Arabic Opinion Mining Using Combined Classification Approach

Alaa El-Halees

ARABIC OPINION MINING USING COMBINED CLASSIFICATION APPROACH

Alaa El-Halees
Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestine
alhalees@iugaza.edu.ps

Abstract:

In this paper, we present a combined approach that automatically extracts opinions from Arabic documents. Most research efforts in the area of opinion mining deal with English texts and little work with Arabic text. Unlike English, from our experiments, we found that using only one method on Arabic opinioned documents produce a poor performance. So, we used a combined approach that consists of three methods. At the beginning, lexicon based method is used to classify as much documents as possible. The resultant classified documents used as training set for maximum entropy method which subsequently classifies some other documents. Finally, k-nearest method used the classified documents from lexicon based method and maximum entropy as training set and classifies the rest of the documents. Our experiments showed that in average, the accuracy moved (almost) from 50% when using only lexicon based method to 60% when used lexicon based method and maximum entropy together, to 80% when using the three combined methods.

Keywords*: Opinion Mining, Sentiment Classification, Combined Classification, Arabic Opinion Mining.*

---

1. INTRODUCTION

People increasingly participate to express their opinions on the Web. That makes researchers more interest in mining opinions from sources such as product reviews, discussion forums and personal blogs, which are collectively called *user-generated contents [1]*. *User-generated content*s are written in natural language with unstructured-free-texts scheme. Manually scanning through large amounts of user- generated contains is time consuming and sometime impossible. In this case, opinion mining is better alternative. Opinion mining is a subtask of text mining that automatically extract knowledge from the various user-generated contains [2]. It has wide range of applications include**:** product reviews, advertising systems, market research, public relations, financial modeling and many others [3].

Most research efforts in the area of opinion mining deal with English texts. Some new works deal with other languages, but in Arabic, which is a language for Millions of people, their is a little work in this area. Arabic is a challenging language for a number of reasons. It has a very complex morphology as compare to English language. This is due to the unique nature of Arabic language. Arabic language is a highly inflectional and derivational language which makes monophonically analysis a very complex task [4, 5]. For example, one word may have more than lexical category in different contexts. In case of user-generated contains, it brings another complexity since most writer express their opinion using local accent instead of standard Arabic language. So, we end with many written accents instead of one formal language. Also, many times writers misspelled the words either by accident or deliberately (e.g. for short).

Opinion mining studies opinions at three different levels: word level, sentence level and document level [6]. In this research we will concentrate at document level, which is the most common one. It is mostly applied to documents, where systems assign positive or negative sentiment for a whole document [7]. Many approaches have been used in opinion mining the

most common ones are lexicon based and machine learning. In lexicon the simplest representation of a text is the bag-of-words approach. Opinion lexicons are resources that associate sentiment orientation and words. It considers a document as a collection of words without considering any of the relations between the individual words. In this approach positive opinion words are used to express desired states while negative opinion words are used to express undesired states [8]. The drawback of this method is that a word that is considered to be positive in one situation may be considered negative in another. Another approach is machine learning which uses classification methods to classify a document as positive or negative. Pang's researches in [2] indicate that standard machine learning methods perform very well, even definitively outperform human classifiers. But it requires an annotated corpus to train a classifier which is not easy to obtain from Arabic corpus. Prabowo and Thelwall in [9] used multiple classifiers in a hybrid manner; the procedure is that if one classifier fails to classify a document, the classifier will pass the document onto the next classifier, until the document is classified.

This paper used a new approach that combined the lexicon based method and machine learning methods. It passes the document from lexicon based method to two classifiers, maximum entropy and k-nearest. The justification of that is using only one approach produces a poor performance. In addition after applying lexicon based method, the classified documents are used as training set for machine learning methods. Then maximum entropy produces accurate results if they can classify the document, using another classifier, k-nearest, will classify the others.

To evaluate our approach we collected three Arabic datasets from three different domains: Education, Sports, and Politics forums.

The rest of the paper is structured as follows: section two for related work, section three contains opinion classification, section four is experimental setup, section five gives the results of experiments and section six concludes the paper.

2. RELATED WORK

In publications, we found three works that mentioned the idea of Arabic opinion mining. First, Almas and Ahmad in [10] used computational linguistics for Arabic, Urdu and English languages. They described a method for automatically extracting specialist terms they called it local grammar. However, in their work they only used financial news data. Also, however they have an acceptable precision of (88.1 %) the performance of the method is very low especially for the recall which is about (17.2%). Second, Abbasi et. al. in [11], proposed sentiment analysis methodologies for classification of web forum opinions in multiple languages, namely from Arabic and English. They used specific feature extraction components that integrated to account for the linguistic characteristics of Arabic. They only classify sentiments relating to hate and extremist groups' forums. They have a very good accuracy which is about (93.62%). However, there domain is very specific since hate and extremist vocabulary is limited and it is not hard to distinguish positive and negative words. Also, they did not use any preprocessing stage which is crucial for Arabic language. Third, Elhawary and Elfeky in [12] showed how to extract the business reviews scattered on the web written in the Arabic language. The mined reviews are analyzed and provided their sentiments. They used Arabic Similarity Graph which is lexicon based method. No evaluation has been made by this work.

3. OPINION CLASSIFICATIONS

Our work is based on document-level sentiment classification. The problem of document-level sentiment classification can be formulized as: Given a set of opinionated documents D, determines whether each document $d \in D$ expresses a positive or negative opinion. To do that, in this section, we present our combined classification approach by applying multiple classifiers in sequence, as depicted in figure 3.1. In the proposed approach, we first applied lexicon-based opinion classifier, then Maximum entropy method and finally k-nearest method.

3.1 LEXICON- BASED OPINION CLASSIFIER

Lexicon based opinion classifier uses opinion words and phrases to determine the sentiment orientation of the whole document. It tries to find out the words or phrases that indicate the sentiment, determine the orientation of the sentiment words or phrases (i.e. positive or negative), then classify the sentence. After that it can classify the whole document. To do that, it uses a dictionary of positive and negative words (e.g., love, hate) [13, 14]. In our work we manually constructed Arabic subjectivity word list from two main sources: SentiStrength project and online dictionary. SentiStrength from [15] employs several novel methods to simultaneously extract positive and negative sentiment strength from short informal electronic text. It uses a dictionary of sentiment words with associated strength measures. It developed through an initial set of 2,600 human-classified comments. In our work we translated the used world list to Arabic language and the same strength is used. To improve the list to be more applicable to Arabic words and phrases some unrelated words are omitted. Then, we used online dictionary to add another common Arabic words, some of the additional words is a synonym of the other words in the dictionary and some others are added manually.

This phase works as follows: It takes un-annotated documents (to be classified), identify all opinion words and phrases (using negations when needed). Then aggregate these words to give a sentiment (positive or negative) to the document. However, some documents did not appoint to any sentiment polarity which is the documents that does not have enough clear opinioned words.

3.2 MAXIMUM ENTROPY

The next phase in the proposed method is to use maximum entropy classifier. The documents that have been classified from the previous step will be used as training set for the classifier. The goal in this step is to classify as much documents as possible that remain form the previous step.

The maximum entropy model estimates probabilities based on the principle of making as few assumption as possible, other than the constrained imposed. The constraints are derived from training process which expresses a relationship between the binary features and the outcome [16] [17].

In opinion mining classification, *p(s,d)* is the probability of document *d* with sentiment *s* can be formulated as :

$$p(s,d) = \frac{1}{Z(d)} \exp\left(\sum_i \alpha_i F_j(s,d)\right) \quad (1)$$

Where *Z(d)* is a normalization function. And the parameter α_i must be learned by estimation. It can be estimated by an iterative way using algorithms such as Generalized Iterative Scaling (GIS) [18], Improved Iterative Scaling (IIS) [19], or L-BFGS Algorithm [20]. $F_j(s,d)$ is a feature function for f_j given document *d* and sentiment *s*, defined as:

$$F_j(s,d) = \begin{cases} 1 & \text{if } f_j \textit{ appears} \text{ in } d \text{ with} \\ & \text{sentiment } s \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

In our work, training set for maximum entropy is the results of lexicon based classifier. The unannotated data set is given to the maximum entropy probability systems. Given certain threshold (we used 0.75) if the sentiment greater than this probability document will be classified if not it will be unclassified document which will pass be to the next step.

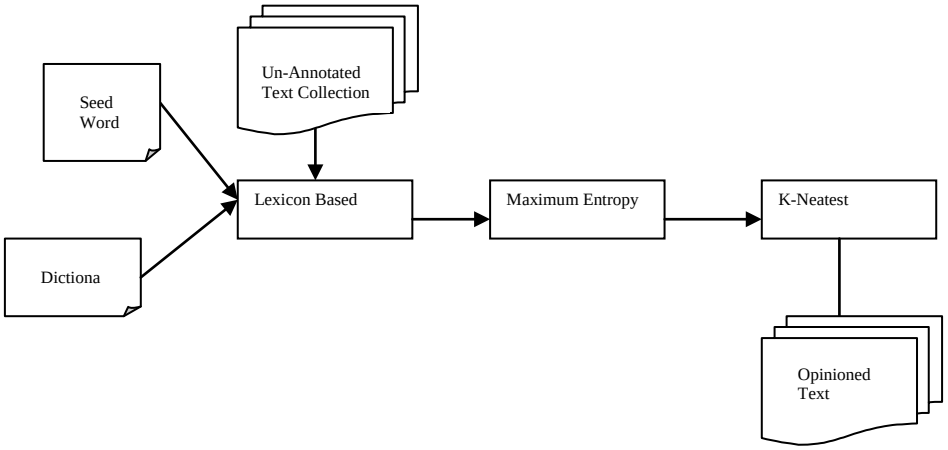

Figure 3.1 Opinion Classification Using Combined Approach

3.3 K-NEAREST NEIGHBOR CLASSIFIER

k- nearest neighbor (kNN) is a simple method to classify document [21]. In our proposed method, given an un-annotated document *d*, the system finds the k nearest neighbors among training documents which are classified in the previous two phases.
The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document. The weighted sum in kNN classification can be written as follows [22]:

$$score(d,s) = \sum_{d_j \in knn(d)} sim(d,d_j)\delta(d_j,s)$$

Where *knn(d)* is the set of k nearest neighbors of document *d*. if d_j belongs to sentiment *s*, $d(d_j,s)$ equal to 1, or otherwise 0. Document *d* belongs to the sentiment s that has the highest score.

4.0 EXPERIMENTAL SETUP

To evaluate our approach, a set of experiments was designed and conducted. In this section we describe the experiments design including the corpus, word list, the preprocessing stage, the used methods and evaluation metrics.

4.1 CORPUS

We collected documents related to opinions expressed in Arabic from three different domains: "education", "politics" and "sports" forum. As depicted in table 4.1, we used total of 1143 posts contain 8793 Arabic statements with average of 7.7 statements in each post.

Table 4.1 Description of Corpus Used in the Experiments

	Number of Files		Number of Statements	
Domain	**Positive**	**Negative**	**Positive**	**Negative**
Education	**204**	**170**	**1166**	**990**
Politics	**205**	**200**	**1829**	**2193**
Sports	**226**	**138**	**1380**	**935**
Total	**635**	**508**	**4375**	**4118**

4.2 WORDS LIST

The lexicon-based opinion classifier needed a word list, initially we used word list included in the SentiStrength software from [15] after translate it from English to Arabic. Advantage of using this list is that the words are scored with sentiment strength not just positive/negative polarity. Since the list is not complete, essential words have been added manually. In addition some unrelated words are deleted. Description of the used list is given in table 4.2.

	Number of Words	Percentage of Words
Positive	**415**	**43.73 %**
Negative	**534**	**56.27%**

Table 4.2 Description of Word List Used in Proposed Lexicon-Based Classifier

4.3 PREPROCESSING

After we collected the data associated with the three domains, we striped out the HTML tags and non-textual contents. Then, we separated the documents into posts and converted each post into a single file. For Arabic scripts, some alphabets have been normalized (e.g. the letters which have more than one form) and some repeated letters have been cancelled (that happens in discussion when the user wants to insist on some words), some of the wrong spelling words are corrected. After that, the sentences are tokenized, stop words removed and Arabic light stemmer applied. We obtained vector representations for the terms from their textual representations by performing TFIDF (Term Frequency–Inverse Document Ffrequency) weight which is a well known weight presentation of terms often used in text mining [23]. We also removed some terms with a low frequency of occurrence.

4.4 EVALUATION METRICS

There are various methods to determine effectiveness; however, accuracy, precision and recall are the most common in this field. Accuracy measures the percentage of the test set that the classifier has labeled correctly. Furthermore, the precision and recall are calculated. Precision is the percentage of predicted documents class that is correctly classified. Recall is the percentage of the total documents for the given class that are correctly classified. We also computed the F-measure, a combined metric that takes both precision and recall into consideration [24].

$$F - measure = \frac{2 * precision * recall}{precision + recall} \quad (4)$$

5.0 EXPERIMENTAL RESULTS

This section presents the results of experiments using the three different domains**.** Evaluation of opinion classification relies on a comparison of results on the same corpus annotated by humans [25].

Therefore, to evaluate our approach, first we manually assigned a label for each user subjective opinion. First, we evaluated the accuracy of the data sets using one classifier. Table 5.1 gives the accuracy of the three domains with methods which are usually used in English opinion mining which are: lexicon based method (Lex), Maximum Entropy (ME), K-nearest neighbor (kNN), Naïve Bayses (NB), and support vector machine (SVM)**.**

	Lex	ME	kNN	NB	SVM
Politics	47.52	43.93	59.09	53.03	48.48
Sports	44.69	67.74	64.52	45.16	58.06
Education	70.43	84.44	68.89	40.00	55.56
Total	50.08	63.00	63.58	46.820	53.75

Table 5.1: Accuracy of Arabic Opinion

Mining Using Various Methods

	Lex	Lex + ME	Lex +ME+ kNN
Politics	47.52	59.72	75.24
Sports	44.69	48.351	81.31
Education	70.43	7.323	84.34
Total	50.08	60.73	80.29

Table 5.2 Accuracy of the Domain When Using One, Two and Three Combined Methods

	Lex			ME			kNN	
	CC	ICC	NC	CC	ICC	NC	CC	ICC
Politics	47.52	10.39	57.91	63.86	15.84	20.3	75.24	37.24
Sports	32.41	7.69	59.9	48.35	10.43	41.22	81.31	18.69
Education	66.16	7.57	26.27	73.23	10.10	16.67	84.34	15.65
Average	48.69	8.55	48.02	61.81	12.12	45.86	80.29	23.86

Table 5.3 Details of Domains Opinions Classification

From the table we can notice that the accuracy of most methods in most domains is low (with exception of Maximum Entropy on education domain). That is mainly because of the complexity of the Arabic language. This is support our suggestion that it is better to use combined methods to classify Arabic documents. Table 5.2 gives the accuracy of applying combined methods. For example, in political domain, when we used lexicon based method only the accuracy of the used test dataset was 47.52%, when applied lexicon based and Maximum Entropy the accuracy increased to 59.72% and when we used the proposed method which a combination of lexicon based approach, maximum entropy and k-nearest the accuracy increased to 75.24%.

It is noticed from the table the clear increase in accuracy as new method is applied. In average it went from 50% when using one method to 60% using two methods and 80% using three methods.

Table 5.3 describes in details the results of each phase. For example in politics domain, after using lexicon based method, it correctly classified (CC) 47.52% of the documents, 10.39% incorrectly classified (ICC) and 57.91% not classified (NC) at all. When using Maximum Entropy 63.86% of politics domain document are correctly classified, 15.84% incorrectly classified and 20.3% not classified. After applying the last method, k-nearest neighbors in the not classified documents, 75.24% correctly classified and 37.24% incorrectly classified.

	Recall		**Precision**		**F-Measure**	
	Pos	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**
Politics	69.15	82.10	81.31	70.27	74.73	75.72
Sports	89.62	69.73	80.50	82.81	84.81	75.70
Education	89.32	78.94	82.14	87.20	85.57	82.86
Average	**82.69**	**76.92**	**81.31**	**80.09**	**81.70**	**78.09**

Table 5.4 The Measures of Recall, Precision and F-Measure for Positive and Negative Documents

From the table we can notice that majority of the education documents classified in the first method, many of the politics domain classified in the second method and many of the sports domain classified in the third method. This is the main advantage of the proposed method where the accuracy of the method is depends on the domain. So, using multiple methods increases the overall accuracy.

In addition to accuracy measure, we used measure of recall, precision and f-measure. Table 5.4 gives the measures of recall, precision and f-measure for the three domains in both the positive and negative polarities.
We can notice that the politics domain has high recall for negative, but low recall for positive.

However, the opposite for precision; where it has high for positive and low for negative. In sports domain, it has high recall for positive and low recall for negative. It has high precision for both negative and positive. In education domain it has an acceptable recall and precision on for both negative and positive documents. However, in average positive documents has f-measure better than negative documents. That is because negation gives more complication to the statements especially in Arabic language.

6. CONCLUSION

We have proposed a combined approach which aims at mining opinions from Arabic documents. The approach used three methods at sequence: First, lexicon based method is used which classifies some documents. The classified documents used as training set for maximum entropy model which subsequently classified some other documents. After that, k-nearest model is used to classify the rest of the documents.

In experiments with 1143 posts contains 8793 Arabic Statements, our system achieved an accuracy of 80.29%. The accuracy almost went from 50% using one method, 60% using two method and 80% using three methods which is a satisfactory performance especially for complex language such as Arabic. The experimental results further show that recall and precision of positive documents are better than the negative one. That means further studies should be done for mining from negation of Arabic statements. Also, in the future, we plan to extend our work to be able to extract features from Arabic opinioned statements.

REFERENCES:

[1] Bing Liu., "Searching Opinions in User-Generated Contents." Invited talk at the *Sixth Annual Emerging Information Technology Conference (EITC-06),* Aug 10-12,, Dallas, Texas, 2006.

[2] Ali Harb and Michel Plantié Gerard Dray. "Web opinion mining: how to extract opinions from blogs?" *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*. ACM New York, NY, USA. 2008

[3] Xiaowen Ding, Bing Liu and Lei Zhang. "Entity Discovery and Assignment for Opinion Mining Applications."

[4] Bassam Hammo, Hani Abu-Salem, Steven Lytinen, and Martha Evens, "QARAB: A Question Answering System to Support the Arabic Language." *In the proceedings of Workshop on Computational Approaches to Semitic Languages*. ACL 2002, Philadelphia, PA, July. p 55-65 2002.

[5] Mostafa Syiam, Zaki Fayed , and Mena Habib. "An Intelligent System For Arabic Text Categorization." *The International Journal of Intelligent Computing and Information*. Volume 6 Number 1 January 2006.

[6] Xiaowen Ding , Bing Liu and Philip S. Yu. "A Holistic Lexicon-Based Approach to Opinion Mining." *In Proceedings of WSDM 2008*.

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques." *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79--86, 2002

[8] Ohana, Bruno, and Brendan Tierney. "Sentiment Classification of Reviews Using SentiWordNet." *In 9th. IT&T Conference, Dublin Institute of Technology*, Dublin, Ireland, 22nd.-23rd. October, 2009.

[9] Rudy Prabowo and Mike Thelwa. "Sentiment Analysis: A Combined Approach." *Journal of Informatics*. Volume 3, Issue 2, April 2009.

[10] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. *ACM Transactions on Information Systems*, Volume 26 Issue 3, Jun. 2008

[11] Yousif Almas and Khurshid Ahmad. "A Note on Extracting 'Sentiments' in Financial News in English, Arabic & Urdu." *In Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages Linguistic Institute*, Stanford, California, USA. July 21-22, 2007.

[12] Mohamed Elhawary and Mohamed Elfeky. "Mining Arabic Business Reviews." *IEEE International Conference on Data Mining Workshops*. 2010.

[13] Bing Liu. "Opinion Mining Encyclopedia of Database Systems," 2008.

[14] Minqing Hu, and Bing Liu. "Mining and Summarizing Customer Reviews." *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, USA, 2004

[15] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai. "Sentiment Strength Detection in Short Informal Text." *Journal of the American Society for Information Science and Technology* vol. 61. issue 12. 2010.

[16] Kamal Nigam, John Lafferty, Andrew McCallum, "Using Maximum Entropy for Text Classification." *In* IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61-67. (1999).

[17] Adam Berger, Stephen Pietra and Vincent Pietra., "A Maximum Entropy Approach to Natural Language Processing." *Computational Linguistics*, Vol., 22. p. 39-71 (1996).

[18] J. Darroch and D. Ratcliff, "Generalized Iterative Scaling for Long_Linear Model." *Annals of Mathematical Statistics*, 43 (5): 1470 - 1480 (1972).

[19] Adam Berger, "The Improved Iterative Scaling Algorithm: A Gentle Introduction." Technical report. (1997)

[20] Robert Malouf, "A Comparison of Algorithms for Maximum Entropy Parameter Estimation." *In Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002).* P 49-55. (2002).

[21] Belur Dasarathy., "Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques." *IEEE Computer Society Press*, 1991.

[22] Songbo Tan and Jin Zhang., "An Empirical Study of Sentiment Analysis for Chinese Documents." *Expert Systems with Applications* Vol 34. 2008.

[23] Gerard Salton and , Christopher Buckley. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management 24 (5):* 513–523. 1988.

[24] John Makhoul, Francis Kubala, Richard Schwartz, Ralph Weischedel. "Performance Measures for Information Extraction." *In Proceedings of DARPA Broadcast News Workshop*, Herndon, VA, February 1999.

[25] Deanna Osman and John Yearwood. "Opinion Search in Web Logs." *Proceedings of the Eighteenth Conference on Australasian Database*, 63. Ballarat, Victoria, Australia. 2007.