Performance Comparison of Different Lexicons for Sentiment Analysis in Arabic

Conference Paper · September 2016

DOI: 10.1109/ENIC.2016.026

Hunaida Ramadan Awwad
Dokuz Eylul University
2 PUBLICATIONS 13 CITATIONS
SEE PROFILE

Adil Alpkocak
İzmir Bakırçay University
80 PUBLICATIONS 544 CITATIONS
SEE PROFILE

Performance Comparison of different Lexicons for Sentiment Analysis in Arabic

Hunaida Awwad
Department of Computer Engineering
Dokuz Eylül University
Izmir, Turkey
hunaida.awwad@ceng.deu.edu.tr

Adil Alpkocak
Department of Computer Engineering
Dokuz Eylül University
Izmir, Turkey
alpkocak@ceng.deu.edu.tr

***Abstract*— This paper deals with the lexicon-based approach in document-level and sentence e-level sentiment analysis (SA) in Arabic. We experimented four different lexicons; a translation of Harvard IV-4 Dictionary (HarvardA), translation of the MPQA subjectivity lexicon developed by Pittsburgh University (HRMA) and two different implementation of MPQA. We evaluated all four lexicons with three datasets from different domains; one of them is about health comments (PatientJo), the second is from Twitter data, and the third is about books reviews (LABR). For sentence-level SA, we suggested six different methods for sentiment values and document polarity. The results show that the HRMA lexicon performs better than other lexicons in LABR while HarvardA perform better in PatientJo dataset. The results show that lexicon-based approach for document-level and sentence-level methods produce similar performance. We observed that giving extra weight for the first and last sentences in sentence-level approach improves the overall performance in terms of accuracy.**

Keywords*— Sentiment Analysis; polarity classification; lexicon-based; Arabic; sentence-level; document-level.

I. Introduction

Sentiment analysis (SA) and opinion mining from text are gaining great attention by researchers over many years as it helps companies and organizations to enhance their products/services. SA is defined as the task of finding the opinions of authors about specific entities. Sentiment analysis deals with the computational treatments of sentiments, subjectivity and opinions within a text [33]. The value of sentiment analysis lies in its ability to convert unstructured text document into structured data that can be qualified and analyzed [31]. In SA, the focus is on analyzing subjective sentences, which contain explicit opinions, beliefs, and views about specific entities [20]. Another common subtask of SA is polarity classification, which reduces the problem to identifying whether the text expresses positive or negative sentiment [33].

Sentiment analysis methods can be grouped into two broad approaches; the supervised learning approach that is called also corpus-based approach and the unsupervised approach that is also called lexicon-based approach. The Supervised approach assumes that there is a finite set of classes into which the document should be classified, and training data available for each class. The simplest case is when there are two classes: positive and negative. A simple extension can be added to a neutral class or to have some discrete numeric scale into which the document should be placed (like the five-stars system used by Amazon). Unsupervised approaches are based on determining the semantic orientation (SO) of specific phrases within the document. If the average SO of these phrases is above some predefined threshold the document is classified as positive, otherwise it is deemed negative. Besides, polarity classification uses two major techniques: firstly, symbolic techniques using a set of rules and manually tuned thresholds. It relies heavily on external lexicons and other structured resources. Secondly, based on machine learning techniques, which are based on supervised methods such as Support Vector Machines, Naive Bayes and Maximum Entropy.

There is a large number of studies for sentiment analysis dealing with English language, while there are a limited numbers of studies about SA in Arabic. However, the statistics in [25] shows that Arabic is considered to be the fastest growing language over the web. Arabic is spoken by more than 360 million people, and is the fastest-growing language on the web (with an annual growth rate of 6,592% in the number of Internet users as of 2015, compared to 3,227% for Russian, 2,080% for Chinese and 520% for English) (www.internetworldstats.com). Therefore, there is an emergent need to more research on SA in Arabic.

Arabic has 28 letters and written from right to left. Letters in Arabic take many shapes depending on the letter location within a word. Arabic has three different variants; the first one is the Modern Standard Arabic (MSA) which can be found in news, TVs, and schools. The second one is the dialects of spoken Arabic which is used for daily communication and might change from one geographic area to the other within the Arab region. The third one is the classic Arabic that is related to Quran which is similar to Modern Standard Arabic but has some characteristics that appear in the Quranic text.

The contribution of this study is of two folds. Firstly, it presents three different lexicons in Arabic; HarvardA, MPQAII, and HRMA. It also provides a comparison between them. The second contribution is to provide a new dataset, PatientJo, for health domain. To the best of our knowledge, PatientJo is the first dataset collected for health domain in Arabic. An Arabic lexicon-based analyzer (ALBA) is created to classify Arabic documents by polarity as positive, negative, and neutral. ALBA is also considered one of the main contributions to this study that will determine document polarity in both document-level and sentence-level.

In this study, unsupervised approach is implemented for SA to find the document polarity in Arabic. Four lexicons are comparatively used in this study HarvardA, MPQA, PMQAII, and HRMA. Beside, we used three datasets during the experimentation study, which are health comments dataset (PatientJo), Tweeter data (TA), and books reviews dataset (LABR). This paper also presents a lexicon-based approach in both sentence-level and document-level.

The rest of this paper is organized as follows: Section II discusses sentiment analysis approaches. Section III presents the state-of-the-art in SA in Arabic. In Section IV, we explain our methodology: datasets, lexicons and methodology in both document and sentence-level and the preprocessing. In Section V, we explain the ALBA analyzer. In Section VI, we explain our experiments and results. Finally, Section VII concludes the paper and gives a look at the future studies in this area.

II. Levels of Sentiment Analysis

Sentiment analysis deals with the text in many ways. It Either considers that there is one opinion that covers the whole document, or there might be multiple opinions within the document.

The first level is to consider the given text as one part; this is called document-level sentiment analysis. On the document-level the analysis assumes that each document contains an opinion.

The second level is the sentence-level sentiment analysis that assumes that each sentence has a single opinion. The overall polarity of a document will depend on the polarity inside the sentences.

The third level is the aspect-based sentiment analysis that is called also feature-based sentiment analysis. In aspect-level the research problem focuses on the recognition of all sentiment expressions within a given document and the aspects to which they refer. In many cases people talk about entities that have many aspects (ie, attributes) and they have a different opinion about each of the aspects. This often happens in reviews about products or in discussion forums dedicated to specific product categories (eg, cars, cameras, smartphones, and even pharmaceutical drugs).

The last level is the comparative level, in many cases users do not provide a direct opinion about one product but instead provide comparable opinions; take for example these sentences derived from the user forums of Edmonds.com: "300 C Touring looks so much better than the Magnum," "I drove the Honda Civic, it does not handle better than the TSX, not even close." Some of words used in the comparative form are: Comparative adjectives adverbs such as: 'more,' 'less,' and words ending with –er (for example, 'lighter'). Superlative adjectives and adverbs such as: 'most,' 'least,' and words ending with –est (for example, 'finest'). Additional phrases such as 'favor,' 'exceed,' 'outperform,' 'prefer,' 'than,' 'superior,' 'inferior,' 'number one,' 'up against.'

This article will cover document-level and sentence-level sentiment analysis in Arabic through lexicon-based approach to find document polarity.

III. Related Works

SA studies in general are categorized based on different criteria such as classification types (positive, negative, or neutral), subjectivity classes (subjective or objective), the classification level (sentence, document, or aspect), or on the used classification techniques (supervised or lexicon-based).

There is little work on sentiment analysis in Arabic, [1], [4], [10], and [29] used the supervised approach to find sentiment analysis in document level. Reference [29] generated a new Arabic corpus for predicting sentiment polarity Opinion Corpus for Arabic (OCA) and achieved an accuracy level of 90% using the SVM and 84% using the Naïve Bayes [29]. On the other hand, some other researchers used supervised approaches (SVM) for SA and used both syntactic and stylistic features in sentence level in Arabic [1], [4], and [10]. Many researchers used Weka tool with SVM, NB, ME, Bayes Net, and J48 classifiers [30], [7], and [18]. Some used SAMAR tool which is a system for subjectivity and sentiment analysis (SSA) for Arabic social media [7]. The best achieved accuracy is 70% [9], using Weka tool with SVM, NB, ME, Bayes Net, and J48. Reference [18] where an in-house system is used to classify whether a document is an Arabic review or not. They evaluate the Arabic Reviews classifier and compare it to the English Reviews classifier. It shows that they were able to achieve a high-precision relatively-high-recall Arabic Reviews classifier.

Some use a combined approach of supervised and lexicon-based [17]. A developed tool was used, the tool presents a combined approach which extracts opinions from Arabic documents using three methods: first method is the lexicon based that will classify as many documents as possible, the second one is the maximum entropy method that will use the results from the first method as a training set to classify other documents, and the third one is the *k*-nearest method that will use the classified documents from the first two methods as a training set and classifies the rest of the documents. The results show that the system achieved an accuracy of 80%.

In the lexicon-based approach, [12], [15], and [8] study the SA in Arabic in the document level. Reference [15] uses lexicon based tool to extract opinion holder. Reference [8] uses rule base: Ara-SATISFI prototype to visualize sentiments in financial news. Reference [12] use local grammar approach LoLo (tVtV). [16] study the SA in the sentence level. They use two approaches to find the tweet polarity: by using the summation or by using Double Polarity (DP). The results show that weighted lexicons with DP summation lead to improvement in polarity identification of Arabic sentiments.

[19] studied the SA on both the document and sentence levels. They used two approaches: a grammatical approach and a semantic approach. The grammatical approach is based on Arabic grammatical structure, and combines the verbal and nominal sentence structures in one general form. The second approach combined syntactic and semantic features. For evaluation of the grammatical approach, only 29 sentences are annotated manually with part-of-speech tags. The results show 89.3% accuracy using an SVM classifier with 10-fold cross-validation. Sentences from 44 random documents are used to evaluate the semantic and syntactic approach using a J48 decision tree classifier. The results show 80% accuracy when the semantic orientation of the words extracted and assigned manually is used, and 62% when the dictionary is used. They also classified the documents by using all sentence features and chunking the document into different parts, reporting 87% accuracy with an SVM classifier. [4] used the lexicon-based approach in document-level. The achieved accuracy is almost 60% when using light stemmer and 63% for Twitter and 70% for Yahoo!-maktoob when no stemmer is used.

Reference [10] designed a tool for Arabic texts in both MSA and colloquial Arabic. The used lexicons are domain-based; which means that there is a lexicon for each domain (Books, Movies, Places, Politics, etc.). The results show that the accuracy reached 90%. It is important to mention that the lexicon-based approach should not be built from the same used dataset. The lexicons here were built from the used corpus and this might justify the high accuracy in lexicon-based approach experiments.

IV. Methodology

The lexicon-based approach is applied to find document polarity in datasets. The datasets contain documents that are collected from different domains (eg, health domain, Twitter and books reviews). In the lexicon-based approach, each word in a document is searched in the defined lexicon to identify word's sentiment value. Each lexicon has two lists: positive words and negative words list. These two lists will be used to identify the polarity of a document. In document-level, each positive word in a document will be given +1 value and each negative word will be given -1 value. The aggregation of these values will consider the sentiment value SV of this document. If the SV is greater than zero, then the sentiment orientation is positive and if the SV is less than 0, then the sentiment orientation is negative. Neutral case appears if the SV is equal to zero. In sentence-level, a proposed approach uses the sentence's position on document polarity. This approach will either drop some sentences or will give different weights to sentences. It is crucial to distinguish between first sentences, middle sentence(s), and last sentence. Punctuation marks will be used for distinguishing.

1. Datasets

During this study, we used three datasets, where TABLE I shows the basic statistics of the three datasets.

PatientJo: It is a health domain dataset collected by the authors of this paper from three Jordanian hospitals that includes 1228 comments (227 positive, 951 negative).

TA: Twitter Dataset for Arabic Sentiment Analysis [15] from the Machine learning repository (UCI): that includes 1000 positive and 1000 negative tweets in both Modern Standard Arabic (MSA) and Jordanian dialect that is released in 2014.

LABR: Large scale Arabic Reviews dataset that includes about 63000 books reviews that collected from goodreads.com [13]. On this site, each review is rated on a scale of 1 to 5 stars, which the authors have mapped to a sentiment polarity. The dataset was then used for the tasks of sentiment polarity classification and rating classification. The positively or negatively annotated reviews were chosen only with total of 51056 reviews with (42832 positive, 8224 negative).

TABLE I. Statistics on Datasets

		Positive	***Negative***
PatientJo	Total comments	227	951
	Total words	2749	29899
	Avg. words per comment	12.1	31.4
	Avg. characters per comment	56.6	133.8
TA	Total comments	1000	1000
	Total words	7633	10678
	Avg. words per comment	7.63	10.68
	Avg. characters per comment	31.2	46.6
LABR	Total comments	42832	8224
	Total words	2178364	479976
	Avg. words per comment	50.9	58.4
	Avg. characters per comment	289.9	255.7

2. Lexicons

The purpose of using the dictionary in Polarity Classification is to help in discovering if a certain word has a positive or negative sentiment. For example, the words clever, correct, love, shiny have positive meanings, while the words confession, awful, difficult, strike have negative meanings. In this study four lexicons are used, where TABLE II shows the statistics of the four lexicons.

1) HarvardA: that is a translation of Harvard IV-4 [23] into Arabic. Harvard IV-4 inquirer dictionary is in English language and has sentiment categories such as positive and negative classes. The dictionary contains 1,915 positive words and 2,291 negative words. Harvard IV-4 dictionary, in fact, contains many categories that can be used in other domains other that sentiment analysis. We translated Harvard IV-4 from English to Arabic. For simplicity the translation of HarvardIV-4 into Arabic will be referred as HarvardA.

2) MPQA: is a translation of the MPQ subjectivity lexicon developed by Pittsburgh University into Arabic. Reference [15] translated MPQA subjectivity lexicon. It contains more than 8000 English words and corresponding Arabic words. This lexicon contains four categories; positive strong, positive weak, negative strong and negative weak. The MPQA includes 1406 positive Arabic words and 2361 negative Arabic words. Since our study is dealing with polarity classification we combined the positive/negative from both strong and weak classes into positive/negtive list class.

3) MPQA II: it is another translation for MPQA using Google translate and is prepared by the researchers. This lexicon includes 1535 positive Arabic words and 2587 negative Arabic words. Since our study is dealing with polarity classification, we combined the positive/negative from both strong and weak classes into positive/negtive class.

4) HRMA: that is a combined lexicon from HarvardA and MPQA II. It includes 2070 positive Arabic words and 3052 negative Arabic words.

TABLE II. Statistics on lexicons

Lexicon	Positive	Negative	Total
HarvardA	1247	1522	2769
MPQA	1406	2361	3767
MPQA II	1535	2587	4122
HRMA	2070	3052	5122

3. *Document-level and sentence-level SA*

The document level of sentiment analysis is dealing with the given text as one part. It means the sentiment words will be considered as either positive or negative sentiment. The sentence level of sentiment analysis has two missions: firstly, to emphasis document sentences through punctuation, capitalization, and emotions. Secondly: to find the sentiment for each sentence in order to be able to find the overall sentiment of the document. Since Arabic has no capitalization and MSA in Arabic does not use emotions, punctuation marks are used to emphasis sentences within a document such as (“.”, “,”, “!”, “?”, “/”). Some Arabic punctuation marks are different from English; such as the comma; Arabic comma is “ ، ” while English comma is “ ,” also the question mark in Arabic is “؟”.

There are many ways to find sentiment value (SV) of a document through sentence-level. In this study we propose three methods. Method1 will find the sentiment value of each sentence by giving +1 weight for each positive word and -1 for each negative one, and then the SV of a sentence is the simple summation of all positive and negative values. Method2 will consider only the first and last sentences in a document to detect its polarity and all middle sentences will be ignored. Method3 will give different weights to sentences depending on the location of the sentence; first and last sentence will have more weight than other sentences. Reference [14] assumes that the first and last sentences in any document are the best estimators of a document polarity. Methods 4, 5, and 6 will drop first or/and last sentences to test the effect of these sentences on document polarity. In method4, first and last sentences will be ignored and only middle sentence(s) will be considered. In method5, first sentence will be ignored and remaining sentences will be considered. In method6, last sentence will be ignored and remaining sentences will be considered.

4. *Preprocessing*

Each word in a document is treated as a separate token; we used bag-of-word model for text representation. The second step is normalization step, since Arabic language is a morphological language; there is a need for the normalization step. In the normalization, different forms of the letter Alif (“أ” “ آ“ “ ا“ “إ”) is normalized into one form (ا) for simplicity to enhance the searching results and finding the polarity class. The third step is stemming; two stemmers are available in ALBA: 1) Khoja is a well-known Arabic stemmer, where it is an open source root-based stemmer [26], 2) Zahero stemmer is another Arabic stemmer that extract the stem (not the root) [24]. The last step in preprocessing is the stop words removal, the selected list is collected from http://code.google.com/p/stop-words/ project.

Both negation and intensification are important for sentiment analysis, in negation the polarity of the word/phrase will be reversed while in intensification the sentiment meaning is stronger than the phrases without intensification. In Arabic there is a list of words that reverses the sentiment in MSA: for example: in the sentence” لم احب الفلم ” that means “I didn’t like the movie” the like has positive sentiment but with the negation (didn’t) the sentiment should be reversed. In the sentence “ فلم رائع جدا” that means “ the movie is very nice”, the word “very” emphasized the sentiment of “nice” and it is stronger than “the movie is nice”. If one of the negation words in Arabic “لا”,”لم”,”ليس”,”لن”,”ما”,”مش”,”مو”,”غير” that means “not” comes before a phrase, the phrase sentiment should be reversed. Intensification words such as “كثيرا”,”فعلا ”,” تماما”,”جدا”,”بافراط” that means “a lot”, “extremely”, “exactly” increase the sentiment meaning of a phrase that comes before them. In this study, both negation and intensification are not considered.

V. Tool Design and Implementation

Processing

In this study we developed a lexicon-based sentiment analyzer called ALBA, in Microsoft Visual studio in C#, and used Microsoft Excel as a dataset repository. We saved lexicons as text files where each lexicon has two lists, one for the positive words and one for the negative words.

ALBA allows users to upload the needed dataset/lexicon for testing. Moreover, it allows user to select the needed level of lemmatizing, stop words removal, sentence-level/document-level, to filter the documents that contain more than two sentences.

The lexicon-based approach is applied to check the performance of the translated Harvard dictionary. The sentiment value (SV) is considered as number of positive words in a document minus number of negative words. If SV is positive it means this document is positive, if SV is negative it means this document is negative and if it is zero it means it is neutral. Figure3 shows the algorithm steps for finding Sentiment Value SV and Document Polarity P in document-level SA while figure4 shows SV and P calculations for sentence-level SA.

Equation 1 and Equation 2 are showing the document-level and sentence-level SV calculation.

$$SV = (+1)\ \sum_{n=1}^{\infty} \mathrm{Pt} + (-1)\ \sum_{n=1}^{\infty} \mathrm{Nt} \tag{1}$$

$$SV = \left(Ptf.wf + ((+1).\sum_{n=1}^{\infty} Ptm.\mathrm{wm}) + Ptl.wl \right) + \left(Ntf.wf + ((-1).\sum_{n=1}^{\infty} Ntm.wm) + Ntf.wl \right) \tag{2}$$

Where *Pt* is the positive token in document, and *Nt* is the negative token in document. *Ptf*, *Ptm*, and *Ptl* are the positive token in first sentence, middle sentence(s), and last sentence, respectively. *wf*, *wm*, and *wl* are the weight of first sentence, middle sentence(s), and last sentence, respectively. *Ntf*, *Ntm*, and *Ntl* are the negative token in first sentence, middle sentence(s), and last sentence, respectively.

VI. Experiments and Results

The mail goal of the experiments is to evaluate the performance of four lexicons using ALBA. In this paper we propose three different lexicons. Additionally we also propose a new sentence-level method and evaluated. Finally, we evaluated the effect of sentence location on document polarity.

To evaluate the tool performance, we used four of the most commonly used performance measures *precision*, *recall*, *f-measure*, and *accuracy*. In order to do this we set up a set of experiments.

Stemmer vs. no stemmer using HarvardA

Two Arabic stemmers are used in this experiment; Khoja is used for tough stemming while Zahero is used for light stemming. This experiment is applied for PatientJo, TA, and LABR datasets with google stop words list. HarvardA lexicon is used in this experiment.

Table III. Stemming effect using HarvardA

Dataset	No stemmer	Khoja	Zahero
PatientJo	37%	39%	52%
TA	16%	30.7%	26%
LABR	43%	52.6%	53%

TABLE III, shows that using stemmer has increased the accuracy from 37% to 52% using Zahero and to 39% using Khoja. Zahero shows better performance compared with Khoja. TA is collected from Twitter in which the tweets are normally short and have dialects; this might justify the low achieved accuracies of TA.

The precision, recall, and F-measure for the positively annotated documents compared to negatively annotated documents are as follows; (0.41,0.31,0.35) and (0.32, 0.24, 0.27) for positive class and (0.92,0.46,0.62) and (0.98, 0.32, 0.49) for the negative class using Zahero and Khoja respectively. The Precision is much higher for the negative class; it reaches 98% compared with 41% for the positive class.

Document-level vs. sentence-level

A study of lexicon richness will be provided in this part of the experiments. The experiment will evaluate HarvardA alone, MPQA alone experiment, MPQAII alone, and the HRMA lexicon that is combination between HarvardA and MPQAII. The results in TABLE IV show that HRMA with LABR achieve the highest accuracy (62.6%). PatientJo performs better with HarvardA rather than other lexicons with 52%. TA performs better with MPQAII with 41.5%.

For the sentence-level, method1, method2, and method3 (mentioned in section VII) are evaluated on the four previously mentioned lexicons. Zahero stemmer is used with Google stop words list.

TABLE IV shows that method3 in sentence-level in which a weight of 1.5 is given to first and last sentences and a weight of 1 is given to middle sentence(s) achieved the better accuracy in the three datasets with the four lexicons except for PatientJo with MPQAII and HRMA.

Sentence location within a document

In this experiment, the impact of sentence location on whole document polarity will be tested. First and last sentences are supposed to carry the opinion more than middle sentence(s). To be able to apply this experiment, documents with more than two sentences are selected for the experiment. PatientJo datasets has 1228 comments; comments with more than two sentences are 213 comments. LABR Dataset has 51056 reviews; reviews with more than two sentences are 24666 reviews.

In LABR dataset, the number of reviews that contain more than two sentences represents 48.3% of the dataset, while in PatientJo the number of comments drops down to 213 comments that represent just 17% of the dataset. In TA dataset, tweets with more than two sentences are 66 out of the 2000 tweets with a percent of 3.3%, therefore TA is not applicable for this experiment.

All the six methods that were previously described in section VII are applied in this experiment on a subset of the datasets (213 for PatientJo and 24666 for LABR).

TABLE V shows that method3- in which a weight of 1.5 is given to first and last sentences and a weight of 1 is given to middle sentence(s)- in sentence-level experiments that applied here to documents that exceeds two sentences achieve better accuracy in the three datasets with the four lexicons except for PatientJo with MPQAII and HRMA. The best achieved accuracy is 71% in HRMA lexicon on LABR dataset. This experiment shows that documents with more than two sentences long perform better than short documents.

TABLE IV. Document-Level And Sentence-Level Accuracy

	HarvardA			**MPQA**			**MPQAII**			**HRMA**		
	PatientJo	***TA***	***LABR***	***PatientJo***	***TA***	***LABR***	***PatientJo***	***TA***	***LABR***	***PatientJo***	***TA***	***LABR***
Document-Level	52%	26%	53%	36%	35%	55%	47%	41.5%	60%	45%	41%	62.6%
Sentence-Level Using Method1	52%	26%	53%	36%	35%	55%	47%	41.5%	60%	45%	41%	62.6%
Sentence-Level Using Method2	51.8%	26%	46.7%	36.4%	34%	50.6%	48.2%	41%	55.5%	46.7%	41%	57.8%
Sentence-Level Using Method3	53.1%	26%	54.4%	37%	35%	60%	48%	41.7%	61.4%	46.3%	41%	64%

TABLE V. Sentence-Level Accuracy For A Subset Datasets

	HarvardA		**MPQA**		**MPQAII**		**HRMA**	
	PatientJo	***LABR***	***PatientJo***	***LABR***	***PatientJo***	***LABR***	***PatientJo***	***LABR***
Method1	60%	59.6%	35%	61.7%	41%	65.8%	37%	68.7%
Method2	59%	46.8%	37%	51.8%	46.5%	56%	45%	58.4%
Method3	66.2%	62.8%	40%	64.8%	44.6%	68%	43.2%	71%
Method4	45%	49.7%	30%	51.2%	34%	56.6%	28%	60%
Method5	48%	55.8%	34.3%	57.7%	36.6%	62.2%	37%	65.3%
Method6	62%	56.2%	36%	58.2%	38%	62.4%	35%	66%

VII. Conclusion and future work

In this study, we provide the stemmer/no stemmer experiments, the results show better accuracy when using stemmer. The accuracy increased from 37% to 52%. Light stemmer such as Zahero is better than the root based stemmer such as Khoja; Zahero 52% while Khoja is 39%. We started with a new stemming approach that we believe that it will lead to better results and we will include it in our next article.

We provide also comparisons of performance of different lexicons on different dataset. To do this we collected some lexicons which is publically available ; we also developed some new lexicons. To achieve this comparison, we used both publically available dataset for SA; besides we developed a new dataset PatientJO in health domain. Experiments have shown that people share their bad experience more in the health sector and share their good experience more in books reviews.

Document-level/ sentence-level experiments for PatientJo dataset and LABR dataset show that the PatientJo dataset best accuracy is 53% in method3 of sentence-level with HarvardA, while LABR dataset best accuracy is 64% in method3 of sentence-level with the HRMA lexicon. The best results achieved in method3 of sentence level that applied on documents with more than two sentences; the best accuracy is 71% (HRMA lexicon with LABR dataset).

References

[1] Abbasi, A., Chen, H. and Salem, A., 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems (TOIS), 26(3), p.12.

[2] Abdulla, N., Mahyoub, N., Shehab, M. and Al-Ayyoub, M., 2013. Arabic sentiment analysis: Corpus-based and lexicon-based. In *Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT).*

[3] Abdulla, N.A., Ahmed, N.A., Shehab, M.A., Al-Ayyoub, M., Al-Kabi, M.N. and Al-rifai, S., 2014. Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and Web Engineering (IJITWE)*, *9*(3), pp.55-71.

[4] Abdulla, N.A., Al-Ayyoub, M. and Al-Kabi, M.N., 2014. An extended analytical study of arabic sentiments. *International Journal of Big Data Intelligence 1*, *1*(1-2), pp.103-113.

[5] Abdul-Mageed, M. and Diab, M.T., 2012, May. AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In *LREC* (pp. 3907-3914).

[6] Abdul-Mageed, M., Diab, M.T. and Korayem, M., 2011, June. 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 (pp. 587-591). Association for Computational Linguistics.

[7] Abdul-Mageed, M., Kuebler, S. and Diab, M., 2012. SAMAR: A system for subjectivity and sentiment analysis of social media Arabic. In *3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), ICC Jeju, Republic of Korea.*

[8] Ahmad, K. and Almas, Y., 2005, July. Visualising sentiments in financial texts?. In *Information Visualisation, 2005. Proceedings. Ninth International Conference on* (pp. 363-368). IEEE.

[9] Ahmed, S., Pasquier, M. and Qadah, G., 2013, March. Key issues in conducting sentiment analysis on Arabic social media text. In *Innovations in Information Technology (IIT), 2013 9th International Conference on* (pp. 72-77). IEEE.

[10] Al-Kabi, M., Gigieh, A., Alsmadi, I., Wahsheh, H. and Haidar, M., 2013. An opinion analysis tool for colloquial and standard Arabic. In *Proceedings of the 4th International Conference on Information and Communication Systems, ICICS* (Vol. 13).

[11] Al-Kabi, M.N., Abdulla, N.A. and Al-Ayyoub, M., 2013, December. An analytical study of arabic sentiments: Maktoob case study. In *Internet Technology and Secured Transactions (ICITST), 2013 8th International Conference for* (pp. 89-94). IEEE.

[12] Almas, Y. and Ahmad, K., 2007, July. A note on extracting 'sentiments' in financial news in English, Arabic & Urdu. In *The Second Workshop on Computation, al Approaches to Arabic Script-based Languages* (Vol. 21, pp. 21-22).

[13] Aly, M.A. and Atiya, A.F., 2013, August. LABR: A Large Scale Arabic Book Reviews Dataset. In *ACL (2)* (pp. 494-498).

[14] Dehkharghani, R., Yanikoglu, B., Saygin, Y. and Oflazer, K., Sentiment Analysis in Turkish: Towards a Complete Framework.

[15] Elarnaoty, M., AbdelRahman, S. and Fahmy, A., 2012. A machine learning approach for opinion holder extraction in Arabic language. *arXiv preprint arXiv:1206.1011*.

[16] El-Beltagy, S.R. and Ali, A., 2013, March. Open issues in the sentiment analysis of Arabic social media: A case study. In *Innovations in information technology (iit), 2013 9th international conference on* (pp. 215-220). IEEE.

[17] El-Halees, A., 2011. Arabic opinion mining using combined classification approach.

[18] Elhawary, M. and Elfeky, M., 2010, December. Mining Arabic business reviews. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on* (pp. 1108-1113). IEEE.

[19] Farra, N., Challita, E., Assi, R.A. and Hajj, H., 2010, December. Sentence-level and document-level sentiment mining for Arabic texts. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on* (pp. 1114-1119). IEEE.

[20] Feldman, R., 2013. Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), pp.82-89.

[21] http://www.claes.sci.eg/coe_wm/Data.htm.

[22] http://www.nhs.uk/aboutNHSChoices/professionals/developments/Documents/2011/data/hospital-comments-and-responses.csv

[23] http://www.wjh.harvard.edu/~inquirer/homecat.htm

[24] https://compilr.com/zahero/arabicstemmer.

[25] Internet world stats, http://www.internetworldstats.com/stats7.htm http://www.internetworldstats.com

[26] Khoja, S. and Garside, R., 1999. Stemming arabic text. *Lancaster, UK, Computing Department, Lancaster University*.

[27] Maamouri, M., Bies, A., Buckwalter, T. and Mekki, W., 2004, September. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools* (Vol. 27, pp. 466-467).

[28] Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A. and Perea-Ortega, J.M., 2011. Bilingual experiments with an Arabic-English corpus for opinion mining.

[29] Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A. and Perea-Ortega, J.M., 2011. OCA: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology*, *62*(10), pp.2045-2054.

[30] Shoukry, A. and Rafea, A., 2012, May. Sentence-level Arabic sentiment analysis. In *Collaboration Technologies and Systems (CTS), 2012 International Conference on* (pp. 546-550). IEEE.

[31] Siegrist Jr, R.B. and Madden, S., 2011. Sentiment analysis turns patients' feelings into actionable data to improve the quality of care. *The Science of Emotion*, pp.27-35.

[32] Wilson, T., Wiebe, J. and Hoffmann, P., 2005, October. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347-354). Association for Computational Linguistics.

[33] Xia, L., Gentile, A.L., Munro, J. and Iria, J., 2009, September. Improving patient opinion mining through multi-step classification. In *Text, Speech and Dialogue* (pp. 70-76). Springer Berlin Heidelberg.