A Generic Approach for Extracting Aspects and Opinions of Arabic Reviews

Shimaa Ismail
Faculty of Computers and Informatics
Benha University, Egypt
+201229069123
shimaa.mustafa@fci.bu.edu.eg

Abdelwahab Alsammak
Faculty of Engineering Shoubra
Benha University, Egypt
+201220588666
asammak1@feng.bu.edu.eg

Tarek Elshishtawy
Faculty of Computers and Informatics
Benha University, Egypt
+20123454723
t.shishtawy@bu.edu.eg

ABSTRACT

New opportunities and challenges arise with the growing availability of online Arabic reviews. Sentiment analysis of these reviews can help the beneficiary by summarizing the opinions of others about entities or events. Also, for opinions to be comprehensive, analysis should be provided for each aspect or feature of the entity. In this paper, we propose a generic approach that extracts the entity aspects and their attitudes for reviews written in modern standard Arabic. The proposed approach does not exploit predefined sets of features, nor domain ontology hierarchy. Instead we add sentiment tags on the patterns and roots of an Arabic lexicon and used these tags to extract the opinion bearing words and their polarities. The proposed system is evaluated on the entity-level using two datasets of 500 movie reviews with accuracy 96% and 1000 restaurant reviews with accuracy 86.7%. Then the system is evaluated on the aspect-level using 500 Arabic reviews in different domains (Novels, Products, Movies, Football game events and Hotels). It extracted aspects, at 80.8% recall and 77.5% precision with respect to the aspects defined by domain experts.

CCS Concepts

Information systems→ Information retrieval → Retrieval tasks and goals → Sentiment analysis

Keywords

Opinion Mining; Sentiment Classification; Feature Extraction; Arabic Sentiment Analysis.

1. INTRODUCTION

The dramatic increase of social network, online news, reviews, forms, and blogs led to the importance of automatically mining the web content for many purposes. Opinion mining is one of the most interesting purposes of mining the web content to summarize the opinions of users from a wide range of reviews, blogs, and tweets. 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. Opinion mining is well-suited to business intelligent systems. It helps customers to make a buy decision through processing hundreds or thousands reviews. They also allow manufacturer or service provider to keep track and manage customer opinions to improve product quality or service performance [8]. 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 decisions concerning product or an event, one needs to know what components and/or aspects of the entity to be sentimentally analyzed. In feature-based approach, the orientations of the extracted features are used to classify the opinions about an entity.

Extracting features can be manual-based, dictionary-based, or corpus-based [22]. In manual-based systems, features are fed directly through predefined lists. Features can be found using language patterns near manually prepared opinion words [16]. The concept was modified by feeding the system with initial seed list of trusted opinions in a given domain with known orientations. Then, the seed list is incrementally expanded as new opinion words are found by searching in a dictionary such as the WordNet for their synonyms and antonyms [17]. The newly found words are added to the seed list for the next iteration. This iterative process will end when no more new words are found. However, the dictionary-based approach faces the problem of context independent for the collected words. Many sentiment words have context dependent orientations. For example, the sentiment orientation of "quiet" is negative for a speaker phone, and positive for a car. The corpus-based approach can help deal with this problem using syntactic patterns that occur together along with a seed list of opinion words to find other opinion words in a domain corpus. However, the issue is more complicated because; opinion analysis should not be limited to specific domains which add extra coverage challenge facing existing ontology-based systems.

In this paper, we followed a different 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 predefined sets of features, nor domain ontology hierarchy. Opinion tags are added to an existing accurate Arabic Root Based Lemmatizer ARBL lexicon [12], 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 bearing 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 word and its polarity at the word level, and then detecting the presence of intensification and/or negation at the sentence level.
2- Exploiting opinion words to extract target noun phrases as candidate aspects or general entity.

INFOS '16, May 09-11, 2016, Giza, Egypt
© 2016 ACM. ISBN 978-1-4503-4062-5/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2908446.2908467

3- Applying predefined syntactic patterns to select the proper aspects of the entity.
4- Estimating sentiment score and attitude of the entity by aggregating the scores of its lemma-form aspects.

The remaining of the paper is organized as follows: The previous work is presented in section 2. The methodology for building the lexicon is described in section 3. The proposed sentiment analysis approach is presented in section 4. Section 5 shows the algorithms for extracting and aggregating the target aspects. The results of test experiments are discussed in section 6. The conclusion of the presented work is given in section 7.

2. RELATED WORKS

Although the sentiment analysis area is emerging, there is a large number of research papers published in this area. Sentiment analysis has been investigated mainly at three levels; document-level, sentence-level and aspect-level [22].

2.1 Document-level

The efforts started by extracting the attitude of the whole entity or subject as positive, negative or neutral. In this approach, the polarity of the whole document is extracted regardless of the details of attributes polarity. The counts of terms or sentences that have positive or negative orientations determine the whole text attitude. An example of this approach is the work of Turney [31], who presented a simple unsupervised learning algorithm for classifying reviews as recommended or not recommended. The algorithm evaluated each two consecutive words from the review if their tags conform to predefined patterns. Another example of document level sentiment analysis was considered using machine learning (ML) techniques [27]. Pang approach was a domain specific that uses indicator words for positive and negative sentiments in movie reviews. Although ML classifiers perform well, their performance dropped on topics or texts that were different from those that they were trained [13]. Pang and Lee presented a comprehensive survey covers challenges, techniques and approaches in this level of sentiment analysis [26].

2.2 Sentence-level

Hu and Liu [16] extended the sentiment classification to be performed at the sentence level and decide whether each opinion sentence is positive or negative. In their work, a small list of seed adjectives was manually created and expanded every time a new adjective was found. Using WordNet for analysis, the presented algorithm achieves an average accuracy of 84% in predicting sentence orientations. Other researchers classify each sentence in the document, and count the number of sentences that have positive or negative orientation [32]. 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 [33] for sentences that have multiple attitudes. Hassan et al. [15] proposed a method to identify attitudes about participants in online discussions. Its first step finds sentences with attitudes using supervised learning. The features were generated using Markov models. Its second step determines the orientation (positive or negative) of the attitudes using a lexicon-based method.

2.3 Aspect-level

Document, sentence, and phrase levels sentiment analysis were not enough to provide people with comprehensive opinions. Therefore, many researchers highlight the importance of sentiment analysis of entity features. For example Hu and Liu [16], defined the ideal opinion mining as a tool that would process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each attribute (poor, fair, good). Liu et al. [21] identified features from Pros and Cons of review format for a given product. In their work, implicit features are not clearly appeared and hard to be identified. Su et al. [29] proposed an automatic identification for implicit product features expressed in the automobile reviews in the context of opinion question answering. In this approach, the implicit features for a specific product are identified by assigning some adjectives in a lexicon to a set of pre-defined product features in a polarity lexicon. The lexicon is then used for finding the relationship between opinion words and the features. Supervised rule mining was used to generate language patterns to identify the features. When the entity to be reviewed is known, a list of its explicit features and adjectives can be utilized to extract feature-based opinions [14]. In this approach, frequent nouns and noun phrases are collected as product features. However, they overcome different writing styles by analyzing extracted phrases to produce patterns. Extracting product features can be done by utilizing patterns and opinion lexicon for specific products [18].

2.4 Arabic Sentiment Analysis

Recently, several efforts have been proposed for subjectivity and sentiment analysis for Arabic documents. Medhat et al. [23] survey the different techniques used for subjectivity and sentiment analysis for Arabic. The attempts started by extracting the features that do not depend on the language itself. Abbasi et al. [1] 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. A predefined list of features was used by Elhawary and Elfeky [10] to produce an Arabic sentiment analyzer in hotel domain using MapReduce by translating the features. This method suffers from the problem of incomplete matching between Arabic and English adjectives. Another domain dependent analyzer was presented by Lazhar and Yamina [20]. The authors identified opinions for Arabic text using domain ontology. In their approach each concept and property are associated to the corresponding labels according to their semantics. Abdul-Mageed et al. [2] built a sentence-level subjectivity and sentiment analysis system for MSA combining language-independent and Arabic morphological features. They proved that using morphology-based features improves the system performance. Abdul-Mageed et al. [4] presented SAMAR, a SVM-based system for subjectivity and sentiment analysis for Arabic social media genres. They manually created a lexicon of 3982 adjectives labeled with one of the following tags {positive, negative, 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. El-Beltagy and Ali [9] highlighted the problems and challenges that face researchers bearing out sentiment analysis of Arabic social media. The addressed problems are the unavailability of colloquial Arabic parsers and sentiment lexicons, the need for person name recognition, and handling compound phrases and idioms. The paper also presents an Egyptian dialect sentiment lexicon.

2.5 Sentiment Lexicons and Corpora

Sentiment dictionaries have a great role in determining the accuracy of sentiment analysis systems. 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. [30], 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 use the WordNet as lexical resource such as SentiWordNet [6] in which 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.

Rushdi-Saleh et al. [28] built an Opinion Corpus for Arabic (OCA) which contains 500 movie reviews, 250 of them considered as positive and other 250 as negative. They use both Support Vector Machines (SVMs) and Naive Bayes (NB) classifiers, reporting 90% F-measure on OCA using SVMs. Abdul-Mageed and Diab [3] 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. It contains 2855 reviews collected from wikipedia talk pages and forums. Morad and Darwish [24] 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. ElSahar and El-Beltagy [11] introduced large multi-domain datasets for Sentiment Analysis in Arabic. The datasets were scrapped from different reviewing websites and consist of a total of 33K annotated reviews for movies, hotels, restaurants and products. Moreover they built multi-domain lexicons from the generated datasets which are publicly available to the scientific community.

3. THE SENTIMENT-ANNOTATED LEXICON

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. The contribution of the presented work started by adding sentiment tags (polarity and score) to the roots and patterns of an existing Arabic Root Based Lemmatizer ARBL lexicon [12]. The lexicon contains 3829 roots, 69 patterns, and a closed set of 346 Arabic words categorized into 16 groups (e.g., prepositions, conjunctions, adverbs, numerals, etc...). During the word-level analysis, a word is identified as an opinion-bearing word, if both of its root and pattern are annotated with sentiment tags in the lexicon. The following two subsections describe the assumptions made during the tagging process of patterns and roots.

3.1 Pattern Tagging Process

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 bearing 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.

In all existing Arabic lexicons, patterns are classified according to their part of speech (POS) tags [19]. We extended the classification to include sentiment tags at the pattern level, as shown in table 1. With the assistance of two Arabic language specialists 39 patterns are tagged as opinion-bearing patterns out of the available 69 patterns collected by the ARBL.

Table 1. Syntactic and Sentiment Patterns

	Pattern Classification	Pattern Class	Pattern form Examples	Word Examples
Syntactic 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 Bearing 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

3.2 Root Tagging Process

Several lexicon based approaches have expressed the semantic orientation as a numerical value range to express the word's strength, [16], and [30]. In our work, we followed another approach, where all opinion words are handled as 'like' or 'dislike' binary opinions, whatever is the strength of vocabulary used in the review. This gives more faithful representation for 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 "love" a stronger weighting than "like", a criticism still raised that the dictionaries are unreliable, as they are either built automatically or hand-ranked by humans [5].
2- The overall sentiment result may be misleading. As an example adapted from Taboada et al. [30], 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 did not have the chance to choose specific opinion word from a closed terms arranged by strength from highly positive to highly negative. Therefore, reviewers express opinions based on their culture background and mode.

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 a positive orientation in hotel domain and a 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 (xyz, "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 this case, we tagged these roots as uncertain or neutral roots.

Following the assumption of "prior polarity" of words [25], we assigned each root a context-independent semantic orientation. The orientation is manually tagged, by two Arabic language experts, and expressed as a numerical value (+1, 0, -1) for positive, neutral, and negative orientation, respectively. In our work, 213 roots are manually marked as positive, 260 roots as negative, and 107 as uncertain oriented roots, out of 3829 roots recognized by ARBL. Examples of the roots are shown in table 2.

Table 2. Examples for positive, negative, and uncertain roots

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

4. SENTIMENT ANALYSIS

The objective of this work is to provide a summarized opinion for generic entities such as topic, person, event or an organization at the level of their detected aspects. Figure 1 shows an overview for the proposed approach. The problem is divided into three tasks: 1) Identifying the opinion-bearing words using sentiment analysis, 2) Extracting the detailed entity aspects and their attitudes, 3) Determining the orientation of the whole text by aggregating the aspects attitudes. In this section, we focus on the first task, which includes the following steps:

1- Word level sentiment analysis to extract syntactic, lexical and opinion tags.
2- Sentence level sentiment analysis to detect negation and intensification.

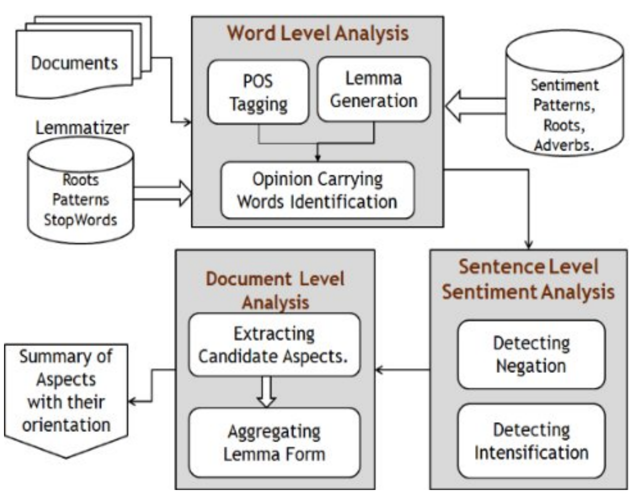

Figure 1. The proposed approach overview.

4.1 Word-level Analysis

All words are analyzed to extract their basic features. The sentiment-annotated ARBL provides the following information: 1) Syntactic information: POS tags (noun, verb, adjective, adverb...). 2) Lexical Information: The word root, pattern, and its lemma form.

In our approach, the word is considered as opinion-bearing if it meets two conditions: 1) Its pattern matches one of the orientation patterns and 2) Its root matches one of the positive or negative roots. Opinion-bearing word is assigned a score of +1 or -1, according to its root polarity. It is important to note that uncertain roots do not affect the aspect extraction process, as both positive and negative roots are used to locate aspects.

4.2 Detecting Intensification and Negation

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 our work, we detect the opinion word and its polarity score at the word level, and then apply the intensification and negation detecting algorithm to update opinion score and/or polarity.

Intensifier parameter assesses the semantic of a word, using some neighboring adverbs like (very, extremely, absolutely, etc...) [7]. In this work, we tagged 27 Arabic intensifier adverbs (e.g. جدا- مطلقا- حقا- دائما- تماما-....). The effect of intensifier words is to increase the score by (1) in its polarity direction.

Negation is an important parameter that affects the orientation of the detected sentiment bearing words. In most cases, negation reverses the word orientation. For example the expression "Service is not good" has the opposite orientation of the expression "Service is good". Examples of Arabic negation words are (لن- بدون- عدم- لم- ليس- لا- ليست). Usually in MSA writing style, negation precedes opinion words. Therefore, starting from the opinion-bearing word, the system scans for the existence of negation word in backward direction within the sentence. Once a negation word is detected, the opinion tag orientation is reversed.

5. ASPECT EXTRACTION AND AGGREGATION

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. As we intend to extract automatically the domain independent aspects or features, the proposed approach does not exploit a predefined set of features, nor domain ontology hierarchy. Instead, the identified opinion-bearing words are used for extracting entity aspects and their orientations.

The presented system analyzes sentences to extract all target noun phrases as candidate aspects. An aspect could appear in different syntactic forms in a review. Therefore, the similar candidate aspects should be aggregated and represented by their lemma-form. The sentiment score and attitude is then calculated for each aspect and the general entity.

5.1 Extracting Candidate Aspects

Each opinion-bearing 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. To extract the target noun phrases, we formulate a set of syntactic rules defining the allowed sequence of n-gram words according to their POS tags. Table 3 shows examples of the syntactic patterns used to extract the target noun phrases as candidate aspects.

Table 3. Syntactic patterns for detecting the candidate aspects

Syntactic Pattern	Candidate Aspect	Example
Prep+DTNN2+Particle+ DTNN1	DTNN1	فى الشقة كانت **الغرف**...
Prep+DTNN2+NN2+NN1	NN2 + NN1	فى الفندق **منتجع صحى**..
DTNN1+Particle	DTNN1	**الإضاءة** كانت..........
DTNN1+Prep+NN	NN	المطعم فى **موقع**........
NN2+NN1+DTNN1	NN2+NN1+ DTNN1	**فريق عمل الفندق**......
NN+DTNN1	NN+DTNN1	**بوفية الإفطار**..........
Particle+DTNN1	DTNN1	إن **القصة**..............
NN	NN	**هاتف**..................

Starting from opinion-bearing word, the system moves backward/forward to locate the occurrence of the nearest syntactic pattern within the sentence as a candidate aspect. 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-bearing word is a verb; else it is backward for all other POS opinions. The algorithm used for extracting the candidate aspects in the backward directions is shown in figure 2.

```
For each opinion-bearing word
  Aspect = "  "
  Check the orientation type (positive/negative)
    Repeat until word_count =0:
      If a noun
        Append the word to Aspect.
        Assign the orientation value to Aspect.
      ElseIf a negation word
        Reverse the orientation value of Aspect.
      ElseIf an intensifier
        Increase/decrease the orientation value of Aspect.
      ElseIf a preposition
        If(word_count !=0)
          Clear the Aspect (Aspect= " ").
      End if
      Decrement word_count.
    End loop
  Save Aspect and its orientation value.
End for
```

Figure 2. The algorithm of extracting the entity aspects in the backward direction

When the phrase is located, its corresponding target candidate aspect is extracted as shown in the second column of table 3. The extracted aspect is assigned the same sentiment score as its base opinion-bearing word (e.g. room service 'خدمة غرفة', +1).

5.2 Aggregating lemma-based Candidate Aspects

The purpose of this task is to group similar candidate aspects and compute their sentiment 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', 'الخدمة بالغرفة', 'الخدمات بالغرف') and would be represented by one aspect as the lemma form (room service 'خدمة غرفة'). 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. The sentiment score of the lemma-based aspect is represented by the sum of sentiment scores of all different lexical forms of the aspect. Thus, the aggregation process produces the non-repeated aspects along with their total sentiment scores (e.g. room service 'خدمة غرفة', +4).

The second problem is the presence of the entity name inside some of the candidate aspects which 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 this 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.

6. EXPERIMENTAL RESULTS

6.1 Testing Data Sets

We used three datasets in different domains 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. It is available to the scientific community for sentiment analysis[A] and is called "Opinion Corpus for Arabic". The second dataset contains 1000 Arabic reviews in the restaurant domain taken from the sentiment datasets[B] introduced by ElSahar and El-Beltagy [11]. The third dataset contains 500 Arabic reviews[C] in different domains: hotels, novels, products, restaurants, and events which collected from different websites (e.g. tripadvisor.com.eg, goodreads.com, unlimit-tech.com, android4ar.com and Al-ahly.com).

6.2 Methods

Two experiments are carried out to evaluate the performance of the proposed approach. The first experiment concerns evaluating the efficiency of the presented algorithm on the entity level in different domains. This experiment is carried out using the three datasets. The results of applying our algorithm are compared with the results obtained by the authors of the datasets. The second experiment concerns evaluating the efficiency of the presented algorithm on the aspect level in different domains. Two domain-oriented human judges are asked to determine the proper aspects of each reviewed entity along with their polarities, because none of the publicly available Arabic datasets are evaluated on the aspect level. The human selected aspects and scores are automatically processed to ensure that there are no redundant aspects in different forms. The processing includes lemma form generation, aggregating similar aspects, and computing sentiment scores for each aspect. The results obtained by applying the proposed algorithm are compared with the results of the human experts.

Precision, Recall, and F-measure metrics are used to measure the accuracy of the proposed system. Precision is an estimate of the probability that a given model identifies an aspect as relevant to a user's aspects. 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. F-measure combines both precision and recall, computing the proportion of true results.

A- http://sinai.ujaen.es/oca-corpus-en/

B- http://bit.ly/1wXue3C

C- http://www.scribd.com/eng.shismail

Given the retrieved aspects by the system (X) and relevant aspects identified by Human judge (Y) as defined by:

Precision = $(X \cap Y) / X$
Recall = $(X \cap Y) / Y$
F-measure = 2*(Precision * Recall) / (Precision + Recall)

6.3 Experiment 1: Entity-Level Evaluation

The objective of this experiment is to measure the efficiency of the proposed algorithm at the entity level in different domains. The first part of the experiment concerns applying our algorithm on the first dataset (OCA). The precision, recall and F-measure are computed and compared to the corresponding values obtained by Pang et al. [27] and OCA [28] using the same dataset. Table 4 shows the results of this comparison.

Table 4. 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

The second part of the experiment concerns applying our algorithm on the second dataset which contains 1000 Arabic reviews in the restaurant domain. The accuracy obtained by applying our algorithm reaches 86.7% compared to 84.6% obtained by applying the SVM classifier and using combined feature vectors [11].

Then we applied our algorithm on the third 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 algorithm with those obtained by two human experts. Table 5 shows the percentage of orientation agreement on the entity level between opinions extracted by the proposed algorithm and the human experts. The results show a high degree of agreement ranging from 81% in event domain to 90.8% in hotel domain with an average agreement 85.9%. This proves that the proposed methodology using the sentiment annotated lexicon can be relied upon to extract opinions in generic domains 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.

Table 5. Percentage of orientation agreement at entity level

Domain	System vs. Expert1	System vs. Expert2	Average Agreement per Domain
Hotels	91.45%	90.2%	90.82%
Novels	82.3%	84.4%	83.35%
Products	83.5%	84.2%	83.85%
Events	82.1%	80.4%	81.25%
Restaurants	88.7%	91.8%	90.25%
Average Agreement			85.90%

6.4 Experiment 2: Aspect-Level Evaluation

In this experiment the aspects extracted by the system are compared with aspects defined by two domain experts. Table 6 shows the Precision, Recall and F-measure values of the extracted aspects in five domains. The results show comparable accuracy values for extracting entities' aspects from reviews in different domains with an average precision 77% and average recall 80%. This proves that the proposed system is generic and able to extract the entity aspects with their orientations for Arabic reviews in different domains. The main problems faced by the proposed system will be discussed in the following section.

Table 6. Precision, Recall and F-measure of extracted aspects

Domain	Precision	Recall	F- Measure
Hotel	0.825	0.853	0.839
Novel	0.754	0.788	0.771
Product	0.775	0.813	0.794
Event	0.702	0.753	0.727
Restaurants	0.817	0.832	0.824
Total Average	**0.775**	**0.808**	**0.791**

6.5 Problems

The proposed system suffers from some problems that have been discovered during the analysis of the experimental results. One of these problems is that the accuracy values of extracting opinions in some domains are slightly lower than others due to the difficulty of extracting their aspects. This difficulty comes from the fact that some reviewers describe their opinions using general terms instead of using the entity aspects as shown in the following review "Windows 8". "الويندوز جميل جدا ومميز, لكن للاسف انا أعمل في صيانة الموبايل, وأشتريت حاسوب يعمل باللمس ومجهز بنظام ويندوز 8 الاصلي, لكن معظم اجهزة فحص المحمول كجهاز بوكس الترنادو لا تعمل على هذا النظام, وأصبت بالخيبه كوني لا اود ان افقد النسخه الاصليه".

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. The overall entity orientation is determined, in our aspect-based opinion extraction, by aggregating the orientations of all entity aspects. This may lead to the wrong decision on the entity level as shown in the following review "cell phone". "هذا الجهاز رائع، ولكن يعيبه الكاميرا الأمامية فهي ضعيفة، وعمر البطارية قصير، مقارنة بجهاز آي فون."

Another problem rises 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"،"المطعم", "غرفة الطعام") .

Also, the reviewed entity may contain Named Entities (NE) such as actors, players, authors, companies …etc. Some of these NEs are Arabic adjectives and may be considered as opinion bearing words which lead to extracting fake aspects as shown in the following review. "قام الكابتن محمد لطيف بالتعليق على مباراة الأهلي والزمالك" .

7. CONCLUSION

In this paper, we present a generic approach for extracting the aspects of entities in Arabic reviews as well as their orientations. The proposed approach relies on the idea that the entity aspects and their opinion-bearing words are usually correlative. These words are used to guide the process of extracting the entity aspects. An Arabic lexicon is annotated with sentiment tags at the root and pattern levels. The sentiment analysis on the word level uses this lexicon to detect the opinion-bearing words. This makes the proposed approach suitable for use in various domains. The system is evaluated on the entity-level using 500 movie reviews and 1000 restaurant reviews with accuracy 96% and 86% respectively. Then the system is tested on the aspect-level using 500 Arabic reviews in different domains. On average, the proposed system achieves a recall 80%, precision 77% and F-measure 79%. Thus, the proposed system proves its ability to rely upon in summarizing Arabic reviews.

8. REFERENCES

[1] Abbasi, A., Chen, H., and Salem, A. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems (TOIS)*, 26, 3, 1-34.

[2] Abdul-Mageed, M., Diab, M., and Korayem, M. 2011. Subjectivity and sentiment analysis of modern standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, Association for Computational Linguistics*, 2, 587–591.

[3] Abdul-Mageed, M. and Diab, M. 2012. AWATIF: a multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. In *Proceedings of LREC,* Istanbul, Turkey.

[4] Abdul-Mageed, M., Diab, M., and Kübler, S. 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28, 1, 20-37.

[5] Andreevskaia, A., & Bergler, S. 2008. When Specialists and Generalists Work Together: Overcoming Domain Dependence in Sentiment Tagging. In *ACL* (2008, February), 290-298.

[6] Baccianella, S., Esuli, A., and Sebastiani, F. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC*. 10(2010, May), 2200-2204.

[7] Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., and Subrahmanian, V. S. 2007. Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone. In *ICWSM*. (2007, March).

[8] Ding, X., Liu, B., and Zhang, L. 2009. Entity discovery and assignment for opinion mining applications. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009, June). 1125-1134.

[9] El-Beltagy, S. and Ali, A. 2013. Open Issues in the Sentiment Analysis of Arabic Social Media : A Case Study. In *Innovations in Information Technology (IIT), 9th International Conference*, IEEE, 215-220.

[10] Elhawary, M. and Elfeky, M. 2010. Mining Arabic business reviews. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference,* (2010, December), 1108-1113.

[11] ElSahar, H. and El-Beltagy, S. R. 2015. Building Large Arabic Multi-domain Resources for Sentiment Analysis. In *Computational Linguistics and Intelligent Text Processing*, Springer International Publishing, 23-34.

[12] El-Shishtawy, T. and El-Ghannam, F. 2012. An Accurate Arabic Root-Based Lemmatizer for Information Retrieval Purposes. *arXiv preprint arXiv:1203.3584*.

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

[14] Ghorashi, S. H., Ibrahim, R., Noekhah, S., and Dastjerdi, N. S. 2012. A Frequent Pattern Mining Algorithm for Feature Extraction of Customer Reviews Extraction of Customer Reviews.

[15] Hassan, A., Qazvinian, V., and Radev, D. 2010. What's with the attitude?: identifying sentences with attitude in online discussions. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, (2010), 1245-1255.

[16] Hu, M. and Liu, B. 2004. Mining opinion features in customer reviews. In *AAAI*. 4, 4, (2004, July), 755-760.

[17] Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. (2004, August), 168-177. ACM.

[18] Khan, K., Baharudin, B., and Khan, A. 2014. Identifying product features from customer reviews using hybrid patterns. *Int. Arab J. Inf. Technol.*, *11*, 3, 281-286.

[19] Khoja, S. and Garside R. 1999, *Stemming Arabic text.* Computer Science Department, Lancaster University, Lancaster, UK.

[20] Lazhar, F. and Yamina, T. G. 2012. Identification of Opinions in Arabic Texts using Ontologies. In *Workshop on Ubiquitous Data Mining*, (2012, August), 61.

[21] Liu, B. 2010. *Sentiment Analysis and Subjectivity*. Invited Chapter for the Handbook of Natural Language Processing, Second Edition, (2010, March).

[22] Liu, B. 2012. Sentiment Analysis and Opinion Mining, *Morgan & Claypool Publishers*, (2012, May).

[23] Medhat, W., Hasan, A., and Korashy, H. 2014. Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal,* 5, 1093–1113.

[24] Mourad, A. and Darwish, K. 2013, 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, 55–64.

[25] Osgood, C. E. 1957. *The measurement of meaning* (No. 47). University of Illinois press.

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

[27] Pang, B., Lee, L., and Vaithyanathan, S. 2002. 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. 10(2002, July), 79-86.

[28] Rushdi-Saleh, M., Mart´ın-Valdivia, M., Ure˜na-L´opez, L., and Perea-Ortega, J. 2011. OCA: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology*, 62, 10, 2045-2054.

[29] Su, Q., Xiang, K., Wang, H., Sun, B., and Yu, S. 2006. 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, 22-30.

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

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

[32] Wilson, T., Wiebe, J., and Hwa, R. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *AAAI*, 4(2004, July), 761-769.

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