



King Saud University  
**Journal of King Saud University –  
Computer and Information Sciences**

www.ksu.edu.sa  
www.sciencedirect.com



# Building an Arabic Sentiment Lexicon Using Semi-supervised Learning



Fawaz H.H. Mahyoub<sup>a,b</sup>, Muazzam A. Siddiqui<sup>a,\*</sup>, Mohamed Y. Dahab<sup>a</sup>

<sup>a</sup> Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

<sup>b</sup> Faculty of Computer Sciences and Information Technology, Taiz University, Taiz, Yemen

Available online 28 September 2014

## KEYWORDS

Sentiment lexicon;  
Sentiment analysis;  
Arabic natural language  
processing;  
Text mining;  
Semi-supervised learning

**Abstract** Sentiment analysis is the process of determining a predefined sentiment from text written in a natural language with respect to the entity to which it is referring. A number of lexical resources are available to facilitate this task in English. One such resource is the SentiWordNet, which assigns sentiment scores to words found in the English WordNet. In this paper, we present an Arabic sentiment lexicon that assigns sentiment scores to the words found in the Arabic WordNet. Starting from a small seed list of positive and negative words, we used semi-supervised learning to propagate the scores in the Arabic WordNet by exploiting the synset relations. Our 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 experiments were conducted on several Arabic sentiment corpora, and we were able to achieve a 96% classification accuracy.

© 2014 King Saud University. Production and hosting by Elsevier B.V. All rights reserved.

## 1. Introduction

Sentiment analysis is the process of determining a predefined sentiment from online texts written in a natural language with respect to a specific subject. The need for sentiment analysis is the product of a sudden increase in opinionated or sentimental texts in the form of blogs, reviews, and discussions (Pang and

Lee, 2008). The idea of processing these comments or reviews has attracted many researchers in the field of text mining, with the aim of extracting a general opinion about one item or theme among the substantial amounts of unstructured data available on the Internet. In this paper, we present an Arabic sentiment lexicon that was developed by exploiting the semantic relations found in the Arabic WordNet. While there are several previous examples of using WordNet to build an English sentiment lexicon (Kim and Hovy, 2004; Esuli and Sebastiani, 2005, 2006), to the best of our knowledge, this is the very first attempt to build an Arabic sentiment lexicon using the Arabic WordNet. The Arabic WordNet is the Arabic version of WordNet and can be seen as a network with a collection of semantically similar words, called synsets, as nodes and a number of semantic and lexical relations as links between the synset nodes. We used a semi-supervised approach to propagate the sentiment scores from a small seed list of positive and negative

\* Corresponding author.

E-mail addresses: [fawazh7@gmail.com](mailto:fawazh7@gmail.com) (F.H.H. Mahyoub), [maasiddiqui@kau.edu.sa](mailto:maasiddiqui@kau.edu.sa) (M.A. Siddiqui), [mdahab@kau.edu.sa](mailto:mdahab@kau.edu.sa) (M.Y. Dahab).

Peer review under responsibility of King Saud University.



Production and hosting by Elsevier

words in the Arabic WordNet. We devised an algorithm that identified the nodes in the Arabic WordNet that contain the words in the seed list and iteratively spread the scores of these words to the neighboring nodes until the entire network was reached. The score for each term was represented as a triplet containing a positive, negative and neutral score. Each of these constituent scores in the triplet for a term was represented as positive numerical values. The scheme is somewhat similar to how scores are represented in the SentiWordNet, but in our case, the scores were unnormalized, i.e., the positive, negative and neutral scores of the term do not sum to one.

The main contribution of this work is the development of an Arabic sentiment lexicon containing 7.5 K terms by exploiting the relations available in the Arabic WordNet. In addition to the sentiment scores, the lexicon also contains the part of speech tag of each term and its diacritized form for lexical disambiguation. For some of the terms, the gloss containing the term definition is also available.

The remainder of this paper is structured as follows: The next section briefly describes the Arabic WordNet. In Section 3, we present the previous major approaches to developing a sentiment lexicon. In Section 4, we describe the development of an Arabic sentiment lexicon. In Section 5, we evaluate the proposed algorithm. Finally, Section 6 is devoted to conclusions and future work.

## 2. What is the Arabic WordNet?

WordNet is a lexical database of the English language. Unlike a dictionary, the words, including nouns, verbs, adjectives and

adverbs, are grouped into sets of synonyms called synsets. These synsets are related to each other through different semantic and lexical relations; hence, the WordNet can be viewed as a directed graph (Fellbaum, 1998). The Arabic WordNet is the Arabic version of the English WordNet. The Arabic WordNet database structure is composed of four principal entity types: item, word, form and link. Items are conceptual entities, including synsets, ontology classes and instances. A word entity is a word sense. A form is a special form that is considered as dictionary information. Links are relations between synsets. They are classified according to the part of speech (POS) of the related synsets (verb, noun, adjective, and adverb) or according to their type (lexical, semantic and lexico-semantic relations). Table 1 presents WordNet and Arabic WordNet statistics (WordNet 3.0 database statistics), (Fellbaum et al., 2006). Table 2 shows different relations in the Arabic WordNet according to their classification type (Mahdi Boudabous et al., 2013).

## 3. Related work

Although plenty of research is available on building sentiment lexicons in English and other languages, Arabic has yet to receive the attention it deserves by researchers in this field. In this section, we will present the most notable studies on building English sentiment lexicons and previous attempts to build Arabic sentiment lexicons. In addition, we will also cover studies that claim language independence.

Hatzivassiloglou and McKeown (1997) developed an algorithm for predicting the orientation of an adjective. Turney and Littman (2002) proposed a method to determine a document's polarity. The method involves issuing queries to a Web search engine. The approach targets adjectives and adverbs; therefore, it relies on the existence of a huge POS-tagged corpus, which is a rarity for the Arabic language. The available POS taggers are not fully qualified to identify all parts of speech and are not able to distinguish between different sentence types (Farra et al., 2010). Lexical resources, such as WordNet (Fellbaum, 1998), are used in Kim and Hovy (2004), Esuli and Sebastiani (2005, 2006), Kamps et al. (2004). These studies started with

**Table 1** WordNet and Arabic WordNet database statistics.

POS	AWN		PWN	
	Word forms	Synsets	Word forms	Synsets
Noun	15,890	7,960	117,798	82,115
Verb	6,084	2,538	11,529	13,767
Adjective	1,243	661	21,479	18,156
Adverb	264	110	4,481	3,621
Total	23,481	11,269	155,287	117,659

**Table 2** Arabic WordNet relation classification (Mahdi Boudabous, 2013).

Type	Relation	Example	Frequency
Semantic relations	Has hyponym	شَرَاب has hyponym ماء, خليط	9352
	Has holo part	أوروبا has holo part فرنسا, إسبانيا	697
	Has subvent	أكل has subvent بُلغ	128
	Has instance	مدينة has instance مرسيليا	1067
	See also	اِكتَسَفَ See also عَرَفَ	192
	Causes	ذَكَرَ Causes تَنَكَّرَ	75
	Has holo member	فرنسا has holo member الاتحاد الأوروبي	334
	Verb group	سَمَخَ – أجازَ	152
	Region term	البلقان Region term حروب اليونان	35
	Category term	جيش categorie termes عقيد, لواء	548
	Has holo made of	هيدروجين, أكسجين has holo made of ماء	60
	Be in state	إمكان Be in state قدر	83
	Usage term	اسم تجاري Usage term أسيرين	3
Lexical relations	Near synonym	فرد near synonym انسان	122
	Near antonym	أب Near antonym أم	722
Lexico-semantic relations	Related to	اِحْتَفَظَ related to اِحْتَفَظَ	4774
	Has derived	أبوي has derived أبوي والذان	178

small hand-crafted seed lists, and by following WordNet relations, they were able to expand the seed lists. Kim and Hovy (2004) used seed lists of 44 verbs (23 positive and 21 negative) and 34 adjectives (15 positive and 19 negative) and subsequently iteratively expanded the seed lists using the WordNet. Synonym and antonym relations were used to expand adjectives, and only synonyms were used to expand verbs. The researchers obtained 5880 positive adjectives, 6233 negative adjectives, 2840 positive verbs, and 3239 negative verbs. Esuli and Sebastiani (2005) used the WordNet to determine the orientation of a term based on the classification of its glosses. The authors assumed that terms with similar orientation tend to have similar glosses. Esuli and Sebastiani (2006) extended their method from Esuli and Sebastiani (2005) to the determination of both term subjectivity and term orientation. Kamps et al. (2004) determined sentiments of adjectives in the WordNet by calculating the relative distance of the term from the two seed words “good” and “bad”. The approach is difficult to adapt in Arabic because the number of relations in the Arabic WordNet is much smaller than its English counterpart. In addition, the glosses for most synsets are not available in the Arabic WordNet.

Elhawary and Elfeky (2010) used a similarity graph to build an Arabic lexicon. A similarity graph is a type of graph whereby two words or phrases have an edge if they are similar in polarity or meaning. The weight of the edge represents the degree of similarity between two nodes. The researchers initially used a seed list of 1600 words (600 positive, 900 negative, and 100 neutral) and subsequently performed label propagation on an Arabic similarity graph. The Arabic lexicon created from the similarity graph consists of two columns, where the first column is the word or phrase and the second column represents the score of the word, which is the sum of the scores of all edges connected to this node (word/phrase). They applied filtering rules to avoid both the sparseness of the data and garbage nodes. They removed nodes with a high number of weighted edges and retained the 25 top-ranked synonyms of the word. This approach depends on a huge Arabic corpus to build the similarity graph, which is not available to us. The entries in the created lexicon are polarity words without scores.

Arabic lexical resources such as Penn Arabic Treebank (Maamouri et al., 2004) and SentiStrength project (Thelwall et al., 2010) are used in Abdul-Mageed and Korayem (2010) and El-Halees (2011), respectively. Abdul-Mageed and Korayem (2010) manually created an Arabic SSL based on the Penn Arabic Treebank. The researchers extracted all adjectives from the first four parts of the Penn Arabic Treebank and manually selected those adjectives that they believed are either positive or negative. Their approach targets only adjectives, and the intensity scores are missing. El-Halees (2011) manually created an Arabic SSL based on two resources: the SentiStrength project and an online dictionary. The researchers translated the English list from the SentiStrength project and subsequently manually filtered it. Common Arabic words were added to the lexicon. The drawbacks of machine translation include the loss of polarity sentiments of some words when translated to other language.

The authors in Elarnaoty et al. (2012) and Abdul-Mageed and Diab (2012) exploited a simple machine translation procedure on an existing English polarity lexicon. Elarnaoty et al. (2012) created an Arabic sentiment lexicon that contains strong as well as weak subjective clues by manually translating

the MPQA lexicon (Wilson et al., 2005). Abdul-Mageed and Diab (2012) used a machine translation procedure to translate available English lexicons, including SentiWordNet (Esuli and Sebastiani, 2006), which is the most famous and most widely used English polarity lexicon (Abdul-Mageed et al., 2011), 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.

El-Beltagy and Ali (2013) created an Egyptian dialect sentiment lexicon. 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 compound negative, 83 compound positive, 3,344 negative, and 772 positive). The work addressed dialectal or slang terms for the Egyptian dialect, which makes it unsuitable for use for other dialects.

#### 4. Building the lexicon

This section presents our algorithm, which assigns sentiment scores to the words found in the Arabic WordNet to build a sentiment lexicon. Starting from a small seed list of positive and negative words, we used semi-supervised learning to propagate the scores on the Arabic WordNet by exploiting the synset relations. We used the relations that were employed in developing the WordNet-Affect (Valitutti et al., 2004) database. These relations include eight semantic/lexical relations {near\_synonym, verb\_group, see\_also\_wn15, has\_derived, related\_to, has\_subevent, causes and near\_antonym}. We used the seed list defined in D. Turney and L. Littman (2002). The seed list contained 14 words {good, nice, excellent, positive, fortunate, correct, superior, bad, nasty, poor, negative, unfortunate, wrong, inferior}. We translated them to Arabic and filtered them based on their availability in the Arabic WordNet. The filtered list contained only four positive and four negative words. Initial runs of our expansion algorithm indicated that with the eight words in the seed list, the algorithm was not able to reach all of the synsets in the Arabic WordNet network. The seed list was extended by randomly choosing new words from the synsets that were unreachable by the previous seed lists and by adding these words to the previous seed lists. The process was repeated until all of the synsets were reached. Table 3 presents the positive seed list, and Table 4 presents the negative seed list.

##### 4.1. Expansion algorithm

The expansion algorithm pseudo-code is shown in Fig. 1 and in Fig. 2. The procedure Expansion Algorithm, presented in Fig. 1, takes three arguments as the input. These include the positive and negative seed lists, the Arabic WordNet database and a special sentiment orientation flag that is used in the process of extending the seed lists. The seed lists are initialized with zero levels and added to the expansion sets (lines 1–3 for positive seeds and lines 4–6 for negative seeds). The procedure in lines 8 and 9 call procedure Orientation Search twice, one time with the positive seed list and the next time with the negative seed list.

**Table 3** Positive seed list.

Word	Buck Walter's transliteration	English gloss
شجاع	\$jAE	Possessing or displaying courage; able to face and deal with danger or fear without flinching
إبتداع	< bodaAE	The ability to think and act independently
أحب	> aHab~a	Find enjoyable or agreeable
سلى	sal~aY	Provide entertainment for
مبتكر	mubotakir	Someone who creates new things
بارع	baAriE	Having or showing knowledge and skill and aptitude
فرح	fariH	Showing or causing joy and pleasure; especially made happy;
سعيد	saEiyd	Enjoying or showing or marked by joy or pleasure or good fortune
أبيض	> boyaD	Being of the achromatic color of maximum lightness; having little or no hue owing to reflection of almost all incident light
جميل	jamiyl	Delighting the senses or exciting intellectual or emotional admiration

**Table 4** Negative seed list.

Word	Buck Walter's transliteration	English gloss
انحطاط	AinoHiTaAT	A condition inferior to an earlier condition; a gradual falling off from a better state
جهالة	jahaAolap	The trait of acting stupidly or rashly
فحيح	quboH	Qualities that do not give pleasure to the senses
فشل	fa\$al	Loss of ability to function normally
عُدوان	EudowaAn	Violent action that is hostile and usually unprovoked
أخطأ	> axoTa > a	To make a mistake or be incorrect
حزن	Hazana	Feel grief; eat one's heart out
انتقد	Ainotaqada	Find fault with; express criticism of; point out real or perceived flaws
فاسد	faAsid	Corrupt morally or by intemperance or sensuality
مقبت	Maqiyt	Dislike intensely; feel antipathy or aversion toward

Procedure orientation search is presented in Fig. 2. This procedure takes a seed list, the Arabic WordNet database, the sentiment orientation relations by which the seed list is extended, and the expansion sentiment orientation flag. A queue structure was used for the priority expansion, where the adjacent seed list was expanded level by level. The queue was initialized with the seed list in line 1. Then, the front node was iteratively removed from the queue to expand its adjacent

(line 3), the Arabic WordNet database was searched for current node neighbors within the predefined relations (line 4), its depth was incremented by 1 for each unvisited neighbor (line 7). Then, if the relation between the current node and the neighbor had the same orientation relations, we added this node to the same orientation expanded set and to the queue for further expansions. If the relation between the current node and the neighboring node had the opposite orientation, we simply added this node to the opposite orientation expanded set. The procedure was repeated until all reachable nodes were visited.

After the expansion algorithm was completed, the sentiment scores for each synset were calculated using the formula described in Eq. (1):

$$\text{Synset}_{\text{Pos,Neg}} = \sum_{i=1}^n \left( \text{Seed}_{\text{score}_i} - \frac{\text{Synset}_{\text{depth}_i}}{\text{Max}(\text{depth}_{\text{Pos}}, \text{depth}_{\text{Neg}})} \right) \quad (1)$$

Here,

$n$  is the number of synsets that reach the current synset;  $\text{Seed}_{\text{score}}$  is the score of the seed word, which we set to 1;  $\text{depth}_i$  is the synset depth starting from the initial seed;  $\text{depth}_{\text{Pos}}$  is the maximum depth reached by the algorithm by positive orientation; and  $\text{depth}_{\text{Neg}}$  is the maximum depth reached by the algorithm by negative orientation.

By using (1), the score of each synset was decreased as a function of depth from a seed word in each iteration by some

#### ProcedureExpansionAlgorithm

##### Input:

Seed<sub>Pos</sub>: a seed list for the Positive category.  
 Seed<sub>Neg</sub>: a seed list for the Negative category.  
 G: an XML object contains the AWN database.  
 SameOrientationRelations: { 'near\_synonym', 'verb\_group', 'see\_also\_wn15', 'has\_derived', 'related\_to', 'has\_subevent', 'causes' }  
 OppositeOrientationRelations: { 'near\_antonym' }

##### Output:

Expansion<sub>Pos</sub>: the expanded set for the Positive category.  
 Expansion<sub>Neg</sub>: the expanded set for the Negative category.

##### Begin:

```
for each node N in SeedPos do
  N.level ← 0
  add N to the ExpansionPos; mark N as visited by Positive
orientation
for each node N in SeedNeg do
  N.level ← 0
  add N to the ExpansionNeg; mark N as visited by Negative
orientation
R ← SameOrientationRelations . union (OppositeOrientationRelations)
OrientationSearch(SeedPos, G, R, +1)
OrientationSearch(SeedNeg, G, R, -1)
```

##### End

**Figure 1** Expansion algorithm (main procedure).

**Procedure**OrientationSearch**Input:**

Seed: a set of nodes in AWN database to be expanded.  
 G: an XML object contains the AWN database.  
 R: a set of relations by which the nodes are expanded.  
 Orientation : The orientation of the expansion either +1 for Positive expansion or -1 for Negative expansion

**Output:**

Expansion<sub>Pos</sub>: the expanded set for the Positive category.  
 Expansion<sub>Neg</sub>: the expanded set for the Negative category.

**Begin:**

```

queue ← Seed //Initialize a queue with seed's nodes
while the queue is not empty do
  CurrentNode ← remove the front node from the queue
  NeighborNodes ← search G for CurrentNode neighbors which their
  relations in R
  for each node N in NeighborNodes do
    if N is unvisited by Orientation
      N.level ← CurrentNode.level+1
      if Orientation>0
        if N.relation is in the SameOrientation
          add N to the ExpansionPos; mark N as visited by Positive
          orientation
          add N to the queue
        if N.relation is in the OppositeOrientation
          add N to the ExpansionNeg
      else if Orientation<0
        if N.relation is in the SameOrientation
          add N to the ExpansionNeg; mark N as visited by Negative
          orientation
          add N to the queue
        if N.relation is in the OppositeOrientation
          add N to the ExpansionPos

```

End

**Figure 2** Expansion algorithm (expansion procedure).

predefined value (Kim and Hovy, 2004; Godbole et al., 2007). The final score of each synset is the sum of the scores received over all paths. We applied (1) to set the scores (positive and negative) for each synset returned by the expansion algorithm. We set all other non-reachable synsets in the Arabic WordNet as neutral words. At the end, the lexicon contained more than 23,000 terms with a score triplet describing the positive, negative and neutral scores for the term. For summary purposes, we assigned a sentiment orientation to each term in addition to the individual positive, negative and neutral scores. The sentiment orientation was assigned by considering the orientation of the sentiment carrying the highest score for the term. Table 5 displays the number of positive, negative and neutral terms thus obtained and categorized by their respective part of speech. The expanded lexicon was manually analyzed for word sense disambiguation, and all of the collocations and multiple senses of the words in the same part of speech were removed. The results of this operation are displayed in Table 6.

**Table 5** Number of positive, negative and neutral terms categorized by part of speech including collocations.

POS	Positive sentiment	Negative sentiment	Neutral sentiment	Total
Nouns	886	475	14,529	15,890
Verbs	841	523	4720	6084
Adjectives	40	36	1.167	1243
Adverbs	2	7	255	264
Total	1769	1041	20,671	23,481

**Table 6** Number of positive, negative and neutral unigrams categorized by part of speech.

POS	Positive sentiment	Negative sentiment	Neutral sentiment	Total
Nouns	473	281	4596	5350
Verbs	375	301	1047	1723
Adjectives	36	31	400	467
Adverbs	1	3	32	36
Total	885	616	6075	7576

## 5. Experimental evaluation

To evaluate the lexicon, we used a task-based evaluation method whereby the scores from the lexicon were incorporated into the features used for a sentiment polarity classification task. The task was carried on two different Arabic corpora, the OCA corpus (Rushdi-Saleh et al., 2011) and a book review corpus. The OCA is a movie review corpus consisting of 250 positive and 250 negative movie reviews in Arabic. Table 7 displays the statistics of the OCA corpus. The book review corpus was developed by crawling several book review websites and manually annotating each review with its sentiment polarity. Table 8 displays the statistics of the book review corpus. Table 9 displays the source websites used to develop the book review corpus. The corpus was annotated by two native Arabic speakers. The inter-annotator agreement computed as a Kappa statistic was 0.95.

**Table 7** Statistics from the OCA corpus.

	Positive	Negative
Total documents	250	250
Total types	27,595	24,283
Total tokens	121,392	94,556
Avg. tokens in each file	485	378
Total sentences	3137	4881
Avg. sentences in each file	13	20

**Table 8** Statistics from the book review corpus.

	Positive	Negative	Neutral
Total documents	330	330	330
Total types	24,317	12,598	7947
Total tokens	75,389	35,998	17,165
Avg. tokens in each file	228	109	52
Total sentences	6361	2734	1719
Avg. sentences in each file	19	8	5

**Table 9** Distribution of reviews crawled from different web pages.

Web page	Positive	Negative	Neutral
<a href="http://www.goodreads.com">www.goodreads.com</a>	288	326	302
<a href="http://www.reading4arab.com">www.reading4arab.com</a>	37	4	28
<a href="http://roaa.me/blog">http://roaa.me/blog</a>	5	0	0
Total	330	330	330

**Table 10** Sequence matching examples.

Sequence 1	Sequence 2	$M^*$	$T^{**}$	$R^{***}$
اتقن	اتقان	4	9	0.889
مبتهج	مبتهجون	5	12	0.833
فرح	يفرحون	3	9	0.667
سعيد	سعداء	3	9	0.667
فساد	الفساد	4	10	0.80
رب	بر	1	4	0.50

\*  $M$  is the number of matches.

\*\*  $T$  is the total number of elements in both sequences.

\*\*\*  $R = 2^*M/T$  is sequences similarity.

We used a vector space model (Salton et al., 1975) to represent the documents in the corpus. In the vector space model, each document is represented as a vector in an  $n$ -dimensional space, where  $n$  is the total number of terms in the corpus. The result is a  $d \times n$  document term matrix, where  $d$  is the number of documents and  $m$  is the number of terms in the corpus. The document vector in the document term matrix can be represented using different weighting schemes, including binary, term frequency (TF), and term frequency-inverse document frequency (TF\*IDF). To convert the text documents in the corpus into the vector representation, the documents were tokenized, and the terms were normalized using a simple letter normalization scheme in Arabic. No stemming or POS tagging was carried out because stemming would make it difficult to find the terms in the lexicon and because the POS tags for the term can be obtained from the lexicon. For a text categorization task such as sentiment polarity classification, feature selection is an important step to remove irrelevant and noisy features. We removed the univariate features, i.e., the features that occurred only once in each category. Removing the univariate features greatly improved the speed and memory requirement but, removing these features could reduce the classification accuracy because it may remove terms that have a sentiment score available in the lexicon. Therefore, we incorporated the lexicon in removing noisy features, where we kept those terms that occur in the lexicon even though they were found only once.

### 5.1. Term retrieval from lexicon

Instead of using exact matching to match document words with lexicon words, we defined an object of the SequenceMatcher class in Python for comparing pairs of sequences (Ratcliff and Metzner, 1988). This object contains a function, called ratio, that returns a measure of the sequences similarity as a float in the range [0, 1]. The ratio can be computed as  $2M/T$ , where  $T$  is the total number of elements in both the sequences and  $M$  is the number of matches. Note that this is 1.0 if the sequences are identical and 0.0 if they have nothing in common. We set the matching ratio to  $>0.80$  and ordered the returned words by matching ratio. We then fetched the score of the word having the first maximum ratio from the ordered list. Table 10 presents some examples of the sequence-matching process.

We used RapidMiner<sup>1</sup>, a data mining tool, to build the sentiment polarity classification model with two machine learning

<sup>1</sup> [www.rapidminer.com](http://www.rapidminer.com).

**Table 11** Results on book review corpus with NB.

Feature weight	Precision		Recall		Accuracy
	Pos	Neg	Pos	Neg	
Binary	0.9773	0.9202	0.9152	0.9788	<b>0.9488</b>
TF	0.9264	0.9162	0.9152	0.9273	0.9213
TF*IDF	0.9408	0.9174	0.9152	0.9424	0.9291
Score	0.9373	0.9091	0.9061	0.9394	0.9232
Binary Score	0.9525	0.9157	0.9121	0.9545	0.9341

Pos = positive; Neg = negative.

**Table 12** Results on book review corpus with SVM.

Feature weight	Precision		Recall		Accuracy
	Pos	Neg	Pos	Neg	
Binary	0.7750	0.7026	0.6576	0.8091	0.7388
TF	0.7621	0.7335	0.7182	0.7758	0.7478
TF*IDF	0.7733	0.7604	0.7545	0.7788	<b>0.7669</b>
Score	0.7508	0.7208	0.7030	0.7667	0.7358
Binary Score	0.7799	0.7464	0.7303	0.7939	0.7632

Pos = positive; Neg = negative.

**Table 13** Results on OCA corpus with NB.

Feature weight	precision		Recall		Accuracy
	Pos	Neg	Pos	Neg	
Binary	0.9838	0.9723	0.9720	0.9840	<b>0.9781</b>
TF	0.9416	0.9671	0.9680	0.9400	0.9544
TF*IDF	0.9529	0.9714	0.9720	0.9520	0.9622
Score	0.9565	0.9676	0.9680	0.9560	0.9621
Binary Score	0.9603	0.9677	0.9680	0.9600	0.9341

Pos = positive; Neg = negative.

classifiers: Support Vector Machine (SVM) and Naïve Bayes (NB). These classifiers were applied to the document term matrices created from the two aforementioned corpora. For each corpus, five different document term matrices were created, representing five different weighting schemes. These include binary, TF, TF\*IDF, score and binary score. The first three schemes did not include the sentiment scores from the lexicon and served as our baseline. The score-weighting scheme incorporated the unnormalized scores from the lexicon into the TF representation by multiplying the sentiment score from the lexicon with the frequency of the term. The binary score multiplied the score with binary representation of the term, with one indicating the presence and zero indicating the absence of the term in the document.

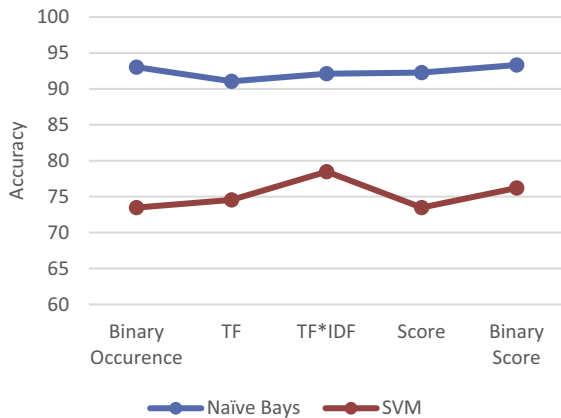
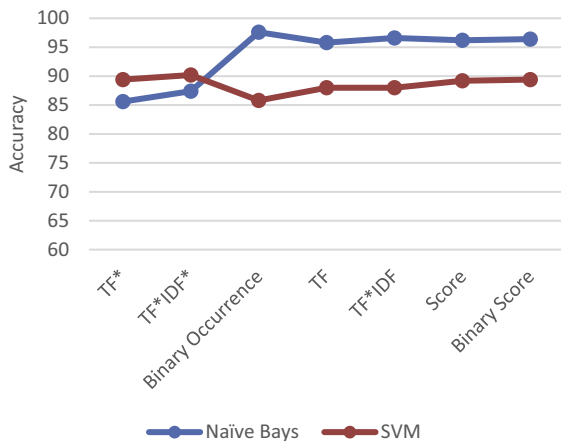
Tables 11 and 12 display the results from applying the Naïve Bayes and the support vector machine classifiers, respectively, on the book review corpus. Tables 13 and 14 display the results from applying the Naïve Bayes and support vector machine classifiers on the OCA corpus. Figs. 3 and 4 plot the same results for the book review and the OCA corpora respectively.

The results show no improvement in the average classification accuracy. One possible explanation could be that incorporating sentiment scores without manipulating other factors, such as position and order of words, may not produce desirable results. Sentiment can be expressed in a subtle manner

**Table 14** Results on OCA corpus with SVM.

Feature weight	Precision		Recall		Accuracy
	Pos	Neg	Pos	Neg	
Binary	0.8571	0.8431	0.8400	0.8600	0.8501
TF	0.8626	0.8992	0.9040	0.8560	0.8809
TF*IDF	0.8593	0.9217	0.9280	0.8480	0.8905
Score	0.8740	0.9118	0.9160	0.8680	<b>0.8929</b>
Binary Score	0.8745	0.9156	0.9200	0.8680	0.8501

Pos = positive; Neg = negative.

**Figure 3** Incorporating word's score using book review corpus.**Figure 4** Incorporating a word's score using movie review corpus. TF\*, TF\*IDF\* are the results produced by the movie review corpus author.

without any ostensible use of negative words. Other factors that make sentiment analysis difficult are that phrases can be expressed with sarcasm, irony, and/or negation.

## 6. Conclusions

In this paper, an Arabic SSL was created with more than 7.5 K terms, with three scores describing the terms being positive, negative or neutral. The created lexicon is context indepen-

dent, and it can be used in any opinion corpora other than book reviews or movie reviews. The SSL was evaluated by incorporating it into a vector space model to apply machine learning classifiers. The experiments show that the accuracy produced by NB is higher than the SVM accuracy. The experiments were conducted on several Arabic sentiment corpora, and we were able to achieve a 97% classification accuracy.

There is still much work that can be performed to develop AWN for sentiment analysis: first, considering different dialect and special regional words; second, considering Franco Arabic and transliteration; and finally, considering compound expressions, phrases and proverbs.

## References

- Alaa El-Halees, 2011. Arabic opinion mining using combined classification approach, the International Arab Conference on Information Technology, pp. 10–13.
- Andrea Esuli, Fabrizio Sebastiani, 2005. Determining the semantic orientation of terms through gloss analysis, Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management, pp. 617–624.
- Andrea Esuli, Fabrizio Sebastiani, 2006. Determining term subjectivity and term orientation for opinion mining, Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics.
- Andrea Esuli, Fabrizio Sebastiani, 2006. SentiWordNet: a publicly available resource for opinion mining, Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), pp. 417–422.
- Christiane Fellbaum et al., 2006. Introducing the Arabic WordNet Project, Proceedings of the 3rd Global Wordnet Conference.
- Fellbaum, Christiane, 1998. *Wordnet, an Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Jaap Kamps, Maarten Marx, Robert J. Mokken, Maarten de Rijke, 2004. Using Wordnet to measure semantic orientation of adjectives, Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, vol. 4, pp. 1115–1118.
- John W. Ratcliff, David Metzener, 1988. Pattern matching: the gestalt approach, Dobb's J., p. 46.
- Muhammad Abdul-Mageed, Mona Diab, 2012. Toward building a large-scale Arabic sentiment lexicon, Proceedings of the 6th International Global WordNet Conference.
- Muhammad Abdul-Mageed, Mohammed Korayem, 2010. Automatic identification of subjectivity in morphologically rich languages: the case of Arabic, Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), pp. 2–6.
- Muhammad Abdul-Mageed, Mohammed Korayem, Ahmed Youssef Agha, 2011. "Yes we can?": subjectivity annotation and tagging for the health domain, Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP, Hissar, Bulgaria.
- Mohamed Elarnaoty, Samir Abdel Rahman, Aly Fahmy, 2012. A machine learning approach for opinion holder extraction Arabic language, in CoRR.
- Mohamed Elhawary, Mohamed Elfeky, 2010. Mining Arabic business reviews, IEEE International Conference on Data Mining Workshops, pp. 1108–1113.
- Mohamed Maamouri, Ann Bies, Tim Buckwalter, Wigdan Mekki, 2004. The Penn Arabic Treebank: building a large-scale annotated arabic corpus, NEMLAR Conference on Arabic Language Resources and Tools, pp. 102–109.
- Mohamed Mahdi Boudabous, Nouha Chaâben Kammoun, Nacef Khedher, Lamia Hadrach Belguith, Fatiha Sadat, 2013. Arabic WordNet semantic relations enrichment through morpho-lexical

- patterns, in *Communications, Signal Processing, and their Applications (ICCSPA)*, 2013 1st International Conference, Sharjah, pp. 1–6.
- Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena, 2007. Large-scale sentiment analysis for news and blogs, *Proceedings of the International Conference on Weblogs and Social Media ICWSM*.
- Noura Farra, Elie Challita, Rawad Abou Assi, Hazem Hajj, 2010. Sentence-level and Document-level Sentiment Mining for Arabic Texts, *ICDM Workshops*, pp. 1114–1119.
- Vasileios Hatzivassiloglou, Kathy McKeown, 1997. Predicting the semantic orientation of adjectives, *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*, pp. 174–181.
- Samhaa R. El-Beltagy, 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.
- Soo-Min Kim, Eduard Hovy, 2004. Determining the sentiment of opinions, *Proceedings of COLING-04, 20th International Conference on Computational Linguistics*, pp. 1367–1373.
- Pang, Bo., Lee, Lillian., 2008. *Opinion mining and sentiment analysis*. *Found. Trends Inf. Retr.* 2 (1–2), 1–135.
- 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.
- Rushdi-Saleh, Mohammed, Martín-Valdivia, María Teresa, Ureña López, Luis Alfonso, Perea-Ortega, José M., 2011. OCA: opinion corpus for Arabic. *J. Am. Soc. Inform. Sci. Technol.* 62 (10), 2045–2054.
- Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. *Commun. ACM* 18 (11), 613–620.
- Thelwall, Mike, Buckley, Kevan, Paltoglou, Georgios, Cai, Di, 2010. Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* 61 (12).
- Theresa Wilson, Janyce Wiebe, Paul Hoffmann, 2005. Recognizing contextual polarity in phrase-level sentiment analysis, *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 347–354.
- Valitutti, Alessandro, Strapparava, Carlo, Stock, Oliviero, 2004. *Developing affective lexical resources*. *Psychology* 2 (1), 61–83.
- WordNet 3.0 database statistics. [Online]. <https://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html#toc>.