Utilizing Artificial Bee Colony Algorithm as Feature Selection Method in Arabic Text Classification

Musab Mustafa Hijazi
Department of Computer Science
Al Ain University, Al Ain, UAE
Musab.hijazi@aau.ac.ae

Akram Zeki
Department of Information Technology,
KICT, IIUM, KL, Malaysia
akramzeki@iium.edu.my

Amelia Ismail
Department of Computer Science,
KICT, IIUM, KL, Malaysia
amelia@iium.edu.my

Abstract: *A huge amount of crucial information is contained in documents. The vast increase in the number of E-documents available for user access makes the utilization of automated text classification essential. Classifying or arranging documents into predefined groups is called Text classification. Feature Selection (FS) is needed for minimizing the dimensionality of high-dimensional data and extracting only the features that are most pertinent to a particular task. One of the widely used algorithms for feature selection in text classification is the Evolutionary algorithm. In this paper, the filter method chi-square and the Artificial Bee Colony (ABC) algorithm were both used as FS methods. The chi-square method is a useful technique for reducing the number of features and removing those that are superfluous or redundant. The ABC technique considers the chi-square method's chosen features as viable solutions (food sources). The ABC algorithm searches for the most efficient selection of features that increase classification performance. Support Vector Machine and Naïve Bayes classifiers were used as a fitness function for the ABC algorithm. The experiment results demonstrated that the proposed feature selection method was able of decreasing the number of features by approximately 89.5%, and 94%, respectively when NB and SVM were used as fitness functions in comparison to the original dataset, while also enhancing classification performance.*

Keywords: *Artificial bee colony, Arabic text classification, wrapper feature selection, feature selection,*

Received April 7, 2023; accepted May 10, 2023
https://doi.org/10.34028/iajit/20/3A/11

1. Introduction

The best method for demonstrating knowledge is through documentation, which indicates that the main information sources are documents. Owing to the internet's rapid expansion, there is a significant increase in the number of E-documents. As a result, it is necessary to find flexible and efficient techniques to access, organize, and extract important information, including text classification and clustering [43]. Classifying or organizing documents into predefined groups or classes using predetermined criteria is called Text Classification (TC). Many applications, including document organization, automated document indexing, spam filtering, and disambiguation of word meaning use text classification [36, 37].

Information overload is a common problem in research, especially with the vast amount of data available on the internet. Strategies for reducing dimensions like feature selection could be applied to extract essential features from high-dimensional data, thus simplifying the analysis process and improving computational efficiency and accuracy. Feature selection is indeed a crucial process for shrinking the dimensionality of high-dimensional data and extracting just the most pertinent features for a certain task [22, 51, 55]. There are two types of FS methods: wrapper and filter methods. Wrapper methods evaluate the performance of a classifier using a particular subset of features and select the best or optimal subset based on the classifier's accuracy. In contrast, filter methods score each feature individually based on a specific metric, and choose the highest-scoring features for the analysis. Both types of methods have their strengths and weaknesses, and the choice of method often rely on the specific characteristics of the dataset and the task at hand. It's important to carefully select the most appropriate feature selection method to ensure accurate and efficient data analysis [22].

Evolutionary algorithms such as genetic algorithms and swarm intelligence algorithms have been widely used in the FS process in text classification tasks [3, 16, 19, 27, 39, 42, 45, 50, 54, 55]. These algorithms work by iteratively selecting a feature subset that improves the performance of the classification model. The Artificial Bee Colony (ABC) algorithm proposed by [41] is a common swarm intelligence algorithm that mimics the foraging behavior of honeybees to find the most suitable solution to an optimization problem. Many optimization problems have been successfully solved using the ABC method, including feature selection for text classification. Using the ABC algorithm as a wrapper FS method can help to identify a subset of relevant features that could improve the performance of the text classification model while diminishing the feature set's dimensionality. This approach can help to overcome the problem of information overload and improve the efficiency of the classification process, especially for Arabic text classification.

2. Literature Review and Background Overview

2.1. Related Works

Many classification techniques are incapable of dealing with high-dimensional feature space problems which are called the curse of dimensionality. Researchers have tried to develop dimension reduction techniques that strive to decrease the feature space dimensionality by eliminating uninformative, insignificant, and noisy data and keeping only informative relevant data [18]. Filter FS methods have been widely used in Arabic text classification researches to decrease the number of features and enhance classification accuracy. These methods typically involve calculating a score for each feature and electing the highest-ranked features based on this value. Commonly used FS selection methods for Arabic text classification include information gain, chi-square, document frequency, and term frequency. These methods have been compared in various studies, and the results have shown that the choice of feature selection method can have a positive effect on classification performance. The most effective method may vary depending on the specific dataset and classification task [1, 2, 6-14, 23-28, 32-34, 37, 38, 43, 47, 48, 53, 57].

Several studies have investigated the impact of wrapper feature selection methods on Arabic text classification, using various optimization algorithms such as Firefly Algorithm, Ant Colony Optimization (ACO), Binary Grey Wolf Optimizer (GWO), Particle Swarm Optimization (PSO), Binary Dragonfly Algorithm, and Simulated Annealing.

Mesleh *et al.* [46] used the ACO feature selection method and found that their method outperformed six state-of-the-art FS methods in Arabic text classification. Furthermore, studies by [5, 19-22, 59] used PSO, Binary Grey Wolf Optimizer, and Binary Dragonfly Algorithm (BDA) integrated with Simulated Annealing, respectively, and found that their proposed FS methods improved Arabic text classification. Moreover, The ABC algorithm was utilized in [17, 29, 35], Firefly Algorithm as FS in [44], and the Rat swarm algorithm in [52]. They found that the proposed wrapper FS method improved the performance of Arabic text classification. These studies highlight the potential of using wrapper feature selection methods with various optimization algorithms to imporve the performance and efficiency of Arabic text classification.

In [49], The semantic classes of the Arabic Hadith were predicted using the Knowledge Graph (KG) and the ACO technique. ACO went over the created graph to select the most pertinent features. The study demonstrated the possibility of predicting the semantic classes of Arabic hadith texts by integrating KG and ACO.

El-Hajj *et al.*[49] Used MFX as an FS method to find the most pervasive and distinctive features (MFX) over all documents falling under the same category. The study demonstrated the potential of MFX as a useful FS technique for text classification. Hadni *et al.* [30] proposed an innovative new metaheuristic method for dimensionality reduction that looks for a more condensed data representation suitable for use with machine learning algorithms.

In [4], they put up a thorough strategy for developing an efficient classification approach for Arabic texts. They collected a large and high-quality dataset. After that, term weighting and feature selection processes were applied to improve the classification performance. Finally, a genetic algorithm was used to select the most efficient subset of features with high quality. A hybrid filter-wrapper FS approach, Principal Component Analysis (PCA) to elect a distinctive subset of features, followed by (GWO) as a wrapper to come up with the most informative features was proposed in [7]. On the hand, Hadni *et al.* [31] proposed an approach using a chaotic sine cosine-based Firefly Algorithm as a hybrid FS method for Arabic text classification. A feature selection method using the Hybrid GWO with PSO was proposed in [56]. The HBGWO approach combines the PSO's search capabilities for the best solutions with the BGWO's local search capabilities.

2.2. The ABC Algorithm as Feature Selection Method

The ABC algorithm proposed by [41] has been applied in various areas of research, such as optimization problems, data clustering, and Feature Selection (FS). In the FS method, ABC is used as a wrapper method, where the feature subset is selected by evaluating the performance of a particular classifier on the selected subset of features. The ABC algorithm starts with an initial random feature subsets population, which is evaluated through a fitness function. The fitness function calculates the quality of the feature subset by evaluating the classification accuracy of the selected features using a classifier. The ABC algorithm uses a combination of local and global searches to explore the search space efficiently. The employed bees perform the local search by searching in the neighborhood of their current solution to find a more effective solution. The global search is performed by the onlooker bees, which select a food source with a high fitness value to explore. The scout bees search for new food sources by randomly selecting a solution from the search space. When reaching a stopping threshold, the algorithm ends, such as a maximum number of iterations or a desired fitness value. The ABC algorithm has shown promising results in feature selection for text classification tasks, and it is an effective technique to identify an appropriate subset of features that can increase classification performance while reducing feature space dimensionality [35, 40, 54]. The ABC algorithm can be summarized into the following steps:

1. Initialization: set the population size (number of employed bees) to the same number as the number of food sources. The locations of the food sources are randomly initialized.
2. Employed bees phase: trying to improve the quality of the food source it is assigned to. The employed bee selects a neighbor food source, modifies its position, and evaluates its quality. The employed bees consider the new food source if its quality is better than the original one.
3. Onlooker bees phase: a food source is chosen by the onlooker bee based on its quality. The better quality food source will have a high opportunity to be selected. The quality of the selected food source will be improved by the onlooker bee in the same way as the employed bees do.
4. Scout bees phase: If a food source has not been improved either by the employed or onlooker bees in a specific number of iterations, it is discarded and a new randomly initialized one will be considered.
5. Memorization: The best food source found by the employed and onlooker bees is memorized and reported as a solution.
6. Termination: when a stopping criterion, such as a maximum number of iterations or the finding of a realistic solution, is satisfied, the algorithm terminates
7. Optimization: the algorithm can be further optimized by adjusting parameters such as the population size, the number of iterations, and the selection mechanism for onlooker bees.

3. Proposed Arabic Text Classification

Text preprocessing, feature extraction, feature selection, and finally the construction of the text classification model using a machine learning classifier are the main phases of the text classification process.

Text preprocessing involves cleaning and preparing the text data by removing unnecessary characters, punctuation, stop words, and stemming. This step helps in reducing the noise in the data and making it easier to work with.

Raw textual data is transformed through feature extraction into a set of numerical features that may be fed into a machine-learning algorithm. There are several techniques for feature extraction in text classification, including bag-of-words, n-grams, and word embedding.

Feature selection involves selecting the most relevant features or words that can distinguish between the different classes in the text data. This step helps in reducing the dimensionality of the data and improving the accuracy and efficiency of the classification model.

Finally, the text classification model is constructed using a machine learning classifier such as NB, SVM, or Random Forest on the preprocessed and feature-selected data.

3.1. Proposed FS Method: Chi-Square and ABC Algorithm as FS Method

This study proposed an FS method based on the chi-square and the ABC algorithm. The chi-square is a statistical method used in feature selection to determine the level of association or dependency between a word (feature) and a predefined class. Features with higher chi-square values are considered more relevant to the classification task and are kept for further analysis, while those with lower values are discarded [15]. The chi-square is an effective way to decrease the number of features and eliminate irrelevant or redundant ones, but it may not be able to capture complex relationships between features and classes. Therefore, it is often used in combination with other feature selection methods, such as wrapper methods based on evolutionary algorithms like the ABC algorithm. The general phases of the proposed FS method are as follows:

1. Preprocessing: the Arabic text is preprocessed to remove stop words, and punctuation marks.
2. Feature extraction: The text is represented in a numerical format using the bag-of-words model. Each unique word in the corpus is represented as a feature.
3. Chi-square FS: the chi-square statistic is calculated for each feature to measure its independence with the class labels. Features with a higher chi-square value are selected as more relevant to the classification task.

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad (1)$$

Where N is the overall number of documents, A is the number of documents containing the feature t and falling under class c, B represents how many documents containing the feature t and not falling under class c, C represents how many documents not containing the feature t and falling under class c, and D represents how many documents not containing the feature t and not falling under class c.

4. ABC feature selection: the selected features by the chi-square are fed into the ABC as potential solutions (food sources). The ABC algorithm explores the optimal/best subset of features that leads to higher classification performance. Figure 1 shows the main phases of the ABC algorithm as the FS method.

   a. Initialization: the forward search strategy is used to create initial food sources; the food source is represented as a bit vector (1 for a feature to be considered or 0 for the feature that will not be considered). A random number is generated for each feature position, and if the number is less than a specified threshold MR, the feature is considered part of the subset. The number of features to be considered is controlled by nFeatures which is a random number between 7 to 50. The following Equation shows the feature position initialization:

$$x_i=\begin{cases}1, & R_i<MR\\ x_i, & R_i\geq MR\end{cases}, \text{ and } 7\leq nFeatures\leq 50 \quad (2)$$

Evaluate the quality of each solution by applying a fitness function to measure the classification accuracy of the corresponding feature subset.

b. After the initial food sources are created, the employed bees search for better food sources by modifying the positions of the features in each food source. The modification is done based on the neighborhood of the current food source. The MR parameter determines the probability of selecting a feature in the neighborhood. The neighborhood of a food source is defined as the set of food sources obtained by flipping the value of a randomly selected feature in the current food source. The nFeatures parameter, which is a random value between 7 and 50, determines how many features will be modified in each step. This helps to prevent large jumps in the number of selected features in each step and ensures a smoother convergence towards the optimal subset of features. Equation 2 is used to modify the position of a feature in a food source. For each position in the current food source, a random number is generated between 0 and 1. If the random number is less than the MR value, the value of the position is flipped, i.e., if the value is 0, it becomes 1, and vice versa. Otherwise, the value of the position remains the same. This process is repeated for each position in the food source, and the resulting food source is added to the neighborhood.

c. After the neighbors are generated, the subset of features for each neighbor is assessed by the classifier. The classifier then calculates the accuracy of each neighbor, which is used as the fitness value for that neighbor. This fitness value represents how good the neighbor solution is in terms of accuracy, and it will be compared to the fitness of the current food source.

d. The employed bees compare the fitness of the new food source (which is generated by modifying the current food source) with the fitness of the current food source. If the new food source has better fitness, it replaces the current food source, and the limit counter is reset to zero. However, if the new food source has a worse fitness, the limit counter is incremented. If the limit counter reaches a MAX_END limit, the current food source is considered exhausted, and a new scout bee is generated to explore a new food source randomly.

e. Onlooker bees gather information about the quality of food sources that have been visited by employed bees. Depending on that information, a food source with better fitness is chosen by the onlooker bees for exploration, and they become employed bees that search for that food source and perform step b. This step is important for maintaining diversity in the swarm, as onlooker bees are not limited to exploring the same food sources as employed bees.

f. The best food source found so far will be memorized. This best food source is the one with the highest fitness value among all food sources visited by the employed and onlooker bees. The memorization of the best food source is important because it represents the current best solution found by the algorithm. This best solution may be used as the final solution or as the initial solution for a subsequent run of the algorithm.

g. When a food source is abandoned, it means that it has reached its limit and cannot be improved anymore. In this case, a randomly newly created food source is assigned to a scout bee, which becomes an employed bee to search for a new solution. The scout bee randomly generates a new solution and evaluates it, if it has better quality than the abandoned food source, it becomes a new food source, otherwise, it continues to search for a better solution. The algorithm continues to execute steps b to g until the termination criteria are met.

Naïve Bases classifier and Support Vector Machine were used in the study as a fitness function.

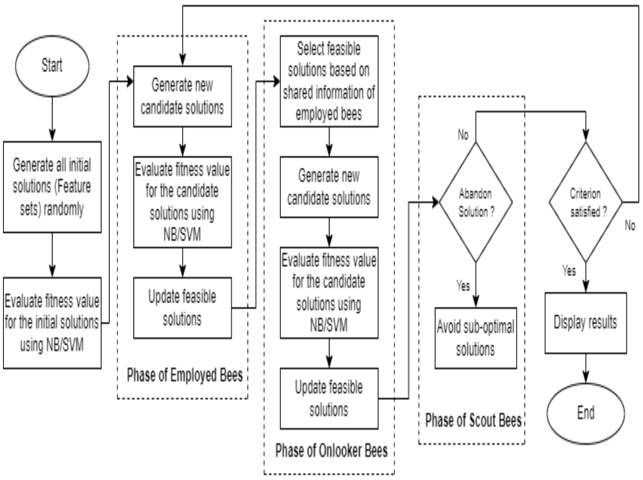
Start
Generate all initial solutions (Feature sets) randomly
Evaluate fitness value for the initial solutions using NB/SVM
Generate new candidate solutions
Evaluate fitness value for the candidate solutions using NB/SVM
Update feasible solutions
Phase of Employed Bees
Select feasible solutions based on shared information of employed bees
Generate new candidate solutions
Evaluate fitness value for the candidate solutions using NB/SVM
Update feasible solutions
Phase of Onlooker Bees
No
Abandon Solution ?
Yes
Avoid sub-optimal solutions
Phase of Scout Bees
Criterion satisfied ?
No
Yes
Display results
End

Figure 1. The main steps of the ABC feature selection method.

- **Naive Bayes classifier (NB)**

It is a probabilistic algorithm used for classification tasks. The basic assumption of NB is that the probability of each feature value is conditionally independent of other feature values given the class label. In other words, NB assumes that the absence or presence of a particular feature is independent of the presence or absence of any other feature in the dataset. To use NB for classification, a training set with labeled examples is required. The classifier learns the probability distribution of each feature given each class label in the training set. The algorithm calculates the prior probability of each class based on the frequency of that class in the training set. Then, for each feature in the test set, the conditional probability of the feature given in each class is calculated using Bayes' theorem. Finally, the highest probability class is chosen as the predicted class for the test example. One of the advantages of NB is its simplicity and efficiency, especially for high-dimensional datasets with a large number of features. However, the assumption of conditional independence may not hold in many real-world applications, which can lead to the suboptimal performance of the classifier [22].

- **Support Vector Machine (SVM)**

It is a powerful machine-learning technique that is widely used for classification and regression tasks. SVM depends on the concepts of minimizing structural risk by mapping input points into a space of higher dimension where a maximal breaking up hyperplane is found. SVM uses a linear dividing or separating method to come up with a classification model that is utilized in the classification of unseen instances. SVM can use kernel methods, such as polynomial, RBF, and sigmoid kernels, to deal with data that is not linearly separable. In text classification tasks, SVM is often used with the linear kernel method, which is suitable for datasets with a large number of features [22].

- **Data Division and Dataset**

Stratified K-fold cross-validation assures having the same distribution of class in each of the K parts of the dataset will be the same in all K parts of the dataset. When we have an imbalanced dataset, Stratified K-fold cross-validation is used to avoid the overfitting problem and it is used in this study [10, 43, 58].

For Arabic text categorization, many datasets were utilized. One of the popular datasets is the BBC dataset. It is free and open to the public, and it has a sufficient quantity of documents for categorization. As a result, it has been frequently employed in past and contemporary studies. It contains 7 classes and 4,763 text documents. The BBC Arabic dataset is utilized in this study to assess the proposed system's categorization performance. It is a big dataset that is non-linearly separable. The distribution of documents per category in the BBC Arabic dataset is shown in Table 1.

Table 1. The number of documents in each Arabic BBC dataset category.

Category	# of documents
Middle East News	2356
World News	1489
Business and Economics	296
Sports	219
Magazine	49
Science and Technology	232
Collection (Art & Culture)	122
Total	4763

- **Performance Evaluation**

It is possible to assess the performance of text classification using a variety of performance metrics. One of the accurate methods for evaluating a test's performance is to use the F1 measure, which is the harmonic mean of accuracy (p) and recall (r). Weighted-F1 score is the score for each class Ci weighted by the number of documents Ni from that class.

$$\text{Precision (P)}=\frac{TP}{TP+FP} \tag{3}$$

$$\text{Recall (R)}= \frac{TP}{TP+FN} \tag{4}$$

$$\text{F1-measure} = \frac{2*P*R}{P+R} \tag{5}$$

$$\text{Weighted F1}= \frac{\sum_{i=0}^{n} F1(C_i)* N_i}{D} \tag{6}$$

4. Experiment Setup and Results

The experiments were conducted using a PC with Core i7 2.7GHz, 16G Ram. The ABC algorithm was implemented in Java language as a feature selection method. Weka 3.9.5 machine-learning software was used for the preprocessing, and classification phases. NB, SVM, and decision tree classifiers were used to evaluate the selected subset of features. Weighted F1-Measure was used as a performance measure.

4.1. Parameters of ABC Algorithm as an FS Method

In the ABC algorithm, the forward search strategy is used. To determine the suitable parameters of the ABC algorithm, experiments were conducted to find out the most suitable parameter values. The Naïve Bayes classifier was used as a fitness function in all experiments. Experiments focused on:

- The number of iterations (Generations).
- Swarm size.
- MAXLIMIT variable value
- MR value.

4.1.1. The Number of Iterations (Generations)

To investigate the effect of the number of iterations (generations) on the performance of the ABC algorithm. Ten sceneries were studied 10, 20, 40, 60, 80, and 100 generations with swarm size=20, MAXLIMIT=3, and MR=0.1. According to the results shown in Figure 2, it appears that there is a positive correlation between the number of generations and the fitness value obtained by the algorithm. Specifically, the lowest fitness value was obtained after 10 generations, while the best fitness value was obtained after 100 generations.

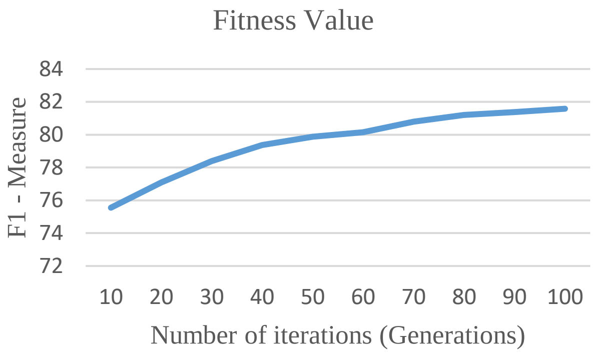

Figure 2. Results of using different generations (iterations) for ABC.

4.1.2. Swarm Size

Due to a large number of features, it was not feasible to have a swarm size equal to the number of features and initialize each food source with a single feature, so different swarm sizes (20, 40, 60, 80, and 100) were tested to find the optimal swarm size for the proposed model. According to the results shown in Figure 3, it seems that increasing the swarm size led to an improvement in the fitness value obtained by the algorithm. Specifically, the largest swarm size of 100 produced the best fitness value of 82.52, while the smallest swarm size of 40 produced the lowest fitness value of 81.51. This suggests that increasing the swarm size can lead to more diversity in the food sources, which in turn can result in better fitness values. It is important to keep in mind that increasing the swarm size might increase the computational cost and the amount of time needed for the algorithm to find a solution. Therefore, it is recommended to balance the benefits of increasing the swarm size with the computational cost and time constraints of the problem being solved.

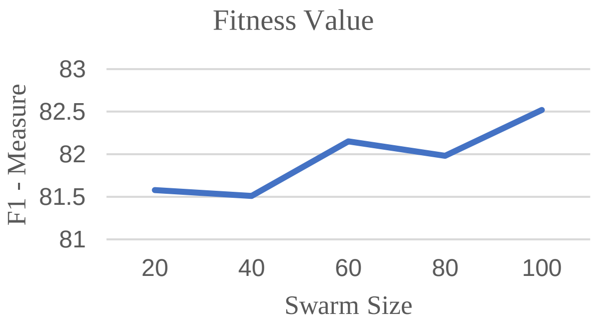

Figure 3. Testing several values of swarm size to find out the best value.

After determining the swarm size = 100, the MAXLIMIT values and MR values will be investigated.

4.1.3. MAX LIMIT Value

The experiment involved testing different values of the MAXLIMIT parameter in a swarm algorithm to determine its effect on the fitness values obtained by the algorithm. Specifically, the values tested were 3, 5, 7, and 10, while the swarm size was varied from 20 to 100, with a Mutation Rate (MR) of 0.1 and 100 iterations. According to the results shown in Figure 4, it seems that the optimal value of the MAXLIMIT parameter depends on the swarm size. For example, when the swarm size was 40, the lowest fitness value was obtained with a MAXLIMIT value of 3. However, when the swarm size was 100, the best fitness value was obtained with a MAXLIMIT value of 10, which allowed the food source to be considered abandoned only after 10 iterations. These results suggest that the optimal value of the MAXLIMIT parameter may depend on the swarm size and the problem is solved. It is important to keep in mind that increasing the MAXLIMIT value might increase the computing cost and time needed for the algorithm to come up with a solution.

4.1.4. MR Value

The experiment tested various mutation rate (MR) parameter values in a swarm algorithm to examine how it affected the fitness values the algorithm obtained. In particular, eight potential MR values (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8) were examined with a fixed swarm size of 20, a MAXLIMIT value of 10, and a 100 iteration number. Figure 5's findings indicate that increasing the MR value had a detrimental effect on the fitness values that the algorithm obtained. In particular, an MR value of 0.1 produced the best fitness value, whereas increasing the MR value resulted in lower fitness values.

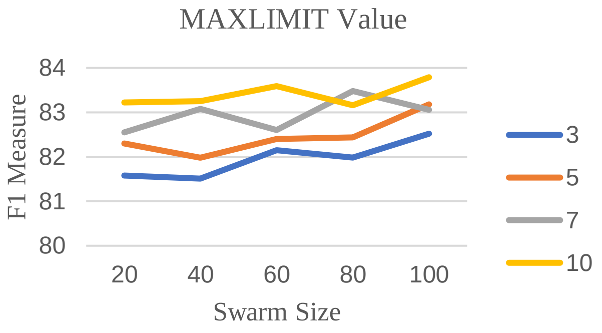

Figure 4. Testing several values of MAXLIMT value to come out with the optimal value.

Based on the several experiments that were conducted to find out the suitable ABC algorithm parameters, the following parameters were used in the proposed model:

- The number of iterations (generations)=100.
- Swarm Size=100.

- MAXLIMIT value=10.
- MR value=0.1

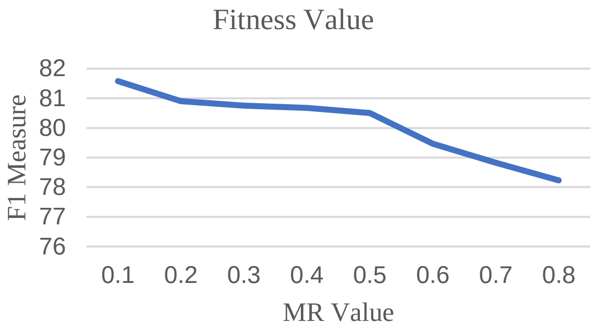

Figure 5. Testing several values of MR value to find out the best value.

4.2. Result of the Proposed FS Method using Naïve Bayes as Fitness Function

The results of applying the chi-square, ABC algorithm as FS method with the Naive Bayes classifier used as a fitness function, and the proposed hybrid approach which combines the chi-square and the ABC algorithm on the dataset are summarized in Table 2. The table shows that the initial dataset has 24253 features. The number of features, after applying the Chi-square, was reduced by approximately 72% to 6817 features. The ABC algorithm was then applied to the original dataset, resulting in a reduction of approximately 82% to 4378 features. Finally, the proposed hybrid approach was used to further reduce the feature set, resulting in a total of 2563 features, which represents a reduction of approximately 89.5% compared to the original dataset. These results suggest that the proposed hybrid approach, which combines the strengths of both filter and wrapper feature selection methods, was able to achieve the highest level of feature reduction, resulting in a more efficient and effective feature subset for the classification task.

Table 2. The number of features in the original dataset and after applying FS methods.

Method	# of Features
Original Dataset	24253
Chi-Square FS Method	6817
ABC algorithm FS Method (NB as fitness function)	4378
Chi-ABC FS Method (NB as fitness function)	2563

Figure 6 displays the results of comparing the weighted F1-measure using the selected features for different feature selection methods, including the Chi-square method, the ABC algorithm, and the proposed hybrid approach. The results show that the proposed hybrid approach achieved higher weighted F1-measure values compared to the other feature selection methods for all classifiers. This is because the proposed approach first used the Chi-square method as a filter selection method to decrease the feature space dimensionality, which removed irrelevant and redundant features. Then, the ABC algorithm was applied to the reduced subset of features to find the optimal set of features that had the minimum number of features and the highest weighted F1-measure value. Moreover, the proposed hybrid approach was able to achieve this optimal set of features in less computation time compared to using the ABC algorithm alone as an FS method. This is because the Chi-square method decreased the dimensionality of feature space, making the search space for the ABC algorithm smaller and more manageable. Overall, these results suggest that the proposed hybrid approach is a more efficient and effective method for feature selection compared to using either the Chi-square method or the ABC algorithm alone.

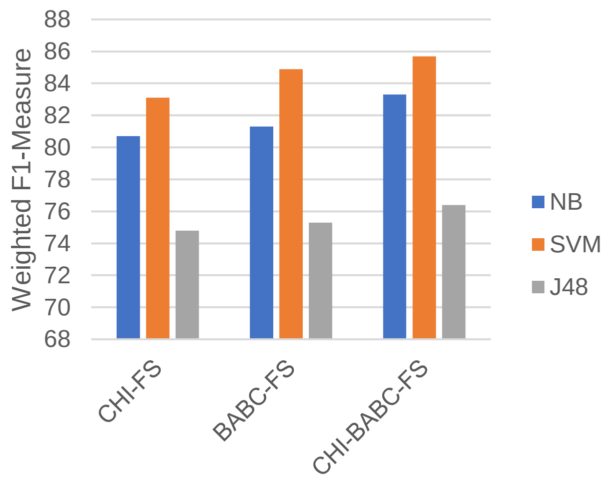

Figure 6. Weighted F1-measure for the proposed method (NB as fitness function) and other methods.

4.3. Result of the BABC proposed FS method using Support Vector Machine as fitness function

This section presents the result of using the proposed ABC FS method-based SVM as a fitness function. The same parameters for the ABC algorithm were used:

- The number of iterations (generations)=100.
- Swarm Size=100.
- MAXLIMIT value=10.
- MR value=0.1

Table 3 summarizes the results of applying three different feature selection methods on the original dataset, including the Chi-square, the ABC algorithm as an FS method with the SVM classifier used as a fitness function, and the proposed hybrid approach which combines both of them. The table shows that the original dataset had a total of 24253 features. After applying the Chi-square, the number of features was reduced by approximately 72% to 6817 features. The ABC algorithm was then applied to the original dataset, resulting in a reduction of approximately 84.4 % to 3789 features. Finally, the proposed hybrid approach was used to further reduce the feature set, resulting in a total of 1476 features, which represents a reduction of approximately 94% compared to the original dataset. These results suggest that the proposed hybrid approach, which combines the strengths of both filter and wrapper feature selection methods, was able to achieve the highest level of feature reduction, resulting in a more efficient and effective feature subset for the classification task.

It is important to note that reducing the number of features can have a significant impact on the performance of the classification model, as it can help to reduce overfitting and improve the model's generalization ability. However, it is also important to strike a balance between the number of features and the classification performance, as using too few features can result in a loss of important information and reduce classification accuracy.

Table 3. The number of features in the original dataset and after applying fs methods.

Method	# of Features
Original Dataset	24253
Chi-Square FS Method	6817
ABC algorithm FS Method (SVM as fitness function)	3789
Chi-ABC FS Method (SVM as fitness function)	1476

Finally, the number of features selected by the proposed FS, when SVM was used as a fitness function, was less than the one achieved when NB was used as a fitness function.

Figure 7 displays the results of comparing the weighted F1-measure using the selected features for different feature selection methods, including the Chi-square method, the ABC algorithm, and the proposed hybrid approach. The results show that the proposed hybrid approach achieved higher weighted F1-measure values compared to the other feature selection methods for all classifiers. Overall, these results suggest that the proposed hybrid approach is a more efficient and effective method for feature selection compared to using either the Chi-square method or the ABC algorithm alone.

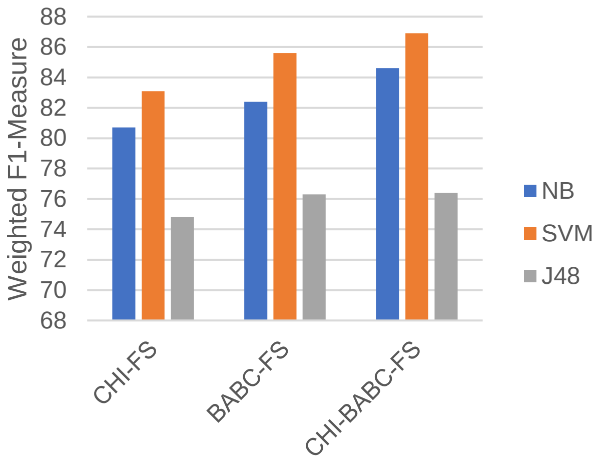

Figure 7. Weighted F1-measure for the proposed method (SVM as fitness function) and other methods.

Based on Figure 8, That is an interesting observation that the SVM algorithm used as a fitness function in the ABC algorithm was able to identify a more optimal set of features compared to the Naive Bayes algorithm. This could be because SVM, compared to Naive Bayes, is a more sophisticated and effective algorithm and may be more suited to detecting relevant features in the dataset.

Based on the findings of this study, it is possible to conclude that the SVM classifier is the most efficient for classifying Arabic text. This is because SVM outperformed all other classifiers in this study in terms of the F1-measure value. In addition, the Naive Bayes classifier successfully classified Arabic text, attaining a respectably high F1-measure value. However, among the classifiers utilized in this study, the C4.5 classifier performed the worst, obtaining the lowest F1-measure value. This could be a result of the difficulty in understanding the Arabic language and the difficulties decision tree-based classifiers have in managing large datasets.

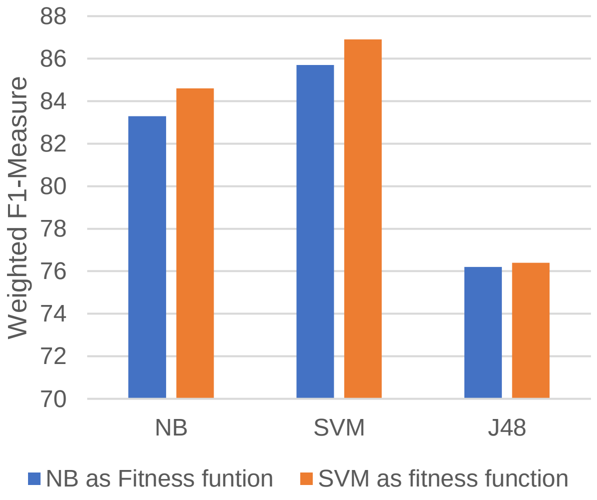

Figure 8. Weighted F1-measure for the proposed methods (NB and SVM as fitness function).

4.4. Comparison of The Proposed Feature Selection Method with Other Feature Selection Methods

A comparison study was conducted with three state-of-the-art wrapper feature selection methods: Particle Swarm Optimization algorithm, Genetic algorithm, and Ant Colony Optimization algorithm, which are implemented in Weka using an SVM classifier. The main objective of this experiment is to compare the performance of the proposed method with other wrapper FS methods.

Figure 9 presents weighted F1-measure values for PSO, ACO, and GA algorithms as wrapper feature selection methods. The results showed that the proposed FS method has outperformed both PSO and ACO methods in Arabic text classification, achieving an F1-measure of 86.9%. This shows that the proposed method can be a valuable addition to the existing FS methods and can be used to effectively reduce the feature space dimensionality for Arabic text classification tasks, leading to improved classification performance.

5. Conclusions

FS method based on Chi-square and ABC was proposed as the FS method to be used in Arabic Text Classification. Two well-known classifiers Naïve Bayes and SVM were used as a fitness function in the proposed method, and three popular machine learning algorithms NB, SVM, and C4.5 were used to evaluate the proposed feature selection method. A set of experiments were conducted to find out the suitable ABC algorithm parameters.

In the first version of the proposed feature selection method, the Naïve Bayes classifier was used as a fitness function to select the optimal/best subset of the features that could improve the classification accuracy and reduce the time complexity. Naïve Bayes, Support Vector Machines, and C4.5 were used to evaluate the proposed subset of features. The results came up that the proposed method was efficient.

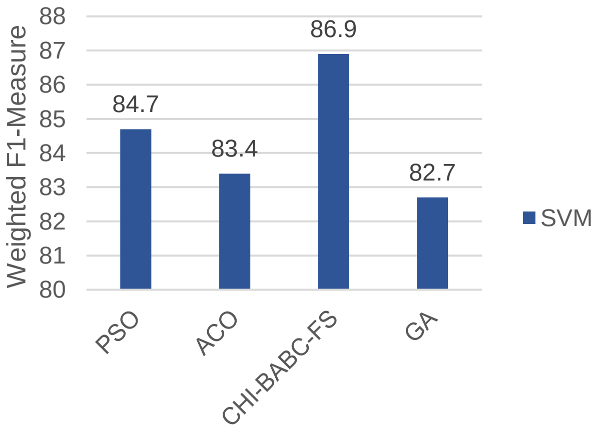

Figure 9. Weighted F1-measure for the proposed method (SVM as fitness function) and other wrapper FS methods.

In the second version of the proposed feature selection method, SVM as a fitness function was used to come up with the optimal/best subset of the features that could improve the classification performance and at the same time decrease the time complexity of the classification task. After that NB, SVM, and C4.5 were used to evaluate the selected subset feature. The experiments showed:

1. Both versions of the proposed method were capable of recognizing the subset of features that decrease the feature vector dimensionality and enhanced the classification performance.
2. The first approach of the proposed method using Naïve Bayes as a fitness function was able to reduce 89.5% of the original feature set.
3. Using SVM as a fitness function in the proposed method has a better classification performance and can reduce 94% of the original feature set.
4. The superior results for SVM compared with the others.
5. In conclusion, the results of this research suggest that SVM and Naïve Bayes classifiers are suitable for Arabic text classification tasks, while decision tree-based classifiers such as C4.5 may not be the best choice.

References

[1] Adel A., Omar N., Albared M., and Al-Shabi A., "Feature Selection Method Based on Statistics of Compound Words for Arabic Text Classification," *The International Arab Journal of Information Technology*, vol. 16, no. 2, pp. 178-185, 2019.

[2] Adel A., Omar N., and Al-Shabi A., "A Comparative Study of Combined Feature Selection Methods for Arabic Text Classification," *Journal of Computer Science*, vol. 10, no. 11, pp. 2232-2239, 2014. doi:10.3844/jcssp.2014.2232.2239.

[3] Aghdam M. and Heidari S., "Feature Selection Using Particle Swarm Optimization in Text Categorization," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 5, no. 4, pp. 231-238, 2015. DOI: https://doi.org/10.1515/jaiscr-2015-0031

[4] Al-Dulaimi A. and Okkalioglu M., "Efficient Arabic Text Classification Using Feature Selection Techniques and Genetic Algorithm," *in Proceedings of the 3rd International Informatics and Software Engineering Conference (IISEC)*, Ankara, pp. 1-6, 2022.

[5] Alhaj Y.A., Dahou A., Al-qaness M., Abualigah L., Abbasi A., Almaweri N., Abd Elaziz M., Damaševičius R., "A Novel Text Classification Technique Using Improved Particle Swarm Optimization: A Case Study of Arabic Language," *Future Internet*, vol. 14, no. 7, pp. 194, 2022. https://doi.org/10.3390/fi14070194.

[6] Alhutaish R. and Omar N., "Arabic Text Classification Using K-Nearest Neighbour Algorithm," *The International Arab Journal of Information Technology*, vol. 12, no. 2, pp. 190-195, 2015.

[7] Alomari O.A., Elnagar A., Afyouni I. Shahin I., Bou Nassif A., Hashem I., and Tubishat M., "Hybrid Feature Selection Based on Principal Component Analysis and Grey Wolf Optimizer Algorithm for Arabic News Article Classification," *IEEE Access*, vol. 10, pp. 121816-121830, 2022. DOI: 10.1109/ACCESS.2022.3222516

[8] Alshaer H.N., Otair M.A., Abualigah L., Alshinwan M., and Khasawneh A. M., "Feature Selection Method Using Improved CHI Square on Arabic Text Classifiers: Analysis and Application," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 10373-10390, 2021.

[9] Al-Thubaity A., Abanumay N., Al-Jerayyed S., Alrukban A., and Mannaa Z., "The Effect of Combining Different Feature Selection Methods on Arabic Text Classification," *in Proceedings of 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Honolulu, pp. 211-216, 2013. DOI: 10.1109/SNPD.2013.89.

[10] Arlot S. and Celisse A., "A Survey of Cross-Validation Procedures for Model Selection," *Statistics Surveys*, vol. 4, pp. 40-79, 2010. DOI: 10.1214/09-SS054

[11] Ayadi R., Maraoui M., and Zrigui M., "Latent Topic Model for Indexing Arabic Documents," *International Journal of Information Retrieval Research*, vol. 4, no. 2, pp. 57-72, 2014. DOI: 10.4018/ijirr.2014040104.

[12] Ayadi R., Maraoui M., and Zrigui M., "LDA and LSI as A Dimensionality Reduction Method in Arabic Document Classification," *in Proceedings of International Conference on Information and Software Technologies*, Druskininkai, pp. 491-502, 2015.

[13] Bahassine S., Madani A., Al-Sarem M., and Kissi M., "Feature Selection Using an Improved Chi-Square for Arabic Text Classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 2, pp. 225-231, 2020. https://doi.org/10.1016/j.jksuci.2018.05.010.

[14] Bahassine S., Madani A., and Kissi M., "An Improved Chi-Sqaure Feature Selection for Arabic Text Classification Using Decision Tree," *in Proceedings of 11th International Conference on Intelligent Systems: Theories and Applications (SITA)*, Mohammedia, pp. 1-5, 2016.

[15] Bao L. and Zeng J., "Comparison and Analysis of the Selection Mechanism in the Artificial Bee Colony Algorithm," *in Proceedings of 9th International Conference on Hybrid Intelligent Systems*, Shenyang, pp. 411-416, 2009.

[16] Basir M.A., Yusof Y., and Saifullah M., "Optimization Of Attribute Selection Model Using Bio-Inspired Algorithms," *Journal of Information and Communication Technology*, vol. 18, no. 1, pp. 35-55, 2019.

[17] Belkebir R. and Guessoum A., "A Hybrid BSO-Chi2-SVM Approach to Arabic Text Categorization," *in Proceedings of ACS International Conference on Computer Systems and Applications (AICCSA)*, Ifrane, pp. 1-7, 2013.

[18] Chandrashekar G. and Sahin F., "A Survey on Feature Selection Methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16-28, 2014.

[19] Chantar H. and Corne D., "Feature Subset Selection for Arabic Document Categorization Using BPSO-KNN," *in Proceedings of 3rd World Congress on Nature and Biologically Inspired Computing*, Salamanca, pp. 546-551, 2011.

[20] Chantar H., Mafarja M., Alsawalqah H., Heidari A.A., Aljarah I., and Faris H., "Feature Selection Using Binary Grey Wolf Optimizer With Elite-Based Crossover for Arabic Text Classification," *Neural Computing and Applications*, vol. 32, no. 16, pp. 12201-12220, 2020.

[21] Chantar H., Tubishat M., Essgaer M., and Mirjalili S., "Hybrid Binary Dragonfly Algorithm With Simulated Annealing For Feature Selection," *SN Computer Science*, vol. 2, no. 4, pp. 1-11, 2021.

[22] Chantar H.K.H., New Techniques for Arabic Document Classification, *PhD Thesis*, Heriot-Watt University, 2013.

[23] Duwairi R.M., "Machine Learning for Arabic Text Categorization," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 8, pp. 1005-1010, 2006. https://doi.org/10.1002/asi.20360

[24] El-Hajj W. and Hajj H., "An Optimal Approach for Text Feature Selection," *Computer Speech and Language*, vol. 74, 2022. https://doi.org/10.1016/j.csl.2022.101364.

[25] Elhassan R. and Ali M., "The Impact of Feature Selection Methods for Classifying Arabic Texts," *in Proceedings of the 2nd International Conference on Computer Applications and Information Security*, Riyadh, pp. 1-6, 2019. DOI: 10.1109/CAIS.2019.8769526.

[26] Elnahas A., Elfishawy N., Nour M., and Tolba M., "Machine Learning and Feature Selection Approaches for Categorizing Arabic Text: Analysis, Comparison, and Proposal," *The Egyptian Journal of Language Engineering*, vol. 7, no. 2, pp. 1-19, 2020.

[27] Ghareb A.S., Bakar A.A., and Hamdan A.R., "Hybrid Feature Selection Based on Enhanced Genetic Algorithm for Text Categorization," *Expert Systems with Applications*, vol. 49, pp. 31-47, 2016.

https://doi.org/10.1016/j.eswa.2015.12.004.

[28] Guru D., Ali M., Suhil M., and Hazman M., "A Study of Applying Different Term Weighting Schemes on Arabic Text Classification," *in Proceedings of Data Analytics and Learning*, Singapore, pp. 293-305, 2019.

[29] Habeeb A., Otair M., Abualigah L., Alsoud A.R., Abd Elminaam D., Abu Zitar R., Ezugwu A., and Jia H., "Arabic Text Classification Using Modified Artificial Bee Colony Algorithm for Sentiment Analysis: The Case of Jordanian Dialect," *Classification Applications with Deep Learning and Machine Learning Technologies*, pp. 243-288, 2022.

[30] Hadni M. and Hassane H., "A New Metaheuristic Approach Based Feature Selection for Arabic Text Categorization," *in Proceedings of the 23th International Arab Conference on Information Technology (ACIT)*, Abu Dhabi, pp. 1-7, 2022. DOI: 10.1109/ACIT57182.2022.9994102

[31] Hadni M. and Hjiaj H., "An Improved Chaotic Sine Cosine Firefly Algorithm for Arabic Feature Selection," *in Proceedings of International Conference on Big Data and Internet of Things*, Tangier, pp. 84-94, 2022.

[32] Haralambous Y., Elidrissi Y., and Lenca P., "Arabic Language Text Classification Using Dependency Syntax-Based Feature Selection," *arXiv Prepr. arXiv1410.4863*, 2014.

[33] Harrag F., El-Qawasmah E., and Al-Salman A., "Comparing Dimension Reduction Techniques for Arabic Text Classification Using BPNN Algorithm," *in Proceedings of 1st International Conference on Integrated Intelligent Computing*, Bangalore, pp. 6-11, 2010.

[34] Harrag F., El-Qawasmeh E., and Pichappan P., "Improving Arabic Text Categorization Using Decision Trees," *in Proceedings of 1st International Conference on Networked Digital Technologies*, Ostrava, pp. 110-115, 2009.

[35] Hijazi M., Zeki A., and Ismail A., "Arabic Text Classification Using Hybrid Feature Selection Method Using Chi-Square Binary Artificial Bee Colony Algorithm," *International Journal of Mathematics and Computer Science*, vol. 16, no. 1, pp. 213-228, 2021.

[36] Hijazi M., Zeki A., and Ismail A., "Arabic Text Classification: Review Study," *Journal of Engineering and Applied Science*, vol. 11, no. 3, pp. 528-536, 2016.

[37] Hijazi M., Zeki A., and Ismail A., "A Review Study on Arabic Text Classification," *in Proceedings of the 23th International Arab Conference on Information Technology*, Abu Dhabi, pp. 1-13, 2022.

[38] Hijazi M., Zeki A., and Ismail A., "Arabic Text Classification: A Review Study on Feature Selection Methods," *in Proceedings of the 22nd International Arab Conference on Information Technology*, Muscat, pp. 1-6, 2021.

[39] Jia D., Duan X., and Khan M., "Binary Artificial Bee Colony Optimization Using Bitwise Operation," *Computers and Industrial Engineering*, vol. 76, pp. 360-365, 2014.

[40] Karaboga D. and Akay B., "A Survey: Algorithms Simulating Bee Swarm Intelligence," *Artificial Intelligence Review*, vol. 31, no. 1-4, pp. 61-85, 2009.

[41] Karaboga D., "An Idea Based on Honey Bee Swarm for Numerical Optimization," Technical report-tr06, 2005.

[42] Karaboga D., Gorkemli B., Ozturk C., and Karaboga N., "A Comprehensive Survey: Artificial Bee Colony (ABC) Algorithm and Applications," *Artificial Intelligence Review*, vol. 42, no. 1, pp. 21-57, 2014.

[43] Khorsheed M. and Al-Thubaity A., "Comparative Evaluation of Text Classification Techniques Using A Large Diverse Arabic Dataset," *Language Resources and Evaluation*, vol. 47, no. 2, pp. 513-538, 2013.

[44] Marie-Sainte S. and Alalyani N., "Firefly Algorithm Based Feature Selection for Arabic Text Classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 3, pp. 320-328, 2020. https://doi.org/10.1016/j.jksuci.2018.06.004.

[45] Meena M.J., Chandran K.R., Karthik A., and Samuel A.V., "An Enhanced ACO Algorithm to Select Features for Text Categorization and its Parallelization," *Expert Systems with Applications*, vol. 39, no. 5, pp. 5861-5871, 2012. https://doi.org/10.1016/j.eswa.2011.11.081.

[46] Mesleh A. and Kanaan G., "Arabic Text Categorization System-Using Ant Colony Optimization-Based Feature Selection," *in Proceedings of 3ed International Conference on Software and Data Technologies*, Porto, pp. 384-387, 2008. DOI: 10.5220/0001892803840387.

[47] Moh'd Mesleh A., "Feature Sub-Set Selection Metrics for Arabic Text Classification," Pattern Recognition Letters, vol. 32, no. 14, pp. 1922-1929, 2011. https://doi.org/10.1016/j.patrec.2011.07.010.

[48] Mohammad A., "Comparing Two Feature Selections Methods (Information Gain and Gain Ratio) on Three Different Classification Algorithms Using Arabic Dataset," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 6, pp. 1561-1569, 2018.

[49] Mosa M., "Feature Selection Based on ACO and Knowledge Graph for Arabic Text Classification," *Journal of Experimental and Theoretical Artificial Intelligence*, pp. 1-18, 2022. DOI: 10.1080/0952813X.2022.2125588.

[50] Naji H., Ashour W., and Al Hanjouri M., "Text

Classification for Arabic Words Using BPSO/REP-Tree," *International Journal of Computational Linguistics Research*, vol. 9, no. 1, 2018.

[51] Prasartvit T., Kaewkamnerdpong B., and Achalakul T., "Dimensional Reduction Based on Artificial Bee Colony," *in Proceedings of 7th International Conference on Intelligent Computing*, Zhengzhou, pp. 168-175, 2012.

[52] Rahab H., Haouassi H., Souidi M., Bakhouche A., Mahdaoui R., and Bekhouche M., "A Modified Binary Rat Swarm Optimization Algorithm for Feature Selection in Arabic Sentiment Analysis," *Arabian Journal for Science and Engineering*, pp. 1-28, 2022.

[53] Saad E., Awadalla M., and Alajmi A., "Dewy Index Based Arabic Document Classification with Synonyms Merge Feature Reduction," *International Journal of Computer Science Issues*, vol. 8, no. 6, pp. 46-54, 2011.

[54] Schiezaro M. and Pedrini H., "Data Feature Selection Based on Artificial Bee Colony Algorithm," *EURASIP Journal on Image and Video processing*, pp. 47, 2013.

[55] Shunmugapriya P. and Kanmani S., "A Hybrid Algorithm Using Ant and Bee Colony Optimization for Feature Selection and Classification (AC-ABC Hybrid)," *Swarm and Evolutionary Computation*, vol. 36, pp. 27-36, 2017. https://doi.org/10.1016/j.swevo.2017.04.002.

[56] Subkhi M., Fatichah C., and Arifin A., "Feature Selection Using Hybrid Binary Grey Wolf Optimizer for Arabic Text Classification," *IPTEK The Journal for Technology and Science*, vol. 33, no. 2, pp. 105-116, 2022.

[57] Syiam M., Fayed Z., and Habib M., "An Intelligent System for Arabic Text Categorization," *International Journal of Intelligent Computing and Information Sciences*, vol. 6, no. 1, pp. 1-19, 2006.

[58] Yousif S., Samawi V., Elkabani I., and Zantout R., "The Effect of Combining Different Semantic Relations on Arabic Text Classification," *World of Computer Science and Information Technology Journal*, vol. 5, no. 1, pp. 12-118, 2015.

[59] Zahran B., Kanaan G., and Sciences F., "Text Feature Selection using Particle Swarm Optimization Algorithm," *World Applied Sciences Journal*, vol. 7, pp. 69-74, 2009.

Musab Mustafa Hijaz is an instructor Department of Computer Science and Software Engineering at College of Engineering at Al Ain University, UAE. His research interest including Data Mining, Text Mining, Artificial Intelligence, Swarm Intelligence, and Natural Language processing.

Akram Zeki is a Professor at Kulliyyah (Faculty) of Information and Communication Technology (KICT) at International Islamic University Malaysia (IIUM). He is a supervisor for more than 25 master and PhD students; he is leading few research grants under the university (International Islamic University Malaysia) or under national grants and international grants. His research interest including Information System Applications, Information Retrieval, Multimedia and Islamic Applications in ICT.

Amelia Ismail is Associate Professor at Kulliyyah (Faculty) of Information and Communication Technology (KICT) at International Islamic University Malaysia (IIUM). She is a supervisor for many master and PhD students. Her research interest including ICT ~ Information, Computer and Communications Technology (ICT) ~ Artificial Intelligence ~ Other Artificial Intelligence n.e.c. - Artificial Immune Systems, Swarm Intelligence and Swarm Robotics Systems, Modelling and simulation of complex system.