New Model of Feature Selection based Chaotic Firefly Algorithm for Arabic Text Categorization

Meryeme Hadni
Computer Science, LAMIGEP, EMSI, Morocco
meryemehadni@gmail.com

Hassane Hjiaj
Department of Mathematics, Faculty of Science, Morocco
hjiajhassane@yahoo.fr

Abstract: *The dimensionality reduction is a type of problem that appear in the most classification processes. It contains a large number of features; these features may contain unreliable data which may lead the categorization process to unwanted results. Feature selection can be used for reducing dimensionality of datasets and find interesting relevant information. In Arabic language, the number of works applies a meta-heuristic algorithm for feature selection is still limited due to the complex nature of Arabic inflectional and derivational rules as well as its intricate grammatical rules and its rich morphology. This paper proposes a new model for Arabic Feature Selection that combines the chaotic method in the Firefly Algorithm (CFA). The Chaotic Algorithm replaces the attractiveness coefficient in firefly algorithm by the outputs of chaotic application. The enhancement of the new approach involves introducing a novel search strategy which is able to obtain a good ratio between exploitation and exploration abilities of the algorithm. In terms In terms of performance, the experiments of the proposed method are tested using classifiers, namely Naive Bayes (NB), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) and three evaluation measures, including precision, recall, and F-measure. The experimental findings show that the combining of CFA and SVM classifiers outperforms other combinations in terms of precision.*

Keywords: *Chaotic method, firefly algorithm, arabic text categorization, feature selection*.

Received April 2, 2023; accepted May 10, 2023
https://doi.org/10.34028/iajit/20/3A/3

1. Introduction

Internet Natural Language Processing (NLP) is considered as a large research domain connection with artificial intelligence fields, computer science and linguistics processing. It includes several topics with different applications in the real world, such as named entity Recognition, Machine Translation, Text Categorization, and Text Summarization.

One of the supervised learning techniques used to assign a predefined category to a given text is the Text Categorization (TC). This technique is used for: sentiment analysis [16], topic warning [9], spam detection [21], and mail filtering applications [19]. Note that text classification is very necessary to reduce space and time consumption.

An essential problem in categorization using Feature Selection (FS) [14, 22] is the fact that learning algorithms are not suitable for dealing with the predominant feature space dimension. In the TC problem, the aim of feature selection is to the improvement of the computational accuracy and efficiency of models by removing repetitive and irrelevant terms from the text. It is also used to feature selection by using enough information concerning the dataset of text.

Some studies have been done on the impact of SF on Arabic TC [23]. For example, in [7], the classification of Arabic texts was obtained using the maximum entropy method and its accuracy was 0.80. Al-Harbi *et al*. [4] have used the Support Vector Machine classifier (SVM) with Chi-square and feature selection appropriate for Arabic Documents classification. Mesleh [15] deduced that: DF, TF-IDF, LSI was outperformed by Stemming and Light Stemming, some of them have been applied successfully in the Arabic text classification.

In our model, we propose a new Firefly Algorithm for feature selection obtained by using the combination of probability theory, hamming distance and categorization SVM, to find the best feature set among the characteristics of each category. By using three classifier methods and the EASC dataset, we will compare the resulting feature sets, and then we will measure the success of feature sets in classifications by using three evaluation metrics to know more about precision, recall and the f1-measure.

This paper is organized as follows: In the section 2, we recall some related works. The section 3 present a description of our feature selection method based on a chaotic firefly algorithm. Then, we assess our method and discuss the results of experiments in section 4. In the last section of this paper, we deduce the conclusion and we present some future work.

2. Related Works

In the literature, some approaches were used to build efficient classifiers using feature selection for the Arabic text.

Ahmad *et al*. [1], the authors have proposed a Feature Selection model for Sentiment Analysis using the method of Ant Colony Optimization (ACO), and to evaluate this proposed model performance, they used the k-NN technique and customer review datasets. Experimental results showed an accuracy of 0. 892. Sarac and Ozel [18], have applied Firefly Algorithm (FA) to improve the precision of document classification. This method was tested by the J48 classifier and the KB web dataset. The results proved that FA improved accuracy and minimized classification time for web documents text.

According to Ahmad *et al*. [1] have proposed a new model of feature selection using Particle Swarm Optimization (PSO). They developed a swarm intelligence technique as a scalable strategy for processed Arabic text synthesis. This method reported an accuracy of 0.67.

The new hybrid Feature Selection model which combines Ant Colony Optimization (ACO) and Trace Oriented Feature Analysis (TOFA) is developed by Alghamdi *et al*. [3]. The results obtained are satisfactory, and the hybrid model (ACO-TOFA) has generated better results than TOFA.

Larabi Marie-Sainte and Alalyani have proposed in [14] a new heuristic model for feature selection of arabic text. This method has been adapted to be applying successfully to different combinatorial issues, and they obtained better accuracy results. Suchanek *et al*. [19], have proposed two approaches: grouping schemes and relation weighting, then tested using Naive Bayes classifier. The results prove that this method outsmart than the statistical Feature Selection methods: Information Gain (IG) and Chi- Square (CHI). They have deduced that the grouping methods reduce feature dimensionality and reinforce classification precision.

Alghamdi and Selamat [2] have proposed a hybrid FS method by combining mutual information, term frequency document and FS method comprising CHI. Also, the authors have used the K- means classifier on some online Arabic newspapers. They proved that the results obtained by hybrid algorithm are better by 28%.

Zhang *et al*. [29] have introduced a new model of Feature Selection by combining (FA) with global promising solutions and Simulated Annealing (SA) enhanced local, and using chaotic-accelerated attractiveness parameters with diversion mechanisms to obtain a better result. They have used 29 classification datasets and 11 regression benchmark data to evaluate the proposed technique and they have compared it with the others methods. They conclude that the results were enhancing by these proposed FA variants.

Hadni and Hassane [10] the authors, proposed a new Firefly Algorithm to solve the feature selection problem by using probability theory and hamming distance to delete the noisy and irrelevant features. The method is validated using SVM, Naive Bayes (NB) and K-Nearest Neighbors (KNN) classifiers, the results achieve a precision value equal's to t98%.

To conclude, the intelligent methods was outperformed by the Firefly Algorithm in different domain, which encouraging us to improve this method by proposing a Modified FA that combining FA and probability theory of some classifiers in Arabic Text Categorization.

3. New Model Based Feature Selection

A significant problem In Text Classifications stems from the huge number of features. Feature selection allows us to choose the relevant words that specifying different classes in the dataset [28]. Thus, FS are the fundamental methods to reduce the dimensionality of features without affecting classification accuracy.

To solve the problem of high data dimensionality, we propose in this paper a new model of meta-heuristic FS algorithm to reduce the feature volume [14] of Arabic text categorization.

However, to deal with TC problem, our input is text, then we transform these entered texts into a vector. First, we generally start with a pre-processing phase of the text before extracting the relevant terms. This method allows us to remove stop words and normalize words by recovering their root, and finally, we transform a text document into a list of features by using Term Frequency-Inverse Document Frequency (TF-IDF) techniques.

3.1. Text Preprocessing

It is fundamental to do natural language processing for the document. Our goal is to create input text based on word extraction/removal, morphological examination and text annotation. This work covers text tokenization [6, 15], stemming [11], and part of vocal word tagging [12]. This task is composed of a few linguistic tools such as:

The tokenization step consists of detecting individual words and removing additional components. Words in the Arabic language are written without short vowels. Text documents contain white spaces, punctuation marks, and several annotations indicating font changes, subdivisions, special characters, and numbers.

Using a set of rules to normalize text is such as: replace letter (“إ أ آ”) with (“ا “) and replacing the letter (“ء ؤ”) with (“ا”.)…etc.

Part of Speech tagging is identifying the parts of speech tag of the existing Arabic term in a sentence with regard to its context. Part of Speech tagging classifies the content words. Tags include nouns, adjectives and verbs [12].

Removes stop words: these are words repeated in each document, so they are considered weak to be distinguished. Such as conjunctions, propositions, adverbs, etc., and often called function words.

Example: إذا , إلى , غالبا , فإن , مثل)

The operation of extracting the kernel word from the original text is called: Stemming, it is done by eliminating prefixes (ك ب، ف و، ،) and suffixes (هن، هم، كن ها،). The stemmer we use is described in [11] which is developed based on light stemming approaches. After the pre-processing stape in which we distinct words that are extracted from Arabic textual documents. The next step we use a new met-heuristic feature selection (CFA) method to extract the features.

3.2. Feature Extraction

Feature extraction is one of significant preprocessing techniques in data mining and text classification that computes features value in documents. Hence, efficient feature extraction techniques like Term Frequency-Inverse Document Frequency (TF-IDF) techniques are mainly utilized in term weighting.

The Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme is a popular unsupervised method for weighting terms in a text corpus. The goal of this method is to provide a weight to each term in a document that reflects its importance for describing the document in the corpus.

The Term Frequency (TF) is calculated as the number of occurrences of a term in a document divided by the total number of terms in the document. The Inverse Document Frequency (IDF) is calculated as the logarithm of the total number of documents in the corpus divided by the number of documents in which the term appears. The TF-IDF weight for a term is calculated as the product of the TF and IDF factor for that term.

$$W_{TF.(t_i)} = TF(, d_j) * \log(\frac{N}{DF(t_i)}) \tag{1}$$

Where:
N is the number of documents in the corpus.
$D(ti)$ corresponds to the frequency of documents that the term
ti appears in the collection.
$T(ti,\ dj)$: the number of occurrences of a term ti in a document dj.

3.3. A New Chaotic Firefly Algorithm (CFA)

The Feature Selection [14] is a fundamental step before categorization. In this paper, we modified a Firefly Algorithm for FS. It involves reducing dimensionality, removing irrelevant features and selecting a subset of important terms for use in the categorization process.

The Firefly Algorithm [25] is a meta-heuristics algorithm inspired by social behavior the insect named firefly. The fireflies are attracted other fireflies regardless of their gender, and move toward them. The less bright fireflies will move towards, the brighter one, and less the distance between two fireflies is mean, more brightness, if no brighter firefly, they move stochastically.

In this paper, the CFA is applied to each document and thefirefly represents one document.
The CFA algorithm is based on three principal rules:

- Firstly, the fireflies' gender is known to be unisex. The firefly can be attracted to other fireflies regardless of their gender.
- Secondly, attractiveness is proportional to distance, which means that if the distance is negligible for any two fireflies, the attractiveness of the fireflies will be high. When a firefly is close to other fireflies, it looks brighter than usual because of this attractiveness. Moreover, less glowing fireflies move towards brighter ones.
- Finally, the brightness of a firefly is the byproduct of the objective function of the problem under scrutiny.

Our New Chaotic FA (CFA) steps are presented as follows:

1. Initialization phase: The first step of the suggested algorithm is to produce an initial population that represents an ensemble of viable and feasible solutions for the problem at hand:
    - The firefly represents one document.
    - Determines the size of this document.
    - Initializes the number of fireflies to create a swarm.
    - Generates the position of the firefly randomly.
2. Discretization phase: after the initialization parameter, the discretization step is applied to selecting the represented words. The discretization step modifies the position of the firefly to a discrete position, using Equation (2).

$$\text{a.}S(x_{ij}) = \frac{1}{2(1+x_{ij})} S_0 \text{with } 0.25 < S_0 < 1 \tag{2}$$

ifS(xij)<rand then the word j is selected in the document i.

Indicating the probability of firefly's position (the ith word of document j) taking 1. a rand is an arbitrary number between 0and1.

3. Light intensity phase (I): the strength of the current firefly's location is estimated using the objective function which is calculated by adding the frequency of the firefly's position, indicating whether the word can be considered relevant. The intensity is defined as follows:

$$Intensity = \sum_{i=0}^{N} Freq(x)_{ij} \tag{3}$$

Where (x) is the frequency of the firefly's position x_{ij}.

4. The proposed attractiveness of a firefly is equivalent to the brightness. The attractiveness function is used

to move the less bright firefly I_1 to the brighter firefly I_2, using Equation (3).

$$\beta(r) = \frac{\beta_0}{1+\log(1+rm)} \text{with m>0} \tag{4}$$

$$\text{And } \beta_0 = 1 \ , r = \sum_{j=0}^{N} \left| x_{ij} - x_{ik} \right| \tag{5}$$

Where β_0 is the attractiveness of a firefly at a distance r=0, with r is the Hamming distance between any two fireflies i and k, where N is the dimension of the firefly.

The novel position of the firefly is estimated by the following Equation (6):

$$x_i = x_i + \beta(r) * (x_j - x_i) + \alpha * \varepsilon_i * \cos(\pi * rand) \tag{6}$$

$$\text{with } \varepsilon_i = \frac{1}{2}|1 - rand| \text{and } \ \alpha = \left(\left| \frac{I_i - I_j}{I_i + I_j} \right| * rand \right) \tag{7}$$

Where x_i *and* x_j are i^{th} and j^{th} fireflies in the population, rand refers to the random numbers generated between [0,1],and ε_i a vector of random values drawn from a uniform distribution or Gaussian distribution, are chosen arbitrarily in the interval [0,1].

Finally, the intensity must be update using Equation (3), and the best location is discretized to 0 and 1.

This new model of feature selection based on modified chaotic firefly algorithm with modified attractiveness function allows us to obtain better results using the "cos" property, and we obtain a better exploitation and exploration abilities for Arabic Text Categorization.

The CFA is presented in Algorithm (1):

Algorithm 1. The Chaotic Firefly Algorithm (CFA) forArabic Text Categorization

```
Input
   - A firefly is a set of words x=(($x_1$,$x_2$, ... , $x_n$))
   - Initialize a random swarm of fireflies.
   - Define the dimension of the firefly's position.
   - Discretize the firefly's positions
   - Initialize light intensity I.
   - Define maximum iterations T.
   Output: best
   solution Set t=1
   While (k<=T)
       Update the best solution found so
       far For i=1 to n do
           For j=i+1 to
             n do If Ij>
             Ii   then
                Determine distance r using (Eq.4)
                Determine attractiveness β using
                (Eq.3)
                Calculate the new position the firefly (Eq.5)//
       Move firefly i toward firefly j
                Assess the new solution by updating the
                intensity
                Eq.2.
           End if End for j
             Give the position of the best firefly. Discretize the
best solution (Eq.1)
       End for i
End
```

3.4. Classifier

The resulting documents from the previous phase are dealt with as an input for the classifier. In this study, we will use threeclassifiers: SVM, Naive Bayes (NB) and K-Nearest Neighbors (KNN). The description ofeach classifier is below:

3.4.1. The SVM Classifier

The machine learning techniques was introduced firstly by Vapnik [14] and in TC by Joachims [23], The Support Vector Machines (SVM) is a relatively new class of machine learning techniques. The (SVM) is Based on the computational learningtheory with the structural risk minimization principle, SVM seeks a decision surface to separate the training data points into two classes: the effective elements in the training set and the non-effective elements, then, makes decisions based on the support vectors that are selected as effective elements.

We consider a set of N linearly separable points S = { $x_i \epsilon \mathbb{R}^n$ | i = 1, 2… L}, where each point x_i belongs to one of the twoclasses, labeled as $y_i \epsilon$ {-1, 1}. SVM build the categorization model on the training data using a linear separating function toclassify unseen instances. For linearly separable vectors, the kernel function is simple. The hyper-plane that has the largest margin, which means that the distance between the nearest vectors to the hyper-plane is maximal, is the optimal separating hyper-plane.

SVMs decide based on the OSH classification during classification in place of all the training set. The technique simply discovers on which side of the OSH the test pattern is located. Unlike other traditional pattern recognition methods, this property makes SVM very competitive in computational efficiency and predictive precision [22].

3.4.2. The Naive Bayes Classifier

Naive Bayes refers to a probabilistic classification model rely on Bayes' theorem. It is a practical and straightforward classifier [9]. NB assumes that the feature values are conditionally independent of the expected classes.

Consider an ensemble of training examples where each example t is represented by a set of feature values {t1,t2,… tn} and target ranking. Let C be a set of classes defining the target function. Taking an example test t, NB assigns the example test to the class with the highest

probability [8].
The probability that the test example t belongs to a specific class Cj can be estimated as follows:

$$P(C_j/t) = \frac{P(t/C_j)*P(C_j)}{P(t)} \tag{8}$$

$P(t/C_j)$is the probability of the class C_j given a test example t .P(t) is equal for all categories, so it can be ignored:

$$P(C_j/t) = P(t/C_j) * P(C_j).$$

Using the assumption from Bayes theorem that affirms: the features are conditionally independent, the probability of class C_j can be rewritten as follows:

$$P(C_j/t) = P(C_j)\prod_{i=1}^{n} P\left(\frac{f_i}{C_j}\right) \tag{9}$$

With n is the number of features (f_i) that form training examples [16]. The NB classifier determines the class of the test instance t:

$$V_{NB} = argmax_{c_j \in C} P(C_j)\prod_{i=1}^{n} P\left(\frac{f_i}{C_j}\right) \tag{10}$$

This is the output of the Naive Bayes classifier, which refers to the class of test instances.

Although the features independence assumption is unrealistic, Naive Bayes has been very effective for many practical applications like text classification and medical diagnosis. This is thanks to its ability to scale with high dimension feature space.

3.4.3. The K-Nearest Neighbor Classifier (KNN)

The K-Nearest Neighbor Classifier (KNN) [24] is an essential classifier for text categorization. It is also one of the simplest methods to implement on a computer yet devised. It isa type of statistic machine learning method. The basic idea of the KNN is to determine the category of a given request not onlyaccording to the document that is nearest to it in the document space, but also according to the types of the K documents that are closest to it. The algorithm uses a vector-based, we calculating the document's similarity like the Vector method byusing the distance-weighted matching function.

4. Evaluating Results

The following section evaluates the performance of our New Chaotic Firefly Algorithm (CFA) for feature selection by using the EASC'S dataset and three classifiers: SVM, NB and KNN.

4.1. Dataset and Preprocessing

The data used in this paper is in the Arabic Natural Language resource: Essex Arabic Summaries Corpus (EASC). There are 153 articles in Arabic, and 765 summaries generated of those articles. These summaries have been generated using http://www.mturk.com/. Some of the important features of EASC are: Names and extensions are formatted to be compatible with current rating systems. Data are available in the following formats: UTF-8 and ISO-8859-6 (Arabic). This corpus is classified into ten categories, as presented in Table 1.

$$P = \frac{TP}{(TP + FP)} \tag{11}$$

$$R = \frac{TP}{(TP+TN)} \tag{12}$$

$$F = \frac{2RP}{(R + P)} \tag{13}$$

Our system is evaluated using: The Precision Equation (10), the Recall Equation (11), and the F-measure (Equation (12). The precision is defined as percentage of correctly classified documents. The True Positive (TP) defined a set of documents in category A that are correctly predicted to be in category A. The True Negative(TN) determines a set of documents that not in category A andwere predicted to be not in the set A. The False Positive (FP) isa set of documents that is not in category A and were predictedto be in A.

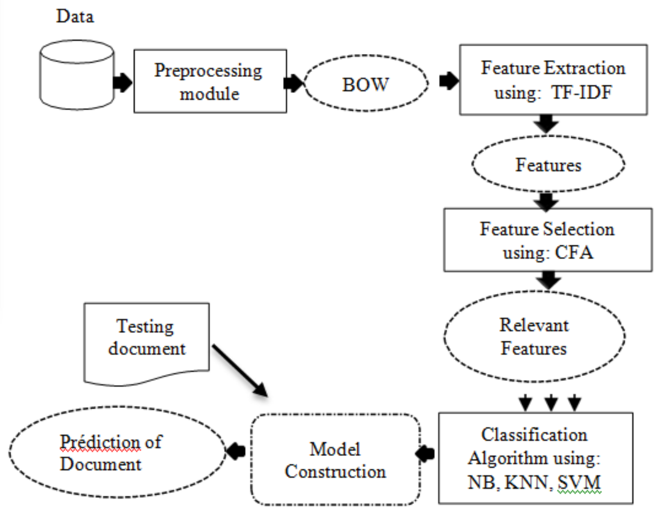

Figure 1. Text Categorization using modified firefly algorithm of feature selection.

The corpus is split into two subsets to construct the training and testing data for system of classification. The training data consists of 80% of documents in every category, and the test data consists of 20% of the documents in every category.

Table I. EASC'S Arabic text corpus.

Categories	Number of Documents
Environment	34
Politics	21
Finance	17
Health	17
Science and Technology	16
Tourism	14
Sports	10
Art and Music	10
Religion	08
Education	07
Total	**153**

For classifying the document, we initially preprocessed the texts by various techniques, the texts must be cleaned by:

- Segmentation: the text is divided into sentences; the For classifying the document, we initially preprocessed the texts by various techniques, the texts must be cleaned by:
- Segmentation: the text is divided into sentences; the sentences are divided into words.
- Deleting the numbers, the punctuation, and the words written in other languages, and any Arabic word containing special characters.
- Deleting diacritics of the words, if it exists.
- Normalizing of documents by replacing the letter ("أآإ") with ("ا"), and replacing the letter ("ؤء") with ("ا").
- Stop-words Removal: removing of stop-words from the documents turns the content of the text to more useful words for the summaries. We used Arabic stop-word list of [18] that contains 1,377 words.
- Stemming: mapping words into their base or common word of [11].

4.2. Methodology

The implementation of the Text Categorization system has been based on the New Chaotic Firefly Algorithm (CFA) for Feature Selection and performance comparison of Firefly Algorithm (FA), Modified Firefly Algorithm (MFA) [10], using NB, SVM and KNN classifiers.

We adopted the different format of terms using Stem and No-Stem as follows:

1. Preprocessing and representation of documents from their textual version to valid inputs understandable by the different algorithms. These valid inputs are vectors or matrices.
2. Perform feature selection with the CFA algorithm as follows:
   - Initializing the CFA parameters population and light absorption coefficient.
   - Initializing the light intensity.
   - Generating a new solution by updating the position of the firefly using Hamming distance.
   - Evaluating new solutions and updating the light Intensity.
   - Classification using Support Vector Machine (SVM), (NB) and (KNN) classifiers.
   - Classification with and without Stem.
3. Comparison of Chaotic Firefly Algorithm (CFA), Firefly Algorithm (FA) and Modified Firefly Algorithm FA (MFA) to Accuracy.

4.3. Results and Discussion

We compared the accuracy of the Chaotic Firefly Algorithm, Firefly Algorithm, and Modified Firefly Algorithm for

Feature Selection techniques of our EASC Dataset. Table 2 shows the accuracy for each method using SVM classifier.

Table 2. CFA compared TO FA AND MFA algorithms using SVM classifier.

Class/ feature selection	Modified Firefly Algorithm			Firefly Algorithm			Chaotic Firefly Algorithm		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Art andMusic	0,94	0,91	0,92	0,91	0.9	0,9	0,98	0,94	0,96
Education	0,88	0,87	0,87	0,88	0.85	0,86	0,9	0,87	0,88
Environment	0,98	0,89	0,93	0,95	0,88	0,91	0,98	0,9	0,94
Finance	0,93	0,87	0,9	0,92	0,92	0,92	0,93	0,87	0,90
Health	0,91	0,93	0,92	0,9	0,92	0,91	0,93	0,93	0,93
Politics	0,98	0,9	0,94	0,88	0,85	0,86	0,98	0,9	0,94
Religion	0,87	0,87	0,87	0,89	0,85	0,87	0,89	0,87	0,88
Science and Technology	0,98	0,92	0,95	0,98	0,87	0,92	0,98	0,92	0,95
Sports	0,97	0,94	0,95	0,87	0,97	0,92	0,97	0,94	0,95
Tourism	0,94	0,9	0,92	0,88	0,91	0,89	0,96	0,92	0,94

From Table 2, it can be observed that the result obtained by CFA was better than the results found by FA, and MFA for the accuracy of the classifier SVM. The classifier accuracy of CFA was almost equal to 0.96, while the result of FA was 0.9, and that of MFA was 0.94.

The New Chaotic Firefly algorithm improved feature selection accuracy and was found to be more efficient.

In Table 3, we compared the accuracy of the new method by using different classifier such as: SVM, NB and KNN for the Arabic Text Classification.

Table 3. The Accuracy using CFA for different classifiers.

Category/Classifier	**NB**	**SVM**	**KNN**
Art and Music	0,75	0,98	0,9
Education	0,81	0,9	0,91
Environment	71	0,98	0,88
Finance	0,65	0,93	0,81
Health	0,78	0,93	0,92
Politics	86	0,98	0,96
Religion	0,92	0,89	0,87
Science and Technology	0,73	0,98	0,98
Sports	0,81	0,97	0,98
Tourism	0,75	0,96	0,82

The above results show that our learning technique based on a Chaotic Firefly Algorithm using SVM is more suitable than NB and KNN classification technique. Moreover, the accuracy using SVM is much higher than NB and KNN based methods, as shown in Table 1 for each dataset category.

To measure the impact of preprocessing on the classification quality, we applied the stemming algorithm defined by [11] to our database. We used the heuristic method proposed on the SVM classifier. The Figure 2 shows the obtained results.

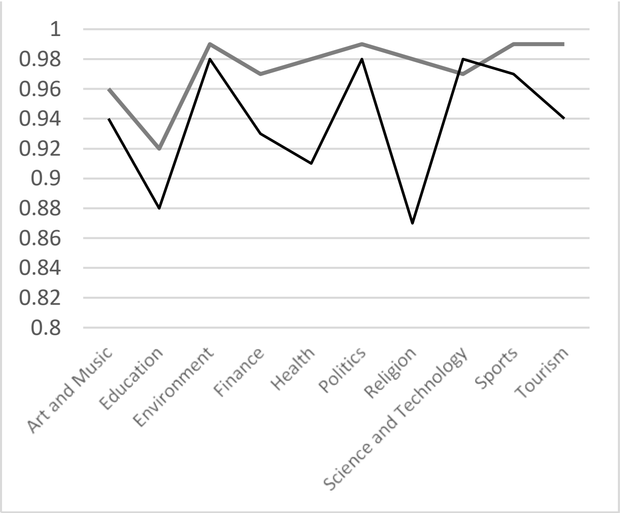

Figure 2. Accuracy for each classifier using CFA with/without stem.

According to the results obtained in Figure 2, the stemming acts directly and positively on the classification quality because the stemming helps remove the inflexion of the Arabic words that have the same root. The documents related to the same theme will have a better chance of ending up in the same class. This constitutes an improvement in an average accuracy of 98% forthe first method (without Stem) and 97.75% for the second (with Stem) for CFA and using the SVM classifier.

5. Conclusions and Future Work

The New Chaotic Firefly Algorithm based Feature Selectionwas proposed to improve classification accuracy. To validate the performance of the new method, the algorithm was tested on EASC'S database, using NB, SVM and KNN classifiers.

The experimental findings show that the suggested methodallows choosing a more discriminant subset of characteristics and increases the categorization accuracy.

This study applied the Chaotic firefly algorithm with continuous variables due to Arabic text classification. Still, thealgorithm can be adopted by hybridizing it with other meta- heuristics algorithms and comparing it to other datasets in the future.

References

[1] Ahmad S., Yusop N., Bakar A., and Yaakub M., "Statistical Analysis for Validating ACO-KNN Algorithm As Feature Selection in Sentiment Analysis," *International Conference on Electronics and Communication System*, vol. 1891, no. 1, 2017. https://doi.org/10.1063/1.5005351

[2] Alghamdi H. and Selamat A., "The Hybrid Feature Selection K-Means Method for Arabic Web Page Classification," *Jurnal Teknologi*, vol. 70, no. 5, pp. 73-79, 2014. DOI: https://doi.org/10.11113/jt.v70.3518

[3] Alghamdi H., Tang H., and Alshomrani S., "Hybrid ACO and TOFA Feature Selection Approach for Text Classification," *in Proceedings of the IEEE World Congress on Computational Intelligence*, Brisbane, pp. 10-15, 2012. DOI: 10.1109/CEC.2012.6252960

[4] Al-Harbi S., Al-Muhareb A., Al-Thubaity M., Khorsheed S., and Al-Rajeh A.," Automatic Arabic Text Classification," *in Proceedings of The 9th International Conference on the Statistical Analysis of Textual Data*, pp. 77-87, 2008.

[5] Al-Zahrani A., and Mathkour H., Abdalla H., "PSO-based Feature Selection for Arabic Text Summarization," *Journal of Universal Computer Science*, vol. 21, no. 11, pp. 1454-1469, 2015. 10.3217/jucs-021-11-1454

[6] Bessou S., Saadi A., and Touahria M., "Un système d'indexation et de recherche des textes en arabe SITRA," *1er séminaire national sur le langage naturel et l'intelligence artificielle LANIA*, pp. 20-21, 2007.

[7] El-Halees A., "Arabic Text Classification Using Maximum Entropy," *The Islamic University Journal*, vol. 15, no. 1, 157-167, 2007.

[8] El-Kourdi M., Bensaid A., and Rachidi T., "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," *in Proceeding of the 20th International Conference on Computational Linguistics*, Geneva, 2004. DOI:10.3115/1621804.1621819

[9] Greene D. and Cross J., "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach," *Political Analysis*, Vol. 25, no. 1, pp. 77-94, 2017.

[10] Hadni M. and Hassane H., "A New Meta-heuristic Approach Based Feature Selection for Arabic Text Categorization," *The International Arab Conference on Information Technology*, Abu Dhabi, pp. 1-7, 2022. doi: 10.1109/ACIT57182.2022.9994102.

[11] Hadni M., El Alaoui S., and Lachkar A., "Effective Arabic Stemmer Based Hybrid Approach for Arabic Text Categorization," *International Journal of Data Mining and Knowledge Management Process*, vol. 3, no. 4, 2013. DOI:10.5121/ijdkp.2013.3401

[12] Hadni M., El Alaoui S., and Lachkar A., Meknassi M., "Hybrid Part-of-SpeechTagger for Non-Vocalized Arabic Text," *International Journal on Natural Language Computing*, vol. 2, no. 6, pp. 1-15, 2013.

[13] Ja'afaru B., SabonGari N., and Zubairu B., "An Analytical Review on the Recent performances of Firefly Algorithm Fa," *Journal of Engineering Research and Application*, vol. 10, no. 4, 2020. DOI: 10.9790/9622-1004032637

[14] Larabi 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, 2018. https://doi.org/10.1016/j.jksuci.2018.06.004

[15] Mesleh A., "Chi-Square Feature Extraction Based SVMs Arabic Language Text Categorization System," *Journal of Computer Science*, vol.3, no. 6, 430-435, 2007. DOI:10.3844/jcssp.2007.430.435

[16] Mohammad S., "Sentiment analysis: Automatically Detecting Valence, Emotions, and Other Affectual States from Text," *Emotion Measurement (Second Edition)*, pp. 323-379, 2021. https://doi.org/10.1016/B978-0-12-821124-3.00011-9

[17] Nahar K., Al-Khatib R., Al-Shannaq M., Daradkeh M., and Malkawi R., "Direct Text Classifier for Thematic Arabic Discourse Documents," *The International Arab Journal of Information Technology*, vol. 17, no. 3, pp. 394-403, 2019. https://doi.org/10.34028/iajit/17/3/13

[18] Sarac E. and Ozel S., "Web Page Classification Using Firefly Optimization," *in Proceedings of the Innovations in Intelligent Systems and Applications* Albena, 2013. DOI: 10.1109/INISTA.2013.6577619

[19] Suchanek F., Kasneci G., and Weikum G., "Yago:A Large Ontology from Wikipedia and WN," *Journal of Web Semantics*, vol. 6, no. 3, pp. 203-217, 2008. DOI:10.1016/j.websem.2008.06.001

[20] Tandra V., Yowen Y., Tanjaya R., Santoso W., and Qomariyah N., "Short Message Service Filtering with Natural Language Processing in Indonesian Language," *in Proceedings of the International Conference on ICT for Smart Society ICISS*, Bandung, pp. 1-7, 2021. DOI: 10.1109/ICISS53185.2021.9532503

[21] Thirumagal D., Nithya S., Sangavi P., and Pugazhendi E., "Email Spam Detection and Data Optimization using NLP Techniques," *International Journal of Engineering Research and Technology IJERT*, vol. 10, no. 8, pp. 38-49, 2021.

[22] Thribhuvan N. and Elayidom S., "Transfer Learning for Feature Dimensionality Reduction," *The International Arab Journal of Information Technology*, vol. 19, no. 5, 2022. https://doi.org/10.34028/iajit/19/5/3

[23] Touati-Hamad Z., Laouar M., Bendib I., and Hakak S., "Arabic Quran Verses Authentication Using Deep Learning and Word Embeddings," *The International Arab Journal of Information Technology*, vol. 19, no. 4, pp. 681-688, 2022. https://doi.org/10.34028/iajit/19/4/13

[24] Wang L. and Zhao X., "Improved knn Classification Algorithm Research in Text Categorization," *in the Proceedings of the 2nd International Conference on Communications and Networks CECNet*, Yichang, pp.1848-1852, 2012. DOI: 10.1109/CECNet.2012.6201850

[25] Yang X., "Firefly Algorithm, Stochastic Test Functions and Design Optimization," *International Journal of Bio Inspired Computation Archive*, vol. 2, no. 2, pp. 78-84 2010. DOI:10.1504/IJBIC.2010.032124

[26] Yang Y. and Liu X., "A Re-Examination of Text Categorization Methods," *in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR'9*9, Berkley, pp. 42-49, 1999. https://doi.org/10.1145/312624.312647

[27] Yoshida M., Ikeda M., Ono S., Sato I., and Nakagawa H., "Person Name Disambiguation by Boot strapping," *in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, pp. 10-17, 2010. https://doi.org/10.1145/1835449.1835454

[28] Yousif S., Sultani Z., and Samawi V., "Utilizing Arabic Word Net Relations in Arabic Text Classification: New Feature Selection Methods," *IAENG International Journal of Computer Science*, vol. 64, no. 4, pp750-761, 2019.

[29] Zhang L., Mistry K., Limc S., and Neoh S., "Feature Selection Using Firefly Optimization for Classification And Regression Models," *Decision Support Systems*, vol. 106, pp. 64-85, 2018. https://doi.org/10.1016/j.dss.2017.12.001

Meryeme Hadni is working as a professor of Computer Science in EMSI, Marrakech, Morocco. His current research interests include: Arabic text mining applications: Arabic Text Classification and clustering, Arabic information and retrieval systems, and medical image application.

Hassane Hjiaj is a Professor at Department of Mathematics, Faculty of Sciences, Tétouan, University Abdelmalek Essaadi, Morocco. he obtained his university habilitation since 2020, His research interests include the study of partial differential equation and it's applications in image processing and Machine learning,.