Arabic Lexicon Learning to Analyze Sentiment in Microblogs

Mahmoud B. Rokaya[1]
Ahmed S. Ghiduk[2]
Department of Information Technology, Taif University
Taif, KSA

Mahmoud B. Rokaya[3]
Department of Mathematics, Faculty of Science
Tanta University
Tanta, Egypt

Ahmed S. Ghiduk[4]
Department of Mathematics and Computer Science, Faculty of Science, Beni-Suef University
Beni-Suef, Egypt

***Abstract*—The study and classifying of opinions distilled from social media is called sentiment analysis. The goal of this study is to build an adaptive sentiment lexicon for Arabic language. Based on those lexicons the sentiments polarity classification can be improved. The classification problem will be stated as a mathematical programming problem. In this problem, we search a lexicon that optimizes the classification accuracy. A genetic algorithm is presented to solve the optimization problem. A meta-level feature is generated based on the adaptive lexicons provided by the genetic algorithm. The algorithm performance is supported by using it alongside n-gram features and Bing liu's lexicon. In this work, lexicon-based and corpora-based approaches are integrated, and the lexicons are produced from the corpus. Five data sets are tested through experiments. The sentiments in all data sets are classified based on five polarity levels. A better understanding of words sentiment orientation, social media users' culture and Arabic language can be achieved based on the lexicons generated by the proposed algorithm. Since stop words can contribute and add to the sentiment polarity, stop words will be considered and will not deleted. The results show that the F-measure is greater than 80 % in three data sets and the accuracy is greater than 80 % for all data sets. The proposed method out-performs the current methods in the literature in two of the datasets. Finally, in terms of F-measure, the proposed methods achieved better results for three datasets.**

Keywords*—Sentiment analysis; sentiment lexicon; social media; twitter; optimization; mathematical programming; genetic algorithm; evolutionary computation; Arabic language

I. Introduction

The subjects of sentiment analysis are the study of opinions and its related concepts such emotions attitudes evaluations and sentiments. For the first time in humanity history, we have that massive volume of recorded data that reflects the opinions, emotions and attitudes of people around the globe. This came from Twitter, reviews, social networks, forum discussions, blogs and microblogs. So, it is natural that the field of sentiment analysis is emerged.

In business, sentiment analysis addresses the problem of studying the customer opinions regarding products through analyzing and extracting opinion from products reviews. However, most current algorithms which developed for the business purpose are not suitable to analyze sentiments in social domain.

The objective of Sentiment Classification task is to take a piece of text written by an author regarding a topic and determine the author general feeling toward this topic, whether this felling be positive or negative.

The current work tries to improve classification of sentiments in microblogs based on building sentiment lexicons. The sentiment classification problem is written as an optimization problem, finding optimum sentiment lexicon is the goal of the optimization process. The solution will be produced based on proposed genetic algorithm to find lexicons to classify text. Then, extraction of a meta-level feature will be done based on it. The experiments are conducted on several Arabic datasets. A better understanding of the Arabic language and culture of Arab Twitter users and sentiment orientation of words in different contexts can be achieved based on the sentiment lexicons proposed by the algorithm.

Since adaptive lexicons are developed in this work, the trends in the ever-changing environment of Twitter can be captured [1]. Updating the lexicons to adapt with the changes in the culture of the users can be done easily. For example, based only on one feature, the results of the proposed method are promising.

Considering real benefits, to understand the social media and their words context in known domains gives the users the ability to use the words in their messages in more effective transmission methods. Similarly, this idea might be used in producing lexicons for languages that do not own one. In analogues with this, this method can be employed to calculate the sentimental scores for same terms in different contexts and websites. The modification of the method for strength and emotion classification will be explored. Based on the method, it is planned to generate lexicons for the Arabic language.

The rest of this paper is organized as follows: Section II presents the related work; Section III presents the methods including. Experiments, results, discussion is presented in Section IV. Finally, the conclusion and main results are presents in Section V.

II. Related Work

In the proposed method, we try to develop an adaptive lexicon for sentiment analysis; the Statistical methods for sentiment analysis, lexicons-based approaches and evolutionary methods are explored.

Statistical methods have been developed based on the following observation. If two words frequently appear together within the same context, they will have the same polarity. So, by calculating a word relative frequency of co-occurrence with special words for a given word, the polarity of this word can be determined. The performance of these algorithms did not give the same or even near results when applied to training data labeled with emotions which has the potential of being independent of domain, topic and time [2].

In that area, many approaches that address different dimensions of opinions, such as subjectivity, polarity, intensity and emotion were proposed to extract sentiment indicators from natural language texts, whether these indicators are at syntactic or semantic levels. Mohammad and Turney, 2013, conducted experiments on how to formulate the emotion-annotation questions and show that asking if a term is associated with an emotion leads to markedly higher inter annotator agreement than that obtained by asking if a term evokes an emotion [3].

T. Wilson, et al., 2005, presented an approach to phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions [4]. M.M. Bradley and P.J. Lang, 2009, developed a set of verbal materials that had been rated in terms of pleasure, arousal, and dominance to complement the existing International Affective Picture System [5].

Despite that classifying manually will give the most accurate results. It is more than difficult to use manual methods in the labeling process for determining the polarity of comments or posts of users in social media. For this reason, some papers use emoticons as labels [6 and 7]. In [8], the author discussed how this method will produce much noise. Using emoticons, Go et al., 2010, distilled 1600,000 tweets from Twitter dataset [7].

Liu et al., 2012, presented a dataset and used a method of labelling that depends on using emoticons and manual classification [9]. Da Silva et al., 2014, created a classifier ensemble for Twitter sentiment classification [10]. Hu et al., 2013, combined the networked data to benefit from emotional spread in sentiment classification [11]. In [12] features that depend on concepts of semantic are combined with the training set [13]. In (Bravo-Marquez et al., 2013), different approach that employs meta-level features for social media sentiment classification is used, namely for twitter. In this method, different features of words are used for polarity and subjectivity classification. Kaewpitakkun et al., 2014, created a lexicon that finds scores for objective and out of vocabulary words, and used a calculation method that depends on weighting scheme for features [14]. A method that depends on distilling patterns of terms and phrases was developed by Saif et al., 2014, for evaluating those terms and phrases on tweet-level and entity-level sentiment analysis [15]. Feature learning approach was introduced by Baecchi et al., 2015. They used this method for classification of tweets. Namely, they targeted posts that might contain pictures [16]. An unsupervised Learning framework was proposed by Hu et al., 2013. In this method, they combined emotional signals, in Twitter datasets [11]. In [17], a sentiment scoring function was used for classification of tweets. Combination of social connections as well as social emotions between users between posts of the same author was employed by Wu et al., 2016 to get better accuracy [18].

Despite, sentiments are implicitly expressed through patterns, dependencies among words in tweets and latent semantic relations, most existing approaches to Twitter sentiment analysis suppose that sentiment is explicitly expressed through affective words. Also, these methods do not consider that words' sentiment orientations and strengths change continuously throughout various contexts in which the words appear.

Sentiment lexicons can be defined as: those groups of terms and phrases that are assigned numeric scores, which give the sentiment emotional value of a term or phrase. Some lexicons, simply, allocate labels for each term or phrase. These labels are either to be positive or negative. For example, we can report Bing Liu's lexicon as the most known lexicon that uses this simple method. Many studies tried to establish lexicons for sentiment analysis [9, 19, 20, 21, 22 and 23].

Lexicon-based approaches to Twitter sentiment analysis becomes more popular because of their simplicity, domain independence, and good performance. These approaches depend on sentiment lexicons, where a list of words is marked with fixed sentiment polarities; for example, [17, 24, 25, 26 and 27]. Arora et al., 2010, and Govindarajan, 2013, used a hybrid of Naive Bayes classifier and genetic algorithm for classification of movie reviews [28].

For Arabic sentiment analysis, Hossam et al., 2015, presented a sentiment analysis based on two lexicons. The first is a lexicon for adjectives and adjective nouns. The other lexicon contains the known idioms. They developed a method to expand the lexicon from seeds or words and idioms. The method reflects a static lexicon with fixed values for the polarity of each term. Also, they depend heavily on a translated version of HU-LUI lexicon [29]. Haidy et. al., 2017, used a hybrid method to determine the sentiment polarity of a tweet. In the first phase they used a lexicon to classify a set of tweets. The result of this phase is the input of the second phase. The lexicon was composed of two parts. The first is a lexicon for words; the second is a lexicon of idioms [29]. Al-Ayyoub and Essa, 2015, presented a sentiment analysis based on lexicon approach was adopted. The polarity of a given word is got from the corresponding English translation. Stop words are deleted with consideration of some stop words that can affect the polarity of a given word. The lexicon words and sentiment expression are stemmed. Using the polarity of the translated terms will reduce the functionality of the words, also neglecting the stop words, which contribute in the total meaning that the author wants to give [30].

Most of these works stated that they follow a supervised or unsupervised leaning approach without mentioning the training phase and testing phases in their works. To say that lexicon-based approach is an unsupervised approach is not correct in general. In this work, no translation will be applied to get the polarity of words. Also, the proposed method builds a dynamic lexicon where the polarity of the words related to the corps.

The polarity of the same word can be different from corpus to another and can be changed for the same topic by adding more and more sentiments. Also, all these works classified the sentiments into two classes, +ve and –ve classes. In the current work the level of polarity is considered, the sentiment polarity can be strong +ve, +ve, netural, -ve and strongly –ve.

III. The Method

Based on one feature, namely AAL (Adaptive Arabic Lexicon), this work tries to find optimized Arabic lexicon. The problem will be written as an optimization problem and the method of optimization will be genetic algorithms. The problem can be stated as: find the lexicon that minimizes the error of polarity classifications for a given set of texts. Suppose the set of lexicons is AL and the set of texts is T. For a given text t_i in T and a lexicon l_j in AL the score of t_i with respect to l_j is the sum of the scores of all words in t_i with respect to l_j. $AAL_{l_j}(t_i) = \sum_{w \in t_i} S_{l_j}(w)$, $S_{l_j}(w)$ is the score of the word w in the lexicon l_j. t_i is classified based on the value of $AAL_{l_j}(t_i)$ according to:

$$Pre dictedClass(t_i, l_j) = \begin{cases} \text{strongly } +\text{ve} & \text{if } \frac{\max_{AAL}}{2} < AAL(t_i, l_j) \le \max_{AAL} \\ +\text{ve} & \text{if } \frac{\max_{AAL}}{4} < AAL(t_i, l_j) \le \frac{\max_{AAL}}{2} \\ \text{neutral} & \text{if } \frac{\min_{AAL}}{4} \le AAL(t_i, l_j) \le \frac{\max_{AAL}}{4} \\ -\text{ve} & \text{if } \frac{\min_{AAL}}{2} \le AAL(t_i, l_j) < \frac{\min_{AAL}}{4} \\ \text{strongly -ve} & \text{if } min_{AAL} \le AAL(t_i, l_j) < \frac{\min_{AAL}}{2} \end{cases}$$

Where

$$max_{AAL} = \max_{t_i \in T} AAL_{l_j}(t_i)$$

And

$$min_{AAL} = \min_{t_i \in T} AAL_{l_j}(t_i)$$

The accuracy $AC_{l_j}(T)$ of a lexicon l_j for the set of texts T is ratio of correctly classified texts in T, $NCCT$, to the total number of texts in T, NT:

$$AC_{l_j}(T) = \frac{NCCT}{NT}$$

So, the optimization problem can be written as: find $l_{best} = \arg\max_{l_j \in AL} A C_{l_j}(T)$

Fig. 1 shows how the above classification works. To solve the optimization problem as a genetic optimization problem, we need to define the fitness function; if we used the accuracy function as the fitness function then the algorithm will try to maximize the value of the accuracy function more than improving the classification accuracy. To get a better approach, the concept of punishment and reward will be used. This means that, if the a given text is classified correctly, then the lexicon will be rewarded by adding a positive value to the fitness function and if it did not classify the text correctly, the lexicon will be punished by adding a negative value to the fitness function. Let the fitness function be $FAAL_{AL}(T)$. The increment function $INC_{l_j}(t_i)$ is given by:

$$INC_{l_j}(t_i) = \begin{cases} |AAL(t_i, l_j)| & \text{if } l_j \text{ correctly classified } t_i \text{ and } |AAL(t_i, l_j)| \neq 0 \\ 1 & \text{if } l_j \text{ correctly classified } t_i \text{ and } |AAL(t_i, l_j)| = 0 \\ -|AAL(t_i, l_j)| & \text{if } l_j \text{ incorrectly classified } t_i \text{ and } |AAL(t_i, l_j)| \neq 0 \\ -1 & \text{if } l_j \text{ incorrectly classified } t_i \text{ and } |AAL(t_i, l_j)| = 0 \end{cases}$$

The fitness function $FAAL_{AL}(T)$ is given by:

$$FAAL_{l_j}(T) = \sum_{t_i \in T, l_j} INC_{l_j}(t_i),$$

where AL_G is the chromosomes of the current generation

Fig. 2 shows an example of the classification of a sentiment based on a given lexicon. In this example the used sentiment is "أردوغان: تعرضنا لمحاولة اغتيال اقتصادي في أغسطس" ("Erdogan: We were hit by an economic assassination attempt in August"). The algorithm distills the polarity of each term from the lexicon, add all values then it classifies the sentiment based on the proportional place of this value between min AAL and max AAL.

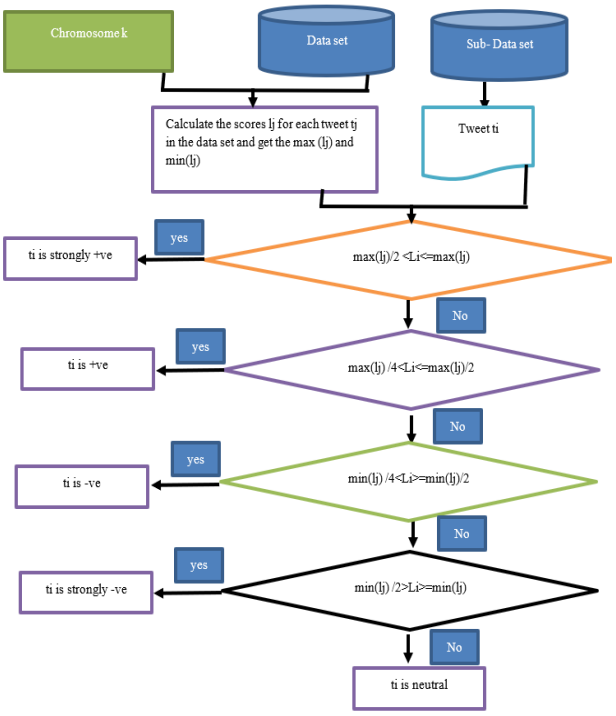

Fig. 1. How to Classify a Given Sentiment.

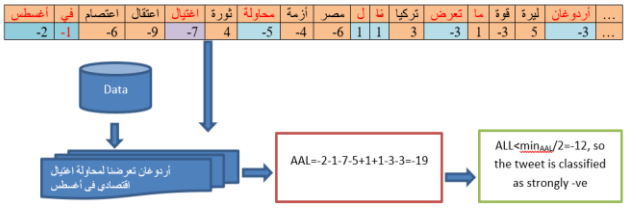

Fig. 2. Example of Classifying a Given Sentiment

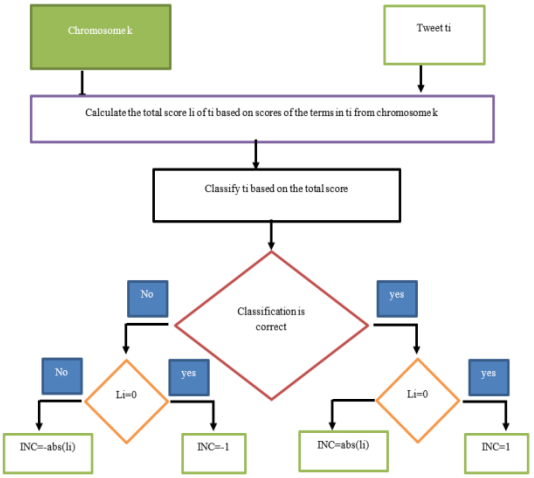

Fig. 3. Calculating INC.

Fig. 3 illustrates how to calculate INC. The following algorithm explains how to calculate the $FAAL_l\ (T)$ function of chromosome $l$in data set T

Algorithm 1 Fitness function of chromosome $l$in data set l

1. Fitness(l,T)
2. f =0
3. for each $t_i$in T
4. $AAL = 0$
5. for each word $w$in t_i
6. $AAL = ALL + S_l(w)$ //the score of $w$in chromosome l
7. end for
8. if the value AALmakes $t_i$to be classified correctly and $|AAL(t_i, l_j)| \neq 0$
9. Then $f = f + |AAL(t_i, l_j)|$
10. if the value of AAL makes t_i to be classified correctly and $|AAL(t_i, l_j)| = 0$
11. Then $f = f + 1$
12. if the value of AAL makes t_i to be classified incorrectly and $|AAL(t_i, l_j)| \neq 0$
13. Then $f = f - |AAL(t_i, l_j)|$
14. if the value of AAL makes t_i to be classified incorrectly and $|AAL(t_i, l_j)| = 0$
15. Then $f = f - 1$
16. end if
17. end for
18. return f

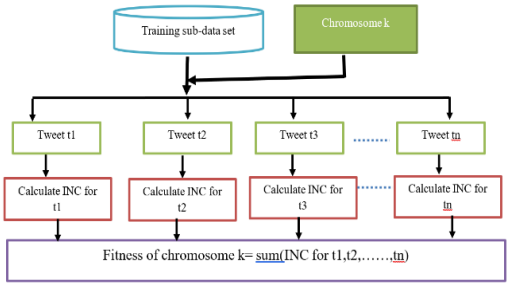

Fig. 4. Calculation of Fitness Function.

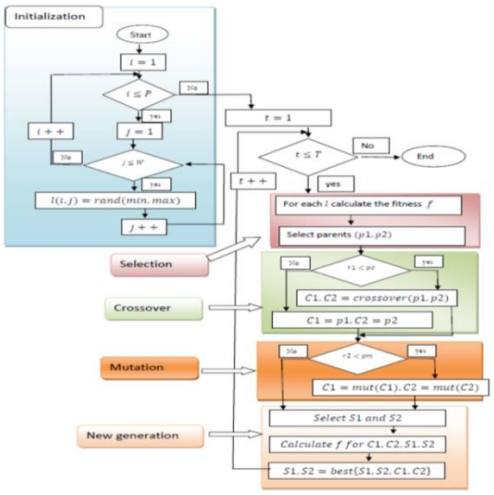

Fig. 5. Genetic Algorithm Flowchart.

Fig. 5 shows the details of the genetic algorithm. The genetic algorithm consists of five main parts. The first part is the initialization part where a random population is chosen. The algorithm will choose random vectors, the length of each vector is equal to the number of the unrepeated words and the values are distributed randomly over the interval (minv, maxv). The algorithm checked many values or minv and maxv during the training phase and kept the values that gave the best results. The second phase calculates the fitness value for each chromosome and chooses the next generation. The last phase includes crossover, mutation and replacement to generate the new generation. Based on the roulette wheel strategy, lexicon with higher fitness values are more likely to be selected. The crossover is implemented randomly. If a selected random number between 0 and 1 is less than a given probability value, then a crossover for the current parents will produce the next children otherwise the new children will be identical to their parents. A mutation is implemented, for a random selected value between 0 and 1; the mutation for the resulting children will be applied if the number is less than a specific probability. Finally, a replacement will be applied; Lexicons with lower fitness values are more likely to be replaced. Fig. 4 gives the details of the calculation of the fitness function.

IV. Experiments

In this section, data sets, parameters and results are introduced. The results of using the proposed method on the datasets are analyzed and reported.

A. Data Sets

AAL was run on five different data sets from tweets of usedrs in Twitter. These data sets were given names, TLC, MBH, NSC, SIE and TRE.

- TLC corpuscontains 701 tweets about the crises of Turkish Lira, 200 +ve , 140 –ve, 130 strongly +v, 131 strongly - ve and 100 neutral tweets.
- MBH contains 1073 tweets about Muslims brotherhood. There was 325 strongly +ve, 311 strongly –ve, 168 +ve, 131 –ve and 138 neutral tweets.

- NSC corpus contains 613 tweets about New High School regulations in Egypt. 121 strongly +ve, 115 strongly +ve, 175 +ve, 111 –ve and 91 neutral tweets.
- SIE contains 608 tweets about the last Egyptian elections. 81 strongly +ve, 211 strongly –ve. 25 +ve, 240 –ve and 51 neutral tweets.
- TRE consists of 982 tweets about the American elections. 95 strongly +ve, 315 strongly –ve, 74 +ve, 357 –ve and 141 neutral tweets. Table I summaries the data sets information.

Fig. 8 shows how the program is running. A program was written to implement the proposed algorithm. A k-fold method was used for the algorithm with k=15. Each time the data sets are divided into 15 subsets and 14 of these subsets were used as the training set, the 15th subset was used as the validation set. This process was repeated 15 times, each time one subset was used as a validation set and the remaining 14 sets were used as the training set. The final result is the average of the 15 running's of the algorithm. The range of terms polarity, crossover range and mutation rate were set as follows:

Fig. 6 shows how the crossover process is applied. Some cells are chosen randomly from each chromosome. The chosen cells from Parent A are replaced by the corresponding cells from Parent B cells in

Fig. 7 shows an example of mutation:

- The range of polarity for each term in the lexicon was set to be between -10 and +10
- A uniform crossover was applied with rate 0.8
- The mutation rate was set to 0.05

The algorithm was run till no improvement can be achieved. Sets of parameters were chosen to run the algorithm on different data sets. Namely, there were two sets of parameters which were used with two different sets of data. The original sets of data were randomly divided into two equal data sets. Equal here means that the number of sentences in each set is equal to the number of sentences in the other set.

B. Results

In this section, we will provide the results of our approach to build adaptative lexicon in terms of F1-measure for our data sets.

TABLE. I. POSITIVE, NEGATIVE AND NEUTRAL TOTAL NUMBER OF TWEETS IN EACH DATASET

Polarity	DATASET				
	TLC	*MBH*	*NSC*	*SIE*	*TRE*
Strongly Positive	131	325	121	81	95
Positive	200	168	175	25	74
Neutral	100	138	91	51	141
Negative	140	131	111	240	357
Strongly Negative	131	311	115	211	315
Total	702	1073	613	608	982

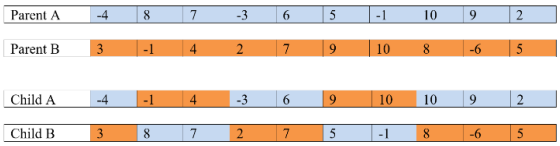

Fig. 6. Example of Crossover.

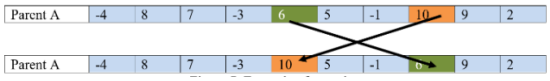

Fig. 7. Example of Mutation.

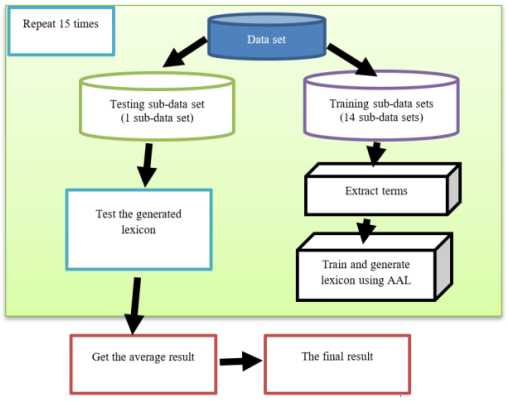

Fig. 8. K-fold Method to Test the Proposed Method.

Table II shows the results of F1-measure and Accuracy values for different mutation and crossover rates on the SIE and TRE datasets. In each case, the best values of crossover and mutation rates were reported.

TABLE. II. THE F1-MEASURE AND ACCURACY VALUES FOR DIFFERENT MUTATION AND CROSSOVER RATES ON THE SIE AND TRE DATASETS

		SIE			
pc	*pm*	*Accuracy*	*F1-Score*	*Accuracy*	*F1-Score*
0.6	0.01	75.51	78.21	81.9	83.92
0.6	0.02	81.98	78.22	80.48	80.38
0.6	0.05	80.83	78.42	84.25	80.74
0.6	0.1	76.34	76.76	83.12	77.9
0.7	0.01	74.65	78.75	82.81	78.53
0.7	0.02	78.69	75.67	81.63	82.08
0.7	0.05	80.13	79.87	83.37	81.86
0.7	0.1	83.44	78.96	82.89	76.79
0.8	0.01	75.99	79.01	80.37	82.39
0.8	0.02	74.87	75.25	77.72	80.72
0.8	0.05	77.81	75.81	81.47	82.02
0.8	0.1	80.67	76.8	81.03	82.13
0.9	0.01	81.38	78.84	81.38	81.24
0.9	0.02	77.54	75.36	79.38	78.07
0.9	0.05	77.26	77.89	81.31	81.16
0.9	0.1	79.81	77.26	83.24	82.77
1	0.01	79.4	80.31	79.91	81.14
1	0.02	80.42	75.51	82.79	75.42
1	0.05	78.83	82.38	81.47	79.62
1	0.1	75.48	77.22	77.66	78.65

For testing mutation and crossover rate settings, we examined different values. For these two datasets Fig. 9 and Fig. 10 show the relation between different parameter values and F-measure. For each dataset and setting, the algorithm was run. The results were reported based on averaging running. From the results we can conclude that the best performance was at values between 0.6 and 0.9 for crossover and at values between 0.05 and 0.1 for mutation. To insure the results independence from crossover and mutation rates, crossover and mutation rates were fixed at 0.8 and 0.06. Reviewing the results in Table III, the proposed method gave good results that outperform the current available methods in many cases. Regarding the number of iterations, a limited number of iterations, 100,000 iterations were enough, and conversion was achieved for small data sets. For big data sets, the conversion was achieved with iterations numbers around 250000 iterations. This leads us to consider iterations number 250000 for all data sets.

TABLE. III. AAL RUNNING RESULTS ON ALL DATA SETS

Method	***Accuracy (%)***	***Precision (%)***	***Recall (%)***	***F1 (%)***	***Precision (%)***	***Recall (%)***	***F1 (%)***	***Average F1 (%)***
TLC-dataset								
Bing=Liu-Lexicon	66.0	72.0	27.9	40.2	66.3	90.6	76.6	58.4
Random-search	51.8	39.3	42.6	40.9	65.9	55.3	60.2	50.5
AAL	83.1	77.1	79.3	78.2	85.5	84.2	84.9	81.5
AAL-SW	79.4	74.6	74.6	74.6	85.0	84.6	84.8	79.7
AAL+1,2,3-grams	86.5	81.9	78.3	80.1	88.7	88.8	88.8	84.4
AAL+lex	84.7	84.1	79.0	81.5	91.5	87.3	89.4	85.4
AAL+lex+1,2,3-grams	86.5	86.1	81.4	83.7	88.7	88.9	88.8	86.2
Best-reported-results-from-the-literature	80.7	75.2	77.6	76.4	85.5	88.9	87.2	81.8
MBL-dataset								
Bing=Liu-Lexicon	69.5	73.8	73.1	73.4	73.8	71.3	72.6	73.0
Random-search	52.9	45.2	39.7	42.3	52.6	60.8	56.4	49.4
AAL	83.1	77.7	88.7	82.8	86.8	78.1	82.2	82.5
AAL-SW	81.1	80.4	84.2	82.3	77.8	79.8	78.8	80.5
AAL+1,2,3-grams	85.3	83.6	87.2	85.3	87.6	84.6	86.1	85.7
AAL+lex	88.0	82.2	85.7	83.9	86.0	89.1	87.5	85.7
AAL+lex+1,2,3-grams	85.9	86.4	89.4	87.9	89.9	88.8	89.3	88.6
Best-reported-results-from-the-literature	97.6	83.9	86.8	85.3	87.1	85.6	86.3	85.8
NSC-dataset								
Bing=Liu-Lexicon	81.6	22.9	33.5	27.2	60.8	83.2	70.3	48.7
Random-search	59.1	24.8	40.0	30.6	60.8	44.4	51.3	41.0
AAL	75.0	39.3	33.6	36.2	65.0	65.7	65.3	50.8
AAL-SW	72.1	39.3	30.8	34.6	59.0	66.4	62.5	48.5
AAL+1,2,3-grams	82.9	61.3	37.4	46.5	63.9	72.5	67.9	57.2
AAL+lex	80.5	42.1	35.5	38.5	64.0	65.0	64.5	51.5
AAL+lex+1,2,3-grams	82.5	62.8	40.9	49.5	66.3	76.5	71.0	60.3
Bes-reported-results-from-the-literature	80.4	55.8	60.4	58.0	65.9	69.7	67.7	62.9
SIE-dataset								
Bing=Liu-Lexicon	69.5	61.2	23.7	34.1	66.4	94.8	78.1	56.1
Random-search	52.7	62.6	61.5	62.1	36.0	40.2	38.0	50.0
AAL	78.9	76.4	71.3	73.8	85.6	87.5	86.5	80.1
AAL-SW	77.5	68.4	65.0	66.7	78.2	82.9	80.5	73.6
AAL+1,2,3-grams	80.0	75.0	71.7	73.3	83.0	86.2	84.6	79.0
AAL+lex	80.4	77.3	68.5	72.6	80.6	84.5	82.5	77.6
AAL+lex+1,2,3-grams	83.7	78.2	74.4	76.2	87.1	90.0	88.5	82.4
Best-reported-results-from-the-literature	82.9	75.9	67.4	71.4	82.8	87.6	85.2	78.3
TRE-dataset								
Bing=Liu-Lexicon	72.8	68.2	91.2	78.1	84.9	57.7	68.7	73.4
Random-search	53.8	55.1	47.6	51.1	49.6	53.3	51.3	51.2
AAL	77.8	78.9	76.2	77.5	79.1	78.0	78.5	78.0
AAL-SW	79.3	82.6	75.0	78.6	78.8	82.8	80.7	79.7
AAL+1,2,3-grams	82.1	80.6	88.6	84.4	86.2	78.7	82.2	83.3
AAL+lex	78.4	83.3	80.8	82.0	81.3	82.1	81.7	81.9
AAL+lex+1,2,3-grams	85.9	83.2	90.7	86.8	88.5	85.6	87.0	86.9
Best-reported-result-from-the-literature	88.0	85.9	91.8	88.8	84.6	89.2	86.8	87.8

TABLE. IV. ACCURACY AND F1 VALUES FOR 0.95 CONFIDENCE INTERVAL FOR ON THE FIVE DATASETS

	TLC	*MBH*	*NSC*	*SIE*	*TRE*
Accuracy-of-AAL	82.26±2.13	82.81±2.04	67.67±2.2	80.81±1.75	69.91±2.64
F1-Score-of-AAL	79.85±2.42	80.73±2.06	68.65±1.68	75.82±2.18	71.95±2.35
Accuracy-of-AAL+lex+n-grams	87.92±2.17	87.26±2.62	64.74±2.2	82.29±1.69	76.04±2.28
F1-Score-of-AAL+lex+n-grams	84.13±2.27	87.66±3.02	69.95±2.64	81.29±1.88	79.81±2.37

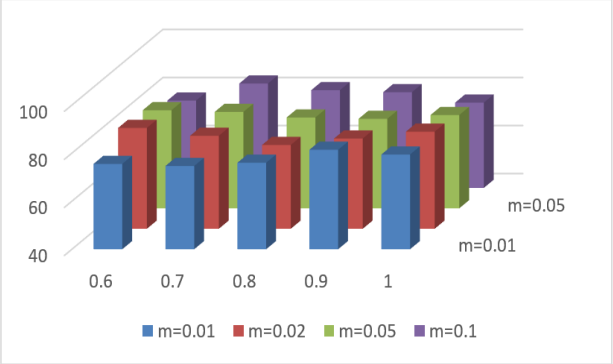

Fig. 9. Values of F1-Measure for Multiple Mutation and Crossover Rates on the SIE Dataset.

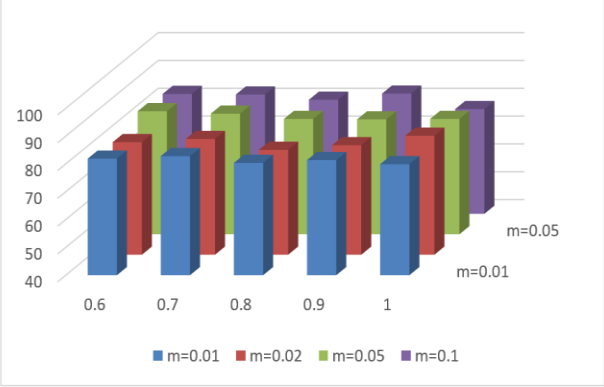

Fig. 10. Values of F1-Measure for Multiple Mutation and Crossover Rates on the TRE Dataset.

C. Discussion

Random search approach and Bing Liu's lexicons are considered the best methods. So, it was natural to compare the performance of the proposed method with these approaches. Table IV shows the comparison results. Best values are bolded. In the random search, based on the representation in the proposed algorithm, a random value is given to initiate a single chromosome. For 250,000 iterations, a neighbor of chromosome is given through changing a single cell in it randomly. If the fitness value of the generated neighbor is higher (based on AAL calculations), the neighbor replaces the original one. A confidence interval is reported since the algorithm is run fifteen times for each fold and it each fold and we have 15 folds. The 0.95 confidence intervals are shown in Table IV. Many variations enhance the AAL performance. AAL-SW is the AAL after removing the stop words. AAL+1,2,3-grams are variations of AAL are the result of applying AAL supported by n-grams features. Enhancing AAL by considering features of meta-level Bing Liu's lexicon produces a modified version of AAL, AAL+lex. Adding n-grams features and metalevel features of Bing Liu lexicon improves the results and makes them better in many measures in the datasets. From Table III, we note that AAL alone could to outperform the other methods in MBH data set. This is due to the clearness of positivity and negativity levels in this data set. However, the worst results of AAL were in NSC data set, also this due to that the level of polarity ambiguity in this data set is the highest among the other data sets. The results reflect a promising result based on using AAL alone. As a classifier, AAL results outperform other classifiers, see [7], [9]. Falsely results in AAL can be explained because of tone of tweet problem. The terms that have low frequency tend to have higher variance when running the algorithm multiple times. Consequently, those terms tend to have improper values. The standard deviation of scores values of sentiment of terms is shown in Table IV.

V. CONCLUSION

In this work, we proposed a genetic algorithm to build an adaptive Arabic lexicon for sentiment analysis. We can report that the F-measure of AAL is 4.13 percentage points better than the average of reported results on the MBH dataset, 3.28 on the TLC dataset, 2.14 on the SIE dataset, and 1.56 on the TRE dataset. AAL achieved accuracy levels better than traditional methods on three data had better accuracy results than state-of-the-art methods on three datasets. For F-measure results, the proposed method achieved better results in four datasets. This work shows that adaptive lexicons can be applied for Arabic language. In fact, the independence of the method from the language is approved. The proposed method can enable better understanding of sentiment words. Since, we did not remove stop words, then this show that all words in Arabic can be considered as sentiment words. In this paper, we approved that writing generating adaptive lexicon as optimization search and applying genetic algorithms to get optimal solution can give an excellent result when applied to Arabic language. It is shown that, AAL can give a high accuracy with small data sets. From the business point of view, the companies can use AAL to create lexicons to help in finding and exploring what users think about. Companies can also use AAL to enrich the knowledge about individual words and their importance; this will increase the effectiveness of manual analysis of sentiments. For example, A supermarket manager can use AAL to create a lexicon for the products and use it for sentiment analysis of their customers behaviors. In this paper, AAL used to analyze the strength of opinions of sentiments. In the future, building a deep net that can apply AAL online with active learning to provide real time adaptive lexicons will be explored.

ACKNOWLEDGMENT

The authors would like to thank Taif University for funding this research under project number: 5607-438-1.

REFERENCES

[1] H. Keshavarz, M. S. Abadeh, "ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs", Knowledge-Based Systems 122, pp. 1–16, 2017.

[2] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification", in: Proceedings of the ACL Student Research Workshop, ACLstudent'05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 43–48, 2005.

[3] S.M. Mohammad, P.D. Turney, "Crowdsourcing a word–emotion association lexicon", Comput. Intell. 29 (3),pp. 436–465, 2013.

[4] W. J. Wiebe, P. Hoffmann, "Recognizing contextual polarity in phraselevel sentiment analysis", in: Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, Canada, pp. 347–354, 2005.

[5] M.M. Bradley, P.J. Lang, "Affective Norms for English Words (ANEW) Instruction Manual and Affective Ratings", Technical Report C-1, The Center for Research in Psychophysiology University of Florida, 2009.

[6] C. L. Sarmento, M.J. Silva, E. de Oliveira, "Clues for detecting irony in user-generated contents: oh...!! it's so easy;-)", in: Proceeding of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion, Hong Kong, China, pp. 53–56, 2009.

[7] Go, R. Bhayani, L. Huang, "Twitter Sentiment Classification using Distant Supervision", Technical report Stanford University, 2010.

[8] B. Marquez, M. Mendoza, B. Poblete, "Meta-level sentiment models for big social data analysis", Knowl. Based Syst. 69,pp. 86–99, 2014.

[9] Liu, W. Li, M. Guo, "Emoticon smoothed language models for Twitter sentiment analysis", in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Canada, 2012.

[10] N.F.F. da Silva, E.R. Hruschka, E.R. Hruschka, "Tweet sentiment analysis with classifier ensembles", Decis. Supp. Syst. 66, pp. 170–179, 2014.

[11] X. Hu, L. Tang, J. Tang, H. Liu, "Exploiting social relations for sentiment analysis in microblogging", in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 537–546, 2013.

[12] H. Saif, Y. He, H. Alani, "Semantic sentiment analysis of twitter", in: Proceedings of the 11th International Conference on The Semantic Web, ISWC'12, Springer-Verlag, pp. 508–524, 2012.

[13] B. Marquez, M. Mendoza, B. Poblete, "Combining strengths, emotions and polarities for boosting twitter sentiment analysis", in: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, 2013.

[14] Y. Kaewpitakkun, K. Shirai, M. Mohd, "Sentiment lexicon interpolation and polarity estimation of objective and out-of-vocabulary words to improve sentiment classification on microblogging", in: Proceedings of 28th Pacific Asia Conference on Language, Information and Computation, pp. 204–213, 2014.

[15] H. Saif, M. Fernandez, Y. He, H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of Twitter", in: Proceedings of the 9th language resources and evaluation conference (LREC),pp. 810–817, 2014.

[16] B. T. Uricchio, M. Bertini, A. Del Bimbo, "A multimodal feature learning approach for sentiment analysis of social network multimedia, Multimed". Tools Appl, pp. 1–19, 2015.

[17] B. P. Arora, S. Madhappan, N. Kapre, M. Singh, V. Varma, "Mining sentiments from tweets", in: Proceedings of the WASSA, 2012.

[18] F. Wu, Y. Huang, Y. Song, "Structured microblog sentiment classification via social context regularization", Neurocomputing, 175, pp. 599–609, 2016.

[19] L. W. Li, M. Guo, "Emoticon smoothed language models for Twitter sentiment analysis", in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, Toronto, Canada, 2012.

[20] F. Nielsen, "A new anew: evaluation of a word list for sentiment analysis in microblogs", in: Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages, Heraklion, Crete, Greece, 2011.

[21] E. F. Sebastiani, "Sentiwordnet: a publicly available lexical resource for opinion mining", in: Proceedings of the 5th Conference on Language Resources and Evaluation, pp. 417–422, 2006.

[22] B. A. Esuli, F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining", in: Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valletta, Malta, pp. 2200–2204, 2010.

[23] M. Thelwall, K. Buckley, G. Paltoglou, "Sentiment strength detection for the social web", J. Am. Soc. Inf. Sci. Technol, 63 (1),pp. 163–173, 2012.

[24] A. H. Chen, A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums", ACM Trans. Inf. Syst, 26 (3), pp. 12-34, 2008.

[25] B. Gómez, N. Luis Mingueza, M.C. García del Pozo, OpinAIS: "An artificial immune system-based framework for opinion mining", Int. J. Artif. Intell. Interact. Multimed. 3, pp. 25-29, 2015.

[26] A. E. Mayfield, C. Penstein-Rosé, E. Nyberg, "Sentiment classification using automatically extracted subgraph features", in: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, 2010.

[27] Govindarajan, "Sentiment analysis of movie reviews using hybrid method of Naive Bayes and genetic algorithm", IJACR, 3 (4), pp. 139–145, 2013.

[28] H. S. Ibrahim, Sherif M. Abdou and Mervat Gheith, "Sentiment Analysis for Modern Standard Arabic and Colloquial", International Journal on Natural Language Computing (IJNLC), 4(2), pp. 95-109, April 2015.

[29] H. H. Mustafa, A. Mohamed, and D. S. Elzanfaly, "An Enhanced Approach for Arabic Sentiment Analysis", International Journal of Artificial Intelligence and Applications (IJAIA), 8(5), pp. 1-14, September 2017.

[30] M. Al-Ayyoub, S. Bani Essa; I. Alsmadi, "Lexicon-based sentiment analysis of Arabic tweets", International Journal of Social Network Mining (IJSNM), 2(2), pp. 1-14, 2015.