovered during the analysis of the experimental results. One of these problems is that the precision and recall values for some domains are slightly lower than others due to the difficulty of extracting their aspects. This difficulty comes from the fact that some reviewers focus on certain problems without stating any of the entity aspects as shown in the following review for "Windows 8".

"الويندوز جميل جدا ومميز، لكن للاسف انا أعمل في صيانة الموبايل، وأشتريت حاسوب يعمل باللمس ومجهز بنظام ويندوز ٨ الاصلي، لكن معظم اجهزة فحص المحمول كجهاز بوكس الترنادو او ما يسمى UFS3 وبوكسMAXKEY لا تعمل على هذا النظام، واصبت بالخيبه كوني لا اود ان افقد النسخه الاصليه، وأعود لنظام ويندوز سيفين"

Another problem faces our aspect-based opinion extraction is that the overall entity orientation is determined by aggregating the orientations of all entity aspects. In some cases, the reviewer may start with a phrase concluding that the entity is excellent followed by many phrases focusing on its malfunctions or comparison with similar entities. This may lead to wrong decisions on the entity level based on the stated aspects' orientations as shown in the following review "cell phone".

"هذا الجهاز ***رائع*** ، ولكنه يعيبه الكاميرا الأمامية فهي ***ضعيفة***، وعمر البطارية ***قصير***، مقارنة بجهاز آي فون"

Another problem arises from the use of synonyms of entity aspects. Although it does not affect the entity-level orientation, it leads to extracting redundant or similar aspects stated in different synonyms in different reviews as shown in the following aspects extracted from hotel reviews.

("بوفيه الافطار"، "وجبة الافطار"، "غرفة الطعام"، "المطعم") & ("الغرفة"، "الشقة"، "الجناح"، "الاقامة").

Also, the reviewed entity may contain Named Entities (NE) such as actors, players, writer, product name…etc. Some of these NEs are Arabic adjectives and may be considered as opinion carrying words which lead to extracting fake aspects as shown in the following review.

"قام الكابتن محمد ***لطيف*** بالتعليق على مباراة الأهلي والزمالك" & "أحرز الهدف اللاعب عماد ***متعب***".

Also, some words that not considered as negation words, but they reverse the orientation and the system couldn't detect it as shown in the next sentences.

"كان من **المفترض** أن يكون أداؤه جيد" & " **المتوقع** من مخرج رائع الحصول على فيلم جيد"

In addition to some special cases that use words with conflict meaning in the negation and/or intensification processes such as

"تمكن أحد المستخدمين من ***اختراق*** نظام الحماية ***بسهولة*** شديدة" ,"***الفوز الصعب*** للنادي الأهلي" , "أداء الجهاز ***ليس ممتازاً***".

In addition to the quad-literal roots that are not included in the used Arabic lexicon such as (هلهل - عبقر).

6.3. Future Work

To enhance the proposed system and overcome the above stated problems, some suggestions for future work can be outlined as follow:

1) Adopting an accurate technique to handle the problem of named entities by using an Arabic NE recognizer.
2) Solving the issue of using different sentiment words to describe the same aspect and also for describing different aspects by adding the semantic level to the proposed sentiment analysis.
3) Complete the sentiment-lexicon for the quad roots of Arabic language.
4) Using the Arabic WordNet to overcome the problem of synonym aspects.

References

References

1) Abbasi, A., Chen, H., & Salem, A. "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums", ACM Transactions on Information Systems (TOIS), Vol. 26, No. 3, (2008), pp. 12.
2) Abdul-Mageed, M., & Diab, M. "AWATIF: a multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis", In Proceedings of LREC, Istanbul, Turkey, (2012).
3) Abdul-Mageed, M., Diab, M., & Kübler, S. "SAMAR: Subjectivity and sentiment analysis for Arabic social media", Computer Speech & Language, Vol. 28, No. 1, (2014), pp. 20-37.
4) Andreevskaia, A., & Bergler, S. "When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging", ACL, (2008), pp. 290-298.
5) Baccianella, S., Esuli, A., & Sebastiani, F. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", LREC, Vol. 10, (2010), pp. 2200-2204.
6) Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., & Subrahmanian, V. S. "Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone", ICWSM, (2007).
7) Brill, E. "Some advances in transformation-based part of speech tagging", arXiv preprint cmp-lg/9406010, (1994).
8) Choi, Y., & Cardie, C. "Learning with compositional semantics as structural inference for subsentential sentiment analysis", In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, (2008), pp. 793-801.
9) Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J., & Vallejo, C. G. "A knowledge-rich approach to feature-based opinion extraction from product reviews", In Proceedings of the 2nd international workshop on Search and mining user-generated contents, ACM, (2010), pp. 13-20.
10) Dichy, J., Krauwer, S., & Yaseen, M. On lemmatization in Arabic, "A formal definition of the Arabic entries of multilingual lexical databases", In ACL/EACL

11) 2001 Workshop on Arabic Language Processing: Status and Prospects. Toulouse, France, (2001).
12) Ding, X., Liu, B., & Yu, P. "A holistic lexicon-based approach to opinion mining", In Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, (2008), pp. 231-240.
13) Ding, X., Liu, B., & Zhang, L. "Entity discovery and assignment for opinion mining applications", In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, (2009), pp. 1125-1134.
14) El-Halees, A. "Arabic opinion mining using combined classification approach", In Proceeding of: 2011 International Arab Conference on Information Technology ACIT, (2011).
15) Elhawary, M., & Elfeky, M. "Mining Arabic business reviews", In Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, IEEE, (2010), pp. 1108-1113.
16) El-Shishtawy, T., & Al-Sammak, A. "Arabic keyphrase extraction using linguistic knowledge and machine learning techniques", arXiv preprint arXiv:1203.4605, (2012).
17) El-Shishtawy, T., & El-Ghannam, F. "An Accurate Arabic Root-Based Lemmatizer for Information Retrieval Purposes", arXiv preprint arXiv:1203.3584, (2012).
18) Esuli, A., & Sebastiani, F. "Determining Term Subjectivity and Term Orientation for Opinion Mining", EACL, Vol. 6, (2006).
19) Farra, N., Challita, E., Assi, R. A., & Hajj, H. "Sentence-level and document-level sentiment mining for Arabic texts", In Data Mining Workshops (ICDMW), 2010 IEEE International Conference, IEEE, (2010), pp. 1114-1119.
20) Gamon, M., & Aue, A. "Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms", In Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing, Association for Computational Linguistics, (2005), pp. 57-64.

21) Ghorashi, S. H., Ibrahim, R., Noekhah, S., & Dastjerdi, N. S. "A Frequent Pattern Mining Algorithm for Feature Extraction of Customer Reviews Extraction of Customer Reviews", (2012).

22) Habash, N. Y. "Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies", Vol. 3, No. 1, (2010), pp. 1-187.

23) Habash, N., Rambow, O., & Roth, R. "MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization", In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, (2009), pp. 102-109.

24) Harb, A., Plantié, M., Dray, G., Roche, M., Trousset, F., & Poncelet, P. "Web Opinion Mining: How to extract opinions from blogs?", In Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, ACM, (2008), pp. 211-217.

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

26) Hatzivassiloglou, V., & Wiebe, J. M. "Effects of adjective orientation and gradability on sentence subjectivity", In Proceedings of the 18th conference on Computational linguistics, Association for Computational Linguistics, Vol. 1, (2000), pp. 299-305.

27) Hripcsak, G., & Rothschild, A. S. "Agreement, the f-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association", Vol. 12, No. 3, (2005), pp. 296-298.

28) Hu, M., & Liu, B. "Mining opinion features in customer reviews", AAAI, Vol. 4, No. 4, (2004), pp. 755-760.

29) Hu, M., & Liu, B. "Mining and summarizing customer reviews", In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, (2004a). pp. 168-177.

30) Hu, M., & Liu, B. "Opinion Feature Extraction Using Class Sequential Rules", In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, (2006), pp. 61-66.

31) Hu, M., & Liu, B. "Opinion extraction and summarization on the web", AAAI, Vol. 7, (2006a), pp. 1621-1624.

32) Kennedy, A., & Inkpen, D. "Sentiment classification of movie reviews using contextual valence shifters", Computational Intelligence, Vol. 22, No. 2, (2006), pp. 110-125.

33) Kim, S. M., & Hovy, E. "Extracting opinions, opinion holders, and topics expressed in online news media text", In Proceedings of the Workshop on Sentiment and Subjectivity in Text, Association for Computational Linguistics, (2006), pp. 1-8.

34) Khan, K., Baharudin, B., & Khan, A. "Identifying product features from customer reviews using hybrid patterns", Int. Arab J. Inf. Technol., Vol. 11, No. 3, (2014), pp. 281-286.

35) Khoja, S., & Garside R. "Stemming Arabic text". Computer Science Department, Lancaster University, Lancaster, UK, (1999).

36) Korayem, M., Crandall, D., & Abdul-Mageed, M., "Subjectivity and sentiment analysis of arabic: A survey", Advanced Machine Learning Technologies and Applications. Springer Berlin Heidelberg, (2012), pp. 128-139.

37) Lazhar, F., & Yamina, T. G. "Identification of Opinions in Arabic Texts using Ontologies". In Workshop on Ubiquitous Data Mining, (2012), p. 61.

38) Liu, B., Hu, M., & Cheng, J. "Opinion observer: analyzing and comparing opinions on the web", In Proceedings of the 14th international conference on World Wide Web, ACM, (2005), pp. 342-351.

39) Liu B., "Tutorial on sentiment analysis" based on Chapter 11 of the book "Web Data Mining - Exploring Hyperlinks, Contents and Usage Data", (http://www.cs.uic.edu/~liub/), (2007).

40) Liu, B. "Sentiment analysis and subjectivity". Handbook of natural language processing, Vol. 2, (2010), pp. 627-666.

41) Medhat, W., Hasan A., & Korashy, H. "Sentiment Analysis Algorithms and Applications: A Survey", Ain Shams Engineering Journal, Vol. 5, (2014), pp. 1093–1113.

42) Mehlitz, M., Bauckhage, C., Kunegis, J., & Albayrak, S. "A new evaluation measure for information retrieval systems", In Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference, IEEE, (2007), pp. 1200-1204.

43) Mourad, A. & Darwish, K. "Subjectivity and Sentiment Analysis of Modern Standard Arabic and Arabic Microblogs", In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Atlanta, Georgia, (2013), pp. 55–64.

44) Osgood, C. E. "The measurement of meaning", University of Illinois press, No. 47, (1957).

45) Pang, B., Lee, L., & Vaithyanathan, S. "Thumbs up? sentiment classification using machine learning techniques", In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Association for Computational Linguistics, Vol. 10, (2002), pp. 79-86.

46) Pang, B., & Lee, L. "Opinion mining and sentiment analysis", Foundations and trends in information retrieval, Vol. 2, (2008), pp. 1-135.

47) Polanyi, L., & Zaenen, A. "Contextual valence shifters", In Computing attitude and affect in text: Theory and applications, Springer Netherlands, (2006), pp. 1-10.

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

49) Qiu, G., Liu, B., Bu, J., & Chen, C. "Opinion word expansion and target extraction through double propagation", Computational linguistics, Vol. 37, No. 1, (2011), pp. 9-27.

50) Quirk, R., Greenbaum, S. L., & Leech, G. G. and Svartvik, J. "A Comprehensive Grammar of the English Language", Harlow: Longman, (1985).

51) Riloff, E., Wiebe, J., & Wilson, T. "Learning subjective nouns using extraction pattern bootstrapping", In Proceedings of the seventh conference on Natural

language learning at HLT-NAACL 2003, Association for Computational Linguistics, Vol. 4, (2003), pp. 25-32.

52) Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. "OCA: Opinion corpus for Arabic. Journal of the American Society for Information Science and Technology", Vol. 62, No. 10, (2011), pp. 2045-2054.

53) Sawalha M., & Atwell, E. "Comparative Evaluation of Arabic Language Morphological Analysers and Stemmers", Coling 2008: Companion volume – Posters and Demonstrations, Manchester, (2008), pages 107–110.

54) Schouten, K., & Frasincar, F. "Finding Implicit Features in Consumer Reviews for Sentiment Analysis", Web Engineering, Springer International Publishing, (2014), pp. 130-144.

55) Somprasertsri, G., & Lalitrojwong, P. "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization", J. UCS, Vol. 16, No. 6, (2010), pp. 938-955.

56) Stavrianou, A. & Chauchat, J. "Opinion Mining Issues and Agreement Identification in Forum Texts", Intelligence and Security Informatics, IEEE, (2007), pp. 51-58.

57) Stone, P., Dunphy, D. & Smith, M. "The General Inquirer: A Computer Approach to Content Analysis", (1966).

58) Su, Q., Xiang, K., Wang, H., Sun, B., & Yu, S. "Using pointwise mutual information to identify implicit features in customer reviews", In Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead, Springer Berlin Heidelberg, (2006), pp. 22-30.

59) Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. "Lexicon-based methods for sentiment analysis", Computational linguistics, Vol. 37, No. 2, (2011), pp. 267-307.

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

61) Versteegh, K. "The Arabic Language", Columbia University Press, (1997).

62) Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. "Development and use of a gold-standard data set for subjectivity classifications", In Proceedings of the 37th

annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, (1999), pp. 246-253.

63) Wilson, T., Wiebe, J., & Hwa, R. "Just how mad are you? Finding strong and weak opinion clauses", AAAI, Vol. 4, (2004), pp. 761-769.

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

65) Zhang, L. "Aspect and Entity Extraction from Opinion Documents", Doctoral dissertation, University of Illinois at Chicago, (2012).

66) Zhang, Y., & Zhu, W. "Extracting implicit features in online customer reviews for opinion mining", In Proceedings of the 22nd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, (2013), pp. 103-104.

67) Zhuang, L., Jing, F., & Zhu, X. Y. "Movie review mining and summarization", In Proceedings of the 15th ACM international conference on Information and knowledge management, ACM, (2006), pp. 43-50.

المراجع العربية

٦٨) شذا العرف فى فن الصرف ، الشيخ أحمد محمد الحملاوى ، ١٨٩٤.

٦٩) معجم اللغة العربية المعاصرة ، للدكتور أحمد مختار عمر، مارس ٢٠١٣.

٧٠) مجمع اللغة العربية "المعجم الوجيز"، الدكتور إبراهيم مدكور ، ١٩٩٤.

APPENDIXES

Appendix A

Samples of Certain Positive Roots

أمن أنق برع بهج بهر برق بسم ثقف ثمر جذب جرؤ جلل جدع جدد جمل حسس حفل حنن حنك حبب حسن حرص حصد خصب خشع ذكي ذهل دهش ربح رحم رقى زخر زهو سخي سمو سلس سهل سعد شسع شمس شغف شمخ شمل شهي شوق صفي صبر صلح ضحك طزج طور عدد عطر عظم عمق فرد فسح فطن فرح قوى كرم كفح كمل لطف لذذ لمع متع متن ميز نجح نسب نظف نظم نوع نبه نضج نقي نمق هذب هدأ وسع وضح وثق وسم وعظ وفي وفر يقن يقظ

Samples of Certain Negative Roots

أسي أجل بأس بخس بخل بشع بطش بغض بهظ بعد بهت ترف تعس تلف تعب ثغر ثقل جحد جحش جرم جشع جهل حبط حرش حرم حطم حقد حنث حجب حسد خبل ختل خدع خرب خسس خسف خطر خجل خشن خطأ خفض دجل دمن ذبل ذعر ذلل ذنب رجم رعب ركد ركك رهب رسب رهق زفت زهق زعج سخط سقم سمم سذج سمن شتت شحح شرس شرر شغب شنع صخب صدم صعب ضعف طمع طفح ظلم عبس عفش عفن عقد عنف غبي غشش غدر غشم غمض فتت فجع فزع فسد فشل فضح فظع قلل قبح قزز كذب كسل لعن لذع مهن ممل نشل نقص نبذ نذل هبب هبط هدر هدم هزل هبل وحش وعر وغد وقح يأس

Samples of Uncertain Positive Roots

بدع برز بسط برر برك بكر توج ثلج جبر جنن جني حبك حذر حدس خفف خلف خلط دلل رتب رهف رخص روع رطب رقق سكر سرع سمك شطر شفف شدد صلب ضحك طري عذب عون فتر قرب كثف كثر كبر لين وجز

Samples of Uncertain Negative Roots

بذل بلغ ثمن جمد حشد خرق صغر ضخم طول ظلل عزل غلي قدم قصر كرر ميس نثر

Samples of Forward Positive Roots

وجد ضمن عمق دعم حوي جود وصي حبب سرع نشط زود وفر عجب نجح روع قنع متع ذهل لزم فرد ميز هتم

Samples of Forward Negative Roots

سوء ضعف عيب فقر خلي ضيع سقط حشو نقص رجع خفض خسر

Samples of Excluded Roots

أكد ضيف صرح خوص خصر درج فرض سبب خبر جرب قبل جمع علم نطق حفظ دخل امر رحل نقط ناس جول قول شحن سبق نجم فرص غيب ريف صدأ عدد سجل بدل فوق شرف بلى ثني وصل فعل حسب شرع رمي

Samples of the sentiment-carrying patterns

افتعال فعالى فعالي فعالية فعالة فعلية فعول فعولة مفعال فعيل فعلان فعلاء فعلة مفعل مفعلة فاعل فاعلة مفالة مفتعل مفاعل مفاعلة متفاعل متفاعلة تفعيل مفعول مفعوله مفعولة فعل متفعل منفعل منفعلة افعال افتعل متفعلة متفعله

Comparator Patterns

افعل فعلى

Negation Words

بل لن بدون عدم لم مش ليس لا مفيش غير لست ليست

Samples of Intensifiers

تماما طبعا للغاية دوما فعلا أبدا مطلقا دائما حقا جدا بحفاوة بجد بالفعل كثيرا ذريعا بالطبع بشدة بالتأكيد

Appendix B

Some POS tagging abbreviations

NN : Noun

DTNN1 : The First Determined Noun found in backward direction

DTNN2 : The Second Determined Noun found.

Prep : Preposition

Particle : Any other POS tagging in Arabic language such as (كان وأخواتها - إن وأخواتها ..)

Appendix C

A form that given to the expert to extract the aspects & orientations.

جامعة بنها – كلية الهندسة بشبرا - قسم نظم الحاسبات رسالة بعنوان استخراج الرأى للنصوص العربية

إستمارة استخراج الخصائص والأراء للنصوص العربية

الاسم :..... الوظيفة :..... إستمارة رقم :.....

تحية طيبة..

من فضلك ضع خط تحت الكلمة التى ترى أنها تمثل خاصية للعنوان الرئيسى. وضع فوق كل خاصية علامة +،++ لو ترى أنها موجبة أو موجبة جدا، وعلامة - ، - - لو ترى أنها سالبة أو سالبة جدا، وفى حالة وصف العنوان نفسة بشكل عام يتم وضع خط تحت العنوان مع وضع العلامة ومع إمكانية وضعها أكثر من مرة فى حالة التكرار. ونشكركم على تعاونكم معنا.

قمة الأهلى والزمالك

١) قمة بلا روح بلا طعم بلا مدربين، هى عنوان مباراة الأهلى والزمالك فى الدوري الممتاز، بعض الحماس والمهارات الفردية أنقذتها من أن تكون مباراة أشبه بمباريات الدرجة الثانية وربما أقل. انتهت القمة بهدف للزمالك وآخر للأهلي والهدفين بالصدفة، بقى الزمالك متصدرا وواصل الأهلى ترنحه في المركز الثالث. أولا: الزمالك سيكتسح الأهلى سيفوز بستة أهداف. مباراة محسومة. توقعات جماهير الأهلى قبل جماهير الزمالك. والحقيقة التي كانت غائبة عن هؤلاء أن محمد صلاح يمتلك كل النجوم ولكن كان خائفا من المقاولون. فكيف يكتسح الأهلى؟ ثانيا: في مباراة القمة، كسر جاريدو أهم عنصر كان كفيل بتفوق الأهلى في السنوات الماضية، عنصر الثقة والثبات فى الملعب، فى القمة الأخيرة ظهر الأهلى ضعيفا مهلهلا مستسلما، لولا بعض مهارات وليد سليمان وتريزيجيه الفرديه ، وثبات نسبي لنجيب وسعد، لظهر الأهلى بمستوى لم يظهر به منذ ١٢ عاما تقريبا. ثالثا: ارتفاع مستوى محمد رزق. هو مشروع لاعب وسط دفاعي مميز جدا. سيظلمه قدره لو بقى غالي على جاريدو. رابعا: أيمن حفنى وأحمد عيد عبدالملك تلاعبا بكل لاعبى الأهلى، تألقهما الفردي أظهر قلة حيلة محمد صلاح، وعدم قدرة مدرب الأهلى إلا على مراقبة اللاعبين فقط! خامسا: سعادة الإدارة وسعادة الجهاز الفني وسعادة اللاعبين في الأهلى بعد التعادل مع الزمالك، تؤكد أن الجميع في الأهلى الآن خارج نطاق الزمن. سادسا: تعادل النادي الأهلى وخسارة نقطتين إضافيتين فى مباراة أدارها طاقم تحكيم أجنبى. نتيجه للتحكيم الخاطئ فى بعض المباريات السابقة، وأيضا نتيجة قلة الإمكانيات الفنية للفريق، وهذا أمر واضح لجماهير الأهلى العظيمة، والذي لم يعتاد عليه الأهلى في الماضي. سابعا : تورط الأهلى في شراء لاعب مصابا، وهو صلاح الدين سعيد الذي يلعب بأقل جهد بدني وفني، ولا يمتلك الأهلى إدارة قوية للإستغناء عنه، قبل أن تبحث عن عرض لبيع موسى إيدان. ثامنا: هدفان ساذجان سكنا مرمى مسعد عوض ومرمى أحمد الشناوي. حارسا المستقبل لمنتخب مصر! تاسعا: لو تعلم جماهير الزمالك أن فوزهم على الأهلى كان سيبقى محمد صلاح مدربا للفريق لما حزنت على التعادل. عاشرا: باستثناء مستوى حسام عاشور الفردى المتطور. ماذا قدم جاريدو للأهلى؟

٢) كانت المباراة ضعيفة جدا، حيث كان لاعبوا نادي الزمالك مشتتين للغاية، لا يتعدى لاعبو دورة رمضانية بيجروا بالكرة يمين وشمال. أما الأهلى أعظم من أن يرشى التحكيم كما يقال. ونسى أن الله كان رؤوف بجماهير نادي الزمالك بعد تصدى العارضه لصاروخ الاهلى. فالتحكيم كان ظالم بسبب عدم طرد أحمد عيد بعد تراجع الحكم عن رأيه.

٣) فريق الزمالك فعليا كسبان، أيمن حفنى لاعب مهارى ولا احد يشكك فى ذلك. فالحكم الاجنبى انقذ النادي الأهلى من هزيمة مؤكده، بعدم احتساب ضربة جزاء صحيحة، وأيضا من انفراد واضح لحفني، وقال أنها تسلل ظالم، مما منع نادي الزمالك من فوز مؤكد، اليوم نادي الزمالك هو الأفضل، والفوز على النادي الأهلى قادم لا محاله، التعادل رغم إنه مزعج ولكنه مفيد، لإن نادي الزمالك هو الأول والمتصدر للدوري، والمباراة القادمة للنادي الأهلى ستكون خاسرة مع فريق إنبى.

٤) هذة هي المرة رقم ٢٠، تنتهي مباراة القمة بين قطبي الكرة المصرية الزمالك والأهلي بنتيجة سوى فوز الزمالك على الأهلي. تعادل الزمالك والأهلي الخميس إيجابيا بهدف في المباراة التي جمعتهما بملعب الدفاع الجوي، في اطار الجولة ١٨ من الدوري الممتاز. وكان آخر فوز حققه الزمالك على الأهلي في مباريات القمة قبل نحو ثمانية سنوات، عندما فاز على الأهلي بهدفين مقابل لا شيء في الجولة ٢٩ من الدوري الممتاز، وهي المباراة التي اقيمت ٢١ مايو من العام ٢٠٠٧. ومنذ ذلك التاريخ التقى الفريقان في ٢٠ مباراة ببطولات الدوري وكأس مصر ودوري أبطال أفريقيا وكأس السوبر المصري. وخلال المباريات العشرين فاز الأهلي في ١١ مباراة مقابل ٩ مباريات انتهت بالتعادل، فيما لم يحقق الزمالك اي انتصار.

٥) هناك الكثير من مجاملات التحكيم المصري والأجنبي لفريق النادى الأهلي، حيث تم احتساب تسلل فاضح على أيمن حفني، والتي كانت ضربة جزاء صحيحة له وهدف مؤكد، والذي كان سينهي المباراة لصالح فريق نادي الزمالك.

٦) فريق الزمالك هذا غير قادر على كسب النادي الأهلي أبدا، العام الماضى النادي الأهلي لاعبهم بفريق ١٧ سنة، ومع ذلك كسبهم وأخذ الدوري أيضا. هو فى فريق كويس فى العالم يعد عشر سنين ما يخدش غير بطولة واحدة، ويعد يخسر من الفريق اللى المفروض إنه بيناقسه، مع عدم وجود اى مناقسة في الأساس، عشر سنين حتى أكسبوا مباراة واحدة اغزوا بيها العين أهو يبقى للتاريخ، ده حتى والأهلى مستواه ضعيف هذا العام، حقق بطولتين منهم بطولة افريقيا، حتى الآن الزمالك اللى قالوا إنه حلو مخدش ولا بطولة.

Appendix D

Movie Negative Orientation

إسم الفيلم: غرفة ٧٠٧،

للأفلام الرومانسية دائما خصوصية وسط تصنيفات الأفلام، فالفيلم الرومانسي فالأغلب لا يقدم قصة أو فكرة جديدة للمشاهد، ففي الأغلب تكن القصة كلاسيكية ومتكررة، ولكن الفرق بين الفيلم الرومانسي الجيد والردئ هي كيفية تناول القصة ومعالجتها وكيفية تقديمها،

فعند تقديم قصة رومانسية أشبه بكلاسيكية روميو وجوليت لايمكن أغفال الكثير من التفاصيل، فكيف نشات قصة الحب وكيف تتطورت ولايمكن أن يختصر ذلك بالحب من أول نظرة، ثم يختصر تتطور العلاقة في مجموعة مشاهد الفوتو مونتاج بخلفية موسيقية، لا تصل بنا أن نقتنع أن الأثنين وصلا إلي ذروة الحب الذي لا يفرقهم الا الموت، ليس هناك ما يميز البطلة التي يقع في هواها البطل، ويستعد أن يفعل اي شيئ من أجل ان يبقيان معا، الا جمالها وأناقتها الذي اعتمد عليه فقط كمبرر لذلك الحب الأسطوري.

لذلك فالتطورات تتم في حالة من المليودراما المفتعلة المليئة بلا مبررات أو خلفيات، فأنت أمام أحداث كثيرة لو فكرت في منطقية اي منها لوجدت نفسك تظل تفكر فيها حتى ينتهي الفيلم، بل وحتى تخرج من السينما دون أن تصل الي اجابة منطقية، وذلك لم يكن فقط في الأحداث، بل في أسم الفيلم الذي قد تعتقد أن له دور في الأحداث، أن هذا الرقم قد يحمل معني أو أن له دور في الأحداث، أو أن السر في الغرفة، وليس فقط أن قبل نهاية الفيلم بعشر دقائق تتركز الأحداث داخل غرفة بمستشفي تحمل ذلك الرقم، دون اي دلالة للرقم أو للغرفه فهل لو كانت الغرفة تحمل الرقم ٩ لكان الفيلم اسمه الغرفة ٩؟ حتى العداء بين عائلة البطلين، بالرغم من أن البطل معيد ومثقف ووالده النائب العام، فأنه لا يكتشف أن حبيبته ابنة رجل الأعمال الفاسد الذي حبسه والده منذ سنوات، ومازال يطارد فساده الي بعد أن يتقدم لها فيعرف منه!

ولأن فكرة التوليفة السينمائية تسيطر علي تفكير صناع السينما في مصر، فيجب أن يحمل الفيلم العديد من الخطوط الدرامية، التي قد لا تفيد الحدث الأساسي، الا في وجود علاقة سطحية تملئ بشخصيات وأحداث لا تفيد الفيلم، بل تزيده ترهلا وبعد عن قصته، لذلك فلا مانع أن يتناول فيلم رومانسي قضية بيع الأعضاء، وفساد المستشفيات بل وفساد رجال الأعمال في محاولة لإضفاء قيمة للقصة بجوار كونها رومانسية ايضا، بلا اي مبرر او منطق،

ولأن التبرير يحكم الفيلم فلا مانع من أن يصاب البطل بطلق نار من عدو له وهو يقود سيارته، ولا مانع من أن مرور الشهور علي البطلة في حالة موت اكلينيكي طويل، لا يغير شكلها بلا تظل بنفس أحمر الشفاه ونفس النضارة، بالرغم من ان لحية البطل طالت ووصلت الي مرحلة متقدمه.

Movie Positive Orientation

اسم الفيلم: عسل أسود،

أفضل ما بالفيلم أنه لم يقدم حلولا غير منطقية أو مقترحات ساذجة لإصلاح مشكلات البلد، فقط عرض المشكلة كما هي، وهو ما يجسد المعنى الحقيقي لرسالة السينما، العرض وليس الإصلاح. فالعاملون بتلك الصناعة من منتجين وممثلين ومخرجين وكُتاب ليسوا مطالبين بالبحث عن حلول للمشكلات التي تعاني منها البلد، ولكن يكفيهم تقديم عمل فني محترم يلمس أوضاع سيئة يعيشها ملايين. وهو ما ركز عليه المؤلف خالد دياب من بداية السيناريو وحتى نهايته، ليقدم لخالد مرعي مادة مميزة أحسن الأخير استغلالها للغاية.

فكرة "عسل إسود" ليست بالجديدة، فقد اعتمد أكثر من فيلم قديم على فكرة الشخص العائد من الخارج الذي يصطدم بالواقع السيء الذي يعيشه أبناء بلده، ولكن تناول الفيلم الجديد للفكرة جاء مختلفا. فنحن لم نشاهد على السبيل المثال، أحمد حلمي يظهر مثل شكري سرحان بفيلم قنديل أم هاشم، والذي يقابل ما يراه من سلوكيات خاطئة بعصبية شديدة وسخرية، ولكن جاء حلمي أكثر هدوءا لأنه بدا وكأنه غير مستوعبا في الأصل ما يحدث أمامه. وهو ما يُحسب للثنائي خالد دياب وخالد مرعي، فمن تلك الفكرة جاءت كوميديا الفيلم، والتي اعتمدت على كوميديا الفكرة وليس كوميديا الموقف، فقد تجد كثيرا من المشاهد متشابهة بفكرتها، ولكنك ستضطر للضحك مع كل مشهد بفضل موهبة أحمد حلمي.

أيضا من مميزات الفيلم بالجانب الكوميدي أنه لم يعتمد على أحمد حلمي فقط لالتقاط الابتسامة من وجه المُشاهد كما يحدث في أغلب الأفلام الكوميدية، وأيضا لم يعتمد على ممثل أو اثنان آخرين للقيام بنفس المهمة كما يحدث في بعض الأفلام، ولكن أكثر مشاهد الفيلم كوميديا كان أبطالها كومبارس، وهو الشيء الذي فعله حلمي من قبل بفيلم ألف مبروك. ويكشف ذلك حُسن اختيار أحمد حلمي للسيناريو الجيد الذي يعتمد على فكرة كوميدية معينة تستمر طوال مشاهد الفيلم، وليس مجرد سيناريو يضم عدد من الأفيهات المستهلكة. الحديث عن أحمد حلمي الممثل لن يحتاج للكثير من الكلام، فالنجم الكوميدي أثبت أنه الأنجح بالساحة الفنية على الإطلاق، وليس كفنان كوميدي فقط. فحلمي منذ أن قام بمرحلة تغيير الجلد بفيلم كدة رضا وهو يقدم سينما راقية مُبهرة كوميدية نظيفة، وهي الخلطة التي فشل كثيرين في تكوينها والاستمرار بها.

ظهر حلمي في عسل إسود أكثر هدوءا وثقة أكثر من أي فيلم مضى، وهو ما استمده من نجاحات أفلامه السابقة، وليس من الضروري عقد مقارنة بين عسل إسود وأي فيلم آخر لحلمي في الثلاث سنوات الماضية، فجميعها أفلامه وجميعها أبدع فيها. وجاء ظهور كل من إدوارد وإيمي سمير غانم مميزا للغاية، بالإضافة للفنانين الكبار لطفي لبيب ويوسف داود وإنعام سالوسة. وكالعادة، احتلت موسيقى الموسيقار الكبير عمر خيرت مكانة مميزة بأحداث الفيلم، ونجح خيرت في وضع موسيقى ملائمة للأحداث خاصة عند عودة حلمي لمصر.

الفيلم نجح في تقديم تجربة واقعية يصطدم بها كثيرون ويعاني منها ملايين، ومن يتهم صُناع الفيلم بنشر الغسيل القذر في أحداثه يتحدث بشكل غير منطقي، فالتلوث والسرقة والفساد وانهيار التعليم وقلة الأدب ليست أشياء سرية تحتاج لفيلم سينمائي ليفضحها.

عسل إسود واحدا من أكثر الأفلام رُقي خلال السنوات الأخيرة، فعندما يدفعك فيلما للضحك الشديد في بعض مشاهده، ثم يصيبك بأعراض اكتئاب بمشاهد أخرى، يكون بالتأكيد فيلما مميزا نجح في توصيل فكرته كاملة للمشاهد.

ملخص الرسالة

تحتوي وسائل الإعلام على شبكة الإنترنت وعلى مواقع التواصل الاجتماعي والمنتديات على ملايين الصفحات التي تستعرض آراء الناس حول مواضيع كثيرة منها الكتب والفنادق وكثير من المنتجات وليس المنتجات فحسب ولكن أيضا على جميع الأحداث الجارية. فسيكون من الجيد الإستفادة من هذة الآراء والخبرات قبل اتخاذ القرارات المتعلقة بهذه الكيانات عن طريق استخراج الخصائص المتعلقة بهذة الكيانات وتحليلها إلي إيجابي أو سلبي أي يوضح المميزات والعيوب لكل منتج أو حدث أو موضوع.

ولتحقيق هذا الهدف قمنا فى هذة الرسالة بتقديم نهجا جديدا لا يختص بمجال معين ولا بقاعدة بيانات بعينها. حيث نقترح نهجا يعتمد على ثلاث خطوات رئيسية ألا وهى على مستوى الكلمة ثم على مستوى الجملة وأخيرا على مستوى الكيان. أولا على مستوى الكلمة يتم تحليل النصوص باستخراج الجذور والأوزان باستخدام القواعد النحوية والصرفية وتحديد "المقتبسات" والتي تستخدم كمدخل للمعجم وتحمل المعنى المقصود للكلمة بدون الاضافات الممكنة لها وفيها يتم وضع بعض القواعد التى تستخدم فى تحديد الكلمات الدالة علي الرأى . ثانيا على مستوى الجملة ويتم فيها عملية استخراج كلمات النفي التى تعكس اتجاة الرأى من وإلى إيجابى وسلبى، أيضا يحدد فيها الكلمات التى تقوى الرأى ثم يتم وضع قيم محددة لهذة الكلمات. أخيرا علي مستوى الملف ككل حيث يتم استخراج الخصائص المتعلقة بالكيان فى اتجاهين إما أمامى أو خلفى من الكلمات الدالة على الرأى باستخدام بعض القواعد النحوية لأنماط الجمل. وفي النهاية يتم الحصول على الخصائص باتجاهاتها سواء إيجابى أو سلبى بقيم ثابتة لمعرفة مميزات وعيوب كل كيان.

تم اختبار النظام على مستويين: أولا على مستوى الملف بأكملة، لتحديد الأتجاة الرئيسى للكيان ككل سواء كان إيجابى أم سلبى. وذلك باستخدام قاعدة بيانات متاحة تضم ٥٠٠ ملف فى مجال الأفلام، منهم ٢٥٠ إيجابى و ٢٥٠ سلبى. وقد حقق النظام كفاءة تصل إلى ٩٦% مقارنة مع ٩٠% لمنتج هذة البيانات. ثانيا على مستوى الخصائص، وقد تم استخدام ٢٠٠ رأى فى أربع مجالات مختلفة ألا وهى (روايات، منتجات، أحداث مبارة كرة قدم وأخيرا فنادق)، تتمثل هذة الآراء فى ١٢ كيان ثلاثة فى كل مجال. وقد تم تقييم النظام فى هذا المستوى من خلال مرحلتين، المرحلة الأولى هى قياس كفاءة النظام فى استخراج الخصائص مقارنة مع اثنين من الخبراء لكل مجال وأخذ المشترك بينهم كمرجع أساسى للنظام، المرحلة الثانية تبين مدى نسبة توافق النظام فى تحديد اتجاه الخصائص المشتركة بينه وبين الخبيرين، وتم أفتراض مؤشر يوضح نسبة التأكد أثناء استخراج الخصائص للكيان . بالنسبة لعملية استخراج الخصائص، لقد حقق النظام متوسط معدل دقة ٨٥% ومتوسط معدل استرجاع ٨٩%. أما بالنسبة الى مدى التوافق فقد حقق نسبة ٩١% بنسبة تأكيد تصل إلى ٩٠%. وبهذا يثبت النظام فعاليته فى استخراج الخصائص وتحديد إتجاهاتها الإيجابية والسلبية.

تنقسم الرسالة إلي ست أبواب: الباب الأول يعرض مقدمة وملخص البحث والدافع لهذا العمل وأهدافه وأهم الاسهامات. الباب الثاني يراجع أهم الطرق المستخدمة فى عملية استخراج الخصائص للكيانات المختلفة وطرق تحديد اتجاهاتها الموجبة والسالبة. الباب الثالث يعرض الوسائل المقتبسة والتى تم إنشائها لاستخدامها فى بناء هذا النظام. الباب الرابع يقدم النهج المقترح لعملية استخراج الرآى من النصوص العربية. الباب الخامس يعرض مجموعة البيانات المستخدمة والتجارب لاختبار وتقييم النظام المقترح. الباب السادس يشمل ملخص عام للرسالة والأعمال المقترحة مستقبليا.

جامعة بنها

كلية الهندسة بشبرا

قسم الهندسة الكهربائية

استخراج الرأى للنصوص العربية

رسالة مقدمة من

المهندسة / شيماء إسماعيل محمد مصطفي

(بكالوريوس هندسة نظم الحاسبات)

للحصول علي

درجة الماجستير في هندسة نظم الحاسبات

(قسم الهندسة الكهربائية)

تحت إشراف

أ.د / طارق الششتاوي

كلية الحاسبات و المعلومات

جامعة بنها

أ.م.د. / عبد الوهاب السماك

كلية الهندسة بشبرا

جامعة بنها

القاهرة – مصر

٢٠١٥