A New Model in Arabic Text Classification Using BPSO/REP-Tree

Hamza Naji[1], Wesam Ashour[2] and Mohammed Alhanjouri[3]

[1]Department of Computer Engineering, Islamic University of Gaza, Palestine.
[2]Department of Computer Engineering, Islamic University of Gaza, Palestine.
[3]Department of Computer Engineering, Islamic University of Gaza, Palestine.

Abstract—Specifying an address or placing a specific classification to a page of text is an easy process somewhat, but what if there were many of these pages needed to reach a huge amount of documents. The process becomes difficult and debilitating to the human mind. Automatic text classification is the perfect solution to this problem by identifying a category for each document automatically. This can be achieved by machine learning; by building a model contains all possible attributes features of the text. But with the increase of attributes features, we had to pick the distinguishing features where a model is created to simulate the large amount of attributes (thousands of attributes). To deal with the high dimension of the original dataset, we use features selection process to reduce it by deleting the irrelevant attributes, words, where the rest of features still contain relevant information needed in the process of classification. In this research, a new approach which is Binary Particle Swarm Optimization (BPSO) with Reduced Error Pruning Tree (REP-Tree) is proposed to select the subset of features for Arabic classification process. We compare the proposed approach with two existing approaches; Binary Particle Swarm Optimization BPSO with K-Nearest Neighbor (KNN) and Binary Particle Swarm Optimization BPSO with Support Vector Machine (SVM). After we get the subset of attributes that result from features selection process, we use three common classifiers which are Decision Trees J 48, SVM and the prepared algorithm REP-Tree (as a classifier) to build the classification model. We created our own Arabic dataset; the BBC Arabic News dataset that are collected from the BBC Arabic website and another one existing is used datasets in our experiments, Alkhaleej News Dataset. Finally, we present the experimental results and showed that the proposed algorithm is missionary in this area of research.

Index Terms—Text classification, BPSO, REP-Tree, Binary Particle Swarm Optimization.

I Introduction

The huge increase of using text in the electronic devices and web sites, in particular, is a motivation for categorizing these texts in automatic manner. That’s because of the insufficiency of human ability to handle them manually. The core task in the categorization is called the Text Categorization or Classification TC. The previous task is the ability of classifying a huge amount of groups of texts; each of them is called a text data-set or Corpora, to some predefined classes. In case of news data-set; for example, the classes can be Sport, Health etc., and other various classes based on their contents.

Text classification process in general consists of two phases. The first one is the preprocessing phase defined as the process that implements on the amount of texts to make some improvements for reducing the unnecessary terms. The preprocessing phase also contains reducing the extra phrases of one term by a process called Stemming. Stemming is the process of eliminating the derived words of one basic word such as the words “making makes” and turning them to their roots as the word "make". Another example of the stemming process are the words (argue, argued, argues, arguing, and argues) turning them to the stem "argu". On the other hands, (argument and arguments) are turned to the stem "argument". The preprocessing phase includes the removing of some prefixes and suffixes from the word instead of extracting the original root.

The second phase of text classification process is the classification step. The process of classifying the preprocessed text in the previous phase and presenting the corpora using a mechanism is called a classifier. To apply such two phases, we need to convert each dataset to a term vector which is the basic of text processing [1]. But how many terms we need in each dataset based on what term we need is a question to be answered. The previous question leads us to add a new step in the text classification process, Arabic Text Classification in this paper.

There is a middle step between preprocessing and classification process called "feature selection" [2], it is a complementary process to the preprocessing stage performed after it to reduce the redundant terms (features) and to keep the sufficient terms to continue the classification process [3]. We demonstrate a combination of Binary Particle Swarm Optimization BPSO and Reduced Error Pruning Tree REP-Tree for the last process of selecting good sets of features for the Arabic TC task. Then we use the second half of the hybridized approach the REP-Tree and use it as a classifier as mentioned above.

The text classification processes can be done easily on the English language due to the smooth environment of it. In contrast, Arabic language is considered a complex language that contains many formations and many different kinds of forms of the word. The aforementioned difficulty in the Arabic language requires greater efforts in dealing with the classification of texts. paper focuses on the classification of the Arabic text which is the difficulty of Arabic expressive style when being employed in alternative languages like Persian, Urdu, Iranian language and alternative regional languages of Pakistan, Afghanistan and Persia. The Arabic language contents constitute a 3% of the web text content with the fourth order in languages ordering on-line [4]. The previous amount of content needs an accurate and effective classification to help the humans to easily use it .Thus, in the last 10 years the need for the effective and accurate classification has quickly been grown.

There are some classification algorithms that can be done in general text classification and can be proposed in Arabic such as: Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor(KNN), Maximum Entropy (ME), Artificial Neural Network (ANN), Decision Tree (DT)and the Rocchio Feedback Algorithm. More recently, Reduced Error Pruning tree REP-Tree is investigated in Arabic TC. RET-Tree is a fast decision tree learning machine and it builds a decision tree based on the information gained or reducing the variance. Also, REP-Tree is a fast decision tree learner which builds a decision/regression tree using information gained as the splitting criterion, and prunes it by using reduced error pruning [5]. REP-Tree was first used in Indian and English text classification in 2015 and 2012 [6], [7].The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 explains BPSO concepts. Section 4 explains the second term of the proposed approach REP-Tree. Section 5 shows proposed work. Section 6 presents the results, and finally, we tend to conclude the paper in Section 7.

II Related Works

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

(Brahimi, Touahria and Tari, 2016) [8] addressed sentiment analysis for tweets in the Arabic language using some approaches with two free available datasets of (2000 tweets). They applied the light and root stemmer as a preprocessing phase and investigated the impact of reducing the size of the dataset by selecting the most relevant features on the classification efficiency and accuracy of three well used machine learning algorithms Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbor (KNN).

(Oraby, El-Sonbaty and El-Nasr, 2013) [9] worked on the impact of Stemming by applying the Khoja stemmer [10], Information Science Research Institute (ISRI) stemmer [11], and Tashaphyne Light Arabic Stemmer [12] on two datasets of the opinion classification problem, the results show that the Khoja stemmer is the best one.

(Shoukry andRafea, 2012) [13] performed the classifiers Support Vector Machine SVM and Naïve Bayes NB on a dataset collected from twitter website. They applied the experiments on 2 documents of Arabic tweets and the results showed that the Support Vector Machine SVM was better than Naïve Bayes NB.

(Al-Thwaib, 2014) [14] used the Sakhr summarizer Sakhr company website 2016 as a feature selector to choose the best words of documents instead of using all words and they used the TF feature. Documents, after using TF for feature selection, are classified using SVM classifier; the data set they used consists of 800 Arabic text documents. It is a subset of 60913-document corpus collected from many newspapers and other web sites. He succeeded to increase the accuracy by using the summarized corpus as input for Support Vector Machine SVM classifier.

(Al-Hindi and Al-Thwaib, 2013) [15] made a comparison between two data-sets, each one contained 1000 Arabic documents.Text summarization was applied on one without the other. Accuracy has not improved much, but there was a difference in the time. When they used summarized documents, less time was needed to build the learning model.

(Abu-Errub, 2014) [16] proposed a method to classify Arabic text by comparing a document with predefined documents categories based on its contents using the Term Frequency Times Inverse Document Frequency TF.IDF method measure. After that the document is classified into the appropriate sub-category using Chi Square measure. The dataset used in this study contained 1090 documents for training and 500 documents for testing, categorized into ten main categories. The results show that the proposed algorithm can classify the Arabic text datasets into predefined category.

(Goweder, Elboashi and Elbekai, 2013) [17] used their developed technique, Centroid-based, to classify Arabic text. The proposed algorithm is evaluated using a dataset containing a 1400 Arabic documents collecting from 7 different classes. The results show that the adapted Centroid-based algorithm can classify Arabic documents without problems. They used some measurements Micro-averaging recall, precision, F-measure, accuracy, and error rates respectively. The measurements factors record a performance percentage of 90.7%, 87.1%, 88.9%, 94.8%, and 5.2% according to the previous order of measurements.

(Abidi and Elberrichi, 2012) [18], in this paper, they presented a comparative study to assess the effect of a conceptual representation of the text. The K-Nearest Neighbor used and feature extraction was achieved via three preprocessing schemes Bag of Words, N grams, and a conceptual representation. The F-measure of Bag of Words is 64%, 68% for N gram's F-measure, and 74% for F-measure conceptual representation. Finally, the conceptual representation was the best one as the results shown.

(Raho,Al-Shalabi, Kanaan and Nassar, 2015) [19] investigated the importance of feature selection in Arabic corpus classification by making a comparison of the performance between different classifiers in different situations using feature selection with stemming, and without using stemming. The dataset collected from BBC Arabic website and the classifiers they used are DT, K nearest neighbors KNN, Naïve Bayesian Model NBM method and Naïve Bayes NB; also they used factors Measurements such as precision, recall, F-Measures, accuracy and time. The results showed the Accuracy of each classifier as the following: (D.T 99.4%, KNN 66.3%, NBM 92%, and NB 91.9%).

(Mohammad, Al-Momani and Alwada, 2016) [20] provided a comparative study of Arabic text classification between three types of classifiers (k-Nearest Neighbor, Decision Trees C4.5, and Rocchio Classifier). These well-known algorithms are applied on a collected Arabic data set. Data set used consists from 1400 documents belongs to 8 categories, the same number of documents was used in the study experiments. They used two types of Measurements precision and recall, and the results of the experiments showed that the K-Nearest Neighbor records an average of 80% for Recall and 83% for precision, While Rocchio Classifier records an average of 88% for Recall and 82% for precision. Both of the previous Classifiers are better than C4.5 with average of 64% for Recall and 67% for precision.

(Kanan and Fox, 2015). [21] This study talks about a new approach in Arabic text classification stemming; they developed a new model called tailored stemming, a new Arabic light Stemmer, with the usage of Support Vector Machine SVM classifier. The experiments were performed under 10-fold cross-validation training type, and gave these results for the predefined classes after using SVM as the following: Art and Culture 91.8%, Economics 93.5%, Politics 91.5% and Society 99.1%.

(Al-Anzi and Abuzeina, 2016) [22] grouped the similar unlabeled document into pre-specified number of topics using Latent Semantic Indexing LSI and Singular Value Decomposing SVD methods. The corpus they used contains 1000 documents of 10 topics, 100 documents for each topic. The results showed that EM method is the best of other methods with an average categorization accuracy of 89%.

(Zubi, 2009). [23] This study is about using the web contents and applies some Arabic classification techniques on it. The general purpose of this study is to compare between two classifiers. The author used the K-Nearest Neighbor KNN Classifier and Naïve Bayes NB Classifier to apply the experiment. As mentioned by the author in his study. A corpus of Arabic text documents was collected from online Arabic newspapers archives, including Al-Jazeera, Al-Nahar, Al-hayat, Al-Ahram, and AlDostor as well as few other specialized websites. He collects 1562 documents classifying it into 6 different categories. After the comparison experiment finished, the results showed that the K Nearest Neighbors KNN with an average of (86.02%) was better than Classifier Naïve Bayesian with accuracy of (77.03%).

(Zrigui, Ayadi, Mars and Maraoui, 2012). [24] They developed a new model based on the Latent Dirichlet Allocation (LDA) and the Support Vector Machine SVM; they used the LDA to sample "topics" of groups of texts. The results showed that the proposed LDA-SVM algorithm is able to achieve high effectiveness for Arabic text classification task (Macro-averaged F 1 88.1% and Micro-averaged F – 91.4%).

III Binary Particle Swarm Optimization BPSO

Before talking about BPSO as a feature selection algorithm, we will first describe the intended of the word "Swarm" in full definition of PSO "Particle Swarm Optimization" algorithm. What is the swarm and where this name came? That's what we got from the final meaning of the definition. Many forms of life in some organisms affected the aspirations of some researchers and invited them to develop some successful theories for solving problems based on this random life. There is a group of successful theories based on this mode of thinking, including the DNA counting, membrane algorithm, Particle Swarm Optimization algorithm, artificial immune systems algorithm, and Ant Colony Optimization algorithm. One of the algorithms is the Particle Swarm Optimization algorithm that was developed in the 1995 by Eberhart and Kennedy [25]. This idea has been built on the basis of the collective behavior of flocks of birds. PSO creates a random optimization algorithm to give solutions, particles, for some positions in the search space. Each of those particles holds an initial random velocity within the search space symbolized by V i = (V i 1 ; V i2 ; ...V iN), and each particle is symbolized by P i = (P i 1 ; Pi2 ; ...; P iN). Update its velocity according to its experience or other particles experiences. For the best particle in the search space, swarm, we called it the best global symbolized by g, and when the velocity has been updated, the particle it finds the new position with the latest velocity according to the following equations [26]

The main equation is:

$$Xid = Xid + Vid \qquad (1)$$

New position = Current position + New velocity.

$$Vid = \omega * Vid + C1 * rand(\quad) * (Pid - Xid) + C2 * rand(\quad) * (Pgd - Xid) \qquad (2)$$

Where

rand () is a random number between (0, 1) [27]. **c1, c2** are acceleration factors. Usually c1 = c2 = 2. **Pgd** = global best. Vid = velocity of particle [28].

Xi is the current position of the particle initialized with random binary values. Where 0 means that the corresponding feature is not selected and by 1 means that the feature is selected. **Pi** is the best previous position of the particle and initialized by the same value of **Xi**.

Vi is the velocity of **Pi**.

What if there was no previous velocity, then particles will navigate to the same position (current position), and that is the (local search). But if we get a new velocity, then particle will extend its search (the global search). Some problems resulted from the previous questions. Inertia weight ω solve these problems by balancing the local and global search. [ ]perform a sequence of experiments to give the best value of **ω** which is 1.2.In Binary Particle Swarm Optimization Binary PSO, particle position is considered as a binary vector, but how binary vectors deal with velocities. [29] provided some equation to deal with velocity, a vector, (with real value in which this value is kept between (0, 1)), provides a group of probabilities. According to the previous we can use the BPSO to select the relative features in the Arabic Text Classification. As mentioned in [30], the probability of bit changing is determined by the following:

$$\boldsymbol{S(Vid) + \frac{1}{1 + e^{-Vid}}} \qquad \mathbf{(3)}$$

$$\boldsymbol{If\ (rand(\quad) < S(Vid))\,then\ Xid = 1;\ \ Else = 0} \qquad \mathbf{(4)}$$

Where rand () is a random number between (0, 1) [27]. c1, c2 are acceleration factors. Usually c1 = c2 = 2. Pgd = global best. Vid = velocity of particle [29].

IV Reduced Error Pruning Tree REP-Tree

More recently, Reduced Error Pruning tree REP-Tree is investigated in Arabic TC [31]. REP-Tree is a fast decision tree learning algorithm and it builds a decision tree based on the information gained or reducing the variance. REP-Tree is a fast decision tree learner which builds a decision/regression tree using information gained as the splitting criterion, and prunes it using reduced error pruning. REP-Tree was first used in Indian and English text classification in 2015 [32] and 2012 [33]. The REP-Tree first starts the training process on the existing dataset, and then builds the training model by decisions, then get a mix results of some instances from the first learning step and from the pruned dataset which is a part of the dataset for post-pruning of the tree, then performing the test process. For a sub-tree of the tree, if replacing it by a node or leaf, which doesn't take more prediction errors on the pruning set than the original set, the tree replaces by a leaf. That means that the REP-Tree prunes each node after the natural classification. If the misclassification error determined for the instances from the pruned data set is not larger than the misclassification error rate computed on the original training data, the misclassification error can be presented in the Figure (1) below.

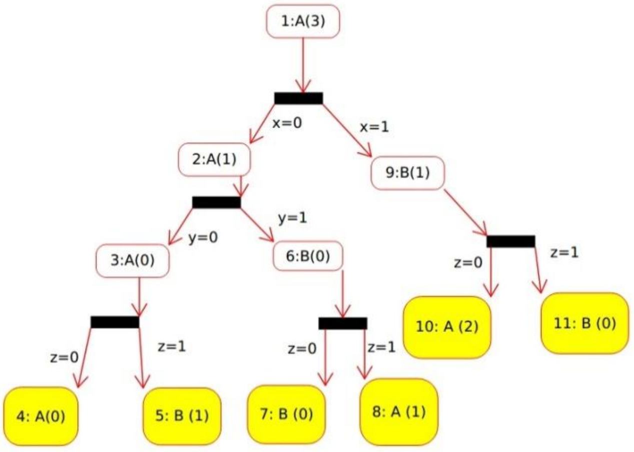

Figure 1 The misclassified detection in the pruning set of REP-Tree (binary sample), [34]

by using a pruning set shown in the following table:

TABLE 1

Contains some samples

Category	X	Y	Z
A	0	0	1
B	0	1	1
B	1	1	0
B	1	0	0
A	1	1	1
B	0	0	0

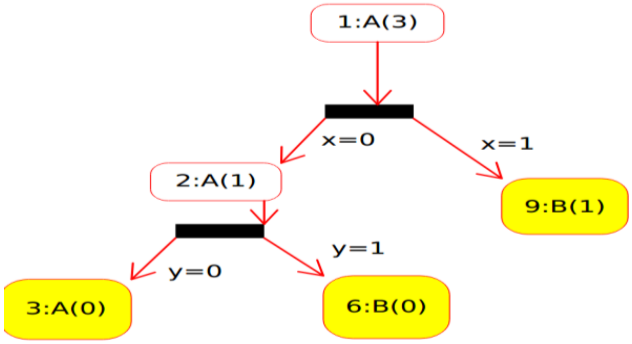

Figure 2 The final REP-tree

The REP-tree begins from bottom from node three. We show that node three can be produced into a leaf which makes the minimum errors, on the pruning set, than as a sub tree. As a sub-tree (the pruned tree) the classification occurs at nodes four and five. One error happened in node five; but no errors happened in node three. The same matter happened in node six and node nine. However, node number two cannot be made into a leaf since it makes one error while as a sub tree, with the newly-created leaves three and six. It makes no errors as shown in Figure (2). The pruning comes as a solution to the sub-tree replication problem that happened with the decision tree starts splitting. The definition of this case as "When sub tree replication occurs, identical sub trees can be found at several different places in the same tree structure" [28].

V Proposed Work

In this section, the whole Arabic text classification process will be explained then it will divide the work into a collection of systems, each system has special combinations to produce the final process of classification after preparing the dataset. These combinations are taken from what has already been explained in previous section.

Arabic Text Datasets

In this subsection, we will present the datasets used in the experiments of our paper. The used datasets are as the following:

BBC-Arabic News Dataset

The first data set contains the number of 4680 documents of BBC-Arabic news, classified into the following predefined categories {'Middle East', 'World News', 'business', 'sport', 'newspapers', 'Science', 'Misc.'}. We choose a random set of existing documents 3000 documents manually; with the knowledge that classifies types in all documents as "single label" classification as mentioned in section (3.2.1 section) "types of text classification". The following table, Table (2), shows the division of the documents into seven preset categories.

TABLE 2
The division of BBC-Arabic news Dataset based on 60% training set.

#	Class	Training Set	Testing Set	Full Dataset
1	Middle East	630	420	1050
2	World News	222	148	370
3	Business	124	82	206
4	Sport	348	232	580
5	Newspapers	234	155	389
6	Science	141	94	235
7	Misc.	102	68	170
Total		**1801**	**1199**	**3000**

Note that BBC-Arabic data-set is collected during our work, and other two datasets are already existing in the literatures (Arabic Corpora - Mourad Abbas.) and (Arabic Corpora - Alj-News.).

Alkhaleej News Dataset

We present the second data set which contains a number of 5690 documents for Alkhaleej News Dataset (Arabic Corpora - Mourad Abbas.), (Arabic Corpora - Alj-News.) that classified into the following predefined categories {'International News', 'Local News', 'Sport', 'Economy'}. We choose a random set (2770 documents) with the knowledge that classifies types in all the documents as a single label classification (Abbas, Smaili 2005). The following table, Table (3) shows the division of the documents into four preset Categories.

TABLE 3
The division of Alkhaleej News Dataset based on 60% training set.

#	Class	Training Set	Testing Set	Full Dataset
1	Local News	630	400	1030
2	International News	480	320	800
3	Economy	264	176	440
4	Sport	300	200	500
Total		**1674**	**1096**	**2770**

The tables above show that data is partitioned into two parts data for learning and data for testing based on 60% of learning; this style existed in Weka tool with many options for this purpose.

The Proposed Systems

In this section, we will give a set of regulations contain some processes that listed in the previous section, and then a comparison will be performed between all the existing combinations in the form of independent systems and extract the results in the next section.

System A: Binary Particle Swarm Optimization and K-Nearest Neighbor.
System A is the first proposed system. It works on the classification of Arabic documents using the three main processes preprocessing, feature selection, and classifications as mentioned. This system contains three processes shown in Figure (3):

(1- Tokenization, Stop words discarding 2- BPSO/KNN, 3- J 48).
(1- Tokenization, Stop words discarding 2- BPSO/KNN, 3- SVM).
(1- Tokenization, Stop words discarding 2- BPSO/KNN, 3- Rep-Tree). .

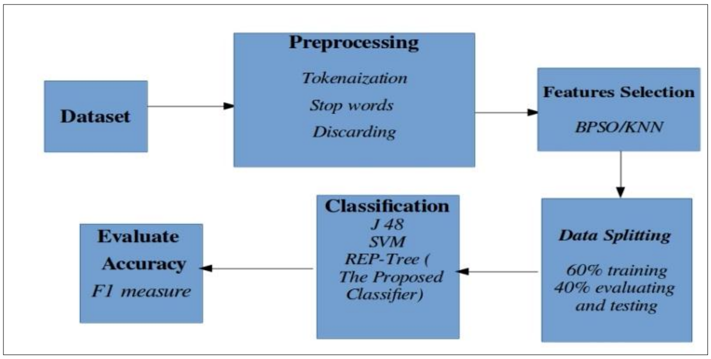

Figure 3. System A.

Figure 3 shows the processes of system A using the BBC-Arabic dataset with the previous processes.

BPSO+KNN Experiment Steps

Step 1. We need to prepare a population of particles in the features space and spread particles randomly.
Xi is the current position of the particle initialized with random binary values. Where0 means that the corresponding feature is not selected and by 1 means that the feature is selected.
Pi is the best previous position of the particle and initialized by the same value of Xi.
Vi is the velocity of Pi.
•According to the evaluation of each particle in the swarm gbest (global best) initializes by the best fitness value of a particle.
Step2. (Determining the fitness).Fitness of subset resulted by particle with the evaluation process occurs after each feature selection iteration. The best fitness is the best accuracy in the evaluation process of the selected subset of features measured by classifiers algorithms (KNN) according to the following equation [27].

$$Fitness = (\alpha * Acc) + \big(\beta * (N - T/N)\big) \qquad (5)$$

Where
•Acc refers to the classification accuracy of the particle using chosen classifier.
•To make a balance between classification accuracy and the dimension of the feature sub set that selected by particles, we use the β and α parameters to do this purpose, with range of [0, 1] for α, and 1- α for β.
- N refers to the all features.
- T refers to the selected features using particle P.
•The fitness now is updated and then the private best of each particle is updated for each particle.
Step 3. (Updating gbest).The gbest is now updated.
Step 4. (Updating position).According to the BPSO velocity equation from section three, we can alter and update both velocity and position for all particles (Mendes, Kennedy and Neves, 2004). Equation (1) and (2).
As mentioned in [25], the probability of bit changing is determined by the following: equations (3) and (4).
Where rand () is a random number between (0, 1) [27]. c1, c2 are acceleration factors. Usually c1 = c2 = 2. Pgd = global best. Vid = velocity of particle [28].
Step 5.If the fitness value is better than the best fitness value (gbest) in history then set current value as the new gbest.
Step 6.Now for evaluation in our case KNN, we use the Euclidean Distance ED to measure the relevancy between current instance and the other instances in the data-set.
Step 7.Define the repository R.
•If the predicted classifications of instances were similar to the predefined classification, increase repository R by 1.
Step 8. Now, we can measure the classification accuracy of particle P by [27].

$$ClassificationAccuracy = \frac{R}{N} \qquad (6)$$

Where R is the group of results after testing the features from all training set N.

The Experiment Parameters (BPSO+KNN).

(1) Inertia weight (ω): in the previous equation (2) is to balance the local search and the global search [27], and from the literature the best value of ω is 1.2.
(2) The swarm dimension is 50 units.
(3) Iterations are 200 iterations.
(4) [0, 1] for α, and 1- α for β. If we use the 1 for α then β = 0 and this mean that the dimension of the features subset is neglected, so we choose a random number between [0, 1] for α (0.70); and β is 1 – 0.70 = 0.30.

System B: Binary Particle Swarm Optimization and Support Vector Machine.
The second system in this also studies inserting the second middle phase (Feature Selection). In this system we will use the BPSO with SVM, and then classify the resultant features by (Decision Trees J 48, Support Vector Machine SVM, and Reduced Error Pruning-Tree Rep-Tree) as shown in Figure (4).

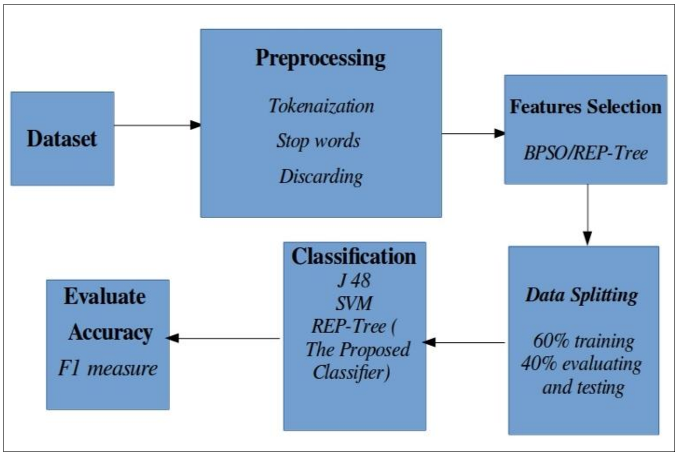

Figure 4 System B.

Figure (4) shows the processes of system B using the BBC-Arabic dataset with the previous processes by adding the BPSO+SVM as a feature selection, and the resultant features will be classified using the three classifiers SVM as classifier, J48, and REP-Tree for Arabic words.

BPSO+SVM Experiment Steps:

Step 1. The same in system A.

Step2. (Determining the fitness).

Here we use the previous equation in system A, (1).

Here we use SVM to measure the classification instead of KNN in the previous system A.

Step 3. (Updating gbest) the same in A.

Step 4. (Updating position) the same in A using the equations (2), (3), and (4).

Step 5. The same in A.

Step 6.Now for evaluation in our case SVM, we use the SVM classifier in Weka tool to measure the relevancy between current instance and the other instances in the data-set.

Then repeat both step 7 and 8 as mentioned in system A. also the same previous parameters in system A. experiments.

System C: Binary Particle Swarm Optimization and Reduced Error Pruning Tree.

The last system in this study also involves inserting the middle phase Feature Selection including the previous processes and contents in system A, and B. In this system we will use the BPSO with Reduced Error Pruning-Tree Rep-Tree where it was not used in Arabic text classification field yet and it was recently used in English news classification. Finally, we will classify the resultant features by Decision Trees (J 48), Support Vector Machine SVM, and Reduced Error Pruning-Tree. Rep-Tree) (As a classifier) as shown in Figure (5).

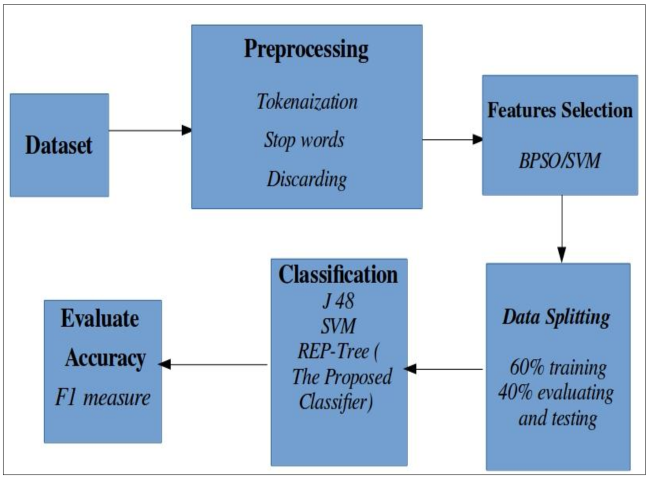

Figure 5 System C.

Figure (5) shows system C with adding the BPSO+REP-Tree as a features selection REP-Tree here (evaluator), and the resultant features will be classified using the three classifiers (SVM, J48, and REP-Tree (as classifier)) for Arabic words.

BPSO+REP-Tree Experiment Steps

Step 1.The same in system A.

Step2. (Determining the fitness).Here we use Reduced Error Pruning-Tree REP-Tree as a feature evaluator to measure the classification accuracy of the particle in a training set, instead of KNN in system A.

Step 3. (Updating gbest) the same in A.

Step 4. (Updating position) the same in A. using the equations (2), (3), and (4).

Step 5. The same in A.

Step 6.Now for evaluation in our case REP-Tree we use REP-Tree classifier in Weka tool to measure the relevancy between current instance and the other instances in the dataset.

Then repeat both step 7 and 8 as mentioned in system A. also the same previous parameters in system A. experiments.

We can alternate the last three steps by measuring the F measure factor to estimate the classification accuracy.

We can list the previous steps in short and general points as the following:

(1) First and after preparing the features ,terms, space and spread particles randomly, we determine the accuracy of the classification (Acc) of a particle P in training data-set by using Reduced Error Pruning-Tree REP-Tree.

(2) Start extracting and filtering the features subset of the training set that selected by particle.

(3) Evaluate the previous extracted features data-set by the

REP-Tree by 60 % training set validation.

(4) Determine the F measure factor that result from the REP-Tree experiment to determine the fitness of the particle.

VI. Experimental Results

In this section, the experimental results of the previous systems are described in last section. We have executed our experiments on two data-sets, the BBC-Arabic news dataset and Alkhaleej News dataset. As mentioned in the previous section, we split the data into 60% for training and 40% for testing, and then display the results in Tables and Figures. After that, we will compare every system with the other in specific graph. We will start presenting the results of system A using the three classifiers which have been previously described in section 4. Then gradually we will review the results of system B, and finally we end with system C.

6.1 System A.A ("BPSO+KNN"/J 48)

The experimental results of system A with J 48 tree are shown by Table (4) and (5) using the previous two datasets:

TABLE 4

System A with J 48 tree applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	67.3	69.7	68.4
World News	81.5	85.4	83.4
Business	72.4	73.4	72.8
Sport	84.2	79.7	81.8
Newspapers	87.3	88.9	88.0
Science	62.7	86.1	72.5
Misc.	83.9	89.6	86.6
Average	77	81.8	79

Table (4) shows the classification of BBC-Arabic documents using BPSO+KNN as a feature selector and J 48 decision tree as a classifier. As it is clear from the table, the results are as the following: the best classification is in "Newspapers" class with precision of 87.3, recall of 88.9 and F1-Measure of 88.0. The second performance rank of classes is the "Misc." with precision of 83.9, recall of 89.6 and F1-Measure of 86.6. There is a convergence in the outcome of both "Word News" and "Sport" with a little outperforming in recall of 85.4 for "Word News" class. The worst two classes were the "Science" and the "Middle East" classes with precision of 62.7, recall of 86.1 and F-Measure of 72.5 for "Science" and the worst precision with 67.3 and F-Measure with 68.4 for "Middle East" class. Then we have the second data-set (Alkhaleej News Dataset) with the same previous experiment, Table (5) shows the results as the following:

TABLE 5

System A with J 48 tree applied on Alkhaleej News Dataset

Table (5) shows the classification of Alkhaleej News Dataset documents using BPSO+KNN as a feature selector and J 48 decision tree as a classifier. The best F-Measure is for "Sport" class with 84.2, and the worst F-Measure is for "Economy" class with 62.8.

Class	Precision%	Recall%	F1-Measure%
Local News	75.8	78.4	77
International News	74.6	72.3	73.4
Economy	65.2	60.7	62.8
Sport	81.3	87.5	84.2
Average	74.2	74.7	74.3

6.2 System A.B ("BPSO+KNN"/SVM)

The experimental results of system A with SVM classifier are shown by Tables (6) and (7) using the previous two datasets (BBC Arabic, and Alkhaleej datasets as the following:

TABLE 6

System A with SVM classifier applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	88.3	79.7	83.7
World News	81.7	87.3	84.4
Business	84.5	92.4	88.2
Sport	87.2	79.7	83.2
Newspapers	86.4	88.2	87.3
Science	81.4	85.6	83.4
Misc.	89.4	95.6	92.3
Average	85.5	86.9	86

Table (6) shows the classification of BBC-Arabic documents using BPSO+KNN as a feature selector and SVM as a classifier. As it is clear from Table (6), the results are as the following: the best classification is for "Misc." class with precision of 89.4, recall of 95.6 and F1-Measure of 92.3. The second performance rank of classes is the "Business" with precision of 84.5, recall of 92.4 and F1-Measure of 88.2. There is a convergence in the F1-Measure outcome of both "Middle East" and "Science" with F1-Measure of 83.7 and 83.4 gradually. The worst class is the "Sport" with precision of 87.2, recall of 79.7 and F-Measure of 83.2. Now we will apply system A (the same previous experiment with SVM) on the second data-set (Alkhaleej News Dataset) and Table (7) shows the results as the following:

TABLE 7
System A with SVM classifier applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	86.1	90.4	88.1
International News	82.4	81.7	82
Economy	91.6	87.8	89.6
Sport	95.3	89.5	92.3
Average	88.8	87.3	88

Table (7) shows the classification of Alkhaleej News Dataset documents using BPSO+KNN as a feature selector and SVM as a classifier. The best F-Measure is for "Sport" class with 92.3, and the worst F-Measure is for "International News" class with 82.

6.3 System A.C ("BPSO+KNN"/REP-Tree)

The third combination of system A is our proposed classifier REP-Tree which has recently been used in English text classification as mentioned previously in the past sections. Here, the REP-Tree is a classifier used to classify a group of feature resulting from the operation of features selection by BPSO+KNN. The experimental results of system A with REP-Tree classifier are shown by Tables (8) and (9) using the previous two datasets BBC Arabic and Alkhaleej datasets as the following:

TABLE 8
System A with REP-Tree classifier applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	87.7	91.5	89.5
World News	85.9	85.7	85.7
Business	86.1	90.6	88.2
Sport	80.3	72.2	76
Newspapers	89.2	88.7	88.9
Science	83.8	87.8	85.7
Misc.	79.2	72.3	75.5
Average	84.6	84.1	84.2

Table (8) shows the classification of BBC-Arabic documents using BPSO+KNN as a feature selector and REP-Tree as a classifier. As it is clear from Table (0), the results are as the following: the best classification is for "Middle East" class with precision of 87.7, recall of 91.5 and F1-Measure of 89.5. The second rank of performance is for classes "Newspapers" with precision of 89.2, recall of 88.7 and F1-Measure of 88.9. We can detect the convergence between the previous class performance and the "Business" class performance with precision of 86.1, recall of 90.6 and F1-Measure of 88.2. The worst performance was the "Misc." class with precision of 79.2, recall of 72.3 and F-Measure of 75.5. As in all previous experiments we'll apply the REP-Tree classifier on the other datasets. Now we will apply system A (the same previous experiment with REP-Tree) on the second data-set (Alkhaleej News Dataset) and Table (9) shows the results as the following:

TABLE 9
System A with REP-Tree classifier applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	88.4	91.5	89.9
International News	93.2	85.2	89
Economy	80.1	83.6	81.8
Sport	92.7	82.7	87.4
Average	88.6	85.7	87

Accuracy results were comparable between REP-Tree and SVM with average F1-Measure of 87% for REP-Tree and 88% for SVM. For more details of the results the best F-Measure is for "Local News" class with 89.9, and the worst F-Measure is for "Economy" class with 81.8.

6.4 System B.A ("BPSO+SVM"/J 48)

The experimental results of system B with J 48 tree are shown by Tables (10) and (11) using the previous two datasets (BBC-Arabic news dataset and Alkhaleej News dataset):

TABLE 10
System B with J 48 tree applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	70.4	72.6	71.4
World News	88.3	83.1	85.6
Business	77.5	71.2	74.2
Sport	87.7	78.5	82.8
Newspapers	85.2	87.3	86.2
Science	61	77.4	68.2
Misc.	82.5	87	84.6
Average	78.9	79.5	79

Table (10) shows the classification of BBC-Arabic documents using BPSO+SVM as a feature selector and J 48 decision tree as a classifier. As it is clear from the table, the results are as the following: the best classification performance is "Newspapers" class with precision of 85.2, recall of 87.3 and F1-Measure of 86.2. The second rank of classification performance is the "World News" with precision of 88.3, recall of 83.1 and F1-Measure of 85.6. We can see that the worst classes are the "Middle East" and the "Science" classes with precision of 70.4, recall of 72.6 and F-Measure of 71.4 for "Middle East" and the worst precision with 61.0 and F-Measure with 68.2 for "Science" class. Here we can be quite sure that the J 48 tree failed in the classification accuracy of "Science" class by 31.8% according to its F-Measure. Now we have the second data-set (Alkhaleej News

Dataset) with the same previous experiment, Table (11) shows the results as the following:

TABLE 11

System B with J 48 tree applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	49.8	52.4	51
International News	93.3	62.4	74.7
Economy	67.1	77.5	71.9
Sport	85.3	69.8	76.7
Average	73.8	65.5	68.5

Table (11) shows the classification accuracy of Alkhaleej News Dataset documents using BPSO+SVM as a feature selector and J 48 decision tree as a classifier. The best F-Measure is for "Sport" class with 76.6, and the worst F-Measure is for "Local News" class with 51. Also here we can be quite sure that the J 48 tree failed in the classification accuracy of "Sport" class by 49% according to its F-Measure.

6.5 System B.B ("BPSO+SVM"/SVM)

The experimental results of system B with SVM classifier are shown by Tables (13) and (14) using the previous two datasets (BBC Arabic and Alkhaleej datasets as the following:

TABLE 12

System B with SVM classifier applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	67.9	88.7	76.9
World News	98.7	90.3	94.3
Business	87.9	89.3	88.5
Sport	60.3	80.7	69
Newspapers	79.8	84.2	81.9
Science	99.2	85.6	91.8
Misc	90.4	98.8	94.4
Average	83.4	88.2	85.2

Table (12) shows the classification of BBC-Arabic documents using BPSO+KNN as a feature selector and SVM as a classifier. As it is clear from Table (12), the results are as the following: the best classification is for "Misc." class with precision of 90.4, recall of 98.8 and F1-Measure of 94.4. The second performance rank of classes is the "World News" with precision of 98.7, recall of 90.3 and F1-Measure of 94.3. The worst class is the "Sport" with precision of 60.3, recall of 80.7 and F-Measure of 69. Now we will apply system B (the same previous experiment with SVM) on the second data-set (Alkhaleej News Dataset), and Table (13) shows the results as the following:

TABLE 13

System B with SVM classifier applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	83.2	88.6	85.8
International News	88.5	85.7	87
Economy	96.6	90.9	93.6
Sport	90.3	89.7	89.9
Average	89.6	88.7	89

Table (13) shows the classification of Alkhaleej News Dataset documents using BPSO+SVM as a feature selector and SVM as a classifier. The best accuracy (F-Measure) is for Economy class with 93.6 and the worst F-Measure is for "Local News" class with 85.8.

6.6 System B.C ("BPSO+SVM"/REP-Tree)

The third combination of system B is our proposed classifier REP-Tree as we mentioned in the previous experiments which has recently been used by (Kalmegh, 2015), (Patel and Upadhyay, 2012) in English text classification and by (Naji and Ashour, 2016) in Arabic text classification (a previous paper related to the existing paper), as mentioned previously in the past sections specifically in the first section. Here the REP-Tree is a classifier, which is used to classify a group of features resulting from the operation of features selection by BPSO+SVM. The experimental results of system B with REP-Tree classifier are shown by Tables (14) and (15) using the previous two datasets (BBC Arabic and Alkhaleej datasets as the following:

TABLE 14

System B with REP-Tree classifier applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	77	89.4	82.7
World News	98.3	96.1	97.1
Business	87.2	78.5	82.6
Sport	79.5	75.8	77.6
Newspapers	88.2	88.9	88.5
Science	85.4	87.1	86.2
Misc	89	69.4	77.9
Average	86.3	83.6	84.6

Table (14) shows the classification of BBC-Arabic documents using BPSO+SVM as a feature selector and REP-Tree as a classifier. As it is clear from Table (14), the results are as the following: the best classification is for "World News" class with precision of 98.3, recall of 96.1 and F1-Measure of 97.1. The second rank of performance is for classes "Newspapers" with precision of 88.2, recall of 88.9 and F1-Measure of 88.5. We can detect the convergence between the "Middle East" class performance and the "Business" class performance with F1-Measure of 82.7 and 82.6. The worst performance was the "Sport" class with precision of

79.5, recall of 75.8 and F-Measure of 77.6. As in all previous experiments we'll apply the REP-Tree classifier on the other datasets. Now we will apply system B (the same previous experiment with REP-Tree) on the second data-set (Alkhaleej News Dataset), and Table (15) shows the results as the following:

TABLE 15

System B with REP-Tree classifier applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	72	78.3	75
International News	89.6	92.2	90.8
Economy	87.3	88.3	87.7
Sport	95.4	87.5	91.2
Average	86	86.5	86.1

From Table (15) we see that the best accuracy of REP-Tree F1-Measure is 91.2 for “Sport” class, and the worst F-Measure is for “Local News” class with 75. We note that the results were comparable with SVM classifier. Now we will apply the REP-Tree on another data-set.

6.7 System C.A (“BPSO+REP-Tree”/J 48)

System C consists of Binary PSO as a feature selector and the proposed REP-Tree as an evaluator to check the best group of features then we use the three previous classifiers (J 48, SVM, and REP-Tree) to build the classification model; the classification in the resultant group of features in the training set to reduce the dimension of the original data-set and then apply the classifiers on the test data-set. We have previously noted that REP-Tree has recently been used by (Kalmegh, 2015), (Patel and Upadhyay, 2012) to classify English text and by (Naji and Ashour, 2016) in Arabic text classification.

The experimental results of system C with J 48 tree are shown by Tables (16) and (17) using the previous two datasets (BBC-Arabic news dataset and Alkhaleej News dataset):

TABLE 16

System C with J 48 tree applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	88.7	83.3	85.9
World News	90.4	87.4	88.8
Business	75.2	70.5	72.7
Sport	84.8	74.2	79.1
Newspapers	80.1	83.8	81.9
Science	79.8	78.3	79
Misc.	77.6	85.7	81.4
Average	82.3	80.4	81.2

Table (16) shows the classification of BBC-Arabic documents using BPSO+REP-Tree as a feature selector and J 48 decision tree as a classifier. As it is clear from the table, the results are as the following: the best classification performance is “World News” class with precision of 90.4, recalling of 87.4 and F1-Measure of 88.8. The second rank of classification performance is the “Middle East” with precision of 88.7, recall of 83.3 and F1-Measure of 85.9. We can note that the worst class was the “Business” class with precision of 75.2, recall of 70.5 and F-Measure of 72.7. Here we can be quite sure that the J 48 tree failed in the classification accuracy of “Science” class by 27.3% according to its F-Measure.

Now we have the second data-set (Alkhaleej News Dataset) with the same previous experiment, Table (17) shows the results as the following:

TABLE 17

System C with J 48 tree applied on Alkhaleej News Dataset

Table (17) shows the classification accuracy of Alkhaleej News Dataset documents using BPSO+REP-Tree as a feature selector and J 48 decision tree as a classifier. The best F-Measure is for “Economy” class with 82.5, and the worst F-Measure is for “Local News” class with 58.4. Also here we can be quite sure that the J 48 tree failed in the classification accuracy of “Local News” class by 47.6% according to its F-Measure.

6.8 System C.B (“BPSO+REP-Tree”/SVM)

Class	Precision%	Recall%	F1-Measure%
Local News	60.3	56.8	58.4
International News	68.6	70.9	69.7
Economy	90.4	75.9	82.5
Sport	84.8	72.5	78.1
Average	73.5	69	72.1

The experimental results of system C with SVM classifier are shown by Table (18) and (19) using the previous two datasets (BBC Arabic and Alkhaleej datasets as the following:

TABLE 18

System C with SVM classifier applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	98.6	94.4	96.4
World News	68.2	88.9	77.1
Business	82.3	85.7	83.9
Sport	64.6	78.5	70.8
Newspapers	81.4	82.8	82
Science	97.2	87.1	91.8
Misc.	92.5	96.9	94.6
Average	83.5	87.7	85.2

Table (18) shows the classification of BBC-Arabic documents using BPSO+REP-Tree as a feature selector and SVM as a classifier. From Table (18) we note the equality in F-Measure average value using the same classifier SVM with a different features selection combination (BPSO+REP-Tree). The current results have been compared with Table (12, 13) (BPSO+SVM features selection). We get here an average F-Measure of 85.2 and 89.6 for SVM (the same classifier but different feature selector). As usual, we will apply system C (the same previous experiment with SVM) on the second data-set (Alkhaleej News Dataset), and Table (19) shows the results as the following:

TABLE 19

System C with SVM classifier applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	97.2	93.7	95.4
International News	94.5	82.9	88.3
Economy	90.3	95.5	92.8
Sport	79.5	80	79.7
Average	90.3	88	89.05

Table (19) shows the classification of Alkhaleej News Dataset documents using BPSO+REP-Tree as a feature selector and SVM as a classifier. The best accuracy (F-Measure) is for "Local News" class with 95.4 and the worst F-Measure is for "Sport" class with 79.7. In this experiment, we note the equality and convergence in the classification process results using the same classifier SVM with a different features selection combination (BPSO+REP-Tree).

6.9 System C.C ("BPSO+REP-Tree"/REP-Tree)

The third combination of system C consists of Binary PSO as a feature selector and the proposed REP-Tree as an evaluator then we use REP-Tree as a classifier, as we mentioned in the previous section System C subsection. The experimental results of system C with REP-Tree classifier are shown by Tables (20) and (21) using the previous two datasets (BBC Arabic and Alkhaleej datasets as the following:

TABLE 20

System C with REP-Tree classifier applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	97.2	95.3	96.2
World News	88.6	78.5	83.2
Business	87.3	88.6	87.9
Sport	79.9	75.9	77.8
Newspapers	86.1	98.4	91.8
Science	80	86.9	83.3
Misc.	82.5	92	86.9
Average	85.9	87.9	86.7

Table (20) shows that the REP-Tree has been effective enough in the classification for BBC-Arabic documents using BPSO+REP-Tree as a feature selector and REP-Tree as a classifier. The results are as following: the best classification is for "Middle East" class with precision of 97.2, recall of 95.3 and F1-Measure of 96.2. Next we have the second classification performance the "Newspapers" with precision of 86.1, recalling of 98.4 and F1-Measure of 91.8. The third classification accuracy is the "Business" with F-Measure of 87.9. We can detect the convergence between the "Science" class performance and the "World News" class performance with F1-Measure of 83.3 and 83.2. The worst performance was the Sport class with F-Measure of 77.8.

As usual, we will apply the REP-Tree classifier on the other datasets. Now, we will apply system C, the same previous experiment with REP-Tree, on the second data-set (Alkhaleej News Dataset), and Table (21) shows the results as the following:

TABLE 21

System C with REP-Tree classifier applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	98	97.4	97.6
International News	91.3	92.5	91.8
Economy	85.7	87.1	86.3
Sport	93.8	89.6	91.6
Average	92.2	91.6	91.8

From Table (21), we see that the best accuracy of REP-Tree (F1-Measure) is 97.6 for "Local News" class, and the worst F-Measure is for "Economy" class with 86.3. The average accuracy of the REP-Tree in this experiment was 91.8.

6.10 Performance of the Three Systems

In this subsection, we will make a comparison between the previous results on the previous two datasets (BBC-Arabic and Alkhaleej) before adding some enhancements to each system in the preprocessing phase. Both Table (22) and Figure (6) show the results of this comparison.

TABLE 22

Comparison between the F-Measure averages of the three systems

Datasets	System A (BPSO+KNN)%	System B (BPSO+SVM)%	System C (BPSO+REP-Tree)%
BBC-Ar (J48)	79	79	81.2
BBC-Ar (SVM)	86	85.2	85.2
BBC-Ar (REP)	84.2	84.6	86.7
Alkhaleej(J48)	74.3	68.5	72.1
Alkhaleej(SVM)	88	89	89

Alkhaleej(REP)	87	86.1	91.8

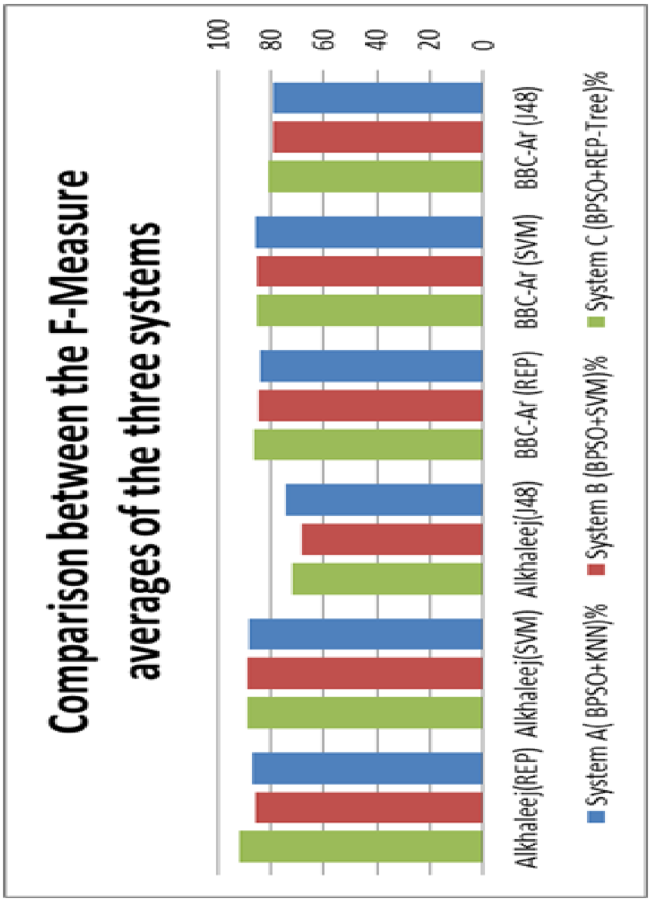

Figure 6 Comparison between the accuracy of the three systems.

From Table (22) and Figure (6), we draw the overall results of all the experiments, calculate the average for F1-Measure values, and compare all the systems with each other.

VII. CONCLUSION

This paper proposed a new feature selection approach to select the best subset of features from the original Arabic document .We showed that the proposed approach works well in this area after extracting the experimental results. The proposed approach can be used in the field of Arabic search engines and classifying huge amounts of Arabic websites pages into hierarchal classes, labels.

We proposed the Reduced Error Pruning-Tree classifier, which was not used with Arabic text classification before for two purposes. The first one is an evaluator to evaluate the subset of features that resulted from the features selection algorithm Binary Particle Swarm Optimization BPSO. To evaluate this approach (BPSO+REP-Tree), we used two Arabic datasets, BBC-Arabic News dataset and Alkhaleej News dataset. The second purpose of the Rep-Tree is to use it as a classifier to build the learning model. We compare the first purpose (BPSO+REP-Tree) with two existing approaches, (BPSO+KNN) and (BPSO+SVM), and the second purpose (REP-Tree classifier) with two well-known classifiers, J 48 and SVM. We named the three features selection approaches with A for (BPSO+KNN), B for (BPSO+SVM), and C for (BPSO+REP-Tree). After we get the experimental results, we concluded that the proposed approach System C is effective. We choose the F1-Measure to estimate the accuracy of the classification process which came from two factors, precision and recall factors.

The values of F1-Measure for system A with the classifier J 48 is in the range of 73% - 79%, with SVM is in 86% - 88% and with the proposed classifier REP-Tree is in the range of 84% - 87%. Next, we get the F1-Measure values of the second system (B) with the same classifiers as the following, with J 48 are in the range of 60.9% - 84.6%, with SVM is in 85.2% - 89.6% and with the proposed classifier REP-Tree is in the range of 84.6% - 89.5% and here is the last two algorithms which are comparable in the accuracy. Finally, we apply the experiments on our proposed approach system (C) in features selection domain and it gave these ranges of accuracy as the following, with J 48 was in the range of 69.5% - 79.6%, and with SVM is in 87% - 89.8% and with the proposed classifier REP-Tree is in the range of 86.7% - 91.8%.

REFERENCES

[1] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval. Information Processing & Management, vol.24, no.5, (1988), pp.513-523. doi:10.1016/0306-4573(88)90021-0

[2] S. Li, R. Xia, C. Zong and C. Huang, "A framework of feature selection methods for text categorization". Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09, (2009), doi:10.3115/1690219.1690243

[3] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics". Bioinformatics, vol.23, no.19, (2007), pp.2507-2517. doi:10.1093/bioinformatics/btm344

[4] M. M. Al-Tahrawi and S. N. Al-Khatibb, (2015).Arabic text classification using Polynomial Networks. Journal of King Saud University - Computer and Information Sciences, vol. 27, no. 4, (2015), pp. 437-449. http://dx.doi.org/10.1016/j.jksuci.2015.02.003

[5] T. Aimunandar and E. Winarko, Regional Development Classification Model using Decision Tree Approach.

International Journal of Computer Applications IJCA, vol. 114, no. 8, (2015), pp.29-34. doi:10.5120/20000-1755

[6] S. Kalmegh, Analysis of WEKA Data Mining Algorithm REP-Tree, Simple Cart and Random Tree for Classification of Indian News. PARIPEX Paripex - Indian Journal of Research, vol. 2, no. 2, (2015), pp. 438-446. doi:10.15373/22501991/feb2015.

[7] Patel, N., & Upadhyay, S. Study of Various Decision Tree Pruning Methods with their Empirical Comparison. International Journal of Computer Applications, vol. 60, no.12, (2012), pp. 20-25. doi:10.5120/9744-4304

[8] Brahimi, M. Touahria, and A. Tari, Data and Text Mining Techniques for Classifying Arabic Tweet Polarity. Journal of Digital Information Management, vol.14, no.1, (2016), pp. 12-19.

[9] S. M. Oraby, Y. El-Sonbaty, M. A. El-Nasr, Exploring the Effects of Word Roots for Arabic Sentiment Analysis. In International Joint Conference on Natural Language Processing, Nagoya, Japan (2013), pp. 471-479.

[10] Khoja, S. (1999).Stemming Arabic Text, Lancaster, U.K, Computing Department, Lancaster University.

[11] K. Taghva, R. Elkhoury, and J. Coombs, Arabic stemming without a root dictionary. Paper presented at International Conference on Information Technology: Coding and Computing (ITCC'05). (2005). doi:10.1109/itcc.2005.90

[12] Tashaphyne, Arabic light stemmer, 0.2. Available at https://pypi.python.org/pypi/Tashaphyne (2010).

[13] A. Shoukry, and A. Rafea, Sentence-level Arabic sentiment analysis. International Conference on Collaboration Technologies and Systems (CTS). (2012). doi:10.1109/cts.2012.6261103.

[14] E. Al-Thwaib, Text Summarization as Feature Selection for Arabic Text Classification. World of Computer Science and Information Technology Journal (WCSIT), vol. 4, no.7, (2014), pp. 101-104.

[15] K. Al-Hindi, and E. A. Al-Thwaib, Comparative Study of Machine Learning Techniques in Classifying Full-Text Arabic Documents versus Summarized Documents. *World of Computer Science and Information Technology Journal (WCSIT), vol. 2*, no. 7, (2013), pp. 126-129. Retrieved August17, 2016, from http://www.wcsit.org/pub/2013/vol.3.no.7/A Comparative Study of Machine Learning Techniques in Classifying Full-Text Arabic Documents versus Summarized Documents.pdf

[16] A. Abu-Errub, Arabic Text Classification Algorithm using TF.IDF and Chi Square Measurements. *International Journal of Computer Applications IJCA, vol. 93*, no. 6, (2014). Pp. 40-45. doi:10.5120/16223-5674

[17] Goweder, A., Elboashi, M., & Elbekai, A. (2013).*Centroid-Based Arabic Classifier.* The International Arab Conference on Information Technology (ACIT'2013), 108(3). Retrieved June 27, 2016, from http://acit2k.org/ACIT/2013Proceedings/108.pdf

[18] Abidi, K., & Elberrichi, Z. (2012). Arabic Text Categorization: A Comparative Study of Different Representation Modes. *Journal of Theoretical and Applied Information Technology, 38*(1), 465-470. Retrieved May 21, 2016, from http://ccis2k.org/iajit/PDF/vol.9,no.5/2983-10.pdf

[19] Raho, G., Al-Shalabi, R., Kanaan, G., & Nassar, A. (2015). Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study. *International Journal of Advanced Computer Science and Applications Ijacsa, 6*(2) 23-28. doi:10.14569/ijacsa.2015.060228

[20] Mohammad, A. H., Al-Momani, O., & Alwada, T. (2016). Arabic Text Categorization using k-nearest neighbor, Decision Trees (C4.5) and Rocchio Classifier: A Comparative Study. *International Journal of Current Engineering and Technology, 6*(2), 477-482. RetrievedMay29, 2016,from http://inpressco.com/wp-content/uploads/2016/03/Paper16477-482.pdf

[21] Kanan, T., & Fox, E. A. (2015).Automated Arabic text classification with P-Stemmer, machine learning, and tailored news article taxonomy. *Journal of the Association for Information Science and Technology J Assn Inf Sci Tec.* doi:10.1002/asi.23609

[22] Al-Anzi, F. S., & Abuzeina, D. (2016).Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. *Journal of King Saud University - Computer and Information Sciences.* doi:10.1016/j.jksuci.2016.04.001

[23] Zubi, Z. S. (2009). *Using Some Web Content Mining Techniques for Arabic Text Classification.* Recent Advances on Data Networks, Communications, Computers, 73-84. doi:10.1109/mmcs.2009.5256696

[24] Zrigui, M., Ayadi, R., Mars, M., & Maraoui, M. (2012). Arabic Text Classification Framework Based on Latent Dirichlet Allocation. *Journal of Computing and Information Technology CIT, 20*(2), 11-14. doi:10.2498/cit.1001770

[24] Kennedy, J., & Eberhart, R. (1995).Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks, 4, 1942-1948. doi:10.1109/icnn.1995.488968

[25] Kennedy, J., & Eberhart, R. (1995).Particle swarm optimization. Proceedings of ICNN'95 - International Conference on Neural Networks, 4, 1942-1948. doi:10.1109/icnn.1995.488968

[26] Yang, Y., & Pedersen, J. O. (1997).A comparative study on feature selection in text categorization. Machine Learning International Workshop Then Conference, 412-420. Morgan Kaufmann Publishers, INC

27 Chantar, H. K., & Corne, D. W. (2011). Feature subset selection for Arabic document categorization using BPSO KNN. Third World Congress on Nature and Biologically Inspired Computing. doi:10.1109/nabic.2011.6089647

[28] Tsai, M., Su, C., Chen, K., & Lin, H. (2012).An Application of PSO Algorithm and Decision Tree for Medical Problem. Neural Comput & Applic Neural Computing and Applications, 21(8), 124-126. Retrieved September 07, 2016, from http://psrcentre.org/images/extraimages/31012565.pdf

[29] Shi, Y., & Eberhart, R. (1995).A modified particle swarm optimization. Proceedings of The 1998World Congress on Computational Intelligence, 6, 69-73. doi:10.1109/icnn.1995.4889684

[30] Kennedy, J., & Eberhart, R. (1997).A discrete binary version of the Particle swarm algorithm. Proceedings of The 1998World Congress on Computational Cybernetics and Simulation, 4, 4104-4108.

[31] Naji, H., Ashour, W. (2016). Text Classification for Arabic Words Using Rep-Tree. International Journal of Computer Science and Information Technology IJCSIT, 8(2), 101-108. doi:10.5121/ijcsit.2016.8208

[32] Kalmegh, S. (2015).Analysis of WEKA Data Mining Algorithm REP-Tree, Simple Cart and Random Tree for Classification of Indian News. PARIPEX Paripex - Indian Journal of Research, 2(2), 438-446. doi:10.15373/22501991/feb2015.

[33] Patel, N., & Upadhyay, S. (2012). Study of Various Decision Tree Pruning Methods with their Empirical Comparison. International Journal of Computer Applications, 60(12), 20-25. doi:10.5120/9744-4304

[34] Mitchell, T. M. (1997). Machine learning.McGraw Hill. Retrieved May 3, 2016, from http://personal.disco.unimib.it/Vanneschi/McGrawHill_-_Machine_Learning_-Tom_Mitchell.pdf