I.J. Education and Management Engineering, **2017, 1, 1-13**
Published Online January 2017 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2017.01.01

Available online at http://www.mecs-press.net/ijeme

Word Clustering as a Feature for Arabic Sentiment Classification

Saud Alotaibi [a,*], Charles Anderson [b]

[a] *Umm Alqura Univerisity, Al Taif Road, Makkah and 24382, Saudi Arabia*
[b] *Colorado State Univesity, 1100 Center Ave, Fort Collins, CO80521, US*

Abstract

Rich morphology language, such as Arabic, requires more investigation and methods targeted toward improving the sentiment analysis task. An example of external knowledge that may provide some semantic relationships within the text is the word clustering technique. This article demonstrates the ongoing work that utilizes word clustering when conducting Arabic sentiment analysis. Our proposed method employs supervised sentiment classification by enriching the feature space model with word cluster information. In addition, the experiments and evaluations that were conducted in this study demonstrated that by combining the clustering feature with sentiment analysis for Arabic, this improved the performance of the classifier.

Index Terms: Sentiment Classification, Polarity Classification, Arabic Natural Language Processing, Arabic Sentiment Sentence Classification, Machine Learning Classifier, Word Clustering.

1. Introduction

With the growth of the Internet as a means of communication between people, many technologically advanced methods have been developed in order to facilitate the use of this form of communication. As a result of this phenomenon, an increasing number of people's opinions and thoughts are being published over the Internet. From forums and websites, through to Twitter and Facebook, numerous opinions, thoughts and sentiments appear online on a daily basis. In addition, user reviews, which are found on many marketing or other websites, may be considered a good source with which to help grow people's imagination on specific topical areas. Over the past decade, the extraction of sentiment from text has attracted a lot of attention, both in industry and in academia. Sentiment analysis attempts to discern an individual's opinion based upon their writing. Many fields are included within this topical area, such as natural language processing, machine learning, and computational linguistics.

There are three main aspects within the sentiment analysis field: Lexicons, Annotated Corpora, and Tools. 'Lexicons' relate to demonstrating words, phrases and patterns that can be used to express subjectivity. 'Tools'

* Corresponding author: Tel.: +966 50 466 3239; fax: +966 12 550 1000
E-mail address: ssotaibi@uqu.edu.sa

include two parts. The first part is the machine learning classifiers that use text classification algorithms. The second part is the NLP tools; which includes the POS tagger, the stemmer, and the morphology tagger. These may be utilized in the pre-processing phase, prior to starting the use classification algorithm in order to obtain specific features of the text. The basic aspect of sentiment analysis is the Corpora, which contains pieces of text annotated with their polarity. These Corpora are then utilized by classification algorithms to determine the sentiment of the new text.

Much of the research on sentiment analysis has been undertaken using the English language, as English is the dominant language utilized within and by scientific researchers. However, due to its complexity and the scarcity of available resources, Arabic natural language processing has become an attractive research topic for researchers. According to Farghaly and Shaalan [1], the field of Natural Language Processing (NLP) in Arabic is currently at an early evolutionary stage, despite concerted efforts being made to-date with the fundamental NLP tools of Arabic.

The Sentiment Analysis (SA) of Arabic is also currently in the early stages, and increased efforts and the reliability of low level tools are required in order to build upon this foundation. Many current approaches in Arabic sentiment analysis rely on the bag-of-word (BOW) model representation to build feature vector models [2-5]. Other types of features are then combined with the baseline model in order to leverage the classifier's performance. However, for a language that is rich in morphology and high flection, such as Arabic, the feature vectors are spare due to the variety of the language vocabulary. Even after utilizing the stemming or lemmatization techniques, this may not be enough to preserve the same sentiment orientation of the word [6]. To resolve the issue of data sparsity in the sentiment analysis of Arabic text, the possibilities of using the word clustering technique to enrich the baseline model of Arabic sentiment analysis was investigated.

The remainder of this article is organised as follows: The second section displays the most important related works that have been conducted within the discipline of Arabic sentiment analysis. This section demonstrates the work has been conducted on the topic of Arabic sentiment corpus, as well as the different features and methods of sentiment analysis. The third section introduces a proposed method that is based upon how the cluster approach is utilised when conducting an Arabic sentiment analysis. This is followed by the Experiments and Results discussion in sections four and five. The Conclusion and proposed future studies are contained within the last section.

2. Related Works in Arabic Sentiment Analysis

As aforementioned, most of the research on sentiment analysis has been undertaken using the English language, as English is the dominant language of scientific researchers. Recently, a number of researchers have concentrated on applying sentiment analysis to other languages, one example is Arabic. This section shows the related works that has been conducted in Arabic focusing on the different aspects of analysis, including the corpora, and the features and methodologies.

2.1. Arabic Sentiment Corpora

The Opinion Corpus for Arabic (OCA) [7] (which is the only corpus that has been published) contains 500 reviews on movies. They are annotated at the document level. Half of these reviews are considered positive and the remainder are negative. Further work that was undertaken to build a multi-genre subjectivity and sentiment corpus for modern standard Arabic is called AWATIF [8]. The domain of this data was extracted from a news wire within different domains (400 documents), Wikipedia talk pages (around 5342 sentences), and web forums (around 2532 threads from seven web forums). The annotation was directed toward the sentence level, and three different conditions were used to annotate the data: (1) Gold Human with Simple Guidelines (GH-SIMP); (2) Gold Human linguistically-motivated and Genre-nuanced (GHLG); (3) Amazon Mechanical Turk with Simple Guidelines (AMT- SIMP) [8]. In addition, an attempt was made to build a labelled social media corpus based on subjectivity and sentiment in the Arabic language (in the SAMAR project) [9]. The data was

collected from four different forms of social media. These included Arabic chatting, Tweets, Wikipedia Talk, and online forums. This corpus was a combination of long and short sentences, as well as MSA and a few aspects of DA. They provided stand-off annotations on top of the Arabic Tree Bank ATB[a] (Part 1 Version 3), which is free of charge for users who subscribe to the LDC[b] since 2003.

2.2. Features and Methods

In a multi-language web forum at document level, Abbasi, et al., [2] proposed a system to be utilised for sentiment analysis tasks. This proposed system depends on an Entropy-Weighted Genetic Algorithm (EWGA) to choose the most relevant features., along with the SVM with linear kernel for sentiment classification. Their method attempts to determine an overlap between language-independent features, including syntactic and stylistic features. The syntactic features include POS (only utilised for the English language, not for Arabic). In order to evaluate the validity of their method, the authors measured the accuracy of the classifier by dividing the number of correctly classified documents by the total number of documents. In this study, however, a more accurate measurement methodology was required to assist in evaluating the method used in both classes. It was found that syntactic features achieved a higher result than the stylistic ones. When the two features were employed together using EWGA, the accuracy result increased to 93.6% (in the Middle Eastern forum domain).

The work of Rushdi-Saleh, et al., [7] focused on investigating two ML classifiers, Naive Bayes and Support Vector Machine. They also used two different weighting schemes (term frequency and term frequency-inverse document frequency) and three n-gram models. The effect of using the stem of the Arabic work was also investigated with different n-gram models. These researchers built their sentiment corpus by collecting approximately 500 movie reviews in Arabic sourced from different websites. They reported an accuracy of 90.6% when using the SVM with the tri-gram model, and with no stemming for document level classification. In addition, they claimed that there was no significant impact when using the TF or TF-ID as a weighting scheme. This makes logical sense because both schemes represent the count of the term over the total document. It may prove useful to compare the presence of the term, versus the term-frequency scheme in other studies.

El-Halees [5] proposed a combined classification approach in Arabic for document level polarity classification. His method applied three different classifiers grouped in a sequential manner: a lexicon-based classifier, a maximum entropy classifier and the K-Nearest Neighbour classifier. The results obtained from one classifier was used as 'training' data for the next classifier. The text was manipulated by removing the stop words, prior to using the first classifier. A few Arabic letters were normalised, and some misspelled words were corrected. A simple stemmer was utilised in this study to generate the stem of Arabic words and the TF-IDF was used as the term-weighting scheme. The F-measure was used as the evaluation metric. The F-measure measurement was between 75% and 84%, depending on the domain of the data. The average of the F-measure was also calculated; this was 82% for the positive documents and 78% for the negatives. A major issue for the purpose of this study was that there were not any more available features that could be added to the classifier that could assist in increasing the classifier's performance and accuracy.

Other studies have attempted to investigate the linguistic features of Arabic, and to combine these with an ML classifier in order to conduct a sentiment analysis. One such study attempted to analyse the grammatical structure of Arabic [10]. This work attempts to analyse sentiment, firstly, at the sentence level, followed by using these results to analyse sentiment at the document level. At the sentence level, two different approaches were compared. The first approach was generalising an Arabic sentence into a general structure that contains both the actor and the action. The second approach utilised semantic and stylistic features. Different classifiers were used for each different approach. The SVM was used for the grammatical classifier, and obtained an accuracy of 89%, while the J48 decision tree was used with the semantic approach, and achieved an accuracy

---

[a] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2005T02
[b] http://www.ldc.upenn.edu/

of 80%, (when the semantic orientation of the words was extracted and manually assigned), and 62% when a dictionary was used.

Another study, which investigated the effect of language-independent and Arabic-specific features on the performance of the classifier, was undertaken by Abdul-Mageed, et al., [4]. The researchers conducted two kinds of sentence level sentiment analysis with two different domains: news and social media domains. The SVM was used to classify both the subjectivity and polarity of sentences with different features, including N-gram, adjective features and a unique feature. All the words occurring fewer than four times were replaced with the token "UNIQUE" (MSA morphological features (person, gender and number)). Different stemming and lemmatization settings with dissimilar types of independent language and Modern Standard Arabic morphology features were used. An F1 result of 72% was achieved for subjectivity, and 96% for the polarity. This was with the stem, the morphology setting, and ADJ features using the newswire domain as the context for study. In SAMAR [9], the effect that the standard and the genre-specific features had on the subjectivity and sentiment classification of the Arabic social media domain was studied.

3. Proposed Method

'Word clustering' is a process used to distribute words that have the same semantic or syntactic relationship within the same group. After the clustering process is complete, the cluster label of the word is used as a feature. This feature achieves improved performance when undertaking different natural language processing (NLP) tasks, such as Name Entity Recognition [11]. This process may also support the classifier in capturing the similarity between words, along with the sentiment orientation of words. In addition, the process may also prove useful in the case of Dialect Arabic, when there is a lack of morphology tools that are adequate to manage this form of Arabic language.

In order to conduct the proposed method, words were grouped into different clusters. The first step is to use the suitable clustering algorithm that works well with the Arabic language. Algorithm options available are reduction dimensionality [12, 13], distributed word embedding [14], and the Brown clustering words algorithm. This last algorithm has been used as a standard technique in solving many NLP problems [15]. Thus, this clustering will be used in the current study due to its simplicity, and the hierarchical nature of its output, along with its easy availability and implementation.

3.1. Word Clustering Using the Brown Clustering Algorithm

This clustering algorithm is considered as a class-based bi-gram language model. It works by maximizing the mutual information of adjacent clusters [16, 15]. The main methodology of the cluster is achieved by grouping words together that have the same distribution as 'neighbour' words. The Brown Clustering algorithm clusters the words depending on their context within the same data set. A word is selected and the algorithm then computes the probabilities of this word occurring within a similar context. For example, in the following cluster the probability of the distribution of neighbour words such as Jeddah, with similar words, such as Denver, for example. The 'inference' of the algorithm is based upon the two words in this example being the names of two cities. Also, the clustering algorithm assumes that the context of these two words will also be similar. It could also be assumed that the sentimental words might also appear within the same context. As a result of this, the Brown Clustering Algorithm would cluster these word together in one group.

The Brown Clustering algorithm is a word cluster-based approach that takes a sequence of words (w_1, w_2, ..., w_n) as an input, and then generates the cluster of these words as a binary tree. The 'leaf' of the tree contains the words, and the internal nodes represents the cluster bit string. An example of an output from this clustering technique is shown in Fig 1.

For example, suppose a batch of data is required to be divided into 50 cluster groups. At the end of the cluster algorithm process, 50 cluster names will be generated, with the output contained within the 'leaf' of the tree. Each group may contain one or more words. Following this, the clusters are grouped into one standard

upper cluster (in the binary manner). This process continues to generate outputs until reaching to the 'root'. In Fig 1, the words (bought and purchased) are grouped in one cluster. Their cluster tag is (100). This tag is called a bit string ID. This ID begins from the root and ends at the leaf. The sibling cluster (101) contains two words: 'run' and 'drive'. From analysing this cluster, it may be inferred that the verbs in group (100) are synonyms with 'buying'. In the other cluster (1010), the meanings of the words are obviously different from those contained within the (100) cluster. In the case of the upper cluster tag (internal node), which is cluster 10, the words are those in all sub-clusters which belong to the 'parent' cluster. In this case, all words in cluster 100 and cluster 101. This may also assist to preserve some of the syntactic or semantic features of these words. All of these words are verbs. More details of the specifications of this algorithm are presented in [16]. The bit string ID of a word cluster will be used to input the information from the cluster into the space vector model. The following section demonstrates how this may be applied when conducting an Arabic sentiment analysis.

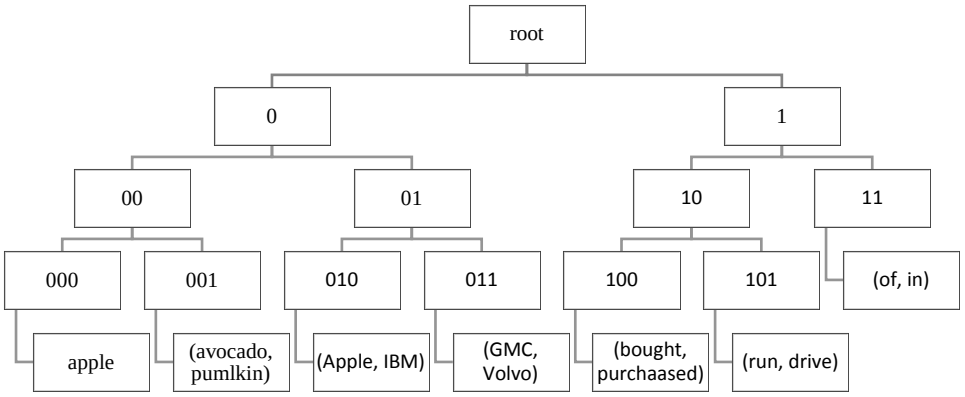


Fig.1. An example of the output from the Brown Word Clustering Algorithm

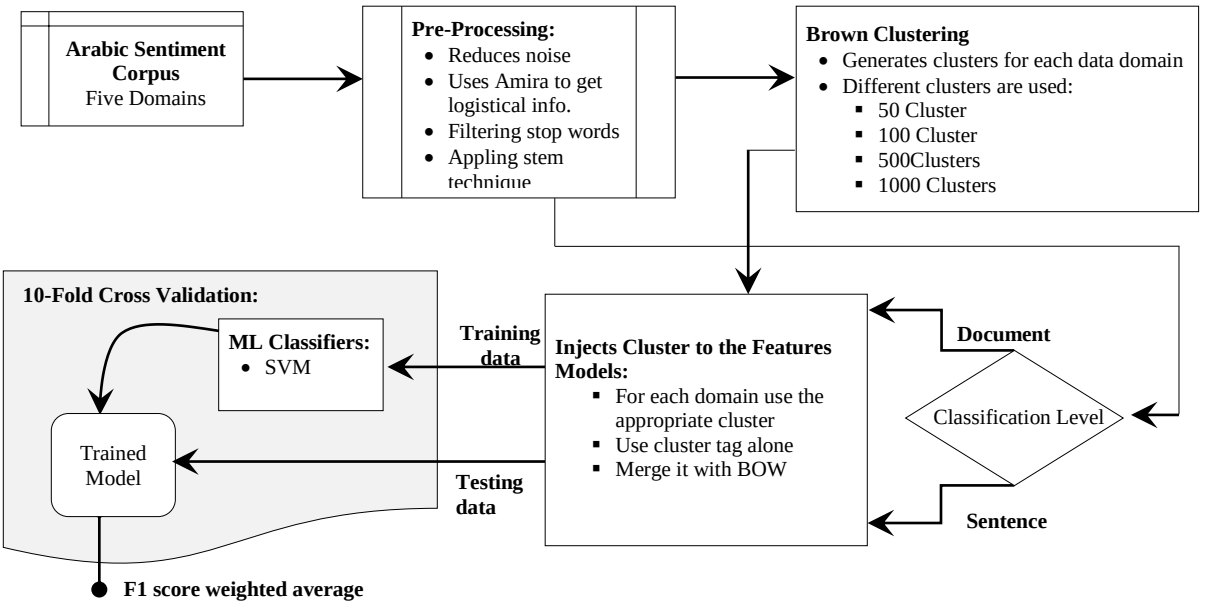


Fig.2. Word Cluster Technique for use in conducting an Arabic Sentiment Analysis

3.2. Injecting Words Clustering with a Feature Model

Fig 2 outlines the steps taken in this experiment. In the first step, the clustering algorithmc processes all data in order to group words into different clusters. Following this, there will be a database of all of the words of the sentiment corpus, along with their cluster tag. The cluster tag indicates which cluster group that a particular word is contained within. There are four different cluster groups for each data domain in our corpus. In the first step to this feature, there is a requirement to create a cluster from the given text. Each of the data domains are processed by the Brown cluster algorithm into four cluster numbers: (50, 100, 500, 1000). The typical cluster number that is used in research is the 1000 numbered cluster [11, 15, 17]. A small number of clusters are used in order to investigate the effect of that in the different types of domains.

The second step in the proposed features determines how this information may be utilized when conducting an Arabic sentiment analysis. The first approach proposed is to use the cluster tag itself as a feature to build a viable feature model. This means that the cluster label is relied upon to develop the feature model. The second method is to inject this feature with the standard BOW model that is the baseline of our experiment. A word is attached to its cluster in the feature model, based on the same methodology as utilized by the POS feature. The last method is to combine the first model with the BOW model.

4. Experiments Setup

This section outlines the experiments which are conducted to investigate and test the features and performance of an ML classifier in an Arabic sentiment analysis. The first part describes the data that has been used. The second part discusses the process that was performed in order to test the proposed methods.

4.1. Arabic Sentiment Corpus

A corpus was developed solely for the purpose of this study. This was required due to the scarcity of sentiment Arabic corpus available. The research corpus was structured and developed from five different genres: news, news reviews, user market reviews, restaurants reviews, and movie reviews. The news data was sourced from the Sabqd website, and from among different domains: local, sport, economics, technology, and social news. The reviews of the news have also been sourced from the same website where individuals can add their comments and feelings about news topics. The Souqe (considered as the “Amazon market place” for Arab countries) was used as the source for market reviews. The restaurant reviews have been sourced from the work of [3], which encapsulates the personal viewpoint of the user concerning different restaurantsf. The movie reviews were taken from a movie review websiteg and is used in [10]. In total, our corpus contains 6268 documents, with more than 33000 sentences. Approximately 7674 positive sentences, 9202 negative sentences, and 3351 neutral sentences were identified.

Two individuals who have been educated in Arabic were chosen to annotate the data. Each annotator was provided with specific guidelines. Firstly, they should determine if the document is subjective or objective in nature. Secondly, they had to establish the polarity of the subjective text among three categories: whether positive, neutral, or negative. Thirdly, the annotator must analyse each sentence in the document, noting its polarity if it is subjective, otherwise the sentence should be determined as being objective. The first step was to train the two annotators, who were then requested to work on the same data-set which contained approximately

---

[c] We relied on the implementation of Liang [15] for the Brown clustering algorithm
[d] http://sabq.org
[e] http://saudi.souq.com/sa-ar
[f] http://www.qaym.com
[g] http://www.filfan.com

33% of the sentences. During this process, the inter-annotator agreement between them was calculated using the Kappa coefficient [18], The result was between 0.72 and 0.84. (If you wish to obtain a copy of these data-sets, please contact the corresponding author)

4.2. Classification Process

The pre-processing phase contains steps that should be undertaken prior to the text being passed on to the classifier. The first step includes the filtering of any irrelevant data that may be found within the text, including single letters or non-Arabic characters. The second step is to normalise any lengthy words that may make some letters redundant. The third step is to use the AMIRA [19] tool kit on all the data in order to develop the speech tag component of the words. The final step involves removing the 'stop' word lists, and modifying them, so as to manage these while constructing the vector space model that represents these words. The stop word lists in [20] were then used. To evaluate the method, a number of experiments were undertaken using a support vector machine (SVM) classifier. This model had linear kernel with 10-fold cross-validation using the Scikit-Learning Library [21]. In this experiment, the data set was divided into ten distinct parts, with equal proportions of samples within each class. Nine of these were used to train the classifier, with the remainder being used to test the model that is generated during the training process. This process will be repeated ten times, as there is ten components of the data set. In each cycle, a new component is used for the testing phase. During every cycle, the F1 metric was calculated. This measures the accuracy of the classifier after calculating his/her precision and recall scores. Default parameters were used in the SVM that comes with the Scikit-Learning Tool, as it was found that these parameters were congruent with the data set in this experiment.

The goal of this experiment was to evaluate whether using the word cluster tag added any sentimental knowledge to the classifier in the context of using Arabic. The experiment was also designed to determine whether the same cluster group has the same sentimental words contained therein. The classification process was performed on different types and levels that included subjectivity, and polarity classification, as well as at the document and sentence level. The subjectivity classification aimed to determine the subjective content of the document. The second step determines the polarity of the text that exhibits whether the text has positive or negative connotations.

5. Results Discussion

The following section discusses the experiments performed, and the results that were obtained, in order to provide an evaluation of the effect of word cluster techniques conducted in performing an Arabic sentiment analysis. This is achieved by comparing the different approaches related to the concept. Each of the experiments are classified at the document level into two different types: subjectivity (subjective and objective), polarity (positive and negative).

5.1. Experimentation using Word Cluster only

Tables 1, and 2 display the experiment of using a cluster method during an Arabic sentiment analysis. The first table shows the results of classification at the document level , and the second at the sentence level. This experiment uses the cluster ID (the 'bit string ID' that was explained in *Section 2.1*) of the words as a feature in order to construct a feature model. Different cluster groups were compared in order to find the best cluster group that might be congruent when conducting a sentiment analysis. The BOW: (Bag of Words) was used as a baseline model to evaluate the performance of the clustering approach. The numbers in boldface within these tables illustrate the best results that were achieved using a particular feature model setting. The SVM classifier is only used to evaluate the effect of the cluster idea. The "NA" symbol in the Table 1 shows that the classification process is not applicable, as there is no objective document within the 'movie review' domain. The sentiment analysis was performed in two different classification types, which were subjectivity and

polarity, as shown by the accuracy of F1 in Tables 1 and 2.

Table 1. Results Using Cluster ID Alone in the Feature Vector Model at Document Level of Classification

		Subjectivity	Polarity
News Reviews	BOW	**88%**	**56%**
	50 Clusters	75%	51%
	100 Clusters	74%	50%
	500 Clusters	79%	55%
	1000 Clusters	81%	54%
Restaurant Reviews	BOW	**96%**	**85%**
	50 Clusters	77%	69%
	100 Clusters	79%	74%
	500 Clusters	91%	74%
	1000 Clusters	93%	77%
Market Reviews	BOW	**93%**	**90%**
	50 Clusters	83%	80%
	100 Clusters	97%	81%
	500 Clusters	89%	87%
	1000 Clusters	90%	87%
Movie Reviews	BOW	NA	**80%**
	50 Clusters	NA	76%
	100 Clusters	NA	71%
	500 Clusters	NA	74%
	1000 Clusters	NA	75%
News	BOW	**63%**	76%
	50 Clusters	45%	71%
	100 Clusters	47%	77%
	500 Clusters	59%	70%
	1000 Clusters	58%	**79%**

Table 2. Results Using Cluster ID Alone in the Feature Vector Model at Sentence Level of Classification

		Subjectivity	Polarity
News Reviews	BOW	**69%**	**58%**
	50 Clusters	44%	50%
	100 Clusters	50%	49%
	500 Clusters	57%	53%
	1000 Clusters	62%	52%
Restaurant Reviews	BOW	**71%**	**83%**
	50 Clusters	55%	55%
	100 Clusters	61%	60%
	500 Clusters	69%	71%
	1000 Clusters	70%	71%
Market Reviews	BOW	**89%**	**88%**
	50 Clusters	80%	77%
	100 Clusters	78%	80%
	500 Clusters	84%	84%
	1000 Clusters	85%	86%
Movie Reviews	BOW	45%	**80%**
	50 Clusters	41%	70%
	100 Clusters	46%	78%
	500 Clusters	**47%**	73%
	1000 Clusters	46%	71%
News	BOW	35%	**80%**
	50 Clusters	36%	61%
	100 Clusters	36%	63%
	500 Clusters	35%	69%
	1000 Clusters	**37%**	69%

Using the cluster ID of words as a feature is not overly useful in most cases when conducting a sentiment analysis when Arabic is used. What is clear is that the BOW baseline feature achieved the best results when compared with the other cluster configurations. For example, the best results were achieved using the BOW model in subjectivity classification for all the domains of the dataset. The F1 score decreased by more than 10% when only the cluster ID was used to construct the features model. However, there are some benefits in using the cluster that gave confidence that by making improvements to the method, it may exhibit improved validity as a method.

Improvements were identified on the two sides in Table 1. The first one relates to increasing the performance of the classifier in one domain, that is Newswire. The F1 score improved by 3% in the case of polarity classification in movie reviews. This infers that the cluster may play a role in the polarity classification process, and may also preserve the sentimental orientation across the different cluster groups. The second direction of improvement was identified with the increase of the F1 score after increasing the number of cluster groups. The F1 score of the 50 cluster group is particularly low when comparing with the BOW. However, by adding more cluster groups, the F1score improved noticeably. For example, the cluster of the 50 grouping achieved a 77% F1 score, then it increased until it reached 93% (when using 1000 clusters during the subjectivity classification process within the restaurant review domain). These two improvements encouraged further investigation by combining the cluster method with the BOW model.

The same approach that was applied when achieving the results in Table 1 was also applied to achieve the results. however, these results are at the sentence-level of classification. These results illustrate similar findings that were recorded at the document level of classification. The performance did not significantly improve in most cases, except that it improved by 2% with the subjectivity classification in the movie reviews and news domains. The trend of improvement with different cluster groups is the same that was identified at the document level of classification. Therefore, the following experiment investigates using the cluster method in a different manner.

5.2. Experimentation using Word Cluster combined with BOW

Tables 3 and 4 show the results of the document classification process using the enhancement cluster approach. For the previous experiment, it was noted that the 1000 clusters achieved the best results. Therefore, this cluster was the only one considered to improve the process of experimentation. An attempt was made to merge the cluster ID of the word with the word itself, as a method of adding the component of speech tag (POS) to the word [22]. It was then compared using this feature combined with the BOW baseline model. (The numbers in bold are the best results).

Table 3. Results Using Cluster ID with BOW Model at Document Level of Classification

		Subjectivity	Polarity
News Reviews	BOW	**88%**	**56%**
	With 1000 Clusters	**88%**	**56%**
Restaurant Reviews	BOW	**96%**	**85%**
	With 1000 Clusters	**96%**	84%
Market Reviews	BOW	93%	90%
	With 1000 Clusters	**94%**	**92%**
Movie Reviews	BOW	NA	80%
	With 1000 Clusters	NA	**81%**
News	BOW	**63%**	76%
	With 1000 Clusters	**63%**	**77%**

Table 4. Results Using Cluster ID with BOW Model at Sentence Level of Classification

		Subjectivity	Polarity
News Reviews	BOW	**69%**	**58%**
	With 1000 Clusters	**69%**	**58%**
Restaurant Reviews	BOW	**71%**	**83%**
	With 1000 Clusters	70%	82%
Market Reviews	BOW	**89%**	88%
	With 1000 Clusters	**89%**	**91%**
Movie Reviews	BOW	**45%**	80%
	With 1000 Clusters	**45%**	**81%**
News	BOW	35%	**80%**
	With 1000 Clusters	**36%**	**80%**

Tables 3 and 4 display the results of using the newly enhanced method, in the second row for each dataset. The record is tagged "With 1000 Cluster". The 1000 numbered clusters were used to perform this method due to the positive outcomes noted in previous experiments. It was found that the newly enhanced model cluster feature achieved a similar or improved result. For example, the F1 score increased by 3% in the polarity classification with the domain of market reviews (Table 4). It was also noted that adding the cluster of 'knowledge' does not affect the performance of the classifier when compared with the baseline score, except in one case which was the polarity classification within the restaurant reviews.

The only issue that was identified with the current method was noted to occur within the restaurant review domain. The result of the polarity classification process was not improved by merging the cluster ID with the word. From this it may be inferred that the domain of 'restaurant review' has a greater overlap between the words than occurs in the other domains. In addition, the cluster was unable to preserve the sentimental orientation of words within the same cluster. Table 4.25 represents examples of the words found within the restaurant reviews, and this also shows that there are a number of different sentimental words within the same cluster. Most of the words in this cluster have a negative connotation, such as (سيء/syyˆ/ bad)[h]. Other, such as (هادئ /hAdy/ quiet) carry a positive meaning, but are actually located within a cluster that contains mostly negative words. This anomaly may affect the classifier when using the 'opposed feature model', and within the restaurant review domain.

Table 5. Some of the Words in Same Clusters of Restaurant Reviews Domain

Cluster ID	Word
1010101111011	(هبطت/ *hbTt* / landed)
1010101111011	(انخفضت / *InxfDt* / decreased)
1010101111011	(سيئ / *syy'*/ bad)
1010101111011	(مزحوم / *mzHwm* / crowded)
1010101111011	(هادئ / *hAdy* / quiet)
1010101111011	(متواضع/ *mtwADς* / humble)

[h] Arabic sentence is represented in three variants: (Arabic sentence / transliteration scheme / English translation) as in [23]

6. Conclusion and Future Work

This investigation determined that by taking the external knowledge of a word cluster into account while analysing for sentiment content within an Arabic text, this may assist and improve the performance of the classifier using a machine-based learning algorithm. This article described and comes developed proposals by investigating two approaches. The results that were achieved in this study demonstrate the potential gain that might be obtained through the inclusion of word clustering as a feature. The results that were achieved during the process of classification were promising. It is intended to continue to further develop and investigate this method in the future.

There are many different directions to take in this field in order to continue the work on word clustering when conducting a sentiment classification on Arabic texts. One direction relates to the method of adding a cluster to the model. Rather than adding a cluster tag to the words, as the approach used in previous experiments (Section 5.2), we plan to use cluster ID to add to the model. This is hoped to merge the BOW model with the model that was proposed in this paper (as outlined in Section 5.1). Working with more clusters may also be another direction for future studies. For example, the words could be clustered into more than 1000 numbered cluster, then the effect of using different high cluster groups could be investigated. Merging the cluster method with other features, such as the POS, could also be an option for further study. This may add more information to the clustering method, and assist to distinguish between the type of words contained within the same cluster. Another direction would involve investigating the process of the clustering from the beginning. In this study, the word clustering algorithm was applied on the data in sequence (i.e. domain by domain). This efficiency of this method may be improved by entering all of the data at once into the word clustering algorithm, or, alternatively, to locate a large Arabic corpus. Then, by applying the word clustering to the corpus, the outputs would be used to conduct an analysis for sentiment within Arabic texts.

References

[1] Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," vol. 8, no. 4, pp. 14:1–14:22, Dec. 2009.

[2] Abbasi, H. Chen, and A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums," *ACM Trans. Inf. Syst.*, vol. 26, no. 3, pp. 12:1–12:34, Jun. 2008.

[3] A. Al-Subaihin, H. S. Al-Khalifa, and A. S. Al-Salman, "A Proposed Sentiment Analysis Tool for Modern Arabic Using Human-Based Computing," in *Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services*, ser. iiWAS '11. New York, NY, USA: ACM, 2011, pp. 543–546.

[4] M. Abdul-Mageed, M. T. Diab, and M. Korayem, "Subjectivity and Sentiment Analysis of Modern Standard Arabic," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, ser. HLT '11. Strouds- burg, PA, USA: Association for Computational Linguistics, 2011, pp. 587–591.

[5] A. El-Halees, "Arabic opinion mining using combined classification approach," in *Proceeding The International Arab Conference On Information Technology, Azrqa, Jordan*, 2011.

[6] Habernal, T. Ptacek, and J. Steinberger, "Supervised sentiment analysis in czech social media," *Inf. Process. Manage.*, vol. 50, no. 5, pp. 693–707, Sep. 2014.

[7] M. Rushdi-Saleh, M. Martin-Valdivia, L. Urena-Lopez, and J. Perea-Ortega. OCA: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, volume 62(10): pages 2045–2054, 2011.

[8] M. Abdul-Mageed and M. Diab. AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. *In Proceedings of the Eight International Conference on Language Resources and*

Evaluation (LREC'12), pages 19–28, Istanbul, Turkey, may 2012.

[9] M. Abdul-Mageed, S. Kubler, and M. Diab. Samar: A system for subjectivity and sentiment analysis of arabic social media. *In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 19–28. *Association for Computational Linguistics*, 2012.

[10] N. Farra, E. Challita, R. A. Assi, and H. Hajj, "Sentence- Level and Document-Level Sentiment Mining for Arabic Texts," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, dec. 2010, pp. 1114 –1119.

[11] M. Tkachenko and A. Simanovsky, "Named Entity Recognition: Exploring features," in *Proceedings of KONVENS 2012*, J.Jancsary, Ed.OGAI, September 2012, pp.118– 127, main track: oral presentations.

[12] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 160–167.

[13] A. Mnih and G. Hinton, "A Scalable Hierarchical Dis- tributed Language Model," in *Advances in Neural Infor- mation Processing Systems*, vol. 21, 2008.

[14] M. Lamar, Y. Maron, M. Johnson, and E. Bi- enenstock, "Svd and clustering for unsupervised pos tagging," in *Proceedings of the ACL 2010 Conference Short Papers*, ser. ACLShort '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 215–219. [Online].

[15] P. Liang, "Semi-supervised learning for natural lan- guage," in *MASTER THESIS, MIT*, 2005.

[16] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Comput. Linguist.*, vol. 18, no. 4, pp. 467–479, Dec. 1992. [Online].

[17] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proceedings of the Thirteenth Conference on Com- putational Natural Language Learning*, ser. CoNLL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 147–155.

[18] Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Comput. Linguist.*, vol. 22, no. 2, pp. 249–254, Jun. 1996.

[19] M. Diab, "Second generation tools (AMIRA 2.0): Fast and robust tokenization, pos tagging, and base phrase chunking," in *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, K. Choukri and B. Maegaard, Eds. Cairo, Egypt: The MEDAR Consortium, April 2009, pp. 285–288.

[20] A. El-Khair, "Effects of stop words elimination for arabic information retrieval: a comparative study," *Inter- national Journal of Computing & Information Sciences*, vol. 4, no. 3, pp. 119–133, 2006.

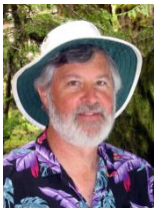
[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cour- napeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008. [Online].

[23] N. Habash, A. Soudi, and T. Buckwalter. On Arabic Transliteration. In A. Soudi, A. d. Bosch, and G. Neumann, editors, Arabic Computational Morphology, volume 38 of Text, Speech and Language Technology, pages 15–22. Springer Netherlands, 2007.

Authors' Profiles

Saud Alotaibi received his Bachelor of Computer Science degree from King Abdul Aziz University, 2000. He worked as an Assistant Lecturer at Umm Alqura University, Makkah, Saudi Arabia (Jan, 2001). He also earned his Master of Computer Science degree from King Fahd University, Dhahran, May 2008. Following this, Saud worked as the Deputy of the IT-Centre for E-Government Affairs in Jan 2009, (Umm Alqura University). In 2015, Saud completed his Ph.D. degree in Computer Science from Colorado State University, Fort Collins, US. Currently, he works at Umm Alqura University, as an Assistant Professor.



Charles Anderson is a professor of computer science at Colorado State University. He is also a faculty member of CSU's School of Biomedical Engineering, Graduate Degree Program in Ecology and the Molecular, Cellular, and Integrative Neurosciences Program. He graduated with a Ph.D. in computer science from the University of Massachusetts, Amherst, in 1986, and worked at GTE Laboratories in Waltham, MA, until he arrived at CSU in 1991. He teaches graduate courses in machine learning and undergraduate courses in programming, data structures, and graphics. His research is in machine learning with a focus on reinforcement learning, EEG pattern recognition, and neural networks.

How to cite this paper: Saud Alotaibi, Charles Anderson,"Word Clustering as a Feature for Arabic Sentiment Classification", International Journal of Education and Management Engineering(IJEME), Vol.7, No.1, pp.1-13, 2017.DOI: 10.5815/ijeme.2017.01.01