Text Classification for Arabic Words Using BPSO/REP-Tree

Hamza Naji, Wesam Ashour, Mohammed Al Hanjouri
Department of Computer Engineering, Islamic University
Gaza, Palestine
hamzan1991@gmail.com, washour@iugaza.edu.ps, mhanjouri@iugaza.edu.ps

ABSTRACT: *Text Classification is the process of grouping documents into categories based on their contents. Many Text Classification (TC) algorithms have been proposed for Arabic TC. The main idea in text Arabic classification is the accuracy of the results, which these results depends on the correctness of the text classification phases which start with the prepossessing phase and end with feature selection and the choosing of the best classifier that can classify text in related groups. In this paper we provide a new system for text classification based on BPSO/REP-Tree hybrid. The first term refers to the "Binary Particle Swarm Optimization" that we use it for the feature selection process and the second term refers the classifier we used "Reduced Error Pruning Tree". We will show the results of the experiments on a data-set collected from the BBC-Arabic website using the Weka tool which specific for data classification.*

Keywords: Data Mining, Text Classification, Text Data Mining, BPSO, Arabic Text Classification, Preprocessing, Binary Particle Swarm Optimization

Received: 11 March 2017, Revised 9 April 2017, Accepted 19 April 2017

1. Introduction

The process of Text-Classification considered as the ability of classifying or clustering the huge amount of text to some classes according to its deep content like classifying news's website to several fields (Weather, Politics, Sport...). Due to the increasing in text amount in web we need an effective classification to do this mission. We can achieve the previous need mission by the classification algorithm which called at the end (The Classifier). In this paper we focus on the classifying of the Arabic text, which the difficulty of Arabic expressive style is the employing in alternative languages like Persian, Urdu, Iranian language and alternative regional languages of Pakistan, Afghanistan and Persia. The Arabic language contents are constituting a 3% of the web text content with the fourth order in languages ordering on-line [1]. The previous amount of content needs an accurate and effective classification to help the humans to use it easy; thus, in the last 10 years the need for the effective and accurate classification has

been grown quickly. There are some classification algorithms can do well in general text classification and can proposed in Arabic such as:(Support Vector Machines (SVM) (Mesleh, 2007[2], Al-Harbi et al., 2008[3], El-Halees, 2008[4], Said et al., 2009[5], Al-Saleem, 2010[6], Al-Saleem, 2011[7], Chantar and Corne, 2011[8] and Khorsheed and Al-Thubaity, 2013[9]), NaÃŕve Bayes (NB) (El-Kourdi et al., 2004[10], Duwairi, 2007[11], El-Halees, 2008[12], Kanaan et al., 2009[13], Al-Saleem, 2010[14], Al-Saleem, 2011[15], Chantar and Corne, 2011[16], Khorsheed and Al-Thubaity, 2013[17], Belkebir and Guessoum, 2013[18] and Sharef et al., 2014[19]), K-Nearest_Neighbor (kNN) (Duwairi, 2007[20], El-Halees, 2008[21], Kanaan et al., 2009[22], Khorsheed and Al-Thubaity, 2013[23] and Ababneh et al., 2014[34]), Maximum Entropy (Sawaf et al., 2001[25], El-Halees, 2007[26]), Artificial Neural Network (ANN) (El-Halees, 2008[27], Belkebir and Guessoum, 2013[28] and Khorsheed and Al-Thubaity, 2013[29]), Decision Tree (DT) (Al-Harbi et al., 2008[30], El-Halees, 2008[], Chantar and Corne, 2011[31] and Khorsheed and Al-Thubaity, 2013[32]) and the Rocchio feedback algorithm (Kanaan et al., 2009[33]). More recently, Reduce Error Pruning tree (REP-Tree) is investigated in Arabic TC [34]. RET-Tree is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP-Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning[35]. REP-Tree was first used in Indian and English text classification in 2015 [36] [37]. The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 shows proposed work. Section 4 presents the results, and finally, we tend to conclude the paper in Section 5.

2. Related Work

In the discussion below, we focus on the works addressing Arabic TC. Since the number and quality of features used to express texts have a direct effect on classification algorithms, the following discusses the main goal of feature reduction and selection and their impact on TC.

Duwairi et al. [38] the study make a comparison between (stemming, light stemming, and word cluster). For training purposes they choose K Nearest Neighbor KNN technique, to show that light stemming achieves the highest accuracy and lowest time of model construction.

Another study [39] compared 3 Feature Subset Selection (FSS) metrics. They carried out a comparative study to examine the effect of the feature selection metrics in terms of precision. The results in general revealed that Odd Ratio (OR) worked better than the others. Some studies focused on other techniques like N-gram and different distance measures and proved their effects on Arabic TC. For example, [40] used a statistical method called Maximum Entropy (ME) for the classification of Arabic words. The author showed that the Dice measures using N-gram outperforms using the Manhattan distance. Similar classifier was used in [41], but different selection and reduction techniques were applied. The author used normalization, stop words removal to increase the ultimate accuracy. Most of related work in the literature used small datasets, and applied one or two classifiers to classify one corpus which is not enough to evaluate Arabic TC.

El-Kourdi et. al. [42] classified Arabic text documents automatically using NB. The average accuracy reported was about 68.78%, and the best accuracy reported was about 92.8%. El-Kourdi used a corpus of 1500 text documents belonging to 5 categories; each category contains 300 text documents. All words in the documents are converted to their roots. The vocabulary size of resultant corpus is 2,000 terms/roots. Cross-validation was used for evaluation.

Sawaf et. al. [43] (2001) used Maximum entropy (ME) to make a classification to News articles. The study gives accuracy about 62.7% .

Al-Zoghby [44] used Association Rules for Arabic text classification, and also he used CHARM algorithm with soft-matching over hard big O exact matching. Data sets consisting of 5524 records. Each record is a snippet of emails having the subject - nuclear. The vocabulary size is 103,253 words.

Harrag et. al. [45] used the feature selection based on hybrid approach for Arabic text classification. He used direct tree algorithm and the accuracy was of 93% for scientific data set, and 90% for literary data-set. Harrag collected 2 data sets; the first one is from the scientific encyclopedia.

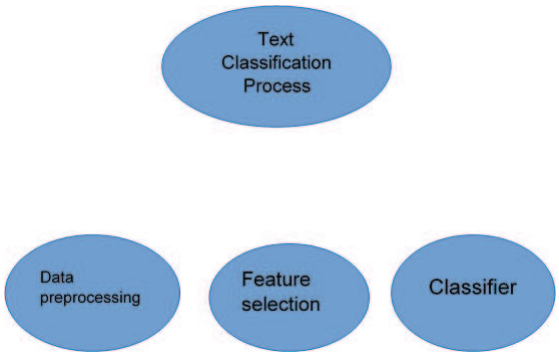

Figure 1. Shows the TC process

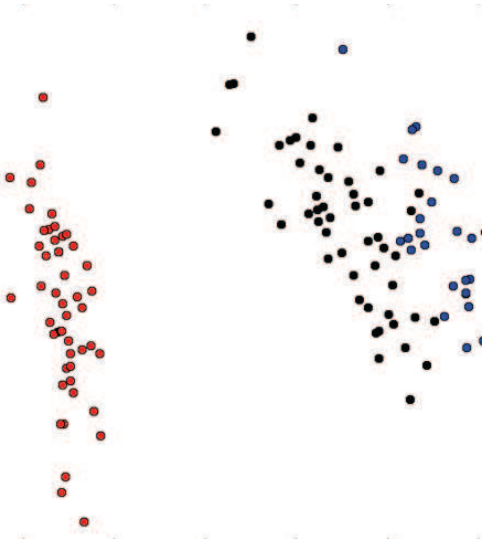

Figure 2. Shows groups of data in the search space

Al-Shalabi ET. Al. use KNN to classify Arabic text [46], by used the weighting scheme tf-idf and stemming and feature selection to give an accuracy of 95%. The authors enplane the lacking freely publically availability of Arabic data sets. Data sets were collected from (Al-Jazeera, An-Nahar, Al-Hayat, Al-Ahram, and Ad-Dostor) and from Arabic Agriculture Organization website. The data set consists of 621 documents belonging to 1of 6 categories (politics 111, economic 179, sport 96, health and medicine 114, health and cancer 27, agriculture 100). They pre-processed the corpus by applying stop words removal and light stemming.

3. Proposed Work

The text classification process consists of these three phases: Data preprocessing, feature selection, and the chosen classifier as shown in figure 1 bellow. In this paper we focus on the last two phases and improve these process by provide a new model consisting of the Binary Particle Swarm Optimization (BPSO) for feature selection purpose and the Reduced Error Pruning Tree (REP-Tree) as a classifier.

3.1 Binary Particle Swarm Optimization (BPSO)

Particle swarm optimization (PSO) was developed by Eberhart and Kennedy in 1995 [47], motivated in part by the social behavior of flocks of birds. This technique proposed that if there are groups of data this algorithm scatter the

initial classes over the data and starts siding towards the top concentrate of the data. We can represent this approach by a swarm of bees, when it finding an area of flowers to suck nectar, bees spread in the place randomly even to choose the densest area and then inform the rest of the flock (swarm) of its new information. The PSO made for data classification problems solution, which deal with the several proposed solutions and called everyone on those solutions a particle. A particle is a position solution in the search space shown below in figure 2. The flock of bee's example can be generalized on the image below; the algorithm can spread some initial classes over the data and each class start search about the high concentrate of data and generalized its data to the other classes to share the classification process.

In binary PSO (BPSO) [48], a particle's position is simply a binary vector, which first seems hard to fit with the notion of having velocities associated with a particle. The previous approach in [49] provides equations used to manage velocities in PSO, with the key difference being that in BPSO a velocity vector (a real-valued vector in which each component is kept between 0 and 1) represents a set of probabilities, one for each component. The proposed binary encoding is natural for a feature selection purpose.

3.2 Reduced Error Pruning Tree (Rep-Tree)

REP-Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning [2]. REP-Tree was first used in Indian and English text classification in 2015 [3] [4]. We will use this tree as a classifier to classify the resulting features from the selection process. This classifier first used by a previous paper in an Arabic data-set collected from the BBC-Arabic website [50] and this paper conceder as an extension of the aforementioned paper.

3.3 The Proposed System

Now we can provide a presentation of our system by these steps as shown below: First we collect the corpus (data-set) from the BBC-Arabic website as we talk previous and changing it into ARRF files to use it with Weka tool (a common tool for classification purposes we talked about it in the previous paper. The second step, we perform the preprocessing phase to the data-set and we get the following results by performing the Stemming process we talked about it in the previous paper:

3.3.1 The Preprocessing Phase

- **Tokenization:** Which converts a text document from a stream of characters into a sequence of tokens (features or terms) by recognizing delimiters such as white spaces, punctuations, special characters, etc.
- Removal of the non-Arabic letters.
- Removal of numbers, diacritics, special characters and punctuations.
- **Removal of Stop Words:** These include pronouns, conjunctions, and prepositions.
- **Stemming:** Reducing an inflected or derived word to its stem. The stem needs not to be a valid morphological root of the word as far as related words map to the same stem. The main advantage of this preprocessing step is to reduce the number of terms in the corpus so as to reduce the computational and storage requirements of TC algorithms [51].

3.3.1.1 The Stemming Process and Effects on a Corpus

This is the core of the preprocessing phase in our data-set which convert the term (word) to its identified root to reduce the number of training set can used in the classification process.

The Stemming process can do the following activities:

1. Remove any punctuation such as $(, - .; “” :)$.
2. Remove stop words.
3. Remove the special prefix for example (ζ).

4. Replace
5. Remove Suffixes.
6. Match the result against a list of Patterns.
7. Two letter roots are checked to see if they should contain a double character; if so, the character is added to the root.

3.4 Features Selection Process

The third process of text classification is the feature selection process, we use the BPSO and feature selector after we stem and preprocessed the ARRF file in weka tool. This approach can do these steps as followed:

1. Make a search space as figure 2 and provides a group of particles on the space which consisting of the followed three variables: A represents the current position of particle, B represents the best previous position- A is initialized with random binary values where 1 means the corresponding feature is selected and 0 means not-selected B is initialized with a copy of A, and V represents the velocity of B.

2. The fitness of the particle as mentioned in [52] we can compute it as the following equation:

$$Fitness = (\alpha \times Acc) + (\beta \times ((N-T)/N))$$

   Where

   - Acc is the classification accuracy of the particle found using K-NN.
   - α and β are two parameters used to balance between classification accuracy and feature subset size, where α is in the range [0,1] and $\beta = 1 - \alpha$.
   - N is the total number of features.
   - T is length of the selected subset of features.

3. Updating the best particle's information (the particle which can arrive to the density first) that's mean updating velocity and position of all particles in the population according to standard approach in BPSO.

4. Do the previous two steps to all particles of the search space.
   - Then we can calculate the accuracy of the features selection by dividing the instances (words) (we get from the previous steps) count to the all training set.

3.5 The Classification Phase

In this subsection we will choose the classifier that does the final classification process for Arabic text. We talked about some classification methods implemented in the English text and do well in the Arabic text. In this paper we will use the Reduce Error Pruning Tree (REP-Tree) as a new classifier in Arabic text classification which implemented by [53] and [54] in the classification of English and Indian news. And more recently has been implemented in our previous paper in Arabic text. We will present the results in the next section with the Weka tool.

4. Experimental Results and Analysis

Experiments had been applied on Arabic dataset collected from BBC- Arabic website (*http://www.BBc.co.uk/arabic/*). The dataset contains 110 text documents belonging to one of the seven categories (Middle East news, Sport, Health, Computer and Technology, Varieties and Communications).For text classification; we use REP-tree with specific training set of 66% cross validation.

We will perform the classification process in the Weka tool and our results factors will be the three factors listed in the Table 1 below (Precision, Recall, and F-Measure).

From Table 1 above we can see the resultant data comes from the classification process using REP-Tree in the Weka

Class	Precision	Recall	F-Measure
Middle East News	94.9	96.2	92.3
Sport	93.1	91	95.6
Health	78.4	92.4	77.3
Technology	99	92.4	96.8
Varieties	99.5	96.7	90.2
Communications	94.3	95.1	93.7
AVG	**93.2**	**93.9**	**90.9**

Table 1. Shows the classification process results using REP-Tree approach

tool, and we can see that the "Varieties" class has the best performance with precision of 99.5, recall of 96.7 and F-Measure of 90.2, not the best F-Measure but do well with the other factors and recorded the highest average. The second best class performance was the "Technology" class with precision of 99, recall of 92.4 and F-Measure of 96.8. The third performance rank of classes is the "Middle East News" with precision of 94.9, recall of 96.2 and F-Measure of 92.3, and the next one with the smaller than the previous with 3 is the "Communications" in rank four with precision of 94.3, recall of 95.1 and F-Measure of 93.7. The worst two classes were the "Sport" and the "Health" classes with precision of 93.1, recall of 91.0 and F-Measure of 95.6 for "Sport" and the worst precision with 78.4 and F-Measure with 77.3 for "Health" class.

Now we want to perform the classification process of the previous data-set on another classifier "J48 tree" to compare the two approaches on the same data-set. Table 2 will show the results from the experiment as followed below:

From Table 2 we perform the TC process on the BBC-Arabic data-set and we found some differences in the re-

Class	Precision	Recall	F-Measure
Middle East News	94.3	96.8	92.7
Sport	93.1	91	95.6
Health	86.2	92.4	84.3
Technology	97.3	90.4	94.8
Varieties	99.5	96.7	90.2
Communications	68.3	76.1	95.7
AVG	**89.5**	**90.5**	**92.1**

Table 2. Shows the classification process results using J48-Tree approach

sults which the J48 recorded some enhancements in performance and in some areas its recorded some degradation in text classification performance like "Communications" class with precision of 68.3, recall of 76.1.

5. Conclusion

In this paper we present the text classification process for Arabic words with the phases of the process the preprocessing, selection of the best feature and the last classification using a classifier. We explain the selection feature process by BPSO and the steps of stemming process to eliminate the amount of rendering text. In the end of the paper we proposed the Rep-Tree as a new Arabic text classifier and we presented the results to show its accuracy. We get the results from the BBC-Arabic website after converting the text to a arrf file and then using this dataset in the Weka tool for data classification purposes

References

[1] Internet world users by language-top 10 languages. (2015). Retrieved June 2015, Available from: http://www.internetworldstats.com/stats7.htm.

[2] Mesleh, A. (2011). Feature sub-set selection metrics for arabic text classification, *Pattern Recognition Letters*, 32 (14) 1922-1929.

[3] Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., Al-Rajeh, A. (2008). Automatic Arabic Text Classification, *In*: JADT 2008 : 9 Journees internationales d'Analys e statistique des Donnees Textuelles.

[4] El-Halees, A.M. (2008). A comparative study on arabic text classification, Egypt. *Comput. Sci. J*., 30 (2).

[5] Said, D., Wanas, N., Darwish, N., Hegazy, N., (2009). A Study of Arabic Text preprocessing methods for Text Categorization, *In*: The 2nd Int. conf. on Arabic Language Resources and Tools, April, 22-23, Cairo, Egypt, p. 230-236.

[6] Al-Saleem, S. (2010). Associative classification to categorize Arabic data sets, *Int. J. ACM* Jordan, 1 (3) 118-127.

[7] Al-Saleem, S.(2011). Automated Arabic text categorization using SVM and NB, Int. Arab J. e-Technol., 2 (2) 124-128.

[8] Chantar, H. K., Corne, D.W. (2011). Feature subset selection for Arabic document categorization using BPSOKNN, IEEE, 546-551, http: dx.doi.org/10.1109/NaBIC.2011.6089647.

[9] Khorsheed, M., Al-Thubaity, A. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset Lang Resour. Eval. Springer, 47 (2) 513-538 http://dx.doi.org/10.1007/s10579-013-9221-8.

[10] El-Kourdi, M., Bensaid, A., Rachidi, T. (2004). Automatic Arabic document categorization based on the Naive Bayes algorithm. *In*: The 20th Int. Conf. on Computational Linguistics, Geneva, August, 27.

[11] Duwairi, R. (2007). Duwairi Arabic text categorization Int. Arab J. Inf. Technol., 4 (2), p. 125-131 http://dx.doi.org/10.1002/asi.20360.

[12] El-Halees, A.M. (2008). A comparative study on arabic text classification Egypt. *Comput. Sci. J*., 30 (2).

[13] Kanaan, G., Al-Shalabi, R., Ghwanmeh, S. (2009). A comparison of text-classification techniques applied to Arabic text, J. Am. Soc. *Inform. Sci. Technol*., 60 (9) 1836-1844. http://dx.doi.org/10.1002/asi.v60:9

[14] Al-Saleem, S. (2010). Associative classification to categorize Arabic data sets, *Int. J. ACM* Jordan, 1 (3 (2010) 118-127.

[15] Al-Saleem, S. (2011). Automated Arabic text categorization using SVM and NB, *Int. Arab J. e-Technol*., 2 (2)124-128.

[16] Chantar, H.K., Corne, D.W. (2011). Feature subset selection for Arabic document categorization using BPSOKNN, IEEE, 546-551 http: dx.doi.org/10.1109/NaBIC.2011.6089647.

[17] Khorsheed, M., Al-Thubaity, A. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset Lang Resour. *Eval. Springer*, 47 (2) 513-538 http://dx.doi.org/10.1007/s10579-013-9221-8.

[18] Belkebir, R., Guessoum, A. (2013). A hybrid BSO-Chi2-SVM approach to Arabic text categorization. *In*: IEEE Computer Systems and Applications (AICCSA), ACS International Conference, 27-30 May, Ifrane, p. 1-7. doi:http://dx.doi.org/10.1109 AICCSA.2013.6616437.

[19] Sharef, B., Omar, N., Sharef, Z. (2014). An automated Arabic text categorization based on the frequency ratio accumulation *Int. Arab J. Info. Technol*., 11 (2) 213-221.

[20] Duwairi, (2007). R. Duwairi Arabic text categorization *Int. Arab J. Inf. Technol*., 4 (2) 125-131 http://dx.doi.org/10.1002/asi.20360.

[21] El-Halees, A.M. (2008). A comparative study on arabic text classification Egypt. *Comput. Sci. J*., 30 (2).

[22] Kanaan, G., Al-Shalabi, R., Ghwanmeh, S. (2009). A comparison of text-classification techniques applied to Arabic text *J. Am. Soc. Inform. Sci. Technol*., 60 (9), p. 1836-1844 http://dx.doi.org/10.1002/asi.v60:9.

[23] Khorsheed, M., Al-Thubaity, A. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset Lang Resour. *Eval. Springer*, 47 (2) 513-538 http://dx.doi.org/10.1007/s10579-013-9221-8.

[24]] Ababneh, J., Almomani, O., Hadi, W., Kamel, N., El-Omari, T., Al-Ibrahim, A. (2014). Vector space models to classify Arabic text *Int. J. Comput. Trends Technol*. (IJCTT), 7 (4) 219-223.

[25] Sawaf, H., Zaplo, J., Ney, H. (2001). Statistical Classification Methods for Arabic News Articles. Arabic Natural

Language Processing, Workshop on the ACL'2001. Toulouse, France, July.

[26] El-Halees, A.M. (2007). A comparative study on Arabic text classification, *Egypt. Comput. Sci. J.*, 30 (2).

[27] El-Halees, A.M. (2008). A comparative study on Arabic text classification, *Egypt. Comput. Sci. J.*, 30 (2).

[28] Belkebir, R., Guessoum, A. (2013). A hybrid BSO-Chi2-SVM approach to Arabic text categorization, *In*: IEEE Computer Systems and Applications (AICCSA), ACS International Conference, 27-30 May, Ifrane, p. 1-7. doi:http://dx.doi.org/10.1109 AICCSA.2013.6616437.

[29] Khorsheed, M., Al-Thubaity, A. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset Lang *Resour. Eval. Springer*, 47 (2) 513-538 http://dx.doi.org/10.1007/s10579-013-9221-8.

[30] Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., Al-Rajeh, A. (2008). Automatic Arabic Text Classification, *In*: JADT: 9 Journees internationales d'Analys e statistique des Donnees Textuelles.

[31] Chantar, H.K., Corne, D.W. (2011). Feature subset selection for Arabic document categorization using BP-SOKNN, IEEE, 546-551. http: dx.doi.org/10.1109/NaBIC.2011.6089647.

[32] Khorsheed, M., Al-Thubaity, A. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset Lang *Resour. Eval. Springer*, 47 (2) 513-538 http://dx.doi.org/10.1007/s10579-013-9221-8.

[33] Kanaan, G., Al-Shalabi, R., Ghwanmeh, S. (2009). A comparison of text-classification techniques applied to Arabic text *J. Am. Soc. Inform. Sci. Technol*., 60 (9), p. 1836-1844 http://dx.doi.org/10.1002/asi.v60:9

[34] Hamza Naji, Weasam Ashour. (2016). Text classification for Arabic words using REP-Tree, *International Journal of Computer Science & Information Technology* (IJCSIT) 8 (2), April.

[35] Srinivasan, Dr. B., Mekala, P. (2014). Mining Social Networking Data for Classification Using REPTree, *International Journal of Advance Research in Computer Science and Management Studies*, 2 (10), October, p-155-160.

[36] Sushilkumar Kalmegh, Analysis of WEKA Data Mining Algorithm REPTree, Simple Cartand Random Tree for Classification of Indian News, February 2015, IJISET.

[37] Nikita Patel, Saurabh Upadhyay. (2012). Study of Various Decision Tree Pruning Methods with their Empirical Comparison, *International Journal of Computer Applications* (0975-8887) 60 (12) December.

[38] Duwairi, R., Al-Refai, M. N., Khasawneh, N. (2009). Feature reduction techniques for arabic text categorization, *Journal of the American Society for Information Science and Technology*, 60 (11) 2347-2352.

[39] Mesleh, A. (2011). Feature sub-set selection metrics for Arabic text classification, *Pattern Recognition Letters*, 32 (14)1922-1929.

[40] Khreisat, L. (2006). Arabic text classification using n-gram frequency statistics a comparative study, *In*: Conference on Data Mining-DMIN'06,p.79. http: dx.doi.org/10.1109/NaBIC.2011.6089647.

[41] El-Halees, A. (2007). Arabic text classification using maximum entropy, The Islamic University Journal (Series of Natural Studies and Engineering), 15, p.157-167.

[42] El-KourdiM, BensaidA, RachidiT. (2004). Automatic Arabic Document Categorization Based on the Naive Bayes Algorithm, *In*: 20th Int.Conf.on Computation alLinguistics, Geneva, Augus

[43] SawafH, ZaploJ, NeyH. (2001). Statistical Classification Methods for Arabic News Articles, *In*: Workshop on Arabic Natural Language Processing, ACL'01,Toulouse, France.

[44] Al-Zoghby, A.,Eldin, A.S., Ismail, N.A., Hamza, T. (2007). Mining Arabic Text Using Soft Matching association rules, *In*: Int .Conf.on Computer Engineering & Systems, ICCES'07.

[45] Harrag, F., El-Qawasmeh, E., Pichappan, P. (2009). Improving Arabic text categorization using decision trees, *In*: 1st Int.Conf.of NDT'09, p.110-115.

[46] Al-Shalabi, R., Kannan, G., Gharaibeh, H. (2006). Arabic text categorization using KNN algorithm, *In*: Proc.of Int . Multi Conf. on Computer Science and Information Technology CSIT06.

[47] Kenedy, J., Eberhart, R. C. (1995). Particle swarm optimization, *In*: Proc, IEEE ICNN (Perth, Australia), IEEE Press, p. 1942-1948.

[48] Kenedy, J., Eberhart, R. C. (1997). A discrete binary version of particle Swarm Optimization algorithm, *In*: Proc, IEEE ICNN (Perth, Australia), IEEE Press, p. 4104-4108.

[49] Kenedy, J., Eberhart, R. C. (1997). A discrete binary version of particle Swarm Optimization algorithm, *In*: Proc, IEEE ICNN (Perth, Australia), IEEE Press, p. 4104-4108.

[50] Hamza Naji, Weasam Ashour. (2016). Text classification for Arabic words using REP-Tree, *International Journal of Computer Science & Information Technology* (IJCSIT) 8 (2) April.

[51] Ashraf Odeh, Aymen Abu-Errub, Qusai Shambour, Nidal Turab. (2014). Arabic Text Categorization Algorithm Using Vector Evaluation Method, *International Journal of Computer Science & Information Technology* (IJCSIT) 6 (6) December.

[52] Chantar, H.K., Corne, D.W. (2011). Feature subset selection for Arabic document categorization using BP-SOKNN, IEEE, 546-551 http: dx.doi.org/10.1109/NaBIC.2011.6089647.

[53] Sushilkumar Kalmegh. (2015). Analysis of WEKA Data Mining Algorithm REPTree, Simple Cartand Random Tree for Classification of Indian News, February, IJISET.

[54] Nikita Patel, Saurabh Upadhyay. (2012). Study of Various Decision Tree Pruning Methods with their Empirical Comparison, *International Journal of Computer Applications* (0975-8887) 60 (12) December.