USING AUTOMATED LEXICAL RESOURCES IN ARABIC SENTENCE SUBJECTIVITY

Hanaa Mobarz[1], Mohsen Rashown [2], Ibrahim Farag [3]

[1,3]Department of Computer Science, Faculty of Computers and Information, Cairo University
[2]Department of Electronics and Communications, Faculty of Engineering, Cairo University

ABSTRACT

A common point in almost any work on Sentiment analysis is the need to identify which elements of language (words) contribute to express the subjectivity in text. Collecting of these elements (sentiment words) regardless the context with their polarities (positive/negative) is called sentiment lexical resources or subjective lexicon. In this paper, we investigate the method for generating Sentiment Arabic lexical Semantic Database by using lexicon based approach. Also, we study the prior polarity effects of each word using our Sentiment Arabic Lexical Semantic Database on the sentence-level subjectivity and multiple machine learning algorithms. The experiments were conducted on MPQA corpus containing subjective and objective sentences of Arabic language, and we were able to achieve 76.1 % classification accuracy.

KEYWORDS

Sentiment analysis, Lexical recourses, Opinion mining, Subjectivity Lexicon, Arabic opinion mining.

1. INTRODUCTION

Subjective text expresses opinions, emotions, sentiment and beliefs, while objective text generally report facts. So the task of distinguishing subjective from objective text is useful for many natural languages processing applications like mining opinions from product reviews [1], summarizing different opinions [2], question answering [3], etc.

Sentiment analysis or opinion mining refers to the general method to extract subjectivity and polarity from text, and semantic orientation refers to the polarity and strength of words, phrases, or texts. Our concern is primarily with the semantic orientation of texts, but we extract the sentiment of words and phrases towards that goal [4].

Lexical Resources for sentiment analysis is collection of words that express the subjectivity in text. In the research literature, theses words are also known as sentiment words, opinion words, opinion-bearing words, and polar words .There are two types of opinion words; positive opinion words which used to express the desired state (for examples "جميل" "beautiful", "رائع" "wonderful", "جيد" "good", and "حب" "love"), and negative opinion words which used to express the undesired state (for examples "رهيب" "terrible", "سئ" "bad", "كره " "hate", and "قبيح" "ugly").

There are three main approaches for generating subjective lexicons; manual approach, dictionary-based approach, and corpus-based approach. Manual approach is very time consuming ([5], [6], [7], and [8]) and thus it is not usually used alone, but combined with automated approaches as the final check because automated methods make mistakes. Also manual approach has a lack of

DOI : 10.5121/ijaia.2014.5601

neutralism that the decision of classification may be affected by (culture, religion, ethics ... etc). Dictionary based approach; one of the simple techniques in this approach is based on bootstrapping using a small set of seed opinion words and a dictionary, e.g., WordNet [9]. The strategy is to first collect a small set of opinion words manually with known orientations, and then to grow this set by searching in the WordNet for their synonyms and antonyms. The newly found words are added to the seed list. The next iteration starts. The iterative process stops when no more new words are found. This approach is unable to find opinion words with domain specific orientations. Corpus based approach; the methods here depend on a seed list of opinion words to find other opinion words in a large corpus and syntactic or co-occurrence patterns. Using the corpus-based approach alone to identify all opinion words is not as effective as the dictionary-based approach because it is hard to prepare a huge corpus to cover all language words. However, this approach has a major advantage that the dictionary-based approach does not have. It can help find domain specific opinion words and their orientations if a corpus from only the specific domain is used in the discovery process [10].

Numerous of researches are available on building sentiment lexicons in English and other languages but there are a very few number of such resources in Arabic. There are some factors that have contributed to reducing the research related to the construction of Arabic lexical resources for sentiment analysis such as:-

- The lack of availability of Arabic semantic database, still consider one of the most difficult problems that face researcher who try to generate an Arabic subjective lexicon.
- Arabic is a high inflectional and derivational language; often a single word has more than one affix such that it may be expressed as a combination of prefix(es), lemma (stem), and suffix(es). The prefixes are articles, prepositions, or conjunctions. The suffixes are generally objects or personal/possessive anaphora.
- The same three-letter root can give rise to different words, different meanings with different orientations. For example (ج, م, ل) may be ("جميل" "beautiful" ,positive) or ("جمل" "camel", objective) .For example (س,هـ ,ل) may be ("سهولة" "easy" ,positive) or ("سهل" "plain", objective). For example (و,ض,ع) may be ("التواضع" " modesty" ,positive) or ("الوضاعة" " rascaldom", negative).
- Arabic letters can be written with different shapes according to their position in the word. *E.g. "Alif"* has four forms (ا,أ,إ,آ), *"Yaa"* has two forms(ى , ي) and *"Taa el marpouta and el haa el marpouta " (ة,ه).*

The main contribution of this work is the development of an Arabic lexical resource of sentiment analysis by exploiting the relations available in the Arabic lexical semantics database. The database archives approximately 150,000 Arabic words, 18,413 semantic fields, and 20 semantic relations, such as synonyms, antonym, hyponymy and causality. In addition, we study the prior polarity effects of each word using our Sentiment Arabic Lexical resource on the sentence-level subjectivity and multiple machine learning algorithms.

The paper is organized as follows: Section 2 mentions related works of subjective lexicon and sentence level subjectivity classification. Section 3 gives an over view of Arabic Lexical Semantic Data Base –RDI which is part of The Research and Development International (RDI) toolkit. It is a main resource for generating our subjective lexicon. Section 4 presents the proposed algorithm for generating Sentiment Arabic lexical semantic database-RDI (SentiRDI) and how to determine the term orientation and subjectivity. Section 5 describes the corpus that is used in sentence subjectivity classifier. Section 6 focuses on tools used in classification and the experimental results. Finally, Section 7 draws our conclusions and future work.

2. RELATED WORK

In this section, we will present the most notable research work on building English sentiment lexicons and previous attempts to build Arabic sentiment lexicons. In addition, we will also present research work performed earlier on subjectivity analysis has been applied on English and Arabic.

Lexical Resources for opinion mining is used in determining term orientation (subjective /objective) and term polarity (positive/negative) out of context by using cognitive knowledge of human beings which called prior polarity lexicons. Various previous works ([11]; [12]; [13]) have already proposed techniques for making dictionaries for those sentiment words. Several sentiment lexicons are available for English such as SentiWordNet [13], Subjectivity Word List [14], WordNet Affect list [15], Taboada's adjective list [16]. SentiWordNet is the most widely used in several applications such as sentiment analysis, opinion mining and emotion analysis. SentiWordNet is an automatically constructed lexical resource for English that assigns a positivity score and a negativity score to each WordNet synset. [17].

For Arabic sentiment lexicon there are few attempts. An Arabic lexicon from similarity graph (A similarity graph is a type of graph where by two words or phrases have an edge if they are similar in polarity or meaning) was created in [18]. Lexicon consists of two columns, the first is the word and the second column represents the score of the word. An Arabic sentiment lexicon was created in [19] that contained strong as well as weak subjective clues by manually translating the MPQA lexicon [14].

A machine translation procedure is used to translate available English lexicons [20], including SentiWordNet [13], which is the most famous and most widely used English polarity lexicon [21], into Arabic. They retrieved 229,452 entries, including expressions commonly used in social media. The authors reported having problems with both coverage and with the quality of some of the entries. They also stated that they have not tested the system for the task of sentiment analysis. An Egyptian dialect sentiment lexicon was created in [22]. The researchers identified a set of lexico-syntactic patterns indicative of subjectivity, used a seed list of 380 manually constructed words, and subsequently performed pattern matching on a data set collected from tweeter. The incorrectly learned candidate terms were manually filtered. They retrieved 4,392 entries (193 compounds negative, 83 compound positive, 3,344 negative, and 772 positive).The work addressed dialectical or slang terms for the Egyptian dialect, which makes it unsuitable for use for other dialects.

Arabic sentiment lexicon that assigns sentiment scores to the words found in the Arabic WordNet (Arabic version of WordNet) was created in [23]. The authors here used lexicon_based approach for generating their lexicon by Starting from a small seed list contains 20 words (10 positive words and 10 negative words); they used semi-supervised learning to propagate the scores in the Arabic WordNet by exploiting the synset relations. The algorithm assigned a positive sentiment score to more than 800, a negative score to more than 600 and a neutral score to more than 6000 words in the Arabic WordNet. The lexicon was evaluated by incorporating it into a machine learning-based classifier.

The sentence level classification considers each sentence as a separate unit and assumes that sentence should contain only one opinion. Sentence-level sentiment analysis has two tasks subjectivity classification and sentiment classification.

Research work performed earlier on subjectivity analysis has been applied only on English and mostly at document-level and word-level. Some methods such as ([24], [25], [26], [27], [28],

[29], [30], [31] and [32]) which concentrated at sentence-level explored the effects of adjective orientation and gradability on sentence subjectivity. Authors in [29] judge the sentence subjectivity based on the adjectives used in the sentence. The idea was extended in [1] by using the polarity of the previous sentence to break classification ties.

Bootstrapping is another approach used in [27] is considered the state of art in sentence subjectivity detection. The idea is to use a high-precision classifier to automatically label a large un-annotated data. The labelled data is then given to a pattern supervised learning algorithm to generate a set of linguistically rich extraction patterns. The learned patterns are used to identify more subjective statements and the bootstrapping process continues. A similar method to automatically construct a corpus of HTML documents with polarity labels was used in [30]. Similar work involving self-training is described in [33] and [34]. A comprehensive survey of subjectivity recognition using different clues and features was presented in [35].

In Arabic, There are few works addressing subjectivity classification problem on sentence level ([18], [35]). Authors in [18] hit subjectivity analysis task on the three levels of document, phrase, and word and also provide polarity for mined sentences. They use lexicon based similarity graph algorithm to obtain sentiment polarities. A two-stage classification approach for subjectivity and polarity classification on sentence level was adapted in [35]. Authors use SVM classifier with a linear kernel to achieve these tasks.

3. ARABIC LEXICAL SEMANTIC DATA BASE –RDI

Arabic Lexical Semantic database (RDI-ArabSemanticDB) is part of The Research and Development International (RDI) toolkit. The Research and Development International (RDI) toolkit mainly consists of Arabic RDI-ArabMorpho-POS tagger [37] and RDI-ArabSemanticDB tool [38]. Arabic Lexical Semantic database of semantic fields [38] based on number of important references ("معجم تصريف الأفعال العربية"-" المعجم الوسيط" –" معجم المكنز الكبير") ("Glossary of Arabic verbs conjugation" – "the intermediary dictionary" – "dictionary thesaurus great"). Applying 20 semantic relations between fields (synonym, antonym, part of, kind of, causality, hyponym ….etc) and morphological analysis of each word into its type, prefix, form root and suffix. Around 2000 general concept are managed under 18,413 semantic fields, and 150,000 words based on stems (multiplied given prefix and suffix). The Database is flexible to be updated specifically for certain domains such as news and broadcasting domains. Some examples of semantic relations in RDI-ArabSemanticDB are given in *Figure 1*. Arabic Lexical Semantics Database RDI toll can do four functions [38]; first, retrieving the Arabic words that fall under a certain semantic field belongs to the category of specific knowledge of semantic fields previously public; known as placement front. Second, attributing any word to one or more of the semantic fields referred to in advance; known as posterior placement. Third, retrieving the semantic relations between any two pairs of semantic fields. Forth, drawing semantic fields associated with any semantic relationship with any particular semantic field.

Synonym	**الديمقراطية, التشاور, التعددية, الانتخابية,.......إلخ** **Democracy, consultation, pluralism, electoral … etc.**
Synonym	**كلمة ''الإرهاب'' و أختيار البحث بالمترادف لتأتي نتيجة البحث متضمنة كل المواضع التي تحتوي كلمة ''الإرهاب'' أو ما يرادفها ''التطرف-.... إلخ** **The word ''Terrorism'' and select searching synonyms to come as a result of search, including all positions that contain the word ''terrorism'' or what synonyms ''extremism -.... etc.**
Include	**كلمة ''الإرهاب'' و أختيار البحث بالاشتمال لتأتي نتيجة البحث متضمنة مواضيع ما يشمله هذا اليوم من '' التفجير- التدمير- التحريم-....إلخ** **The word ''terrorism'' and select searching inclusion comes as a result of the search, including what topics covered in this day of ''bombing- Destruction- prohibition -.... etc.**

Figure1. Examples of semantic relation

4. BUILDING AN ARABIC SUBJECTIVE LEXICON (SENTIRDI)

SentiRDI is a Sentiment Arabic lexical semantic database-RDI which determines both subjectivity and orientation for all semantic fields (18,413) and all words (150,000) which are covered by Arabic Lexical Semantic database. The algorithm determines both subjectivity and orientation for each semantic field. Our algorithm begins with chosen seed sets. Initially, we used the seed list defined in D.Turney and L. Littman [39]. The seed list contained 14 words {good, nice, excellent, positive, fortunate, correct, superior, bad, nasty, poor, negative, unfortunate, wrong, inferior}.We translated them into Arabic and used the 14 translated words in the initial run but it couldn't reach to all the semantic fields. The seed sets were extended by randomly choosing new words from the semantic fields that were unreachable by the previous seed sets and by adding these words to them. Also, we added an objective seed set in order to reach to all semantic fields in Arabic lexical semantic database. *Figure 2*, presents the three seed sets of our algorithm (positive, negative, and objective).

Positive Seed Set (S_pos)	["الألفة", "التواضع", "الشجاعة", "الشهامة", "الكرَم", "الجودة", "العطاء", "الضحك", "الجمال", "الصدق", "السهولة ", "الذكاء", "الاجتهاد", "التهذيب", "الأمل"] *["familiarity," "humility," "courage," "magnanimity", "generosity", "quality", "tender", "laugh", "Beauty", "honesty", "easy," "intelligence," "diligence, "" polite, "" hope "]*
Negative Seed Set (S_neg)	["الخوف", "الفقر", "التمرد", "التشاؤم", "الضعف", "الفساد", "الحزن", "الكره", "البخل", "الإبكاء", "التسول", "الذنب", "القبح", "الجهل", "الكدر"] *["Fear," "poverty," "rebellion," "pessimism," "weak," " Corruption," "sadness," "hatred," "greed," " crying," "beg," "guilt," "ugliness, " " ignorance, "" Chagrin "].*
Objective Seed Set(S_obj)	["الآلة", " الصوت", "الرجل", "الأرض"] *["Machine", "Earth", "man", "sound"].*

Figure2. Positive, negative, and objective seed sets

The expansion algorithm is divided into two parts; first, the expansion algorithm of objective seed set which explained in (*Figure* 3). Second, the expansion algorithm of positive and negative seed sets which explained in (*Figure 4*).

```
Function: Expand_Objective_Seedset
Input:
        - S_obj: objective seed set
        - Obj_Rel{ Totality, Part_of, Inclusion_K, KindOf  Circumstantial_Place , Locality_Place,
            Circumstantial_Time , and Locality_Time}
        - Q: an object contains the Lexical Semantic database RDI.
        - Current_level: number of current iteration ; initialized by 0
Output:
        - Obj_list : list of objective semantic fields
Begin

For each Semantic field in S_obj
            Add Semantic field in obj_ queue  // initialize obj_queue
While (obj_queue is not empty)
            Current_level ← Current_level+1
            Cur_Node_obj ← remove the top node from obj_queue
            If  Cur_Node_obj(not found) in Obj_list   //to ensure that the added node isn't visited
                        Cur_node_obj.level ← Current_level
                        Cur_node_obj.repetation ← 1
                        Add Cur_node_obj to Obj_list
            Else
                        Cur_node_obj.repetation ← Cur_node_obj.repetation +1;

            For each R in Obj_Rel search Q with Cur_node_obj to get Related_Nodes
                        For each N in Related_Nodes
                                    If (N not found in obj_queue AND N not found in Obj_list)
                                                Add N to Obj_queue
End
```

Figure3. Expansion algorithm of objective seed set

Figure 3 explains the Procedure Expand_Objective_Seedset. The procedure takes objective seed set (*Figure 2*), set of semantic relations were needed to expand the objective seed set, the Lexical Semantic database RDI, and the Current_level variable which initialize by zero and incremented by one in each iteration. Queue is initialized by semantic fields in objective seed set. Before putting any semantic field in Obj_list, we ensure that it wasn't added before, set its level with the value of Current_level, set its repetition value equal one. If the semantic field putted in Obj_list before, its repetition value increment by one. All objective relations were applied over each new semantic field in order to expand the set. The function stops when there isn't any new semantic fields are added (queue is empty).

Figure4 explains the Procedure Expand_Pos_Neg_Seedset. The procedure takes positive and negative seed sets (*Figure 2*), set of semantic relations were needed to expand two seed sets, the Lexical Semantic database RDI, and the Current_level variable which initialize by zero and incremented by one in each iteration. This procedure is the same as the pervious but it has two differences; the initialization of two queues and types of semantic relations are needed to expand positive and negative seed sets. Positive and negative queues are initialized by the synonyms of words in positive and negative seed sets. For the relations; there are two types of relations; relations are applied to the semantic field, the resulted semantic fields have the same orientation {Causality, Causative, hyponymy, and Hypernym} and relations are applied to the semantic field, the resulted semantic fields have the opposite orientation {Antonym}.

```
Function: Expand_Pos_Neg_Seedset
Input:
            -   S_pos , S_neg: positive and negative seed set
            -   PN_Same_Rel { Causality, Causative,  hyponymy, and Hypernym }
            -   PN_Opposite_Rel { Antonym}
            -   Q: an object contains the Lexical Semantic database RDI.
            -   Current_level: number of current iteration ; initialized by 0
Output:
            -   Pos_list  : list of positive semantic fields
            -   Neg_list :list of  negative semantic fields
Begin

For each Word in S_pos Get Synonyms
        Add Synonyms in  pos_ queue  // initialize pos_queue

For each Word in S_neg Get Synonyms
        Add Synonyms in  neg_ queue   // initialize neg_queue

While ((pos_queue is not empty) OR  (Neg_queue is not empty) )
        Current_level ← Current_level+1
        Cur_Node_pos ←remove the top node from pos_queue
        If  Cur_Node_pos (not found) in Pos_list  //to ensure that the added node isn't visited
                Cur_node_pos.level← Current_level
                Cur_node_pos.repetation← 1
                Add Cur_node_pos to pos_list
        Else
        Cur_node_pos.repetation← Cur_node_pos.repetation +1;

        Cur_Node_neg← remove the top node from neg_queue
        If  Cur_Node_neg( not found) in neg_list
                Cur_node_neg.level← Current_level
                Cur_node_neg.repetation← 1
                Add Cur_node_neg to neg_list
        Else
                Cur_node_neg.repetation← Cur_node_neg.repetation +1;

        For each R in PN_Same_Rel  search Q with Cur_node_pos to get Related_Nodes
          For each N in Related_Nodes
                If (N not found in pos_queue AND N not found in Pos_list)
                        Add N to pos_queue

        For each R in PN_Opposite_Rel search Q with Cur_node_pos to get Related_Nodes
          For each N in Related_Nodes
                If (N not found in neg_queue AND N not found in Neg_list)
                        Add N to neg_queue

        For each R in PN_Same_Rel search Q with Cur_node_neg to get Related_Nodes
         For each N in Related_Nodes
                If (N not found in neg_queue AND N not found in Neg_list)
                        Add N to neg_queue

        For each R in PN_Opposite_Rel search Q with Cur_node_neg to get Related_Nodes
         For each N in Related_Nodes
                If (N not found in pos_queue AND N not found in Pos_list)
                        Add N to pos_queue
End
```

Figure4. Expansion algorithm of positive and negative seed sets

After more than hundred trials found that number of terms in seed set are thirty four (15 positive terms, 15 negative terms and 4 objective terms) which cover 98.64% of semantic fields of Database [40]. The rest of semantic fields (250) were annotated manually. Number of positive semantic fields are 3156, negative semantic fields are 4169 and objective semantic fields are 10,839.

4.1 Experiments analysis

The evaluation of semantic fields in SentiRDI is based on the exact match between identified polarity of semantic fields and the manually annotated semantic fields according to three properties (positive, negative, and objective). We used precision, recall and F-measure in evaluation.

Testing SentiRDI is done by two ways; the first method by taking a subset of Arabic semantic fields as a “gold standard” and annotated it manually. The subset contained 7612 semantic fields was annotated manually by 5 annotators. The second method, by translating the Micro-WNOp [41] (by Google translator) gold standard that is used to evaluate the SentiWordNet which contains 1.105 synsets .

Table 1 shows the results of two tests that explained in the previous section. For each test, we calculate Recall, Precision, and F_measure for all semantic fields (All_SF), objective semantic fields (Obj_SF), negative semantic fields (Neg_SF), and positive semantic fields (Pos_SF) were included in each test. The results of first test are better than the second one in terms of recall and precision, and F-measure for all and objective semantic fields. The results of second tests are better than the first in terms of Recall, Precision, and F-measure for negative semantic fields.

Table1. Testing SentiRDI Results

	First Test				**Second Test**			
	All_SF	**Obj_SF**	**Neg_SF**	**Pos_SF**	**All_SF**	**Obj_SF**	**Neg_SF**	**Pos_SF**
Recall	**87.32**	**89.3**	83.16	**86.44**	84.67	89	**85**	79,43
Precision	**87.72**	**93.09**	79.95	74.76	84.95	87.93	82,45	**80**
F-measure	**87.52**	**91.16**	81.52	**80.18**	84.81	88.46	83.7	79,71

From results , we conclude that the direct translation from one language (English) to another language (Arabic) doesn't give accurate results due to the drawbacks of machine translation including the loss of polarity sentiments of some words when translated to other language (the same term may have a lot of meaning with different polarities). *Table 2* shows some terms that are found in the Micro-WNOp [41] gold standard and give different meaning with different polarities.

Table2. Samples in Micro-WNOp gold slandered.

Word	**Different Meaning**		
	Positive	**Negative**	**Objective**
Ball	*a picnic* نزهة *Shindig* حفلة راقصة	—	*a game of ball games* لعبة من ألعاب الكرة *Football* الكرة *a bullet* رصاصة
Shark	—	Crook نصاب Swindler محتال	Shark سمك قرش Sea dog كلب البحر
Humble	Humble متواضع Meek وديع	Lowly وضيع Servile ذليل	—
Cure	Healing شفاء	—	Curing fish تقديد السمك Salting meat تمليح اللحم

5. SUBJECTIVITY CLASSIFIER

5.1 Polarity determination of An Arabic word using SentiRDI

The simplest case that input Arabic word (from text) may be exact matching with one of pre-defined polarity of lexicon's semantic fields. However, in most cases , input word might be derivational or fall under a certain semantic field ,so, Arabic RDI-ArabMorpho-POS tagger and RDI-ArabSemanticDB tool were aiding to get stem , root, and synonyms (one or more semantic field) of the word in order to find the corresponding pre-defined polarity of lexicon semantic fields. If the input word not found, we can classify it manually and add it to our lexicon.

```
Function: determining polarity of An Arabic Word
Input:
        -   Word
        -   Subjective lexicon (SentiRDI)
Output:
        -   Pol_word : word polarity (positive, negative, objective)
Begin
    1- Inputs are word, its stem and its root (get stem and root using RDI which uses A* search
       algorithm to find the correct root for the word and perform its morphological analysis
       with 96.8% accuracy).

    2- Search for the inputs in the SentiRDI, if it is found then get its polarity (pol_word) and
       exit.

    3- Else get the synonyms of the input word and the polarity of each one.

    4- Get number of positive synonyms (P), number of Negative synonyms (N) and number of
       objective synonyms (O).

    5- Get M=max( P,N,O)     // Get maximum of three number
          {
              if (M=P) then  pol_word=  positive
              if (M=N) then  pol_word= negative
              if (M=O) then  pol_word= objective
          }
       else if (N=P) then  pol_word=   objective
       else if (N=O) then  pol_word=  negative
       else if (P=O) then  pol_word=  positive

        6-   If the word couldn't be found in semantic database in this case we determine its
             polarity manually and add it with its set polarity to SentiRDI.
End
```

Figure5. Determining polarity of an Arabic word algorithm

Figure5 explains how to determine the prior polarity of an input Arabic word using SentiRDI. The procedure takes two inputs word and SentiRDI. First, stem and root of input word are gotten by using RDI tools. Search for Word, root and stem in SentiRDI if found set the polarity of word with the founded polarity then exit (step 2). Else, get the synonyms of the word and the polarity of each one. Count number of positive synonyms (P), number of negative synonyms (N), and number of objective synonyms (O) (step 4) then get the maximum of the three numbers and set it to (M). Set the polarity of input word with the polarity of (M) (step 5). If the input word isn't covered by the semantic database, in this case set the polarity of word manually and add it to SentiRDI.

5.2 Corpus

Corpus[1] is Translated MPQA corpus provided in [42] containing subjective and objective sentences of Arabic language .for English, MPQA corpus[2] consisting of 535 English-language news articles from 187 different foreign and U.S., manually annotated for subjectivity [32]. The corpus consists of 9700sentences, 55% of them are labelled as subjective, while the rest are objective.

5.3 Text Pre-processing

Simple text pre-processing was executed in order to remove special characters and non Arabic characters. More advanced text pre-processing was executed in order to prepare it for input to different learning algorithm.

- First step was extracting names from documents: this step reduced calculations for determining the prior polarity for tokens because all named entities are objective. Technique used in this step was Arabic Named Entity Recognition (ANER) [43] which was an integration approach between two machine learning techniques, namely bootstrapping semi-supervised pattern recognition and Conditional Random Fields (CRF) classifier as a supervised technique. This technique extracted approximately 59,772 named entity from subjective corpus and 50,787 named entity from objective corpus.

- The Second step that was assigning Part Of Speech tags (POS). The tool that used to assign POS tags was the Research and Development International (RDI) tool.

- Third step to obtain the prior polarity of words by using SentiRDI (Section 5.1).

Table3. Statistics from the MPQA Arabic opinion corpus

	Subjective corpus	Objective corpus
Total tokens	129.169	90,357
Total sentences	5335	4365
Total NER	59772	50787

[1] http://www.cse.unt.edu/~rada/downloads.html#msa
[2] http://www.cs.pitt.edu/mpqa

5.4 Features selection

Feature execration involves extracting tokens that are relevant to detecting sentiment in the sentence. The features proposed here were similar as in [44]:-

1. POS is the Part of Speech (POS) tagging of the word**.** RDI-ArabMorpho-POS tagger was used [38] and by using Prior polarity lexicon (SentiRDI) for determining polarity of each word.

1.1 Number of positive noun.
1.2 Number of positive verb.
1.3 Number of negative noun.
1.4 Number of negative verb.

2. Average Polarity of sentence = $\frac{1}{n}\sum_{i=1}^{n} \mathbf{P}_{\mathbf{wi}}$ (1)

Where *n* is number of words in sentence, *Pwi* polarity of word *i* in sentence that is specified before from prior polarity database (SentiRDI) such that

$$p_{wi} = \begin{cases} -1 \; negative \\ 0 \; objective \\ 1 \; positive \end{cases} \qquad (2)$$

3. Average Term Frequency: Inverse Sentence Frequency (TF-ISF) for sentence (*Si*) can be computed by the following equation:-

$$Avg\ TF\_ISF_{si} \frac{1}{||Si||} \sum_{t=1}^{||si||} \left(TF_{t,si} * ISF_{t}\ \right) \qquad (3)$$

Where *TF* presents the number of occurrences of each term within the sentence and can be normalized by dividing it by size of sentence.

$$TF_{t,s} = \frac{N_{t,s}}{||S||} \qquad (4)$$

Where *Nt,s* is the number of occurrences of term *t* in sentence *S*. ||*S*|| is the number of words in sentence *S*. *ISF* is used for terms that appear in the small number of sentences. This factor is useful because numbers of subjective terms are small compared with neutral (objective) ones.

$$\mathrm{ISF} = \log \frac{S}{S_i} \qquad (5)$$

6. Experimental Analysis

We use more than thirty classifiers with different types (Bayes, Rules, Tress and Function). The experimentation was executed using 10 Cross-Fold-Validation with two prior polarity subjective lexicons (MPQA subjectivity lexicon which is proposed by [19] and SentiRDI Lexicon).
The evaluation is based on the exact match between identified subjectivity sentence and the manually annotated true subjective sentence. We used Precision, Recall and F-measure to evaluate.

$$\mathrm{Precision} = \frac{\mathrm{True\ subjective\ sentences}}{retreived\ subjective\ sentesnces} \qquad (6)$$

$$\mathrm{Recall} = \frac{\mathrm{True\ subjective\ sentences}}{relevant\ subjective\ sentesnces} \qquad (7)$$

$$\mathrm{F_{measure}} = 2 * \frac{\mathrm{Precision} * Recall}{\mathrm{Precision} + Recall} \qquad (8)$$

Subjectivity classifiers results by using SentiRDI and MPQA subjectivity lexicon is presented in Table 4. From our experiment we find that most classifiers give better results with SentiRDI except Naïve Bayes that gives higher precision with MPQA subjective lexicon. FT classifier gives the best result with selective features that is give accuracy 76.1% with using SentiRDI as a prior polarity subjectivity lexicon.

Table4. Subjectivity classifiers results

Classifier Name		MPQA Subjectivity Lexicon	SentiRDI
NaiveBayes	*F_measure*	60.8	70.55
	Precision	**71.9**	70.6
	recall	65.5	70.5
Ridor	*F_measure*	71.2	73.82
	Precision	71.7	75.4
	recall	71.8	72.3
FT	*F_measure*	72.4	**76.1**
	Precision	72.5	**76.2**
	recall	72.6	**76**
LMT	*F_measure*	72.4	76
	Precision	72.6	76
	recall	72.6	76

7. Conclusion and Future Work

This paper investigates the method for generating Sentiment Arabic lexical Semantic Database by using dictionary based approach. Also, we study the prior polarity effects of each word using our Sentiment Arabic Lexical Semantic Database on the sentence-level subjectivity and multiple machine learning algorithms. Our main contribution is to acquire the sentiment Arabic lexical semantic database (SentiRDI) having very large number of Arabic words with different derivational forms and part of speech tags. The created lexicon is context independent and it can be used in any opinion corpora.

In the future, we are going to extend the database depending on further analysis of exiting opinion mining Arabic corpora. Also, we are going to try more classifiers to improve results more than this and apply contextual polarity to get good examples can be added to our lexicon and improve the classification results.

References

[1] Minqing Hu and Bing Liu 2004." Mining opinion features in customer reviews," Proceedings of the National Conference on Artificial Intelligence, 755–760.

[2] Kavita Ganesan , ChengXiang Zhai and Jiawei Han,2010. "Opinosis:a graph-based approach to abstractive summarization of highly redundant opinions," Proceedings of the 23rd International Conference on Computational Linguistics,340–348

[3] Alexandra Balahur Ester Boldrini , Andrés Montoyo and Patricio Martínez-Barco, 2009. "Opinion and Generic Question Answering systems: a performance analysis," Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 157–160.

[4] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," Computational linguistics 37, no. 2, pp. 267-307, 2011.

[5] S. R. Das and M. Y. Chen, “Yahoo! for Amazon: Sentiment extraction from small talk on the Web,” Management Science, vol. 53, pp. 1375–1388, 2007.

[6] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, “Mining product reputations on the Web,” Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 341–349, 2002.

[7] R. M. Tong, “An operational system for detecting and tracking opinions in on-line discussion.”Proceedings of the Workshop on Operational Text Classification (OTC), 2001.

[8] YI, J., NASUKAWA, T., BUNESCU, R. AND NIBLACK, W. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques, In Proceedings of the 3rd IEEE International Conference on Data Mining, 427-434.

[9] C. Fellbaum, ed., Wordnet: An Electronic Lexical Database. MIT Press, 1998.

[10] B. Liu, "Sentiment Analysis and Subjectivity", Handbook of Natural Language Processing. Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010.

[11] Pang Bo, Lee Lillian, and Vaithyanathan Shivakumar,2002. "Thumbs up? Sentiment classification using machine learning techniques" In Proceedings of EMNLP, pages 79–86.

[12] Janyce Wiebe and Rada Mihalcea , 2006. "Word sense and subjectivity," In Proceedings of COLING/ACL-06. Pages 1065--1072.

[13] Esuli Andrea and Sebastiani Fabrizio,2006. SentiWordNet:" A publicly available lexical resource for opinion mining," In Proceedings fLREC.

[14] Wilson Theresa, Wiebe Janyce and Hoffmann Paul ,2005. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," In Proceedings of HLT/EMNLP 2005, Vancouver,Canada.

[15] Carlo Strapparava and Alessandro Valitutti ,2004. "WordNet-Affect: an affective extension of WordNet," In Proceedings of LREC 2004, pages 1083 – 1086, Lisbon, May .

[16] Kimberly Voll and Maite Taboada., 2007". Not All Words are Created Equal: Extracting SemanticOrientation as a Function of Adjective elevanc," In Proceedings of the 20th Australian Joint Conference on Artificial Intelligence. pages. 337-346.

[17] Amitava Das and Sivaji Bandyopadhyay,2010. "Towards The Global SentiWordNet," In the Workshop on Model and Measurement of Meaning (M3), PACLIC 24, November 4, Sendai, Japan.

[18] Elhawary, M., and Elfeky, M. (2010). “Mining Arabic Business Reviews.” IEEE International Conference on Data Mining Workshops

[19] M. Elarnaoty, S. AbdelRahman, and A. Fahmy. A Machine Learning Approach For Opinion Holder Extraction Arabic Language. CoRR, abs/1206.1011, 2012.

[20] Muhammad Abdul-Mageed, Mona Diab, 2012. Toward building a large-scale Arabic sentiment lexicon, Proceedings of the 6th International Global WordNet Conference.

[21] M. Abdul-Mageed, M. Korayem, and A. YoussefAgha. “Yes we can?”: Subjectivity annotation and tagging for the health domain. In Proceedings of the International Confrence Recent Advances in Natural Language Processing RANLP, Hissar, Bulgaria, 2011.

[22] Samhaa R. El-Beltagym, Ahmed Ali, 2013. Open issues in the sentiment analysis of Arabic social media: a case study, Proceedings of 9th International Conference on Innovations in Information Technology (IIT), pp. 215–220.

[23] Mahyoub, F. H., Siddiqui, M. A., & Dahab, M. Y. (2014). Building an Arabic Sentiment Lexicon Using Semi-Supervised Learning. Journal of King Saud University-Computer and Information Sciences.

[24] Vasileios Hatzivassiloglou and Janyce M. Wiebe, 2000." Effects of adjective orientation and gradability on sentence subjectivity," in Proceedings of the International Conference on Computational Linguistics (COLING).

[25] Janyce Wiebe and Theresa Wilson,2002 . "Learning to disambiguate potentially subjective expressions," in Proceedings of the Conference on Natural Language Learning (CoNLL), pp. 112–118.

[26] Hong Yu and Vasileios Hatzivassiloglou, 2003. "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[27] Ellen Riloff and Janyce Wiebe ,2003. "Learning extraction patterns for subjective expressions, "in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[28] Pang, B. , and Lee, L. ,2004. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in Proceedings of the Association for Computational Linguistics (ACL), pp. 271–278.

[29] Kim, S.M. , and Hovy, E. ,2005. "Automatic detection of opinion bearing words and sentences, "in Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP).

[30] Kaji N. , and Kitsuregawa, M., 2006. "Automatic construction of polari-tagged corpus from HTML documents," in Proceedings of the COLING/ACL Main Conference Poster Sessions.

[31] Wilson , T. , Wiebe, J. , and Hwa, R. ,2006. "Just how mad are you? Finding strong and weak opinion clauses," in Proceedings of AAAI, pp. 761–769, 2004. (Extended version in Computational Intelligence, vol. 22, no. 2, pp. 73–99.

[32] Breck, E. , Choi, Y. , and Cardie, C. ,2007. "Identifying expressions of opinion in context," in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad, India.

[33] Janyce Wiebe and Ellen Riloff, 2005. "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts," In Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics, pages 486–497, Mexico City, Mexico.

[34] Riloff, E. , and Wiebe, J. , and Wilson, T., 2003. "Learning subjective nouns using extraction pattern bootstrapping, "in Proceedings of the Conference on Natural Language Learning (CoNLL), pp. 25–32.

[35] Wiebe, J. , Wilson, T. , Bruce, R. , Bell, M. , and Martin, M. ,2004. "Learning subjective language, "Computational Linguistics, vol. 30, pp. 277–308.

[36] M. Abdul-Mageed and M. T. Diab. Subjectivity and sentiment annotation of modern standard arabic newswire. In Proceedings of the 5th Linguistic Annotation Workshop, LAW V '11,pages 110–118, 2011.

[37] M.Attia, and M.Rashwan (2004). "A Large Scale Arabic POS Tagger Based on a Compact Arabic POS Tag Set and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words", NEMLAR.

[38] M.Attia, M.Rashwan, A.Ragheb, M.A.Al-Basoumy, and S.Abdou, 2008."A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields ," Proceedings of the 6th international conference on Advances in Natural Language Processing.

[39] Peter D. Turney, Michael L. Littman, 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus, Technical Report EGB-1094, National Research Council Canada.

[40] Hanaa B. Mobarz, , Mohsen Rashwan, and Samir AbdelRahman, 2011, "Generating lexical Resources for Opinion Mining in Arabic language automatically," The Eleventh Conference on Language Engineering ESOLEC', Cairo-Egypt, http://esole-eg.org/index.php/en/conferences, Sept.

[41] Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. ,2007. Language resources and linguistic theory: Typology, second language acquisition, English linguistics (Forthcoming), chapter Micro-WNOp: "A gold standard for the evaluation of automatically compiled lexical resources for opinion mining,"Franco Angeli Editore, Milano, IT.

[42] J. Wiebe and E. Riloff, 2005. "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts," In Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics, pages 486–497, Mexico City, Mexico.

[43] Samir AbdelRahman, Mohamed Elarnaoty, Marwa Magdy and Aly Fahmy, 2010. "Integrated Machine Learning Techniques for Arabic Named Entity Recognition, "IJCSI International Journal of Computer Science, pp. 1694-0784.

[44] Samir AbdelRahman ,Hanaa B. Mobarz, , Mohsen Rashwan, and Ibrahim Farg, 2014, " Arabic Phrase-Level Contextual Polarity Recognition to Enhance Sentiment Arabic Lexical Semantic Database Generation," IJACSC International Journal of Advanced Research in Artificial Intelligence, Vol 5,No.10.