OCA: Opinion Corpus for Arabic

Mohammed Rushdi-Saleh, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, and José M. Perea-Ortega
SINAI Research Group, Computer Science Department, University of Jaén, 23071, Spain.
E-mail: msaleh@ujaen.es; maite@ujaen.es; laurena@ujaen.es; jmperea@ujaen.es

Sentiment analysis is a challenging new task related to text mining and natural language processing. Although there are, at present, several studies related to this theme, most of these focus mainly on English texts. The resources available for opinion mining (OM) in other languages are still limited. In this article, we present a new Arabic corpus for the OM task that has been made available to the scientific community for research purposes. The corpus 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. Furthermore, different experiments have been carried out on this corpus, using machine learning algorithms such as support vector machines and Naïve Bayes. The results obtained are very promising and we are encouraged to continue this line of research.

Introduction

The proliferation in the use of the World Wide Web and the rise of blogs and forums have paved the way for increased exposure of individual comments and sentiments. The growth of participation in the Internet fortifies the importance of public opinion as well as the use of public polls for different topics that many websites already employ. These opinions can be about different issues such as electronic products, politics, movies, books, cars, and many others. The idea of processing these comments or reviews has automatically attracted many researchers in the field of text mining, the aim being to be able to extract a general opinion about one item or theme among the huge unstructured data available in the Internet. This new task of analyzing and detecting the orientation of some data is given different names: opinion mining (OM), sentiment analysis, subjectivity analysis, or sentiment orientation.

On the other hand, the rapid growth of e-commerce has increased the number of reviews enormously. Nowadays, it is possible to find a variety of reviews for almost all the products in several merchants websites such as Amazon[1] or CNET[2]. When customers need to purchase laptops, cameras, cars, etc., they usually consult comments about that product and learn from other people's experiences. Summarized opinions could facilitate the task of Internet users and help them make the best choice by giving them a general idea about a product, without the need to explore the crowd data. These opinions are interesting not only for customers but also producers, who can obtain feedback through these reviews to more effectively adapt their products to customers' needs.

The tracking of the many reviews posted on different web pages is a challenging task for researchers. However, although comments in the web are expressed in any language, especially after the explosion of the Web 2.0 and the social web, most research in this field has focused on English texts (Pang & Lee, 2008), mainly because of the lack of resources in other languages. For example, despite the fact that Arabic is one of the top 10 languages most used on the Internet, according to the Internet World State[3] rank (see Figure 1) and is spoken by hundreds of millions of people, there is no reference corpus with sentiments or opinions. This is the main reason that has motivated the generation of an opinion corpus for Arabic in this work.

The Arabic language is becoming very interesting for many researchers in the field of text mining and information retrieval (Ahmed & Nürnberger, 2009; Kanaan, Al-Shalabi, Ghwanmeh, & Al-Ma'adeed, 2009). Several studies have been realized in this context, and there are different corpora, resources, and tools available for testing and implementing applications like text classification (Duwairi, 2006; Duwairi, Al-Refai, & Khasawneh, 2009) or name entity recognition (Shaalan & Raza, 2009). However, Arabic resources that focus on analyzing and mining opinions and sentiments are very difficult to find.

In this article, we present a new opinion corpus for Arabic (OCA) collected from a variety of web pages about movie

Received February 7, 2011; revised May 24, 2011; accepted May 24, 2011

 DOI: 10.1002/asi.21598

[1] http://www.amazon.com
[2] http://www.cnet.com
[3] http://www.internetworldstats.com

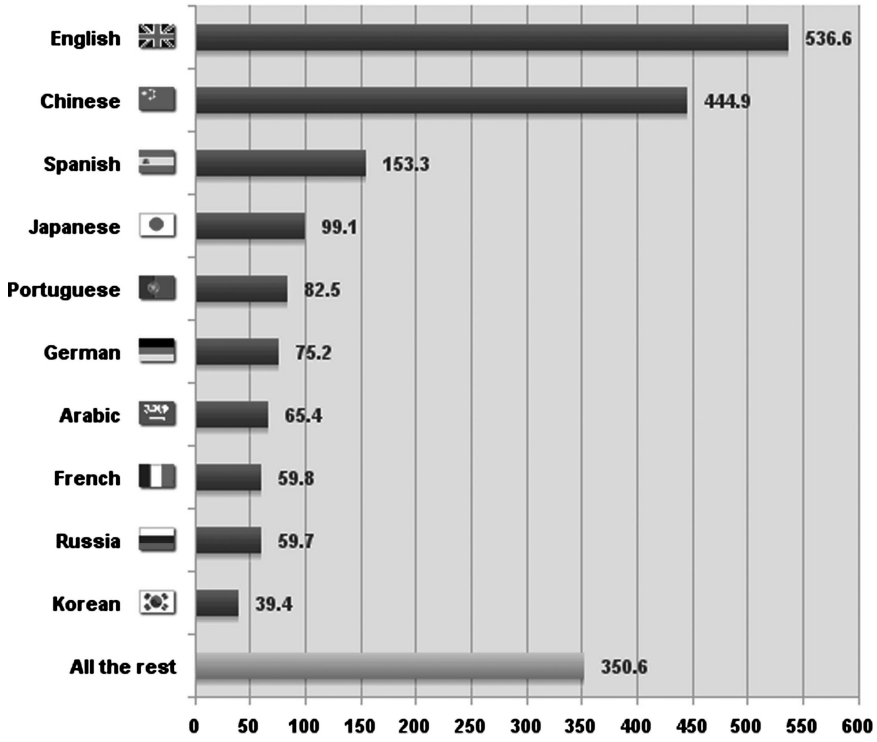

FIG. 1. Top 10 languages on the Internet in 2010 (in millions of users).

reviews in the Arabic language. In addition, we have carried out some experiments on the corpus, using machine learning algorithms to train an opinion classifier. Specifically, we have used the support vector machine (SVM) and Naïve Bayes (NB) algorithms to determine the opinion polarity of the reviews.

Background: Related Work

OM is a discipline that involves several interesting tasks. For example, opinion extraction, a specialization of information extraction, can be considered a specialization of the information extraction task. Its aim is to detect expressions denoting the key components of an opinion within a sentence or document. Another popular OM task focuses on detecting the subjectivity in a document, i.e., whether the document or part of the document is subjective or objective (informative). One of the most widely studied tasks is that of determining the polarity of a document, sentence, or feature (positive or negative) and measuring the degree of the polarity expressed in it. In this article, we train a classifier using SVM to determine whether an Arabic review is positive or negative. Next, we present an overview of the most important research and methods used in this area. In addition, we present a summary of the main work related to OM using non-English languages.

Related Work on Polarity Classification

Different approaches have been applied in the field of polarity or sentiment classification. Two main methodologies can be distinguished in this domain: On the one hand, there is a lot of work based on the semantic orientation approach, which represents the document as a collection of words. Then the sentiment of each word can be determined by different methods, for example, using a web search (Hatzivassiloglou & Wiebe, 2000) or consulting a lexical database like WordNet[4] (Kamps, Marx, Mokken, & Rijke, 2004). On the other hand, machine learning techniques are more extensively used for the classification of reviews. With this approach, the document is represented by different features that may include the use of n-grams or defined grammatical roles like, for instance, adjectives or other linguistic feature combinations, and then a machine learning algorithm is applied. Machine learning algorithms commonly used are SVMs, maximum entropy (ME), or NB.

Regarding methods that consider some linguistic features such as adjectives and adverbs, we can find many studies in the literature (Hatzivassiloglou & McKeown, 1997; Wiebe, 2000; Turney, 2002; Kamps et al., 2004; Hu & Liu, 2004; Ding & Liu, 2007). Another interesting approach is that of Esuli and Sebastiani (2005). They propose a new method based on the assumption that terms with similar orientation tend to have similar glosses. They use a semi-supervised learning algorithm to classify terms as positive or negative. In another study, Ding and Liu improved the previous system proposed by Hu and Liu by assigning a score to opinion words located near the feature. The score depends on the distance between the opinion word and the feature, with a low

[4]http://wordnet.princeton.edu

DOI: 10.1002/asi

score given to the opinion words far from the feature. A common approach in sentiment analysis is to employ supervised machine learning methods to acquire prominent features of sentiment. However, the success of these methods depends on the domain, topic, and time-period represented by the training data.

On the other hand, Pang, Lee, and Vaithyanathan (2002) applied machine learning methods such as NB, ME, and SVM on movie reviews to determine their polarity. The data were downloaded from the Internet Movie Database (IMDb)[5]. They used 700 negative and 700 positive reviews. To apply machine learning algorithms on the documents, the standard bag of features framework was used in this work, predefining a set of features that could appear in a document. They also treated the effect of the negation by adding the negation prefix "*not.*" The word position and the part-of-speech (POS) were also taken into account. They performed several experiments using different n-grams techniques, and the results showed that the use of unigram was the most effective method. In addition, they found that SVM outperforms NB and ME algorithms.

Mullen and Collier (2004) worked on the same dataset used by Pang et al. (2002). They calculated the average rating for the whole collection; the reviews under this average rating were classified as negative and those above the average rating were classified as positives. They investigated several features including various combinations of the Turney value, the three text-wide Osgood values (Osgood, Suci, & Tannenbaum, 1957), word unigrams, or lemmatized unigrams. In addition, they performed experiments on a movie reviews corpus downloaded from the Pitchfork Media[6]. In this case, they extracted the same features and extra features based on the movie domain. The machine learning algorithm used was SVM. They concluded that the combination of unigrams and lemmatized unigrams outperforms the models that do not use this kind of information.

Finally, Prabowo and Thelwall (2009) applied SVM with combined methods to classify reviews from different corpora. One of these datasets was the same as that used by Pang and Lee (2004) and it included 1,000 positive and 1,000 negative samples. Several classifiers were used: General Inquirer Based Classifier (GIBC), Rule-Based Classifier (RBC), Statistics Based Classifier (SBC), and SVM. They accomplished a hybrid classification, whereby if one classifier fails to classify a document, then the classifier passes the document unto the next classifier until the document is correctly classified or no other classifier remains. The results indicated that SBC and SVM improve their effectiveness in the hybrid classification.

Non-English Sentiment Analysis

Most research in OM has focused on English texts, and there is little work using other languages. The main reason for this is the lack of resources oriented to analysis sentiments in other idioms. Generating these resources is very time-consuming and labor-consuming. However, the number of comments, opinions and reviews in all languages is increasing exponentially on the Internet.

According to Mihalcea, Banea, and Wiebe (2007), there are two main approaches in the context of multilingual sentiment analysis:

- Lexicon-based approach, in which a target-language subjectivity classifier is generated by translating an existing lexicon into another idiom.
- Corpus-based approach, in which a subjectivity-annotated corpus for the target language is built through projection, training a statistical classifier on the resulting corpus.

There are some interesting papers that have studied the problem using non-English collections. For example, Denecke (2008) worked on German comments collected from Amazon. These reviews were translated into English using standard machine translation software, and then the translated reviews were classified as positive or negative, using three different classifiers: LingPipe[7], SentiWordNet (Esuli & Sebastiani, 2006b) with classification rule, and SentiWordNet with machine learning. Denecke worked on three different corpora to compare the results:

- The multiperspective question answering (MPQA) corpus[8], in English.
- 1,000 positive and 1,000 negative reviews in English from IMDb.
- 100 positive and 100 negative reviews in German from Amazon.

The experiments carried out for German language were based on translating the reviews into English and then classifying them. They used the IMDb corpus as training data and the dataset translated into English as testing data. Zhang, Zeng, Li, Wang, and Zuo (2009) applied Chinese sentiment analysis on two datasets. In the first one, euthanasia reviews were collected from different websites, while in the second dataset, six product categories were collected from Amazon (Chinese reviews). The euthanasia dataset was manually reviewed and classified into 502 positive and 349 negative articles for training. All the articles were used for testing sentiment analysis approaches, and the standard 10-fold cross-validation was chosen for evaluation. The Amazon dataset was distributed as 310,390 positive and 29,540 negative opinions for the six products. They randomly selected 200 positive and 200 negative reviews for each product to balance the distribution of two classes (positive/negative) for the training dataset. From the remaining comments, 500 positive and 500 negative reviews from each category were randomly selected for testing. The experiments were run using rule-based and machine learning approaches (SVM, NB, and decision tree). Ghorbel and Jacot (2010) used a corpus with movie reviews

[5]http://www.imdb.com
[6]http://www.pitchforkmedia.com
[7]http://alias-i.com/lingpipe
[8]http://www.cs.pitt.edu/mpqa/databaserelease

DOI: 10.1002/asi

in French. They applied a supervised classification combined with SentiWordNet to determinate the polarity of the reviews. Agić, Ljubešić, and Tadić (2010) presented a manually annotated corpus with news on the financial market in Croatia.

Regarding the OM in a multilingual framework using several languages, Ahmad, Cheng, and Almas (2006) performed a local grammar approach for three idioms using financial news: Arabic, Chinese, and English. They selected and compared the distribution of words in a domain-specific document with the distribution of words in a general corpus. Abbasi, Chen, and Salem (2008) accomplished a study for sentiment classification on English and Arabic inappropriate content. Specifically, they applied their methodologies on a U.S. supremacist forum for English and a Middle Eastern extremist group for Arabic language. Boldrini, Balahur, Martínez-Barco, and Montoyo (2009) aimed to build up a corpus with a fine-gained annotation scheme for the detection of subjective elements. The data were collected manually from 300 blogs in three different languages: Spanish, Italian, and English. Text was collected on three different topics, gathering 100 texts for each topic, with a total of 30,000 words approximately for each language.

OCA

In this article, we present OCA, a new Arabic resource made available to the scientific community that can be used in sentiment analysis[9]. First, we explain the difficulty of finding Arabic opinions because of the lack of websites that include reviews and comments using this language. Second, the process followed to generate the OCA corpus is expounded.

Difficulty in Arabic Websites

Despite the importance of the Arabic language on the Internet, there are very few web pages that specialize in Arabic reviews. In fact, our first attempt to build an Arabic corpus aimed at obtaining opinions for typical objects such as electronic products or cars, but, unfortunately, we had little success because of the lack of websites like Amazon or Booking[10] using Arabic. The most common Arabic opinion sites on the Internet are related to movies and films, although these blogs also present several obstacles to their being used in sentiment analysis tasks. Some of these difficulties are stated below:

- Nonsense and nonrelated comments. Many reviews in different web pages are not related to the topic. People attempt to comment on anything, even with unrelated words or nonsense. For instance, instead of comment an item, the user just types a word:

  Thaaaaaaanks = مشكووووور

TABLE 1. Different variants of Roman alphabet transcriptions.

English	*Qatar is a great country*
Arabic	قطر دولة عظيمة
Roman alphabet 1	*Qatar dawla athema*
Roman alphabet 2	*Qatr dawlah 3 athema*
Roman alphabet 3	*9atar dawlah 3 athemah*

- Romanization of Arabic. Many comments use the Roman alphabet. Each phoneme in Arabic can be replaced by its counterpart in the Roman alphabet. This can be because of nonuse of Arabic keyboards for people who comment on Arabic topics from abroad. For instance, Table 1 shows a fragment explaining the problem of commenting on a topic using the Roman alphabet. There are also possible variants in the case of Romanization of Arabic for the above example, taking into account the diacritics in the Arabic language. However, a native speaker could still understand this sentence.
- Comments in different languages. It is also possible to find international languages in Arabic web pages, so you could read comments in English, Spanish, or French mixed with Arabic sentences.

Corpus Generation

To generate the OCA we have extracted the reviews from different web pages about movies. OCA comprises 500 reviews in Arabic, of which 250 are considered as positive reviews and the other 250 as negative opinions. This process involved collecting reviews from several Arabic blog sites and web pages using a simple bash script for crawling. Then, we removed HTML tags and special characters, and spelling mistakes were corrected manually. Next, a processing of each review was carried out, which involved tokenizing, removing Arabic stop words, and stemming and filtering those tokens whose length was less than two characters. Specifically, we have used the Arabic stemmer from the Rapid Miner[11] software. Rapid Miner includes two implementations of Arabic stemming: the basic Arabic stemmer, which is based on Khoja Arabic stemmer (Khoja & Garside, 1999), and the light Arabic stemmer developed by Larkey, Ballesteros, and Connell (2007). In our experiments, we have used only the basic Arabic stemmer of Rapid Miner and the Arabic stop word list provided by the same software. Finally, three different n-gram schemes are generated (unigrams, bigrams, and trigrams) and cross validation is applied to evaluate the corpus. Figure 2 shows the different steps followed in our approach. Table 2 shows an example of generation of unigram, bigrams, and trigrams for a fragment from an original review of the OCA corpus, using the Rapid Miner software and removing the stop words previously with the same tool.

Table 3 presents the number of reviews according to negative or positive classification from each web page, the name of the web page, and the highest score used in the rating system. On the other hand, Figure 3 shows an excerpt from a

[9]The OCA corpus is freely available at the SINAI website http://sinai.ujaen.es/wiki/index.php/OCA_Corpus_(English_version)

[10]http://www.booking.com

[11]http://rapid-i.com

DOI: 10.1002/asi

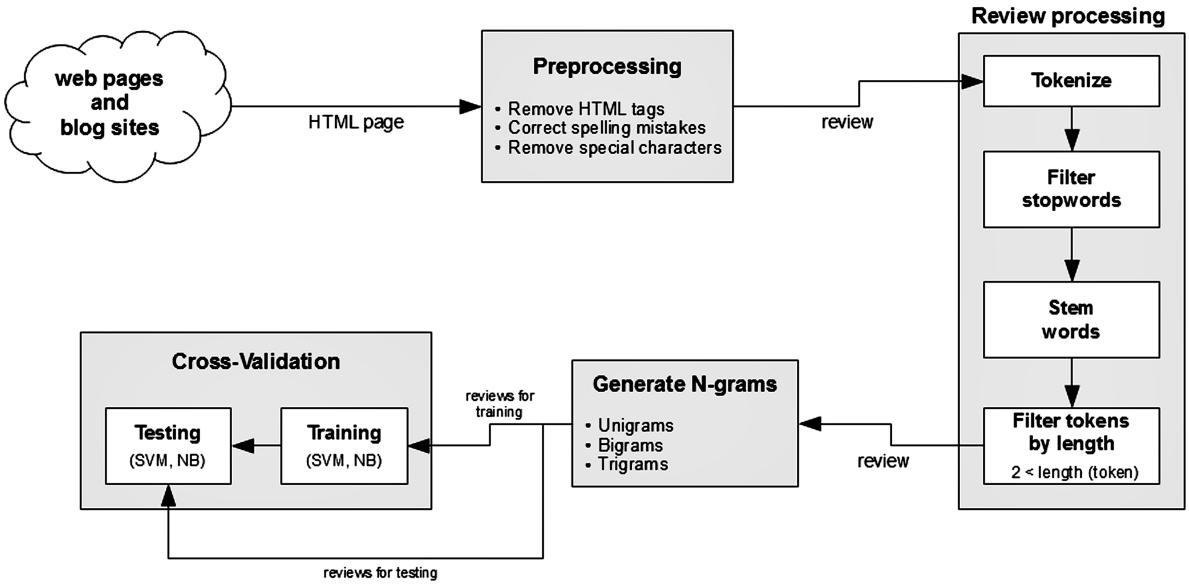

FIG. 2. Steps followed in the generation and validation of the OCA corpus.

TABLE 2. Examples of generation of unigram, bigrams, and trigrams for a fragment from an original review of the opinion corpus for Arabic.

Fragment from an original review	أداء الممثلين كان رائعا.. من قامت بدور هايدي تميزت جدا وأبدعت، ومن قامت بدور دون أبدعت أيضا، مع أنه لوحظ على أدائها التكلف، لكنه هامشي جدا إذا ما قورن بعمر الطفلة.
Unigram	أداء الممثلين رائعا قامت بدور هايدي تميزت وأبدعت قامت بدور أبدعت لوحظ أدائها التكلف هامشي قورن بعمر الطفلة.
Bigrams	وأساسها_أداء أداء أداء_الممثلين الممثلين الممثلين_رائعا رائعا رائعا_قامت قامت قامت_بدور بدور بدور_هايدي هايدي هايدي_تميزت تميزت تميزت_وأبدعت وأبدعت وأبدعت_قامت قامت قامت_بدور بدور بدور_أبدعت أبدعت أبدعت_لوحظ لوحظ لوحظ_أدائها أدائها أدائها_التكلف التكلف التكلف_هامشي هامشي هامشي_قورن قورن قورن_بعمر بعمر بعمر_الطفلة الطفلة الطفلة_الموسيقى.
Trigrams	العائلة_وأساسها العائلة_وأساسها_أداء وأساسها وأساسها_أداء وأساسها_أداء_الممثلين أداء أداء_الممثلين أداء_الممثلين_رائعا الممثلين الممثلين_رائعا الممثلين_رائعا_قامت رائعا رائعا_قامت رائعا_قامت_بدور قامت قامت_بدور قامت_بدور_هايدي بدور بدور_هايدي بدور_هايدي_تميزت هايدي هايدي_تميزت هايدي_تميزت_وأبدعت تميزت تميزت_وأبدعت تميزت_وأبدعت_قامت وأبدعت وأبدعت_قامت وأبدعت_قامت_بدور قامت قامت_بدور_أبدعت بدور بدور_أبدعت بدور_أبدعت_لوحظ أبدعت أبدعت_لوحظ_أدائها لوحظ لوحظ_أدائها لوحظ_أدائها_التكلف أدائها أدائها_التكلف_هامشي التكلف التكلف_هامشي التكلف_هامشي_قورن هامشي هامشي_قورن_بعمر قورن قورن_بعمر قورن_بعمر_الطفلة بعمر بعمر_الطفلة بعمر_الطفلة_الموسيقى الطفلة الطفلة_الموسيقى الطفلة_الموسيقى_والموسيقى.

DOI: 10.1002/asi

TABLE 3. Distribution of reviews crawled from different web pages.

	Name	Web page	Rating system	Positive reviews	Negative reviews
1	Cinema Al Rasid	http://cinema.al-rasid.com	10	36	1
2	Film Reader	http://filmreader.blogspot.com	5	0	92
3	Hot Movie Reviews	http://hotmoviews.blogspot.com	5	45	4
4	Elcinema	http://www.elcinema.com	10	0	56
5	Grind House	http://grindh.com	10	38	0
6	Mzyondubai	http://www.mzyondubai.com	10	0	15
7	Aflamee	http://aflamee.com	5	0	1
8	Grind Film	http://grindfilm.blogspot.com	10	0	8
9	Cinema Gate	http://www.cingate.net	bad/good	0	1
10	Emad Ozery Blog	http://emadozery.blogspot.com	10	0	1
11	Fil Fan	http://www.filfan.com	5	81	20
12	Sport4Ever	http://sport4ever.maktoob.com	10	0	1
13	DVD4ArabPos	http://dvd4arab.maktoob.com	10	11	0
14	Gamraii	http://www.gamraii.com	10	39	0
15	Shadows and Phantoms	http://shadowsandphantoms.blogspot.com	10	0	50
Total				250	250

ليس هناك الكثير من الإهتمام الذي يستطيع الفيلم بثّه لمشاهديه. إنه مثل مقال حول خطاب مهم مكتوب بلغة لا تجسّد تلك الأهمية. مليء بالمشاهد الدالّة لكنها غير المؤثرة خصوصاً وأن الفيلم لا يُشيد زوجين سعيدين فعلاً من البداية، فيبقى الحب بينهما مسألة نظرية او افتراضية.

FIG. 3. Example of an excerpt from a comment of the OCA corpus.

comment of the OCA corpus, which could be translated as follows:

> There is not much of interest in the film, which can be broadcasted for viewers. It is like an article on an important speech written in a language that does not reflect that importance. The movie is filled with scenes, but it is not influential, especially since the film does not describe a happy couple from the beginning, and love remains between them a theoretical or hypothetical issue.

The selection of the web pages was based on the quality of the language used, because many sites use slang, making understanding difficult for many Arabic speakers. Most of Arabic dialects can be understood in different Arabic countries except some specific cases such as some Moroccan dialects. Therefore, for generating the OCA corpus, we have used the reviews provided by the web pages shown in Table 3, without discarding or filtering any comment from them. However, previously, we carried out an in-depth analysis of these blogs to ensure that the dialects used in all comments were understandable by Arabic native speakers. On the other hand, there are important issues that must be taken into account in these blogs:

- Rating system. We found that there is no common system of rating among these blogs. Some of them use a rating scale of 10 points, so reviews with less than five points are classified as negative, while those with a rating between 5 and 10 points are classified as positive. Other blogs use a 5-rating scale. In these cases, we considered the movies with three, four and five points as positive, while those with less than three points were classified as negative. This classification was based on a deep study of the reviews that were rated as neutral. Finally, we also found binary classifications such as *good* or *bad*.
- Cultural and political emotions. We noticed that the culture in Arabic countries could also affect the behavior of the reviewers. For instance, an "Antichrist" movie is rated with 1 point out of 10 in one of the Arabic blogs (clearly, a negative opinion), while the same movie on the IMDb is rated at 6.7 out of 10.
- Movie and actor names in English. There are different ways of naming movies and actors in the reviews. In some cases, the names are translated into Arabic, while others keep the names in English and the reviews in Arabic.

Finally, another important factor in preparing this corpus was the richness of the text. We tried to select reviews that have more tokens than short text reviews. Table 4 shows some statistics on the OCA corpus.

Experimental Study Using OCA

Several experiments have been accomplished to evaluate the OCA corpus. We have used cross-validation to compare the performance of two of the most widely used learning algorithms: SVM and NB. Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a

DOI: 10.1002/asi

TABLE 4. Statistics on the opinion corpus for Arabic.

	Negative	Positive
Total documents	250	250
Total tokens	94,556	121,392
Avg. tokens in each file	378	485
Total sentences	4,881	3,137
Avg. sentences in each file	20	13

model and the other used to validate the model (Manning & Schutze, 1999). The basic form of cross-validation is k-fold cross-validation. In k-fold cross-validation, the data are first partitioned into k equally sized segments or folds. Subsequently, k iterations of training and validation are performed so that within each iteration a different fold of the data is held out for validation, while the remaining k-1 folds are used for learning. In our experiments, the 10-fold cross-validation (k=10) has been used to evaluate the classifiers.

On the other hand, evaluation has been carried out on three main measures: precision (P), recall (R), and accuracy (Acc):

$$precision(P) = \frac{TP}{TP + FP}$$

$$recall(R) = \frac{TP}{TP + FN}$$

$$accuracy(Acc) = \frac{TP + TN}{TP + FP + FN + TN}$$

where TP (true positives) are those assessments in which system and human expert agree for a label assignment, FP (false positives) are those labels assigned by the system that does not agree with the expert assignment, FN (false negatives) are those labels that the system failed to assign as they were given by the human expert, and TN (true negatives) are those nonassigned labels that were also discarded by the expert. The precision tells us how well the labels are assigned by our system (the fraction of assigned labels that are correct). The recall measures the fraction of expert labels found by the system. Finally, accuracy combines both precision and recall, calculating the proportion of true results (both true positives and true negatives; Sebastiani, 2002).

Machine Learning Algorithms

In our experiments, we used two different machine learning algorithms: NB and SVM.

NB is a method of classification based on the Bayes theorem. The major idea of the NB is to use the assumption that predictor variables are independent random variables. This assumption makes it possible to compute probabilities required by the Bayes formula from a relatively small training set. Despite its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, NB-based text categorization still tends to perform surprisingly well (Lewis, 1998). Indeed, Pazzani and Domingos (1997) show that NB is optimal for certain problem classes with highly dependent features. Esuli and Sebastiani (2006a) used NB to determine term subjectivity and term orientation for OM. They also applied other learning algorithms such as SVM or Rocchio, but better results were obtained using NB.

On the other hand, SVM have been shown to be highly effective in traditional text categorization, generally outperforming NB (Joachims, 1998). SVM have been applied successfully in many text classification tasks because of their principal advantages: First, they are robust in high dimensional spaces; second, any feature is relevant; third, they are robust when there is a sparse set of samples; and, finally, most text categorization problems are linearly separable. In addition, SVM have achieved good results in OM and this algorithm has overcome other machine learning techniques (O'Keefe & Koprinska, 2009).

Experiments and Results

For the experiments, we used the Rapid Miner[11] software with its text mining plug-in, which contains different tools designed to assist in the preparation of text documents for mining tasks (tokenization, stop word removal, and stemming, among others). Rapid Miner is an environment for machine learning and data mining processes that includes a cross-validation process to estimate the performance of several learning operators such as SVM or NB. As mentioned above, the 10-fold cross-validation was used to test the classifiers. We applied the Arabic stemming algorithm included in Rapid Miner to reduce words to their common root or stem. The Arabic stop words list included in Rapid Miner was also applied to the texts of the corpus to remove those words without relevant meaning.

On the other hand, a study of different n-gram schemes was also carried out to analyze its influence on the corpus generated. For this reason, we applied several n-gram models (unigram, bigrams, and trigrams) for each learning algorithm in the cross-validation process. In addition, we have evaluated the use of two different weighting schemes in the validation process: tf–idf (term frequency–inverse document frequency) and tf (term frequency). These schemes are often used in information retrieval and text mining. The impact of using stemming in the text preprocessing was also analyzed. Therefore, a total of 24 experiments were carried out on OCA corpus, 12 experiments using tf–idf as weighting scheme and the other ones using tf:

- Unigram, bigrams, and trigrams using SVM or NB as learning algorithms with stemmer,
- Unigram, bigrams, and trigrams using SVM or NB as learning algorithm without stemmer.

Table 5 and Table 6 show the results obtained in the validation process using tf–idf and tf weighting schemes respectively. Comparing the two learning algorithms used in the cross-validation process, SVM slightly improves on the performance of NB. The improvement between the best

DOI: 10.1002/asi

TABLE 5. Ten-fold cross-validation results using term frequency–inverse document frequency as weighting scheme.

		Precision		Recall		Accuracy	
n-gram model	Stemming	SVM	NB	SVM	NB	SVM	NB
Unigram	Yes	0.8614	0.8106	0.8800	0.8880	0.8680	0.8380
	No	0.8699	0.8274	0.9480	**0.9520**	0.9020	0.8740
Bigrams	Yes	0.8685	0.8353	0.9080	0.9040	0.8840	0.8600
	No	0.8738	0.8525	0.9520	0.9480	0.9060	0.8900
Trigrams	Yes	0.8721	0.8361	0.9120	0.9080	0.8880	0.8620
	No	**0.8738**	**0.8525**	**0.9520**	0.9480	**0.9060**	**0.8900**

Note. SVM = support vector machine; NB = Naïve Bayes.

TABLE 6. Ten-fold cross-validation results using term frequency as weighting scheme.

		Precision		Recall		Accuracy	
n-gram model	Stemming	SVM	NB	SVM	NB	SVM	NB
Unigram	Yes	0.8701	0.7999	0.9440	0.8560	0.9000	0.8180
	No	0.8690	0.8104	0.9320	**0.9360**	0.8940	0.8560
Bigrams	Yes	0.8710	0.8275	0.9520	0.8880	0.9040	0.8460
	No	0.8690	0.8404	0.9320	0.9240	0.8940	0.8720
Trigrams	Yes	**0.8710**	0.8275	**0.9520**	0.8880	**0.9040**	0.8460
	No	0.8535	**0.8434**	0.9360	0.9240	0.8860	**0.8740**

Note. SVM = support vector machine; NB = Naïve Bayes.

accuracy results of both models is 1.8% for SVM using tf–idf as weighting scheme and 3.43% using tf. This behavior is similar to that obtained by Pang et al. (2002). Regarding the n-gram model, we can note clearly that trigram and bigram models overcome the unigram model. According to the SVM results, it should be noted that for bigram and trigram models there are no differences using stemming and the tf weighting scheme. Identical behavior is observed when we use tf–idf but without applying stemming. The use of a stemmer in the preprocessing phase will depend on the weighting scheme used. For tf–idf, it is clear that the best solution is not to stem the words. However, for tf, it depends on the learning algorithm selected. If we use SVM, we will always achieve better results by applying stemming, while if we use NB, then the best option is not to use stemming. Finally, the comparison between both weighting schemes is not relevant. tf–idf slightly improves the best result achieved by tf regarding accuracy measure (0.22%). On the other hand, the high values obtained for accuracy during the validation process show the good quality of the corpus proposed (0.90 using both weighting schemes and SVM with trigram model).

According to Kanaan et al. (2009), the results of applying different text classification techniques using Arabic language are comparable to the results obtained for English and other languages. To contrast the results obtained with OCA, we have compared them with similar experiments using the corpus generated by Pang et al. (2002). This corpus is also a collection of 1,400 samples (700 positive and 700 negative) of movie reviews. Table 7 shows the results obtained with Pang's corpus using 10-fold cross-validation and SVM, compared with our best results obtained with OCA using tf–idf, SVM and without applying stemmer in the preprocessing phase.

TABLE 7. Pang corpus 10-fold cross-validation results compared to OCA corpus best results (using tf–idf, SVM, and without stemming).

Corpus	n-gram model	Precision	Recall	Accuracy
Pang	Unigram	0.8493	0.8390	0.8445
	Bigrams	0.8583	0.8450	0.8515
	Trigrams	0.8619	0.8450	0.8535
OCA	Unigram	0.8699	0.9480	0.9020
	Bigrams	0.8738	0.9520	0.9060
	Trigrams	**0.8738**	**0.9520**	**0.9060**

Note. OCF = opinion corpus for Arabic; tf–idf = term frequency–inverse document frequency; SVM = support vector machine.

Analyzing the best results obtained with both corpus, related to the accuracy measure and 10-fold cross-validation, we can observe that the best result (0.90) using SVM over the OCA improves on the best result obtained with the Pang corpus (0.8535), using trigrams to generate the word vectors. This improvement is 5.45%. Moreover, it should be noted that for both corpora, the use of the trigram and bigram models overcomes the use of unigram model.

Conclusions and Further Work

In this work, we have generated a new Arabic corpus for predicting sentiment polarity. Nowadays, it is difficult to find a corpus designed for implementing sentiment analysis application and, more specifically, for the Arabic language.

DOI: 10.1002/asi

Few blogs are oriented to expressing opinions in Arabic. Finding web pages in Arabic about topics such as electronic products, books, or cars is almost impossible. The data for the proposed corpus were collected from several blogs of movies reviews, obtaining a total of 500 comments (250 positive and 250 negative). Some experiments were also carried out on the proposed corpus to evaluate classifiers trained for determining the polarity of a review. The results obtained were very promising.

For further work, we will continue in this line of research by improving our corpus using techniques such as enlarging or fine-grained annotation. Moreover, we will focus on some linguistic features (adjectives, nouns, etc.) using WordNet for Arabic along with English resources like SentiWordNet. Furthermore, it would be worthwhile to translate this corpus into English using standard machine translation software and evaluate it with SVM and NB to analyze the results.

Acknowledgments

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), project TEXT-COOL 2.0 (TIN2009-13391-C04-02) from the Spanish Government, a grant from the Andalusian Government, project GeOasis (P08-TIC-41999) and Geocaching Urbano research project (RFC/IEG2010). Also, another part of this project was funded by Agencia Española de Cooperación Internacional para el Desarrollo MAEC-AECID.

References

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems, 26(3).

Agic, Z., Ljubešic, N., & Tadić, M. (2010). Towards sentiment analysis of financial texts in croatian. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, ... D. Tapias (Eds.), Language resources and evaluation (LREC). Paris: European Language Resources Association.

Ahmad, K., Cheng, D., & Almas, Y. (2006). Multi-lingual sentiment analysis of financial news streams. Proceedings of Science (GRID '06). Retrieved from http://cdsweb.cern.ch/record/964964/files/001GRID2006_001.pdf

Ahmed, F., & Nürnberger, A. (2009). Evaluation of n-gram conflation approaches for Arabic text retrieval. Journal of the American Society for Information Science and Technology, 9(2), 1448–1465.

Boldrini, E., Balahur, A., Martínez-Barco, P., & Montoyo, A. (2009). Emotiblog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. In R. Stahlbock, S.F. Crone & S. Lessmann (Eds.), DMIN (pp. 491–497). Las Vegas, NV: CSREA Press.

Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. ICDE Workshops (pp. 507–512). Washington, DC: IEEE Computer Society.

Ding, X., & Liu, B. (2007). The utility of linguistic rules in opinion mining. In W. Kraaij, A.P. de Vries, C.L.A. Clarke, N. Fuhr, & N. Kando (Eds.), Proceedings of the 30th International Conference on Research and Development in Information Retrieval (ACM SIGIR '07) (pp. 811–812). New York: ACM Press.

Duwairi, R.M. (2006). Machine learning for Arabic text categorization. Journal of the American Society for Information Science and Technology, 57(8), 1005–1010.

Duwairi, R., Al-Refai, M.N., & Khasawneh, N. (2009). Feature reduction techniques for Arabic text categorization. Journal of the American Society for Information Science and Technology, 60(11), 2347–2352.

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. In O. Herzog, H. Schek, N. Fuhr, A. Chowdhury, & W. Teiken (Eds.), Proceedings of the 14th ACM International Conference on Information and Knowledge Management (pp. 617–624). New York: ACM Press.

Esuli, A., & Sebastiani, F. (2006a). Determining term subjectivity and term orientation for opinion mining. In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL '06) (pp. 193–200). East Stroudsburg, PA: Association for Computational Linguistics.

Esuli, A., & Sebastiani, F. (2006b). SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC '06) (pp. 417–422). Paris: European Language Resources Association (ELRA).

Ghorbel, H., & Jacot, D. (2010, June). Sentiment analysis of French movie reviews. Paper presented at the Fourth international Workshop on Distributed Agent-based Retrieval Tools (DART '10), Geneva, Switzerland.

Hatzivassiloglou, V., & McKeown, K.R. (1997). Predicting the semantic orientation of adjectives. In Proceedings of the Joint ACL/EACL Conference (pp. 174–181). Morristown, NJ: Association for Computational Linguistics.

Hatzivassiloglou, V., & Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING '00) (pp. 299–305). Morristown, NJ: Association for Computational Linguistics.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 168–177). New York: ACM Press.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. European Conference on Machine Learning (ECML '98) (pp. 137–142). London: Springer-Verlag.

Kamps, J., Marx, M., Mokken, R.J., & Rijke, M.D. (2004). Using WordNet to measure semantic orientation of adjectives. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC '04) (pp. 1115–1118). Paris: European Language Resource Association.

Kanaan, G., Al-Shalabi, R., Ghwanmeh, S.H., & Al-Ma'adeed, H. (2009). A comparison of text-classification techniques applied to Arabic text. Journal of the American Society for Information Science and Technology, 60(9), 1836–1844.

Khoja, S., & Garside, R. (1999). Stemming Arabic text (Tech. rep.). Computer Department, Lancaster University, Lancaster.

Larkey, L., Ballesteros, L., & Connell, M. (2007). Light stemming for Arabic information retrieval. In A. Soudi, A. Van den Bosch, & G. Neumann (Eds.), Arabic computational morphology (Vol. 38, pp. 221–243). Heidelberg, Germany: Springer.

Lewis, D.D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nedellec & C. Rouveirol (Eds.), Proceedings of the 10th European Conference on Machine Learning (ECML '98) (pp. 4–15). Heidelberg, Germany Springer-Verlag.

Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. Cambridge, MA: MIT Press.

Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In Proceedings of the Association for Computational Linguistics (ACL '07) (pp. 976–983). East Stroudsburg, PA: Association for Computational Linguistics.

Mullen, T., & Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '04) (pp. 412–418). Morristown, NJ: Association for Computational Linguistics.

O'Keefe, T., & Koprinska, I. (2009, December). Feature selection and weighting methods in sentiment analysis. Paper presented at the 14th Australasian Document Computing Symposium, Sydney, Australia.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). The measurement of meaning. Urbana, IL: University of Illinois Press.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings

DOI: 10.1002/asi

of the 40th Annual Meeting on Association for Computational Linguistics (pp. 271–278). Morristown, NJ: Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), pp. 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '02) (pp. 79–86). Morristown, NJ: Association for Computational Linguistics.

Pazzani, M., & Domingos, P. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29 (2–3), 103–130.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2), 143–157.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1.

Shaalan, K.F., & Raza, H. (2009). NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 60(8), 1652–1663.

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 (ACL '02) (pp. 417–424). Morristown, NJ: Association for Computational Linguistics.

Wiebe, J. (2000). Learning subjective adjectives from corpora. Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence (AAAI '00) (pp. 735–740). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Zhang, C., Zeng, D., Li, J., Wang, F.-Y., & Zuo, W. (2009). Sentiment analysis of Chinese documents: From sentence to document level. Journal of the American Society for Information Science and Technology, 60(12), 2474–2487.

DOI: 10.1002/asi