Bilingual Experiments with an Arabic-English Corpus for Opinion Mining

Mohammed Rushdi-Saleh
SINAI research group
University of Jaén
msaleh@ujaen.es

M. Teresa Martín-Valdivia
SINAI research group
University of Jaén
maite@ujaen.es

L. Alfonso Ureña-López
SINAI research group
University of Jaén
laurena@ujaen.es

José M. Perea-Ortega
SINAI research group
University of Jaén
jmperea@ujaen.es

Abstract

Recently, Opinion Mining (OM) is receiving more attention due to the abundance of forums, blogs, e-commerce web sites, news reports and additional web sources where people tend to express their opinions. There are a number of works about Sentiment Analysis (SA) studying the task of identifying the polarity, whether the opinion expressed in a text is positive or negative about a given topic. However, most of research is focused on English texts and there are very few resources for other languages. In this work we present an Opinion Corpus for Arabic (OCA) composed of Arabic reviews extracted from specialized web pages related to movies and films using this language. Moreover, we have translated the OCA corpus into English, generating the EVOCA corpus (English Version of OCA). In the experiments carried out in this work we have used different machine learning algorithms to classify the polarity in these corpora showing that, although the experiments with EVOCA are worse than OCA, the results are comparable with other English experiments, since the loss of precision due to the translation is very slight.

1. Introduction

Nowadays, the interest in Opinion Mining (OM) has grown significantly due to different factors. On the one hand, the rapid evolution of the World Wide Web has changed our view of the Internet. It has turned into a collaborative framework where technological and social trends come together, resulting in the over exploited term Web 2.0. On the other hand, the tremendous use of e-commerce services has been accompanied by an increase in freely available online reviews and opinions about products and services. A customer who wants to buy a product usually searches information on the Internet trying to find other consumer analyses. In fact, web sites such as Amazon[1], Epinions[2] or IMDb[3], can affect the customer decision. Moreover, the automatic Sentiment Analysis (SA) is useful not only for individual customer but also for any company or institution. However, the huge amount of information makes necessary to accomplish new methods and strategies to tackle the problem.

Thus, SA is becoming one of the main research areas that combines Natural Language Processing (NLP) and Text Mining (TM). This new discipline attempts to identify and analyze opinions and emotions. It includes several subtasks such as subjectivity detection, polarity classification, review summarization, humor detection, emotion classification, sentiment transfer, and so on [9]. However, most of works related to OM are oriented to use English language. Perhaps due to the novelty of the task, there are very few papers analyzing the opinions using other languages different to English. In this paper, we present the experiments accomplished with an Opinion Corpus for Arabic (OCA) collected from different web pages with comments about movies. In addition, we have used automatic machine translation tools to translate OCA corpus into English. We have generated different classifiers using Support Vector Machine and Naïve Bayes in order to determinate the polarity of the opinions. The experiments carried out with the English Version of OCA (EVOCA) show that, although we lost precision in the translation, the results are comparable to other works using English texts. So, we can use this procedure in order to determine the polarity of an Arabic corpus by using English translation. This is important because most of resources are in English and we can take advantage of this situation.

The paper is organized as following: Next section presents some papers about OM using non-English language. Section 3 and Section 4 describe the OCA

[1] http://www.amazon.com
[2] http://www.epinions.com
[3] http://www.imdb.com

Proceedings of Recent Advances in Natural Language Processing, pages 740–745,
Hissar, Bulgaria, 12-14 September 2011.

corpus and its English version (EVOCA), respectively. In Section 5, accomplished experiments are showed and results are analyzed. Finally, conclusion and future work is presented.

2. Related works

Although opinions and comments in the Internet are expressed in any language, most of research in OM is focused on English texts. However, languages such as Chinese, Spanish or Arabic, are ever more present on the web[4]. Thus, it is important to develop resources for helping researcher to work with these languages.

There are some interesting papers that have studied the problem using non-English collections. For example, Denecke [5] worked on German comments collected from Amazon. These reviews were translated into English using standard machine translation software. Then the translated reviews were classified as positive or negative, using three different classifiers: LingPipe7, SentiWordNet [6] with classification rule, and SentiWordNet with machine learning.

Zhang et al. [12] applied Chinese sentiment analysis on two datasets. In the first one euthanasia reviews were collected from different web sites, while the second dataset was about six product categories collected from Amazon (Chinese reviews).
Ghorbel and Jacot [7] used a corpus with movie reviews in French. They applied a supervised classification combined with SentiWordNet in order to determinate the polarity of the reviews.

Agić et al. [2] presented a manually annotated corpus with news on the financial market in Croatia.
Boldrini et al. [4] 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.

Regarding opinion mining for Arabic language, Ahmad et al. [3] performed a local grammar approach for three languages: Arabic, Chinese and English using financial news. They selected and compared the distribution of words in a domain-specific document to the distribution of words in a general corpus.

Finally, Abbasi et al. [1] 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.

3. OCA: Opinion Corpus for Arabic

Despite the importance of the Arabic language on the Internet, there are very few web pages which specialize in Arabic reviews. The most common Arabic opinion sites in 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 non related 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= مشكووووووور

- **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 due to non-use 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.

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 alphabet2	*Qatr dawlah 3athema*
Roman alphabet3	*9atar dawlah 3athemah*

- **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.

In order to generate the Opinion Corpus for Arabic we have extracted the reviews from different web pages about movies. OCA consists of 500 reviews in Arabic, of which 250 are considered as positive reviews and the other 250 as negative opinions. This process has consisted of collecting reviews from several Arabic blog sites and web pages. Table 2 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.

[4] http://www.internetworldstats.com

Table 2. Distribution of reviews crawled from different web pages

	Name	web page	Rating system	PR	NR
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	Mzyon-dubai	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

We have removed HTML tags and special characters as well as spelling mistakes were corrected manually. Next, a processing of each review was carried out which consisted of tokenizing, removing Arabic stop words, stemming and filtering those tokens whose length was less than two characters. Figure 1 shows the different steps followed in our approach in order to generate the OCA corpus and Table 3 shows some statistics on such corpus.

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 five 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 which were rated as neutral. Finally, we also found binary classifications such as *good* or *bad.*

Table 3. Statistics on the OCA opinion corpus

	Negative	Positive
Total **documents**	250	250
Total **tokens**	94,556	121,392
Total **sentences**	4,881	3,137

- **Cultural and political emotions**. Culture in Arabic countries can also affect the behavior of the reviewers. For instance, an "Antichrist" movie is rated with 1 point from 10 in one of the Arabic blogs, while the same movie on 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.

4. EVOCA: English Version of OCA

In order to compare the experiment for Arabic and English, we have translated OCA into English using an automatic Machine Translation (MT) tool freely available. Specifically, we have used the online translator provided by PROMT[5].

The processing followed to carry out the translation consisted of splitting the text of the reviews in blocks of 500 characters to fit with the maximum length allowed by the online translator. Secondly, after the translation, extra UTF-8 invalid characters were removed and, finally, the translated reviews were generated from the blocks belonging to each of them. Figure 2 summarizes the processing followed to generate the EVOCA corpus.

The new corpus EVOCA contains the same number of positive and negative reviews that OCA corpus, with a total of 500 reviews. Table 4 shows some statistics for the EVOCA corpus.

Table 4. Statistics on the EVOCA opinion corpus

	Negative	Positive
Total **documents**	250	250
Total **tokens**	122,135	153,581
Avg. tokens per review	488.54	614.32
Total **sentences**	5,030	3,483
Avg. sentences per review	20.12	13.93

[5] Available at http://translation2.paralink.com

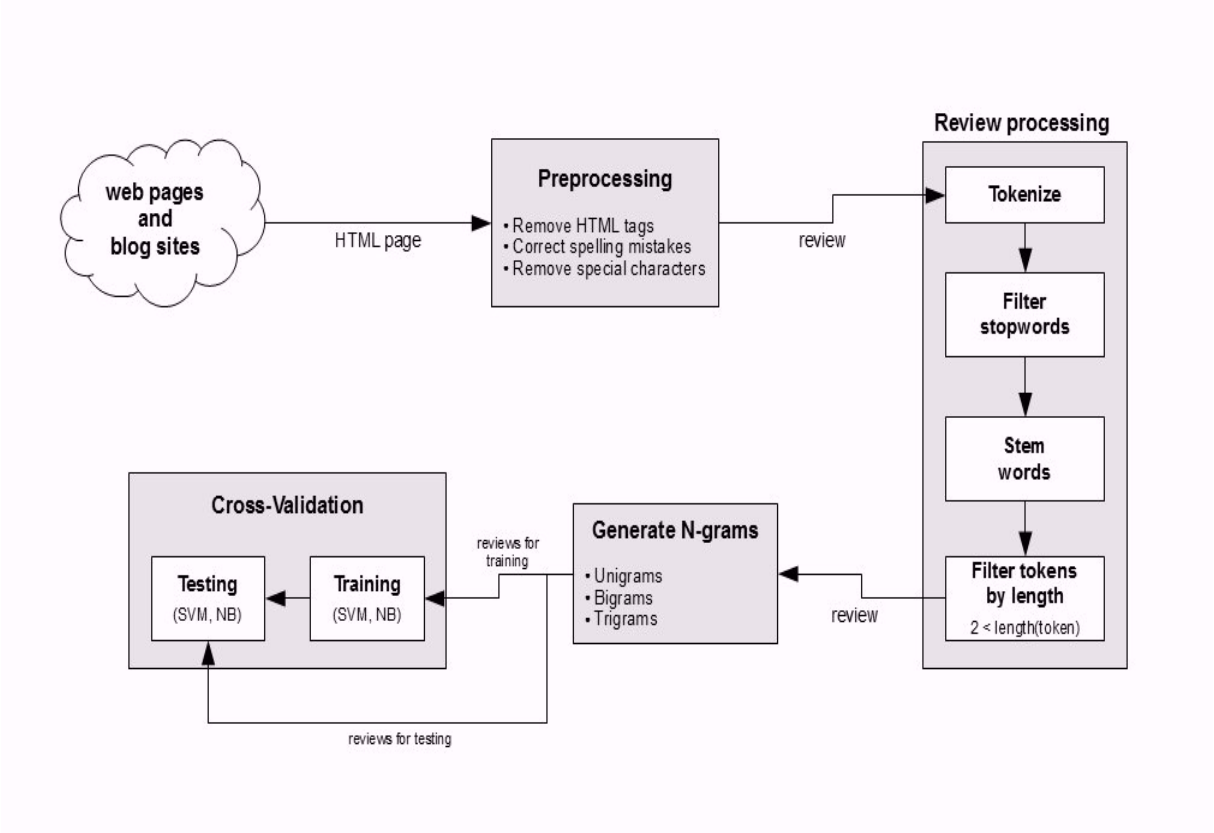

Figure 1. Steps followed in the generation and validation of the OCA corpus

5. Experiments and Results

For the experiments, we have used the Rapid Miner[6] 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.

We have applied two of the most used classifiers: Support Vector Machines (SVM) and Naïve Bayes (NB).

SVM [11] is based on the structural risk minimization principle from the computational learning theory, and seek a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set.

On the other hand, NB algorithm [8] is based on the Bayes theorem. Due to its complex calculation, the algorithm has to make two main assumptions: first, it considers the Bayes denominator invariant, and second, it assumes that the input variables are conditional independence.

In our experiments, the 10-fold cross-validation has been used in order to evaluate the classifier. This evaluation has been carried out on three main measures: precision (P), recall (R) and F1 measure [10].

Moreover, for each machine learning algorithm, we have analyzed how the use of stemmer affects the experiments. TF·IDF has been used as weighting scheme. We have also accomplished several experiments using different n-grams models. However, the obtained results with bi-grams and trigrams were very similar to unigrams. For this reason we have only shown the best results obtained with unigrams. Results for SVM and NB are shown in Table 5 and Table 6, respectively.

As we can see, taking into account the F1 measure, all the experiments with OCA overcome EVOCA except when we use SVM and stemmer. In fact, this is the only case where stemmer obtains a better result although the improvement is very slight (+1.54%). Anyway, the best result is achieved using SVM without stemmer over the OCA corpus with 0.9073 of F1 measure.

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

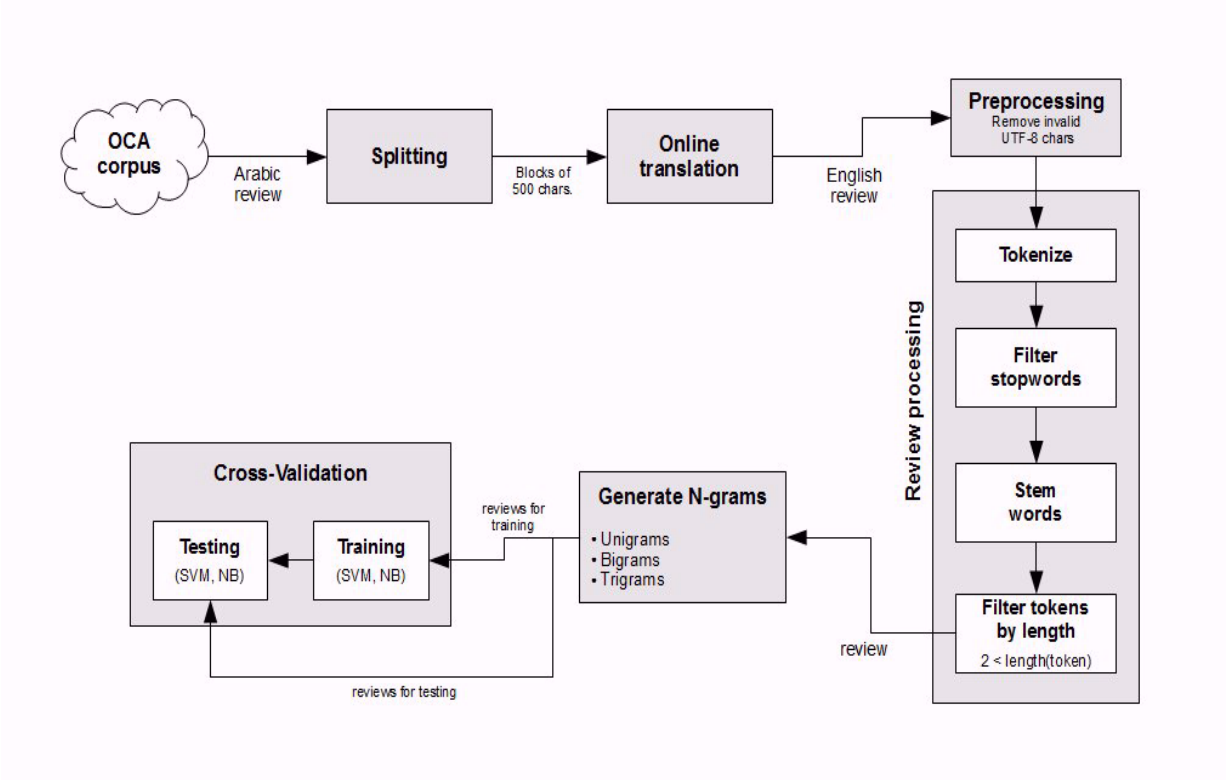

Figure 2. Processing followed to generate and validate the EVOCA corpus

However, it is interesting to note that, in the SVM experiments, the loss of precision due to the translation is very little. The highest difference is 4.31% when we do not apply stemmer, while it is 1.54% when the stemmer is applied. In general, the results with EVOCA, near to 90%, are very good comparing them with other works using SVM and English corpora [9].

Table 5. Results with SVM

	Stem	P	R	F1
OCA	Yes	0.8614	0.8800	0.8706
	No	0.8699	0.9480	**0.9073**
EVOCA	Yes	0.9007	0.8680	0.8840
	No	0.8561	0.8840	0.8698

Table 6. Results with NB

	Stem	P	R	F1
OCA	Yes	0.8106	0.8880	0.8475
	No	0.8274	0.9520	**0.8853**
EVOCA	Yes	0.7100	0.8320	0.7662
	No	0.7323	0.8640	0.7927

As regard the machine learning algorithm, it is clear that SVM works better in all cases. Taking into account the best results on the OCA corpus, SVM improves 2.49% the result obtained with NB (both without applying stemmer). On the EVOCA corpus the difference is higher for SVM +15.37% and +9.73%, using stemmer and without using it, respectively. Although the differences between SVM and NB over the OCA corpus are small, when they are applied over EVOCA, NB loses too much precision. In this case, the translation is affecting highly the results.

Finally, we have analyzed the impact of the stemmer in the experiments. As can be observed in both Table 5 and Table 6, in all cases the stemming process gets worse results except when we use SVM on the EVOCA corpus (+1.63% for stemming). For the OCA corpus, not use the stemmer always improves the results when we apply it (+4.22% using SVM and +4.46% using NB), while we obtain an improvement of 3.46% on the EVOCA corpus using NB.

6. Conclusion

In this paper we have presented an Arabic corpus for opinion mining along with its English translation. OCA and EVOCA corpora are freely available for the research community[7]. The OCA corpus is composed of Arabic reviews obtained from specialized Arabic web pages related to movies and films. Then, we have generated the EVOCA corpus, which is the English translation of the OCA corpus using an automatic machine translation tool. Both corpora include a total of 500 reviews, 250 positives and 250 negatives. In

[7] OCA and EVOCA corpora are freely available at http://sinai.ujaen.es/wiki/index.php/Recursos

addition, we have accomplished several experiments over the corpora using two different machine learning algorithms (SVM and Naïve Bayes) and applying a stemming process. The results obtained show that, although the precision with the EVOCA are lower, they are comparable with other sentiment analysis researches using English texts. This loss of precision due to the translation is very slight (-4.31% when stemmer is not applied) and therefore it is very interesting for the future because we could apply English resources for opinion mining such as SentiWorNet in order to improve the results. On the other hand, we have shown that the use of the stemming process is not recommended to work with these corpora.

7. 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 a grant from the Instituto de Estudios Giennenses, project Geocaching Urbano (RFC/IEG2010). Also, another part of this project was funded by Agencia Española de Cooperación Internacional para el Desarrollo MAEC-AECID.

8. References

[1] Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Trans. Inf. Syst. 26 (3).

[2] Agić, Z., Ljubešić, 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, M. Rosner & D. Tapias (Eds.), Language Resources and Evaluation (LREC). European Language Resources Association.

[3] Ahmad, K., Cheng, D., & Almas, Y. (2006). Multilingual sentiment analysis of financial news streams. Proceedings of Science (GRID2006).

[4] 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). CSREA Press.

[5] Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. ICDE Workshops (pp. 507–512). IEEE Computer Society.

[6] Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the 5th Conference on Language Resources and Evaluation (LREC) (pp. 417-422).

[7] Ghorbel, H., & Jacot, D. (2010). Sentiment analysis of French movie reviews. Proceedings of the 4th international Workshop on Distributed Agent-based Retrieval Tools (DART 2010). Geneva, Italy.

[8] Mitchell, T. (1997). Machine Learning. McGraw-Hill.

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

[10] Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 1.

[11] Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, New York.

[12] 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 (JASIST), 60(12), 2474–2487.