ISSN 2007-9737

Improving Arabic Sentiment Classification Using a Combined Approach

Belgacem Brahimi[1], Mohamed Touahria[2], Abdelkamel Tari[1]

[1] University of Bejaia,
Department of Computer Science,
Algeria

[2] University of Setif,
Department of Computer Science,
Algeria

belkacem.brahimi@yahoo.fr, mtouahria@univ-setif.dz, tarikamel59@gmail.com

Abstract. The aim of sentiment analysis is to automatically extract and classify a textual review as expressing a positive or negative opinion. In this paper, we study the sentiment classification problem in the Arabic language. We propose a method that attempts to extract subjective parts of document reviews. First, we select explicit opinions related to given aspects. Second, a semantic approach is used to find implicit opinions and sentiments in reviews. Third, we combine the extracted aspect opinions with the sentiment words returned by the lexical approach. Finally, a feature reduction technique is applied. To evaluate the proposed method, support vector machines (SVM) classifier is applied for the classification task on two datasets. Our results indicate that the proposed approach provides superior performance in terms of classification measures.

Keywords. Text mining, opinion mining, sentiment classification, supervised learning, review extraction, combined approach.

1 Introduction

In recent years, many people express and share their opinions and evaluations on the internet. Knowing this evaluative content is so interesting. For example, in marketing, companies could use the consumers' opinions and reviews to build their strategies. For individual users, opinions and commetns available on social media about consuming products and services are often consulted before purchasing them.

However, with the rapid growth of web 2.0, the amount of opinionated texts available on the internet is becoming huge which makes its manual exploitation by users a heavy task. For this reason, an automatic mining of such useful information is indispensable.

In this context, the goal of sentiment analysis is to automatically extract and classify document reviews, which express sentiments, emotions and evaluations about various topics such as products, services and persons. Opinion mining is widely applied to social media for a variety of applications like marketing and customer service. Pang et al. [24] cited other applications of using opinion mining such as recommendation systems and document filtering.

Recently, researchers have proposed different methods in order to automatically classify opinionated texts as positive or negative [23, 24, 30]. This recent research domain is known as opinion mining or sentiment analysis. The reader can find a detailed overview on this research area in the work of [18].

Basically, two approaches in the literature are proposed for sentiment analysis: machine learning approaches [24] and semantic (linguistic) ones [31]. In machine learning approaches, we find supervised methods that utilize a sentiment corpus to train and construct predictive models.

These techniques are extensively proposed and evaluated due to their superior performance results compared with the semantic ones [24]. In

the supervised strategy, there are generally two main steps: (i) feature extraction /selection and (ii) sentiment learning/classification [24].

On the other hand, semantic approaches consider the review text as a set of words. These methods use lexical resources, dictionaries like SentiWordNet [13] and rules with the aim of identifying the polarity of the word. Then, the text review is classified as positive or negative by comparing the number of positive and negative opinions words in the document review. Such methods provide generally modest results, as they are domain dependent and have low coverage.

Of course, there are methods that combine both machine learning and semantic approaches to increase sentiment classification performance. We can cite some papers performed in Arabic that merge both methods like [1, 4, 10, 11].

In his doctoral thesis [6], showed through an example of a movie review that the global opinion is generally based on selection criteria. Therefore, the extraction of comments for a film allows selecting the most relevant comments, and it permits to eliminate noise by neglecting extra-subject comments.

In addition, the task of mining movie reviews poses a particular challenge. In fact, before expressing the opinion about a given movie, the author often describes the story of the movie in question including informative texts related to facts and events on that story. Such narrative content have to be excluded as it does not serve for the sentiment classification [19].

This paper is concerned with extracting and predicting sentiments expressed in Arabic movie reviews. The goal of this work is to explore the effect of selecting subjective parts of reviews on opinion classification.

More specifically, we investigate the impact of extracting users' opinions and comments toward different movie aspects such as director and actor on review classification. In addition, we study the contribution of the lexicon approach in opinion prediction. To improve sentiment classification performance, we proposed a combined approach that integrates the two proposed methods.

This article is organized as follows. In the next section, we provide background information about sentiment analysis. Section three gives an overview of the state of the art on Arabic sentiment analysis. In the fourth section, we describe our proposed approach including document pre-processing and aspect comments extraction. Section five provides experimental environment and the conducted experiments (datasets used, algorithm trained, etc.), and the obtained results are discussed. In the last section, we give our conclusion and future work.

2 Sentiment Analysis

Sentiment analysis (called also opinion mining) is the field that uses text mining techniques including natural language processing and data mining algorithms for the task of automatic extraction and classification of opinionated texts as expressing a positive or negative opinion or sentiment. Opinion mining is named sentiment polarity classification or polarity classification [27].

The main task of opinion mining is polarity classification. However, we can find some other subtasks in this research domain, we can cite for example: subjectivity detection, emotion detection and classification, measuring sentiments intensity and opinion summarization.

Additionally, there exist three different levels to study sentiment analysis: word (feature) level, sentence level and document level. The term (word) level consists of identifying the orientation of each word in the text (positive or negative). The sentence level aims to classify each sentence in the document. Finally, the analysis performed on document level tries to recognize the polarity of the whole document as positive or negative.

In our paper, we use the supervised approach for classifying movie reviews in the Arabic language at the document level meaning that we consider the entire document as the basic information unit.

3 State of the Art

Most of scientific studies on sentiment analysis focus principally on the English language, while similar studies conducted in the Arabic language and some other Morphologically-Rich Languages (MRL) are still limited. This is due to the lack of

public lexical resources and the complexity of the automatic processing of Arabic language [1, 19]. Indeed, the agglutinate nature of Arabic makes it more complex compared with the other languages. For instance, one word in Arabic could be a combination of prefixes, stems and suffixes [1].

In this section, we present the most important research articles that have been performed in Arabic sentiment analysis and in particular in Arabic movie review classification.

To the best of our knowledge, the study of [3] is among the first articles that addressed the sentiment analysis problem in the Arabic language. The authors proposed a system that extracts sentiment-bearing patterns related to financial texts by use of local grammar. The performance rate of their proposed system was between 60% and 75%.

The studies of [24, 30] are considered among original research papers related to the movie reviews domain in English. The authors in [30] proposed an unsupervised learning algorithm for classifying opinions as recommended (thumbs up) or not recommended (thumbs down). This algorithm extracts phrases containing adjectives or adverbs, calculates their semantic orientations, and then classifies the review based on the average semantic orientation of the phrases. The author found that the task of classifying movie reviews is difficult compared with banks and automobiles reviews. The accuracy percentage of automobile reviews was about 74%, whereas this value in movie reviews was 66%.

On the other hand, [24] used standard machine learning algorithms to classify movie reviews. In their work, they found that standard machine learning techniques outperform human generated baselines. The researchers used popular classifiers in machine learning namely, Naive Bayes (NB), Maximum Entropy (MA) and Support Vector Machines (SVM). Experimental results showed that SVM is the best classification algorithm.

For classifying Arabic movie reviews, the researchers in [27] collected Opinion Corpus for Arabic (OCA) from websites and blogs related to movies. The size of this corpus is small; it includes 250 positive reviews and 250 negative ones.

The authors used different pre-processing tasks comprising manual spelling correction, stop-words removal, filtering tokens by length, stemming and n-grams word generation. The experiments indicated that using the SVM classifier provides the superior performance results among the other algorithms.

In their next work, the researchers in [28] translated the OCA corpus into English by using an automatic machine translation tool to build an English version called EVOCA. The authors employed two classifiers which are SVM and NB to classify movie reviews in their corpus. Their results showed that the performance in the translated corpus EVOCA was lower compared with the Arabic corpus OCA.

The authors in [22] conducted a study in order to investigate the impact of selecting roots of words on the sentiment classification task. Their experiments were done on two datasets (the PATB, Part 1 v 4.1) [16] and OCA corpus. Results indicated that for the PATB corpus, the Khoja stemmer [15] with word unigrams was the best solution, whereas using the Tashaphyne stemmer [32] recorded the best performance in the OCA movie dataset.

The study of [19] aimed to investigate supervised sentiment classification in the Arabic language. This work was performed at the document level. The researchers employed two collections of documents. The first was ACOM (Arabic Corpus for Opinion Mining) collected by the researchers including two datasets: DS1 consists of 594 movie reviews and DS2, which contains 1492 sport comments. The second dataset tested was OCA. The result of this paper showed that SVM and NB were efficient; whereas the outputs obtained from k-Nearest Neighbor (k-NN) depended on the documents collection.

The researchers in [7] investigated different factors that could impact sentiment classification of Arabic reviews such as text representation, feature selection and the algorithm used for predicting reviews. The outcomes of this study indicated that their pre-processing techniques improve sentiment classification performance.

Another scientific study presented in [21] that explored the effect of feature selection techniques and machine learning algorithms on Arabic sentiment analysis.

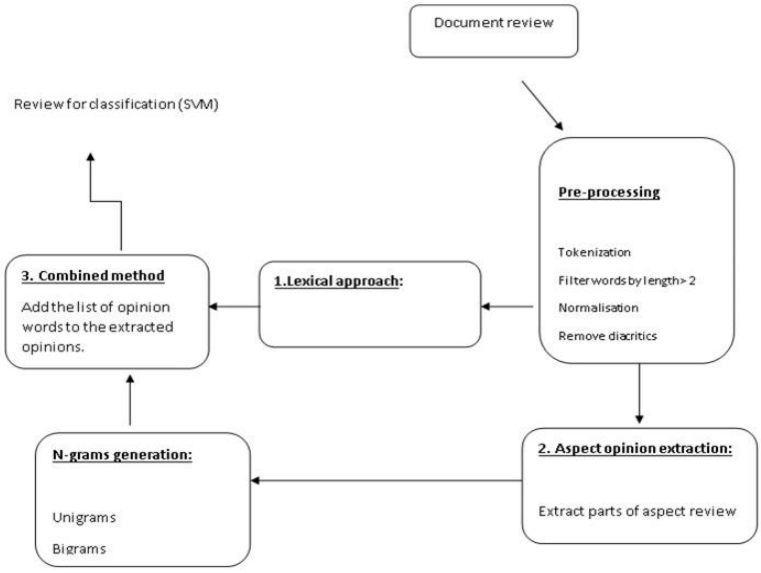

Fig. 1. Steps followed in our combined approach

The work compared seven feature selection methods. In addition, three supervised classifiers (SVM, NB and k-NN) were evaluated. The experiment results of this comparative study showed that the SVM classifier coupled with the SVM-based feature selection method was the best option.

The work of [12] addressed the problem of lack of resources in the Arabic sentiment analysis domain by generating large datasets in different domains such as movies and products. The results obtained revealed that the best performance was provided by SVM, and the best feature models were the lexicon based feature combined with other features such as term frequency inverse document frequency (TFI-IDF) and count.

A suggestion of a combined approach that merges semantic techniques and machine learning ones was presented in the paper of [4]. Another research study [8] described a framework for Arabic opinion classification. The researchers compiled a text review dataset including 2591 tweets/comment, and used three algorithms (NB, SVM and k-NN) to predict the polarity of reviews.

The results of their experiments indicated that the three models (binary model, term frequency and TF-IDF) give similar results. In last years, several advanced studies employed deep learning DL models for opinion classification in Arabic. For example, the authors in [5] presented a study that explored DL algorithms for sentiment classification at the sentence level. The authors tested deferent DL models such as Deep Belief Networks (DBN), Deep Auto Encoders (DAE) and Recursive Auto Encoders (RAE). The authors found that RAE was the best model for classifying sentences.

More recently, the authors of [1] proposed a distant supervision algorithm to create a corpus from Twitter called 'TEAD', they employed in their approach emojis and sentiment lexicons. Another recent study in [20] that explored the effect of the pre-processing task on opinion classification results in Tunisian dialects. The researchers indicated that integrating named entity recognition NER improves opinion classification.

From this brief overview, we can note that datasets compiled to perform Arabic sentiment analysis about movies are of modest size when we compare them with the English ones. In this

Table 1. List of words used for movie aspects

Aspect	Example of aspect words
Movie (الفيلم)	الفلم- فيلم- العمل film, movie, work.
Story (القصة)	قصة- نص- حبكة , tale, text, plot
Actor (الممثل)	ممثل- دور actor, role.
Director (المخرج)	مخرج - مشهد , director, scene.
Music (الموسيقى)	موسيقى- غناء music, song.

Table 2. Example of an excerpt from a movie review and some corresponding comments aspects.

Movie review			
ثلاثة عوامل مهمّة تجتمع لهذا **الفيلم**: **سيناريو** محكم من نِك هورنبي، **إخراج** جيّد من لون شرفيك و**تمثيل** جيد من كاري موليغان في **دور** الطالبة *التي عليها الاختيار بين أكثر من طريق لحياتها المستقبلية في فترة الستينات حيث حرية المرأة كانت لا تزال تواجه عقبات*. أحد أهم **أفلام** العام الماضي.			
Translation in English: Three important factors meet for this **movie**: An **airtight screenplay** from Nick Hornby, **good direction** of Lone Scherfig, and **good acting** by Carrie Mulligan in the **role** of a student *who has to choose between more than one way to for her future life in the 1960s where the freedom of women still faces obstacles. One of the most important* **movies** of last year.			
Movie aspect	**Wind size**	**Comment aspect**	**Translation in English**
Movie (فيلم)	**7**	مهمّة تجتمع لهذا **الفيلم**: **سيناريو** محكم من	meet for this **movie**: An airtight screenplay
Director (مخرج)	**7**	من نِك هورنبي، **إخراج** جيّد من لون	Nick Hornby, good **direction** of Lone Scherfig,
Actor (ممثل)	**7**	من لون شرفيك و**تمثيل** جيد من كاري	Scherfig, and good **acting** by Carrie Mulligan
Lexical approach outputs:		جيد- جيد – محكم in English: good, good, airtight.	

perspective, we decided to collect another dataset including 1000 reviews for movies. The second remark is concerning the pre-processing task, which is heavily investigated in Arabic research studies due to its impact on review classification results. The third observation is about the SVM classifier that is proved to be superior in Arabic sentiment classification in terms of performance results.

4 Proposed Approach

This section presents the details of our methodology for the task of sentiment prediction of movie reviews that includes the following sub-tasks: text pre-processing and representation, lexical approach applied to the entire text review, aspect comment extraction as well as the combination of the proposed methods. Figure 1 illustrates the steps of the proposed approach.

4.1 Text Representation

In the present subsection, we describe the pre-processing step performed to map unstructured texts into document vectors. This is necessary for text classification, as data mining algorithms cannot be applied directly on textual data. In the pre-processing stage, some optional tasks such as stemming and stop word removal may affect the classification performance.

After collecting movie reviews from the web, we perform the pre-processing task that includes deleting unrelated parts of reviews, tokenizing which splits the text into sequences of characters based on whitespaces.

We use additional punctuation marks in Arabic such as the comma '،', question mark '؟' and semicolon '؛'.

In addition, we retain only Arabic words by removing non-Arabic letters, symbols and numbers.

Finally, we perform normalization of Alif and Taa, and we remove diacritics like tanween and shedda. The output of this operation is a set of words.

4.2 Extraction of Aspect Opinions

This part of the study aims to investigate the impact of extracting explicit movie aspect comments on movie review classification. The term explicit opinion means that the target is commented directly. For example, in the following comment "هذا الفيلم جيد", meaning in English "this movie is good", the target is the movie (الفيلم) and the opinion is (جيد), good in English.

We filter irrelevant parts of reviews by extracting explicit opinions and evaluations of different movie aspects.

For aspect opinion extraction, we use the list of aspect words utilized in the work presented in [25], and translate it to Arabic by use of Google translator available on (https://translate. google.com). These aspects are movie, actor, story, director and music.

Each movie aspect is detected by using a words list. Table 1 shows the 5 movie aspects and some corresponding words. To extract explicit movie comments on different aspects of a given movie, we propose to create a window of size n words to extract opinions for each aspect. Initially, the proposed system receives as input a text review, which is pre-processed by applying tokenization, eliminating non Arabic words, etc.).

Next, a window of size n words is created to extract opinions and evaluations on each movie aspect (e.g. movie, actor). For example, in the case of a window of 7 words, 3 words before the word aspect and 3 words after that word are extracted. In our experiments, we use different values of the window size to determine empirically which value is the best for movie review classification.

Finally, the extracted comments of different aspects for each movie review are collected to form a unique review. The collected opinions related to movie aspects can be considered as a review summary. Table 2 illustrates an example of a text review including the extracted aspect comments detected by the window technique. Aspect words are highlighted in boldface, while their related opinions are underlined.

Table 3. Description of the used datasets

	OCA		**ARMD**	
	positive	negative	positive	negative
# docs	250	250	500	500
# tokens	121,392	94,556	108,454	108,024
Average of tokens per doc	485	378	216.91	216.05

After this, we reduce word forms by performing stemming. Popular stemming techniques for Arabic are stemming (Khoja) [15] and light stemming. In this study, we used the Khoja stemmer to stem words.

Concerning features, we employed n-grams of words [29]. By definition, n-grams of words is a succession of n words in a given text. An n-gram of length 1 is referred to unigrams, and two successive words is a bigrams, 3 words is called a trigrams, and so on. The use of n-grams word is useful for capturing valuable information like negation (for instance, ليس جيدا, in English: not good) and relevant opinions about the target (movie) like,فيلم جيد (good movie, in English).

Regarding weighting schemes, in the literature, common used ones are Boolean weighting, Term Frequency TF and Term Frequency Inverse Document Frequency (TF-IDF). We employed TF-IDF since it is a very popular weighting scheme in text mining tasks such as information retrieval and text classification.

Feature reduction/selection is a common technique used to reduce the size of the textual data to be processed. Among the feature reduction techniques, we find Term Frequency Thresholding, a common pruning technique that eliminates features appearing frequently and words that have low frequencies in the dataset.

The reason is that, generally, words having high frequencies in texts are common, and thus

they are not distinctive for the task of classification. In the same way, words with small frequencies have a marginal contribution in text classification since they are not enough observed in the learning phase, and thus they do not help supervised algorithms to build their classification model.

4.3 Lexical Method

In this method, we use a sentiment dictionary to extract opinion words that exist in movie reviews. We receive the entire review as an entry. After this, the text review is tokenized, and for each word of the review, we look if that word exists in the lexicon. The result is a set of opinion words. Table 2 illustrates the output of this method applied on an example of a review text.

4.4 Combined Approach

In order to enhance our system classification accuracy, we merge the proposed method of aspect opinion extraction with the lexicon one. The fusion of the approaches is performed at the feature level, i.e., the list of sentiment words found by the lexicon method is added to the opinions of movie aspects obtained by the comment extraction method.

Our idea is that movie reviews do not contain only explicit opinions related to movie aspects, but also implicit sentiments such as the conclusion opinion. The contribution of the lexicon approach is to capture these implicit sentiments which are missed by the aspect comment extraction method.

After explaining the proposed method that combines opinion extraction and lexical features to ameliorate opinion classification, the next step is to validate the proposed approach, and this is the object of the next section.

5 Experimental Study

The present section describes in detail the experimental part of this work; it includes tests and experiments conducted to validate the proposed approach. The methodology for text processing described in Section 4 was applied. In the experiments, we used the Weka tool for the task of pre-processing and classification. Weka is an open source tool for data mining applications that supports different tasks related to text mining like text pre-processing, clustering, classification and prediction [14].

5.1 Datasets

In our tests, we employed OCA dataset (Opinion Corpus for Arabic) [27] which is publicly available for research studies, it was gathered by the authors comprising 500 reviews in the movie domain.

In addition, to validate the results, we compiled a second review dataset we called AMRD (Arabic Movie Review Dataset). The corpus was created from different web pages and blogs related to movie reviews in Arabic. For selecting reviews, the methodology followed by the researchers in [27] was adopted meaning that we accepted only comments that were expressed in MSA and Arabic dialect. Reviews were mainly gathered from the two popular websites (elcinema.com) and (cairo360.com). In Table 3, we provide a basic description of the collections.

5.2 Used Lexicon

As explained in our methodology (section 4), to increase our system classification performance, we employed the lexicon approach to enrich our summary review (comments on movie aspects) by adding sentiment words. To achieve this goal, we used the Nile University's Arabic sentiment Lexicon NilULex presented in [9].

This dictionary contains Egyptian Arabic and Modern Standard Arabic (MSA) sentiment words and their orientation (positive, negative). The lexicon is suitable for our task since Arabic movie reviews are written in MSA and Arabic dialect. However, this dictionary is for global use and must be adapted to the movie review context. To keep only subjective words in the lexicon, we removed opinion words expressing facts and events such as (اختطف in English to kidnap, سجن which means prison in English).

These words may impact opinion classification results. We also eliminated words describing

movie genre such as رعب (horror) and جريمة which means crime in English. Finally, terms related to movie aspects like acting (تمثيل) and starring (بطولة) were excluded from the lexicon. After the pre-processing step of a given document review, we searched for each word of the review if it exists in our adapted lexicon. The result of this task is a set of sentiment words.

5.3 Classifier Employed

Regarding the classification task, we utilized Sequential Minimum Optimization (SMO) [26], an implementation of the SVM classifier. Indeed, this algorithm is found to be the best algorithm in Arabic sentiment classification [12, 21, 27]. We performed several experiments in order to evaluate the proposed classification system. Moreover, 10-fold- cross validation was used to average the performance results [17].

In addition, we employed F-measure (FM) to measure the performance of our system.

6 Results

In this section, we present experimental results in order to explore the contribution of our model in sentiment classification. As weighting schemes, we employed TFI-DF. We note that, for example, *Win 7* denotes the method that extracts comments using a window size of 7 words, while *7+ lexical* means comments extracted by using a window size of 7 words combined with the lexical method that extracts opinion words from the text review.

The experimental results of applying the review models on OCA dataset are shown in Table 4. To extract opinion segments, we varied the size of the window to obtain the best value for sentiment classification. The best results of performance are highlighted in boldface. As it can be observed from Table 4, using bigrams as features improves slightly opinion classification.

In addition, there is not a significant difference in performance when using 9 and 11 words as a window size, while applying a window size of 7 words provides the worst result of classification. This behavior is expected since this length (three words before and after the aspect word) is not

Table 4. OCA sentiment classification results

Review model	Tokens	
	Unigrams	**Bigrams**
Wind 7	87.40	88.80
Wind 9	89.20	91.40
Wind 11	89.40	91.60
7 + lexical	89.40	92.20
9 + lexical	91.00	**92.40**
11+ lexical	91.60	92.00
entire text	91.20	91.40

Table 5. OCA sentiment classification results applying feature reduction method on entire reviews

Feature size	Stemming (Khoja)	
	Unigrams	**Bigrams**
1,000	90.40	90.00
2,000	89.60	93.00
3,000	90.40	**93.60**
4,000	90.40	93.20

Table 6. OCA opinion classification results using comment extraction method (window size=7) and feature reduction technique

Feature size	Stemming (Khoja)	
	Unigrams	**Bigrams**
1,000	86.20	87.60
2,000	87.00	87.80
3,000	87.00	**89.20**
4,000	87.00	**89.20**

Table 7. OCA opinion classification results using comment extraction method (window size=9) and feature reduction technique

Feature size	Stemming (Khoja)	
	Unigrams	**Bigrams**
1,000	88.20	90.40
2,000	89.40	92.60
3,000	89.40	**94.40**
4,000	89.40	91.40

Table 8. OCA opinion classification results using comment extraction method (window size=11) and feature reduction technique

Feature size	Stemming (Khoja)	
	Unigrams	**Bigrams**
1,000	88.20	91.80
2,000	89.80	92.80
3,000	89.80	92.60
4,000	89.80	**96.00**

Table 9. OCA opinion classification results using the combined method (window size=11) and feature reduction technique

Feature size	Stemming (Khoja)	
	Unigrams	**Bigrams**
1,000	89.00	91.80
2,000	90.10	93.40
3,000	89.40	94.10
4,000	89.40	**96.00**

Table 10. AMRD opinion classification results

Method	Optimal size	FM
Entire document	-	82.10
Entire document, F-reduction	2,000	83.80
Wind 7, F-reduction	4,000	83.00
Wind 9, F-reduction	3,000	81.80
Wind 11, F-reduction	3,000	83.00
Wind 11, lex, F-reduction	3,000	**84.60**

able to capture sufficient number of sentiment words, which are surrounding the aspect words.

We can also note that adding lexical features ameliorates slightly classification results, and utilizing the window size 9 combined with the lexical method yield the best results (92.40%).

Table 5 illustrates the obtained results when we applied the frequency based feature reduction method on the entire text reviews of OCA. The best result is obtained when selecting 3,000 as a feature size of words and bigrams of terms.

In next experiments, we investigate the contribution of combining the pruning technique with our method that extracts the most opinionated bodies of document reviews. Table 6 indicates the classification results using the length of 7, as a window size. We can see that 3,000 and 4,000 as threshold value yield the best classification results.

Next, the outcomes of selecting 9 words around the aspect term are depicted in Table 7, which demonstrate that the optimal value of FM is recorded when 3,000 of bigrams are determined as a threshold value.

It can be seen from Table 8 that the size of window 11 for extracting aspects comments (5 words before the aspect word, and 5 words after that word), combined with bigrams is the optimal setting for reaching the best classification results of FM (96%).

Finally, we provide in Table 9 the results of the combined method that extracts sentiment words by using the lexicon and integrates them to the comments about different aspects in movies.

We selected 11 as a window size since it is the best parameter for sentiment classification in the previous tests.

Our remark is that the combined method provides similar results compared with those of Table 8. The best value of FM recorded is 96%.

The results about the second data collection AMRD are summarized in Table 10. The employed features are bigrams of stemmed words. We recall that wind 7 means the method that extracts comments using a window size of 7 words, while wind 11+ lexical is the combined method. F-reduction denotes feature reduction.

The first remark is about opinion classification results, which are lower compared with those of the OCA dataset.

Results in Table 10 confirmed that the combined method (Wind11, lex, F-reduction) outperforms the others as it provides the highest classification outputs in FM (84.60).

From these results, we can conclude that the presented technique for extracting movie aspect comments allows us to mine useful opinions. The best value of window length was 11. We can also say that using bigrams of terms, applying stemming and feature reduction contribute positively in opinion classification.

To show the contribution of this work, we compare here our obtained results with those

reported in the literature concerning the OCA dataset. The authors of [27] reported 90% in F-measure, while the approach proposed by [19] recorded up to 93% in F-measure. We can say that our proposed methods for extracting relevant opinions provide superior results (96% in F measure).

To summarize the obtained results from the experimental study, we concluded the following findings:

Regarding the method for extracting aspect comments of movies, the best value of window size is 11, while the worst length of comment is 7.

Merging the extracted aspect comments of a movie with the sentiment words of lexical method enhances slightly the performance of opinion classification.

Combining the technique of review extraction with the feature reduction technique ameliorates opinion classification results. In addition, the contribution of bigrams is clearly positive. The optimal feature size is 4,000, and the improvement in f-measure is clear, 4% in OCA dataset.

Moreover, performance classification in our combined method depends mainly on the lexicon employed, we think that enriching and adapting the sentiment dictionary could enhance sentiment classification results.

The last remark regarding movie reviews, where negative comments contain sarcastic and ironic sentences comprising positive words used in a negative sense.

The enhancement of the sentiment classification system implies resolving this problem and other related issues such as negation and comparative opinions.

7 Conclusion and Future Work

This article addressed sentiment classification for movie reviews expressed in Arabic. The contribution of this work is to study the impact of selecting subjective parts in reviews and filtering movie reviews from their irrelevant parts including objective information and movie events description. To achieve this goal, we proposed to extract only explicit opinions and subjective parts related to the movie aspects such as film and actor.

In order to improve sentiment classification, we combined the extracted comments with the list of sentiment words found by the lexicon method. The contribution of the lexicon method is to extract implicit sentiments and opinions, which are not directly related to the movie aspect such as the conclusion opinion. In addition, we used Khoja stemmer and a feature reduction technique based on term frequency with the aim of improving the proposed system for review classification.

For the experiments, we applied our methods on two datasets. One of them is OCA, a free collection available on the web, while the other data collection is AMRD collected by the authors. As feature schemes, we used unigrams and bigrams of words.

The classification task was performed by using SVM, and the performance results proved the effectiveness of our combined approach.

As future work, we intend to apply our classification system on other domains such as social and marketing. Moreover, we plan to enhance Arabic sentiment classification by enriching sentiment lexicons, building specific dictionaries and text collections for different domains and aspects in Arabic. We also think that using deep learning models in the proposed system could augment opinion classification performance.

References

1. **Abdellaoui, H. & Zrigui, M. (2018).** Using Tweets and Emojis to Build TEAD: an Arabic Dataset for Sentiment Analysis. *Computación y Sistemas***,** Vol. 22, No. 3. DOI:10.13053/CyS-22-3-3031.
2. **AbdelRahman, S., Elarnaoty, M., Magdy, M., & Fahmy, A. (2010).** Integrated machine learning techniques for Arabic named entity recognition. *International Journal of Computer Science Issues,* Vol. 7, pp. 27–36.
3. **Ahmad, K., Cheng, D., & Almas, Y. (2007).** Multilingual sentiment analysis of financial news streams. *1st International Workshop on Grid Technology for Financial Modeling and Simulation,* Vol. 26, pp. 01–08. DOI:10.22323/1.026.0001.

4. **Aldayel, H.K., & Azmi, A.M. (2016).** Arabic tweets sentiment analysis–a hybrid scheme. *Journal of Information Science*, Vol. 42, No. 6, pp. 782–797. DOI:10.1177/0165551515610513.
5. **Al-Sallab, A., Baly, R., Hajj, H., Shaban, K.B., El-Hajj, W., & Badaro, G. (2017).** Aroma: A recursive deep learning model for opinion mining in Arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, Vol. 16, No. 4, pp. 1–20. DOI:10.1145/3086575.
6. **Duthil, B. (2012).** *De l'extraction des connaissances à la recommandation.* Thèse de doctorat. Ecole Nationale Supérieure des Mines d'Alès.
7. **Duwairi, R. & El-Orfali, M. (2014).** A study of the effects of pre-processing strategies on sentiment analysis for Arabic text**.** *Journal of Information Science,* Vol. 40, No. 4, pp. 501–513. DOI: 10.1177/0165551514534143.
8. **Duwairi, R.M. & Qarqaz, I. (2016).** A framework for Arabic sentiment analysis using supervised classification. *International Journal of Data Mining, Modelling and Management,* Vol. 8, No. 4, pp. 369–381. DOI:10.1504/IJDMMM.2016.081247.
9. **El-Beltagy, S.R. (2016).** NileULex: A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic*. LREC.*
10. **El-Beltagy, S.R., Khalil, T., Halaby, A., & Hammad, M. (2016).** Combining lexical features and a supervised learning approach for Arabic sentiment analysis. *International Conference on Intelligent Text Processing and Computational Linguistics,* pp. 307–319. DOI:10.1007/978-3-319-75487-1_24.
11. **El-Halees, A.M. (2011).** *Arabic Opinion Mining Using Combined Classification Approach*.
12. **ElSahar, H. & El-Beltagy, S.R. (2015).** Building large Arabic multi-domain resources for sentiment analysis. *International Conference on Intelligent Text Processing and Computational Linguistics,* pp. 23–34. DOI:10.1007/978-3-319-18117-2_2.
13. **Esuli, A. & Sebastiani, F. (2006).** Sentiwordnet: A publicly available lexical resource for opinion mining. *LREC,* Vol. 6, pp. 417–422.
14. **Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009).** The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, Vol. 11, No. 1, pp.10–18. DOI:10.1145/1656274.1656278.
15. **Khoja, S. & Garside, R. (1999).** *Stemming Arabic text.* Computing Department, Lancaster University, UK.
16. **Maamouri, M., Bies, A., Jin, H., & Buckwalter, T. (2010).** The Penn Arabic Treebank*. Computational Approaches to Arabic Script-Based Languages: Current Implementations in Arabic NLP***,** *CSLI NLP Series.*
17. **Manning, C.D. & Schütze, H. (1999).** *Foundations of Statistical Natural Language Processing.* MIT press.
18. **Liu, B. & Zhang, L. (2012).** *A Survey of Opinion Mining and Sentiment Analysis***.** pp. 415–463, Springer, Boston, MA.
19. **Mountassir, A., Benbrahim, H., & Berrada, I. (2013).** Sentiment classification on Arabic corpora. *Document Numérique,* Vol. 16, No. 1, pp. 73–96. DOI:10.3166/dn.16.1.73-96.
20. **Mulki, H., Haddad, H., Bechikh-Ali, C., & Babaoğlu, I. (2018).** Tunisian dialect sentiment analysis: A natural language processing-based approach. *Computación y Sistemas*, Vol. 22, No. 4. DOI:10.13053/CyS-22-4-3009.
21. **Omar, N., Albared, M., Al-Moslmi, T., & Al-Shabi, A. (2014).** A comparative study of feature selection and machine learning algorithms for Arabic sentiment classification. *Asia Information Retrieval Symposium*, pp. 429–443, Springer, Cham.
22. **Oraby, S., El-Sonbaty, Y., & El-Nasr, M.A. (2013).** Exploring the effects of word roots for Arabic sentiment analysis. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 471–479.
23. **Pang, B. & Lee, L. (2008).** Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, Vol. 2, No. 1-2, pp. 1–135. DOI:10.1561/1500000011.
24. **Pang, B., Lee, L., & Vaithyanathan, S. (2002).** Thumbs up?: sentiment classification using machine learning techniques*. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Vol. 10, pp. 79–86. DOI:10.3115/1118693.1118704.
25. **Parkhe, V. & Biswas, B. (2016).** Sentiment analysis of movie reviews: finding most important movie aspects using driving factors. *Soft Computing,* Vol. 20, No. 9, pp. 3373–3379. DOI:10.1007/s00500-015-1779-1.
26. **Platt, J. (1999).** Fast training on SVMs using Sequential Minimal Optimization. In: **Scholkopf, B., Burges, C., & Smola, A. (Ed.),** Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA.
27. **Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., & Perea-Ortega, J.M. (2011).** OCA:

Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology,* Vol. 62, No. 1, pp 2045–2054. DOI:10.1002/asi.21598.

28. **Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., & Perea-Ortega, J.M. (2011).** Bilingual experiments with an arabic - english corpus for opinion mining. *Proceedings of the International Conference Recent Advances in Natural Language Processing,* pp. 740–745.
29. **Shannon, C.E. (1948).** A mathematical theory of communication. *The Bell System Technical Journal*, Vol. 27, No. 3, pp. 379–423. DOI:10.1002/j.1538-7305.1948.tb01338.x.
30. **Turney, P.D. (2002).** Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,* pp. 417–424. DOI:10.3115/1073083.1073153.
31. **Turney, P.D. & Littman, M.L. (2002).** *Unsupervised learning of semantic orientation from a hundred-billion-word corpus. arXiv cs*/0212012.
32. **Zerrouki, T. (2012).** *Arabic Light Stemmer*. https://pypi.python.org/pypi/Tashaphyne/0.2.

Article received on 24/02/2019; accepted on 26/07/2020. Corresponding author is Belgacem Brahimi.