Sentiment Analysis of Social Media Micro Blogs using Power Links and Genetic Algorithms

Mahmoud Rokaya
College of Computers and Information Technology, Taif University
Taif, Saudi Arabia
Faculty of Science, Tanta University
Tanta, Egypt

ABSTRACT
In the current years, social media became one of the most important sources of data for different data analytic purposes. One of the most important issues is how to map different trends in social media and define the relation between different groups based on their sentiment or interests. In this paper, a two-phase approach is used for clustering a set of blogs. At the first phase, the approach builds a lexicon that provides the polarity of each word. In the second phase, the approach clusters the blogs bases on the polarity features and the Power Link features. The output of the second phase is used as the input of the first phase to get an improved lexicon. This process will continue in a loop between phase one and phase 2 till a stable set of clusters is gotten. The approach aims to develop a non-supervised cluster agent that can correctly cluster micro blogs and define different interests of different groups of people. The results of the approach are expressed in terms of precision, recall and F-measure.

General Terms
Artificial Intelligence. Natural Language Processing. And Text Analysis.

Keywords
Sentiment Analysis, Evolution Calculation, Genetic Algorithms, Power Links, Social Media, Micro Blogs.

1. INTRODUCTION
Classifications and clustering are two basic tasks in machine learning and data science [1]. Classifications are used when a set of labels are known, and it is needed to define the suitable label for a specific item [2]. Supervised learning is related to classification problem. In the supervised learning, a set of items with known labels and this set is used to train the machine hoping that it can give the write label for items that their labels are unknown. [3]. Clustering means to group similar items in groups, there this no known labels in advance [4]. Clustering is related to unsupervised learning algorithms. In the clustering problem, there is a set of items and it is needed to group each similar set of items in one group. In classifying and clustering, the concept of features and extraction of the features are essential tasks [5]. In classifying, the training aims to train the machine to combine the features in some way to decide to which label the item best probably could be mapped. In clustering, the machine depends on the features and the distance function. Distance function computes the similarity between items using the distance function. Similar items will have shorter distance [6].

In social media sentiment analysis, using clustering and classification is popular [7]. The classifications approaches can be Lexicon-based and corpus-based methods [8] The approach of Lexicon-based methods depends on predefined static lexicons to determine the polarity of individual words in a corpus, then using a polarity function to determine the polarity of each blog. There are many works that adopted this approach, for example [9]. The corpus-based approach depends on the semantic relation of the words. This make the semantic approach is language dependent [7]. There are many works that adopted this approach, for example [10, 11]. The second approach is language dependent and the first approach is independent from language. Also, dependency of the lexical approach on a fixed lexicon seems to be unsuitable with the highly speed chaining of the content in social media. Also, depending on the language properties seems to be not much effective, considering that the social media users rarely respect the grammar or spelling of the language , also, many times, it is found that the users uses mixed languages or even used the alphabet of another language to express their own words. However, in terms of accuracy the semantic approach gave better results. Cluster methods used in social media are metadata-based approaches, lexical semantics approaches or hybrid. The lexical semantic approaches depend on the meta data of the blogs not on the words on the words in the blog. From the other hand, the contextual semantics approaches depend on the context properties of the blogs [12]. For example [13, 14] used the external semantic properties to cluster hashtags. [15, 16] used the contextual sematic properties. [12] used a hybrid approach.

The mentioned methods either use the supervised learning, which become not preferred these millions of tons of data that appear every day, or they are related to the language features or to sematic features of the content or the meta-sematic and this means that it is needed to develop a different method for each group of languages. Depending on quantitative data a clustering method will solve the two problems. For sure adding the languages features or the meta-structure of the language will improve the performance. In this paper, a clustering method for micro blogs that is not directly dependent on language properties will be presented.

Genetic Algorithm has its origin as a method to solve the optimization problem regardless the nature of the optimization function or the constrains [17]. Genetic algorithms adopt the natural evolution [18]. [7,8] used the genetic algorithms to optimize lexicons for sentiment analysis. [8] explained the uses of genetic algorithms uses in general and for specific purpose of sentiment analysis and building lexicons.

A lexicon is the vocabulary of a person, language, or branch of knowledge. Lexicons in most cases are represented by two columns, the first column, presents the words, the second columns present the function of the lexicon. For sentiment analysis, the second column presents the polarity value of the

word. Some works tried to build lexicons based on solving an optimization problem such as [7, 8].

Power Links concept was suggested by Rokaya and Atlam [19]. Power links has many applications in test extraction and summarization [20, 21]. Power links was used in context spelling [22, 23] and information retrieval [24].

This paper aims to develop a method for clustering micro blogs. The method is dependent from the language's properties, so it can be applied for any language. The method adopts a non-supervised approach. The method can be divided into two main phases, the lexicon building phase and the clustering phase. The output of the first phase is the output for the second phase, so, there is a loop that can be repeated until reaching accepted results. In the second phase, the clustering process goes in two steps, the first step is a clustering based on Power Links. Each cluster resulting from the power link clustering step is further classified into two groups based on the polarity of each blog the cluster. The polarity is calculated based on the polarity of the words. Polarity of words is gotten from the generated lexicon in the first phase.

The remaining sections of the paper are as follows. Section 2 presents the general architecture of the proposed method. It also gives the basic definitions of Power Link (adopted in this work) and the genetic algorithm general architecture. Section 2 gives the details of building the lexicon based on the genetic algorithm and the details of the clustering algorithms. Section 3 presents the experiments that are used to test the method. Finally, section 4 presents the results and the future work.

2. THE METHOD

Fig1 shows the architecture of the clustering algorithm. The first phase used to produce the lexicon. The second phase uses the lexicon beside the Power Link features to produce the clusters of the blogs. Again, these clusters will be used as the new base to rebuild the lexicon and again the lexicon will be used beside the Power Links to produce the new clusters. This process will be repeated until the clusters become stable. The initial lexicon is any available one. The method depends on developing the ability of the method through improving the lexicon through training loops as well as adding new words with proper links that contribute in giving the correct class of a blog. The method depends on considering the current class are the base to classify a new set of blogs. Adding the new set will modify the values of the polarity of each word in the current lexicon to improve the lexicon ability to classify new blogs with a higher accuracy. This process will continue and by the time the chance of improving the method performance will never stop. The experiments will show how this process can proceed.

Let w_i be a word in a blog b_j from a set of blogs S_k. Suppose a lexicon l_m is used and the polarity score of w_i in l_m is given by $pwl(w_i)$ then the polarity score of b_j is given by:

$$PSB(b_j, S_k, l_m) = \sum_{w_i \in b_j} pwl(w_i)$$

To decide the polarity class of a blog b_j, the function $PC(b_j, l_m)$ is used

$$PC(b_j, l_m) = \begin{cases} R1_+ & \text{if } PSB(b_j, S_k, l_m) > 0 \\ R1_- & \text{if } PSB(b_j, S_k, l_m) \le 0 \end{cases}$$

The accuracy of l_m is given its ability to map blogs to the correct class. Let $A(l_m, S_k)$ be the accuracy function, then

$$A(l_m, S_k) = \frac{\text{number of blogs in } S_k \text{ that } l_m \text{ predicted their class correctly}}{\text{number of blogs in } S_k}$$

Now, one can write the problem of finding the best l_m as an optimization problem.

For S_k , $l_{opt} = \arg\max_{l_m} A(l_m, S_k)$

According to [8], it is better to use a penalty function as a fitness function rather than using the accuracy function itself. The penalty function can be stated as:

$$P(S_k, b_i, l_m) = \begin{cases} \left| \frac{PSB(b_j, S_k, l_m)}{\theta} \right| & \text{if } PSB(b_j, S_k, l_m) \text{ failed to classify } b_j \text{ correctly} \\ -\left| \frac{PSB(b_j, S_k, l_m)}{\theta} \right| & \text{if } PSB(b_j, S_k, l_m) \text{ classified } b_j \text{ correctly} \end{cases}$$

Where θ is the penalty factor. The lexicon will get a positive penalty if it failed to classify a blog to its correct class and it will get a negative penalty if the lexicon succeeded to classify a given blog to its correct class. The fitness function for a lexicon l can be written as the sum of penality values for each blog classifies using l. Fig. 2 illustrates the steps of how to use the penalty $P(S_k, b_i, l)$ function as a fitness function $F(S_k, l) = -\sum_{b_i \epsilon S_j} P(S_k, b_i, l)$

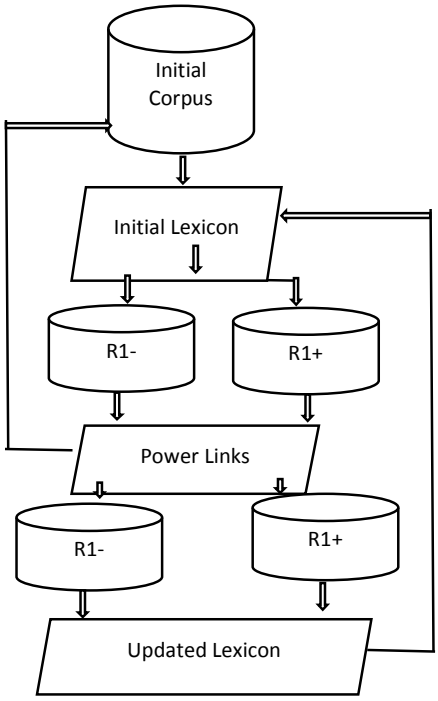

Fig.1 System Overview

The genetic algorithm works through five basic steps, initialization, selection, crossover, mutation and replacement. The initialization step means to initiate the values of all chromosomes randomly. In this case the chromosomes are available lexicons in a given range. Two values, maxpol and minpol are chosen then the polarity of each word is set randomly to be any integer between maxpol and minpol. The selection step is done through the random selection based on roulette wheel strategy. The crossover is achieved by replacing the values in two chromosomes in predefined places with a specific probability. Also the mutation is achieved by

selecting some chromosomes in random places in the same chromosome and for a specific probability. The fitness function determines which set of chromosomes will be selected in the next evolution. Fig.3 shows the detail of the algorithm

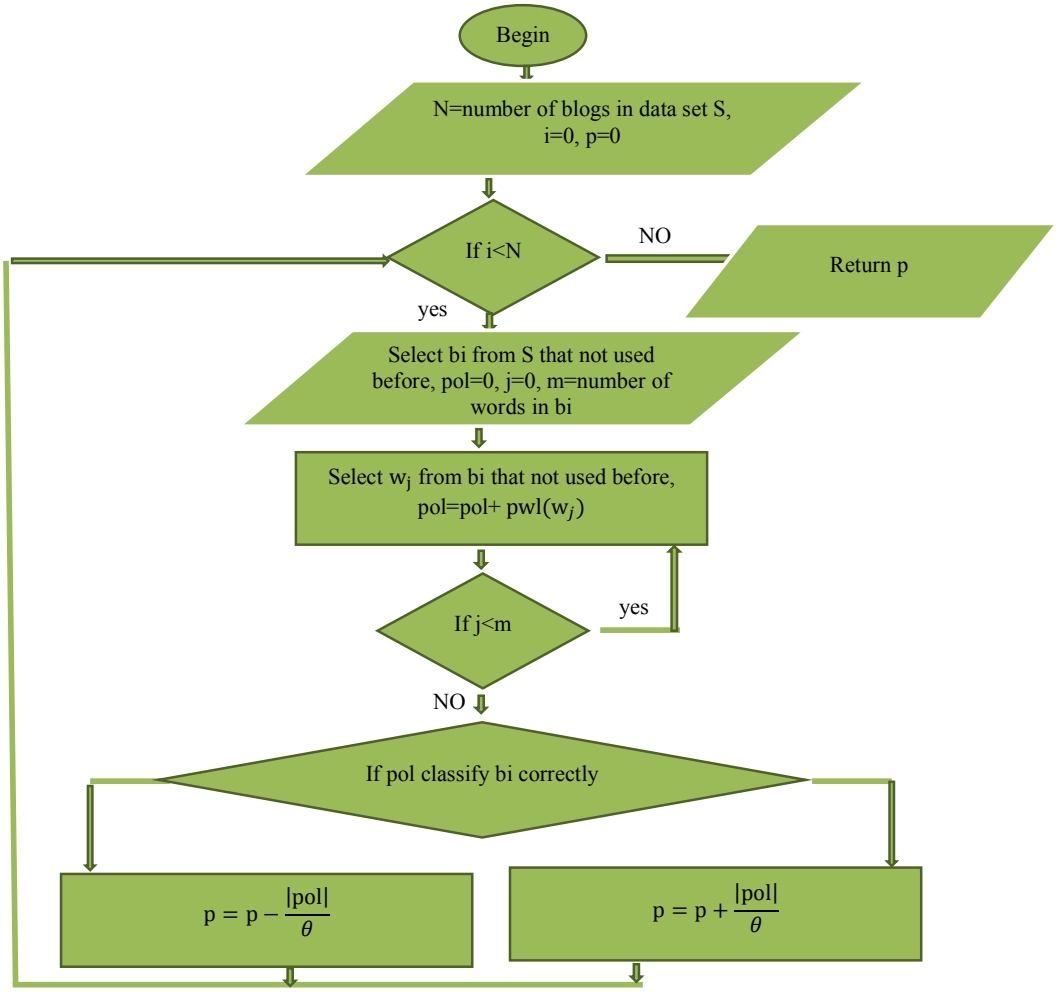

Fig 2 Fitness function calculation

3. POWER LINKS

Power links different from the ordinary frequency, instead of counting the frequency of individuals words, one counts the cooccurrence of two words. Power links is simpler the 2-gram counting, since the algorithm calculates the Power Link based on the appearance of the words in the same document, sentence or micro-blog. The relative distance between the words in the micro-blogs might affect the relation between the words, however, to reduce the calculation overload and since the length of the blogs is short, the Power Link is independent from the distance between words. The distance between words will not considered in the calculation.

In this paper, the Power Link is defined for two different effective words as the number of blogs that the two words appeared in divided by the total number of blogs. Effective words mean non-stop list words.

Let w_1 and w_2 are non-stop list words, where w_1 and w_2 appears simultaneously in $N_{w_1w_2}$ blogs and N is the total number of blogs in a given corpus, then the Power Link between w_1 and w_2,

$$\begin{cases} PL(w_1, w_2) = \frac{N_{w_1w_2}}{N} & \text{if } w_1, w_2 \text{ polarity of the same type} \\ 0 & \text{if } w_1, w_2 \text{ has different polarity} \end{cases}$$

Since, the lengths of the blogs are similar, no need to normalize the Power Link based on the lengths of the blogs. For each word w the algorithm calculate the Power Link between w and other words in the current corpus. This will form a matrix M of dimension $N \times N$, where N is the total number of words.

The lexicon that will be used will be produced in the first phase of the algorithm.

3.1 Classifying based on power links.

Classifying of blogs based on the Power Links is a clustering algorithm. The algorithm begins by selecting one of the blogs then calculates the distance between this blog and each blog in both classes R1+ and R1-. The blog will be added to the class with minimum average distance. This process will continue till all blogs are exhausted.

3.2 Distance between two blogs

For each blog b, define a vector blog vb, the length of vb is N. The element k in vb is the average of the corresponding row in M if $w_k \in b$, else the value is 0. The distance between two blogs b_1, b_2 is the Euclidian distance between the corresponding blog vectors vb_1, vb_2. Fig4 illustrates the details of clustering algorithm based on Power Links.

4. EXPERIMENTS.

4.1 Data sets

In the experiments, a diversity data sets and multilingual is used. Five sets in English. These sets are Healthcare reform (HCR), SemEval dataset, Sanders –Twitter sentiment corpus, The Stanford Twitter dataset (STS dataset) and Obama-McCain debate (OMD) dataset [8] and another five data sets in Arabic. These sets are the crises of Turkish Lira (TLC), Muslims brotherhood (MBH), New High School regulations in Egypt (NHS), Egyptian elections (SIE) and American elections (TRE) [7]. Table 1 shows the distribution of each set for a negative and positive blog.

4.2 Experimental Steps.

In our experiments, the value of minpol and maxpol are -10 and 10 respectively. The crossover and mutation are done based on 0.01 probability. The stop words list is calculated based on the frequency of each word. If a word appeared in all data sets and in each data set this word appeared in more than the half of the blogs then this word will be considered as a stop list word

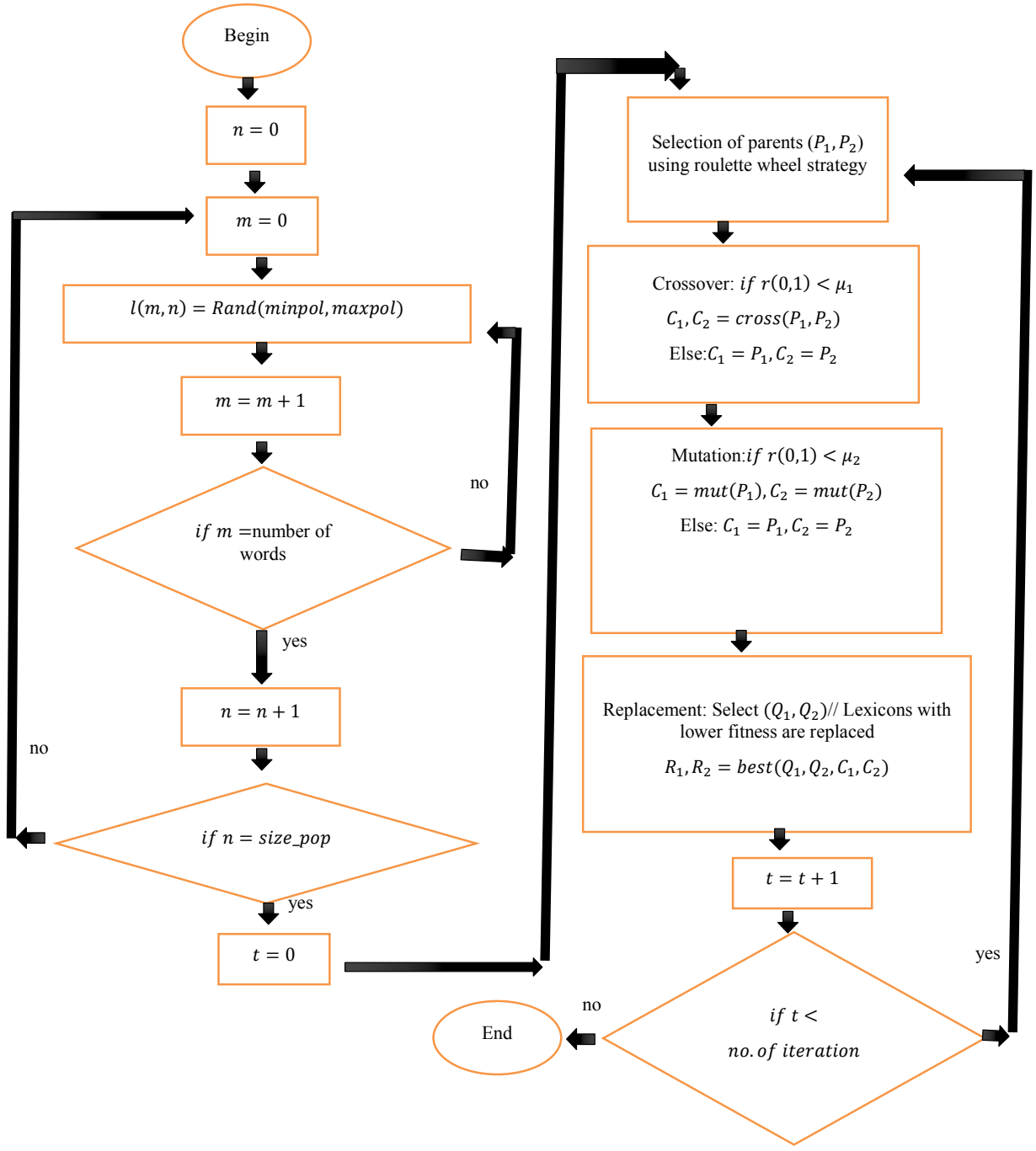

Fig.3 Genetic algorithm to optimize the accuracy of the lexicon

Table1. Negative, positive and total number of blogs in the data sets

	+ve	-ve	Total
HCR	917	369	1286
SemEval	3640	1458	5098
STS	182	177	359
Sanders	570	654	1224
OMD	710	1196	1906
TLC	431	271	702
MBH	631	442	1073
NHS	387	226	613
SIE	157	451	608
TRE	310	672	982

The experiments go according to the following steps:

1. One of the data sets is chosen randomly and the blogs in this data set are classified based on Bing Lui lexicon.
2. The genetic algorithm is used to produce the next lexicon
3. The Power Link clustering algorithm is used to redistribute the blogs
4. The genetic algorithm is used again to produce a refined lexicon
5. The current lexicon is used to classify a new data set
6. The genetic algorithm will be used to produce a refined lexicon with one restriction, the algorithm will work to calculate the polarity of new words that never appear in the previous calculations
7. The steps 3, 4 and 6 are repeated to classify and refine the lexicon till all data sets are exhausted
8. The steps from 1 to 8 are iterated and in each iteration, begin with a different data set. In each iteration, the precision, recall and F-measure are calculated.
9. The iteration that gave the best precision, recall and F-measure is chosen to be the standard lexicon.

The results of these experiments will be tested based on repeating the experiments according to the following:

A. Performing a classifying based on Bing-Lui lexicon only (BLC) [8]

B. Performing a classifying based on random search method (RS) [8]

C. Performing steps from 1 to 9 without using the Power Link clustering (GLO)

D. Refining the results of 1 using the Power Link clustering (BLCP)

E. Performing the experiments based on steps from 1 to 9 (GLOP)

Note that A and B are a supervised method. In these methods it is needed to assign a set for training and a set for testing. In these experiments, a 10-fold method is used. The whole data sets are divided randomly into 10 parts. In each fold, one of the folds is used as a test set and the others are used as a training set. The results below report for the average results for each data sets across the 10 testing iterations.

5. RESULTS AND DISCUSSION

In this section, results for the experiments sets A to E are reported.

Tables 2 to 11 reports the results of the proposed 5 methods for the 10 data sets. Also, the genetic algorithm was implemented using a different value of crossover and mutation probability, μ_1 and μ_2 respectively.

Table 2. Recall, Precision and F values of the data set HCR

	Positive Class			negative Class			AVG F
Method	R	P	F	R	P	F	
BLC	45.57%	45.77%	45.67%	75.94%	77.40%	76.66%	61.17%
RS	50.88%	50.11%	50.49%	65.36%	64.96%	65.16%	57.83%
GLO	94.95%	93.67%	94.31%	85.64%	86.22%	85.92%	90.12%
BLCP	54.31%	51.74%	52.99%	81.80%	79.40%	80.58%	66.79%
GLOP	66.40%	67.12%	66.76%	75.03%	74.40%	74.71%	70.74%

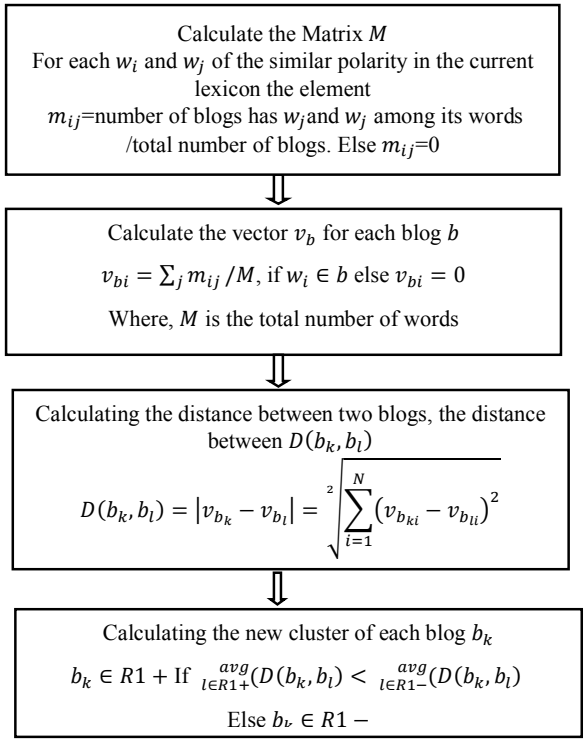

Fig4. Power Link clustering steps

The bold values show the values of the method that outperformed the other methods for the same data set based on F-measure. The results show that GLOP method outperformed all other methods in eight of the data sets. The values of μ_1 and $\mu_2$affects the convergence rate but these values should be less than 0.1 else the algorithm will never converge.

Table 3. Recall, Precision and F values of the data set SemEval

	Positive Class			negative Class			AVG F
Method	R	P	F	R	P	F	
BLC	71.06%	71.91%	71.48%	40.26%	42.39%	41.30%	56.39%
RS	50.42%	48.97%	49.68%	38.71%	39.74%	39.22%	44.45%
GLO	72.42%	74.20%	73.30%	87.54%	85.20%	86.35%	79.83%
BLCP	69.63%	70.95%	70.28%	66.54%	68.70%	67.60%	68.94%
GLOP	83.46%	81.03%	82.23%	86.09%	85.64%	85.86%	84.05%

The number of iterations for convergence was between 150000 iterations and 250000 iterations.

Table 4. Recall, Precision and F values of the data set STS

	Positive Class			negative Class			AVG F
Method	R	P	F	R	P	F	
BLC	84.83%	82.63%	83.72%	68.41%	66.29%	67.33%	75.52%
RS	66.61%	66.87%	66.74%	71.57%	71.37%	71.47%	69.10%
GLO	74.25%	74.07%	74.16%	55.03%	54.10%	54.56%	64.36%
BLCP	71.44%	69.91%	70.67%	87.69%	84.88%	86.26%	78.46%
GLOP	95.03%	93.46%	94.24%	87.66%	89.11%	88.38%	91.31%

It is noted that the number of iterations increase when the value of μ_1 and μ_2 decrease. Also, it is noted that merging the Power Link clustering method improved the results of the methods in average.

The main results that can be reported here is that a non-supervised method could present a performance that equal to or even overcome the performance of a supervised methods.

It is easily to note that the results of English data sets were better than the results of Arabic data sets. This is due to the quality of the initial used lexicons in each case and also due to the difficulty in classification of many Arabic blogs to a correct +ve or -ve class manually. Many of the blogs gave no clear sentiment or gave +ve and -ve sentiment in the same time.

Table 5. Recall, Precision and F values of the data set Sanders

	Positive Class			negative Class			AVG F
Method	R	P	F	R	P	F	
BLC	50.44%	48.38%	49.39%	35.00%	36.63%	35.80%	42.59%
RS	77.67%	75.44%	76.54%	57.85%	56.99%	57.42%	66.98%
GLO	76.38%	77.68%	77.02%	56.49%	57.26%	56.87%	66.95%
BLCP	58.11%	56.93%	57.51%	65.70%	65.16%	65.43%	61.47%
GLOP	88.16%	89.62%	88.88%	92.80%	95.55%	94.15%	91.51%

Table 6. Recall, Precision and F values of the data set OMD

	Positive Class			negative Class			AVG F
Method	R	P	F	R	P	F	
BLC	57.12%	58.00%	57.56%	38.71%	87.08%	53.60%	55.58%
RS	54.87%	53.03%	53.93%	38.28%	60.60%	46.92%	50.43%
GLO	86.53%	87.49%	87.01%	61.77%	84.35%	71.32%	79.16%
BLCP	82.57%	85.10%	83.82%	46.80%	87.68%	61.03%	72.42%
GLOP	77.69%	78.76%	78.22%	55.48%	82.26%	66.27%	72.24%

Table 7 Recall, Precision and F values of the data set TLC

	Positive Class			negative Class			AVG F
Method	R	P	F	R	P	F	
BLC	74.75%	77.64%	76.17%	39.21%	36.22%	37.66%	56.91%
RS	74.24%	72.93%	73.58%	45.02%	43.39%	44.19%	58.88%
GLO	83.05%	84.25%	83.65%	55.63%	57.38%	56.49%	70.71%
BLCP	69.11%	67.25%	68.17%	72.41%	72.09%	72.25%	70.21%
GLOP	85.20%	84.65%	84.92%	79.17%	77.07%	78.11%	80.88%

Table 8 Recall, Precision and F values of the data set MBH

	Positive Class			negative Class			AVG F
Method	R	P	F	R	P	F	
BLC	75.26%	75.39%	75.32%	43.40%	41.16%	42.25%	58.79%
RS	72.85%	75.65%	74.22%	42.60%	45.28%	43.90%	59.06%
GLO	79.22%	80.37%	79.79%	53.97%	53.89%	53.93%	66.86%
BLCP	50.82%	52.91%	51.84%	58.72%	59.78%	59.25%	55.54%
GLOP	81.80%	80.26%	81.02%	75.34%	76.55%	75.94%	78.48%

Table 9 Recall, Precision and F values of the data set NHS

	Positive Class			negative Class			AVG F
Method	R	P	F	R	P	F	
BLC	57.55%	58.82%	58.18%	42.60%	43.32%	42.96%	50.57%
RS	73.30%	70.82%	72.04%	59.22%	57.13%	58.16%	65.10%
GLO	53.46%	56.27%	54.83%	61.34%	62.15%	61.74%	58.29%
BLCP	79.30%	77.24%	78.26%	77.03%	77.06%	77.04%	77.65%
GLOP	79.62%	79.52%	79.57%	71.02%	70.31%	70.66%	75.12%

Table 10 Recall, Precision and F values of the data set SIE

	Positive Class			negative Class			AVG F
Method	R	P	F	R	P	F	
BLC	50.02%	51.63%	50.81%	62.68%	63.45%	63.06%	56.94%
RS	41.43%	41.46%	41.44%	67.48%	68.35%	67.91%	54.68%
GLO	77.08%	79.76%	78.40%	59.89%	58.55%	59.21%	68.80%
BLCP	69.72%	68.42%	69.06%	84.62%	84.12%	84.37%	76.72%
GLOP	82.39%	82.84%	82.61%	83.05%	84.09%	83.57%	83.09%

Table 11 Recall, Precision and F values of the data set TRE

	Positive Class			negative Class			AVG F
Method	R	P	F	R	P	F	
BLC	62.47%	60.19%	61.31%	58.69%	56.16%	57.40%	59.35%
RS	70.60%	72.55%	71.56%	36.83%	38.97%	37.87%	54.72%
GLO	67.31%	65.75%	66.52%	59.67%	61.73%	60.68%	63.60%
BLCP	54.60%	52.80%	53.68%	50.78%	51.09%	50.93%	52.31%
GLOP	89.39%	89.66%	89.52%	91.67%	93.86%	92.75%	91.13%

The quality of the classification of the proposed method is affected by the initial lexicon that should be prepared in advance, however regardless the quality of this lexicon the algorithm was able to present an accepted results even with using a random lexicon at the beginning.

6. CONCLUSION

In this work unsupervised method to determine the polarity of blogs were presented. The method overcome the performance of other methods either supervised or unsupervised methods in most of the data sets. The performance of the algorithm is affected by the initial lexicon and the language of the blogs. The future work will concentrate on developing the algorithm to work without the need of initial lexicon. Also the algorithm will be expanded to increase the number of polarity classes.

7. REFERENCES

[1] Sarangi, S. K., Jaglan, V. and Dash, Y., 2013. A Review of Clustering and Classification Techniques in Data Mining, International Journal of Engineering, Business and Enterprise Applications (IJEBEA), 4(2), March-May, , pp. 140-145

[2] Aucouturier, J.-J. and Pachet, F., 2004, "Improving Timbre Similarity: How high's the sky?," in Journal of Negative Research Results in Speech and Audio Sciences.

[3] Kumar, V., and Rathee, N., 2011, ITM University, “Knowledge discovery from database Using an integration of clustering and classification”, International Journal of Advanced Computer Science and Applications, (March 2011) Vol. 2, No.3.

[4] Sharma, N., Bajpai, A. and Litoriya R., 2012, “Comparison the various clustering algorithms of weka tools”, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, (May 2012), Volume 2, Issue 5.

[5] Devijver, P.A. and Kittler, J., 1982, Pattern Recognition: A Statistical Approach. PrenticeHall.

[6] H. Liu and H. Motoda, 1998, Feature Extraction, Construction and Selection: A Data Mining Perspective. Kluwer Academic.

[7] Rokaya, M., Ghiduk A. S., 2019, Arabic Lexicon Learning to Analyze Sentiment in Microblogs, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 8, 592-599

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

[9] Haj mohammadi, M.S., Ibrahim and Selamat R. A., 2014, Cross-lingual sentiment classifica-tion using multiple source languages in multi-view semi-supervised learning, Eng. Appl. Artif. Intell. 36, 195–203 .

[10] Cui, Z., Shi, X. and Chen, Y., 2015, Sentiment analysis via integrating distributed representations of variable-length word sequence, Neurocomputing, 126–132 .

[11] Weissbock, J. and Inkpen, D., 2014, in: Combining Textual Pre-Game Reports and Statis- tical Data for Predicting Success in the National Hockey League, Advances in Artificial Intelligence, Springer International Publishing, pp. 251–262 .

[12] Javed, B.S., 2018, Hybrid semantic clustering of hashtags, Online Social Networks and Media 5 (2018) 23–36

[13] Vicient , A. M., 2014, Unsupervised semantic clustering of Twitter hashtags, Proceedings of the 21st European Conference on Artificial Intelligence, pp. 1119–1120 .

[14] Javed , B.S., 2016, Sense-level semantic clustering of hashtags in social me- dia, in: Proceedings of the 3rd Annual International Symposium on Informa- tion Management and Big Data, pp. 140–149 .

[15] Muntean, C.I., Morar , G.A., Moldovan, D., 2012, Exploring the meaning behind Twitter hashtags through clustering, Lect. Notes Bus. Inf. Process. 127, 231–242

[16] Bhulai, S., Kampstra, P., Kooiman, L., Koole, Deurloo, G., M. and CCing, B.K., 2012, Trend visualization on Twitter: what's hot and what's not?, Proceedings of the 1st International Conference on Data Analytics, Springer-Verlag, pp. 43–48

[17] Goldberg, D. E., 1989, Genetic Algorithms in Search, Optimization, and Machine Learning. Reading: Addison-Wesley

[18] Pesina, S. and Yusupova, L. G., 2014, Words Functioning in Lexicon, 2nd Global Conference On Linguistics And Foreign Language Teaching, Linelt-, Dubai – United Arab Emirates, December 11 – 13, 2014

[19] Rokaya, M and Atlam, E. S., 2010, Building of field association terms based on links', Int. J. Computer Applications in Technology, Vol. 38, No. 4, pp.298–305.

[20] Rokaya, M, 2013, Automatic Text Extraction Based on Field Association Terms and Power Links, International Journal of Computer and Information Technology (IJCIT), Volume 02– Issue 06 (November 2013), pp 1049- 1053

[21] Rokaya, M, 2013, Automatic Summarization based on Field Coherent Passages. International Journal of Computer Applications 79(9) (October 2013),38-44,. Published by Foundation of Computer Science, New York, USA

[22] Rokaya, M and Aljahdali, S., 2013, Building a Real Word Spell Checker Based on Power Links, International Journal of Computer Applications, Vol. 65 No 7,(March 2013), PP 14-19.

[23] Rokaya, M, Nahla, A. and Aljahdali, S., 2012, Context-Sensitive Spell Checking Based on Field Association Terms. IJCSNS International Journal Of Computer Science And Network Security. Vol. 12 No. 3 pp. 64-68.

[24] Rokaya, M, 2014, Improving Ranking of Search Engines Results Based on Power Links, IPASJ International Journal of Information Technology (IIJIT),Volume 2, Issue 9, September.