11-17-2022

Hybrid feature selection based on principal component analysis and grey wolf optimizer algorithm for Arabic news article classification

Osama Ahmad Alomari
University of Sharjah

Ashraf Elnagar
University of Sharjah

Imad Afyouni
University of Sharjah

Ismail Shahin
University of Sharjah

Ali Bou Nassif
University of Sharjah

See next page for additional authors

Recommended Citation

Alomari, Osama Ahmad; Elnagar, Ashraf; Afyouni, Imad; Shahin, Ismail; Nassif, Ali Bou; Hashem, Ibrahim Abaker; and Tubishat, Mohammad, "Hybrid feature selection based on principal component analysis and grey wolf optimizer algorithm for Arabic news article classification" (2022). *All Works*. 5471.
https://zuscholars.zu.ac.ae/works/5471

Author First name, Last name, Institution
Osama Ahmad Alomari, Ashraf Elnagar, Imad Afyouni, Ismail Shahin, Ali Bou Nassif, Ibrahim Abaker Hashem, and Mohammad Tubishat

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Hybrid feature selection based on principal component analysis and grey wolf optimizer algorithm for Arabic news article classification

OSAMA AHMAD ALOMARI [1], ASHRAF ELNAGAR [2], (Senior Member, IEEE), IMAD AFYOUNI[2], ISMAIL SHAHIN [3], (Member, IEEE), ALI BOU NASSIF [4], (Member, IEEE), IBRAHIM ABAKER HASHEM [2], MOHAMMAD TUBISHAT [5],

[1] MLALP Research Group, University of Sharjah, Sharjah P.O. Box 27272, United Arab Emirates
[2] Computer Science Department, University of Sharjah, Sharjah P.O. Box 27272, United Arab Emirates
[3] Electrical Engineering Department, University of Sharjah, Sharjah P.O. Box 27272, United Arab Emirates
[4] Computer Engineering Department, University of Sharjah, Sharjah P.O. Box 27272, United Arab Emirates
[5]College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates

Corresponding authors: Ashraf Elnagar (ashraf@sharjah.ac.ae) and Osama Ahmad Alomari (oalomari@sharjah.ac.ae)

ABSTRACT

The rapid growth of electronic documents has resulted from the expansion and development of internet technologies. Text-documents classification is a key task in natural language processing that converts unstructured data into structured form and then extract knowledge from it. This conversion generates a high dimensional data that needs further analusis using data mining techniques like feature extraction, feature selection, and classification to derive meaningful insights from the data. Feature selection is a technique used for reducing dimensionality in order to prune the feature space and, as a result, lowering the computational cost and enhancing classification accuracy. This work presents a hybrid filter-wrapper method based on Principal Component Analysis (PCA) as a filter approach to select an appropriate and informative subset of features and Grey Wolf Optimizer (GWO) as wrapper approach (PCA-GWO) to select further informative features. Logistic Regression (LR) is used as an elevator to test the classification accuracy of candidate feature subsets produced by GWO. Three Arabic datasets, namely Alkhaleej, Akhbarona, and Arabiya, are used to assess the efficiency of the proposed method. The experimental results confirm that the proposed method based on PCA-GWO outperforms the baseline classifiers with/without feature selection and other feature selection approaches in terms of classification accuracy.

INDEX TERMS

Arabic text classification, Feature selection, Grey wolf optimizer, Principal component analysis, Logistic regression

I. INTRODUCTION

The global spread and rapid development of internet technologies has led to a massive amount of natural language text documents that are accessible through different repositories such as WORLD WIDE WEB, digital libraries, and electronic publications. However, these documents are presented in a scattered manner, thus organizing these documents in any form of user interaction is an impractical and very time-consuming process. Text classification is a process of allocating documents from a large-scale corpus or repository into predefined labels or categories [1], [2]. Text classification has a substantial influence on different applications such as web page classification [3], sentiment analysis [4], [5], bioinformatics [6]–[8], author identification [9], dialect detection [10], spam e-mail filtering [11], SMS spam filtering [12], and topic detection [13].

In the literature, most of the research conducted on text classification has targeted English and Chinese text document corpora with a minimal efforts put into Arabic language research. Arabic language holds excellent importance since it is considered the fourth most used language on the internet and the sixth official language worldwide by the United Nations [14]. The reasons that may attribute to the limited research work on Arabic language are the lack of high-quality and large Arabic corpora prepared for complex classification tasks. Arabic language has a rich morphology and complex orthography, and the available datasets cannot be freely downloaded [15].

Recently, a huge effort has been exerted in constructing new corpora by collecting news from popular sources [16], namely SANAD (Single-label Arabic News Articles Dataset) and NADIA (multi-label News Articles Dataset in Arabic). Similar to English text documents corpora, Arabic corpora still needs document processing tasks. In practice, the text document format is basically converted into a term-frequency vector, where each word's frequency is considered as a feature in the vector space [17], [18]. Such representation generates a high dimensional feature space that negatively affects the process of text classification due to the existence of irrelevant and redundant features and the increasing computation time. Therefore, a process of reducing the number of features to improve the efficiency of text classification tasks and to optimize the CPU time cost and memory size [19], [20] is needed.

Feature selection is an intelligent process that eliminates the irrelevant and redundant features and recognizes features or a subset of features that can be more informative and representative for real-world datasets [21] including epileptic seizure recognition [22], Functional MRI (fMRI) [23], [24], COVID-19 [25], medical diagnosis [26], [27], chemo-genomic data analysis [28], and CT liver tumor diagnosis [29].

Conventionally, the feature selection techniques are divided into two folds: filter-based and wrapper-based approaches. The filter approach evaluates the features within a short amount of time because it executes its calculation based on intrinsic characteristics of the training without using any machine learning algorithms. Examples of filter approaches are Chi-square [30], Kullback-Leibler [31], ReliefF [32], Minimum Relevancy Maximum Redundancy (MRMR) [33], and Robust MRMR (rMRMR) [31], Principal Component Analysis [34]. Wrapper-based approaches deal with the feature selection as an optimization search problem [35],by employing search techniques to produce candidate feature subsets. They then assess them by recreating the datasets based only on each single feature subset and then applying machine learning methods to the new dataset with reduced dimension to obtain the classification accuracy. Wrapper approaches yield higher classification accuracy when compared to filter approaches, but they suffer from expensive processing costs. A hybrid approach is an integration between filter and wrapper approaches, where it gains the benefits of both approaches. Hybrid approaches have been ensured to be more practical and efficient for high dimensional data such as text data [36], [37], image data [38], [39], EEG data [40] and microarray data [41].

As aforementioned, in the way of text classification, the massive presence of irrelevant features is intractable because the number of candidate feature subsets is grown exponentially with the increase of the number of features. In the wrapper feature selection approach, many researchers adopt metaheuristics methods to find a near-optimal feature subset to produce an efficient and accurate automatic text classification. Examples of feature selection techniques that use filter or wrapper or a combination of both are: document frequency and the term frequency with binary poor and rich optimization algorithm (DFTF-HBPRO) [36], Chi-square [42], Firefly algorithm [43], binary particle swarm optimization and KNN (BPSO-KNN) [19], information gain and principal component analysis with genetic algorithm (IG-PCAGA) [34]. However, most of these methods still have problem with local optima stagnation problem. Therefore, a powerful search method for finding the most informative features/terms that may provide more robust and accurate automatic text classification is required.

The grey wolf optimizer (GWO) is one of the widespread swarm-based optimization method inspired by the life cycle of grey wolf and their behavior in searching for prey (i.e. hunting strategy). The optimization process of GWO consists of three main phases. First, cooperative searching for finding the prey zone is performed, that reflects as an exploratory search mode. Encircling the prey zone and then attacking the prey are the second and third phases, respectively. This process is interpreted as an exploitative search mode. The GWO merits make it widely used due to its simple adaptation to any type of optimization problems, ease of use by the naive optimizers, parameter-free nature, and high flexibility. The GWO gained popularity and attracted the attention of researchers as a robust and effective solution for diverse optimization problems derived from different fields. Examples of these applications are engineering [44]–[46], machine-learning [47], image processing [48], scheduling [49], [50], Electroencephalography [51], networking [52], and Security [53]. Due to the excellent results and interesting merits of GWO, this research is motivated to use GWO as feature selection technique for Arabic text classification.

In this paper, a new hybrid filter-wrapper feature selection method for Arabic text classification is proposed. The proposed method adopts PCA as a filtering-based approach and GWO as a search method for feature subset generation in the wrapper approach. In classification process, Decision Tree (DT) [54], Random Forest (RF) [55], Support Vector Machine based with the popular Radial Basis Function (SVM-RBF) that frequently used in the nonlinear mapping of svm [56], Logistic Regression (LR) [57], and AdaBoost boosting (AB) [58] classifiers are carried out using three News datasets including Alkhaleej, Akhbarona, and Arabiya [16]. The performance of the machine learning methods is compared with and without PCA feature selection. The best classifier is assigned as an elevator for candidate feature subsets generated by GWO. Results show that the GWO-LR method yields the best classification accuracy when compared against machine learning baseline methods and machine learning with the PCA feature selection technique. The main contributions of this work are summarized as follows:

- A new hybrid filter-wrapper feature selection based on PCA as a filter and GWO as a wrapper.
- GWO is converted to a binary version using sigmoid function.
- The performance of the proposed method is tested on a real-world Arabic text data collected from popular Arabic news portals.
- GWO perform better or similar when compared with other optimization feature selection algorithms in all experimented datasets

The remaining of the paper is organised in the following sections: Sect. II presents the related work, Sect. III describes GWO's research background. The proposed method, which illustrates how the GWO is adapted for feature selection, and the datasets used in this research are provided in Sect. IV. The experiment setting and result are presented in Sect. V. The paper is concluded, and suggestions for future work is given in Sect.VI.

II. RELATED WORK

Text classification is not a new problem; it has been studied extensively in natural language processing literature. Most of the research works are applied to English text documents. Despite the importance of Arabic, there has been little research into applying and enhancing existing natural language algorithms for Arabic text classification. In [59], two popular classifiers, including support vector machine (SVM) and decision tree C5.0 were used to classify Arabic document texts, and they were experimented on seven Arabic datasets. The results demonstrated that C5.0 managed to surpass SVM by achieving 78.42% average classification accuracy. In another study, [49], eleven machine learning algorithms including Logistic Regression (LR), Multinomial NB (MNB), DT, SVM, XGBoost Classifiers (XGB), Multilayer Perceptron (MLP), KNN, Nearest Centroid Classifier (NC), AB, and Ensemble/Voting Classifier (VC), are utilized to classify Arabic text data. In this study, two large datasets were extracted from different Arab newspapers, where the articles in these datasets include diverse domains (including Sports, Technology, Business, and the Middle East). The results demonstrated that SVM and XGBoost yielded the highest classification accuracy on the first and second datasets, respectively. Three classifiers, including distance-based, KNN, and Naive bayes for classifying Arabic text, were investigated in [60]. The classifiers experimented on an in-house Arabic corpora, and the results exhibited that Naive Bayes outperformed other classifiers. Harrag et al. [61] investigated multi-classifiers including Decision trees, Naive Bayes, and Maximum entropy on data extracted from the Arabian scientific encyclopedia. The results demonstrated that Decision trees resulted the highest classification accuracy with 93%. The early work of [62], performed Arabic text classification, where Document Frequency threshold (DF) was used in the prepossessing stage, and KNN and SVM were used in the classification stage. The experiments concluded that KNN outperformed SVM by achieving higher precision results by 0.95%. The experiments were conducted on five Arabic newspaper text documents including Al-Dostor, Al-Ahram, Al-Nahar, Al-Jazeera, and Al-hayat. In [63], a comparative study is conducted on three classifiers, including SVM, Decision tree (C4.5), and Naive Bayes (NB) for Arabic text classification. The Arabic text documents are extracted from different sources such as Islamic topics, Poems, etc. The results demonstrated that the highest classification accuracy was obtained by SVM, followed by the C4.5, and NB. An efficient feature selection method on the basis of information gain and document frequency for Arabic text classification was introduced in [64]. In this study, Rocchio was employed as a classifier, and the text data used in the experiments was extracted from Egyptian newspapers, including El-Gomhoria, El-Akhbar, and El-Ahram. The proposed method was evaluated on the basis of three measurements, including recall, precision, and classification accuracy. The results revealed that Rocchio produced better classification accuracy than KNN. In [65], the authors suggested Arabic text classification system using Ant colony optimization (ACO) as a feature selection technique and SVM to perform classification. The performance of the proposed method was evaluated based on macro-averaging F1 measures, precision, and macro-averaging recall. The ACO feature selection technique exhibited better performance when compared against six state-of-the-art feature selection methods. Zahran and Kanaan [66] proposed an intelligent feature selection method for Arabic text classification using Particle Swarm Optimization (PSO) and Radial Basis Function Neural Network. The performance of the proposed method was experimented on Arabic corpora extracted from Arabic newspapers websites (including Al-Dostor, Al-Ahram, Al-Jazeera, and Al-Hayat), and was evaluated based on three measurements which are Precision, Recall, and F-score. The results demonstrated the efficiency of the proposed method when compared against Chi-square, TF-IDF, and document frequency algorithms. The authors in [19] implemented an intelligent method where PSO was used in the feature selection stage and KNN in the classification stage. The classification results showed the applicability of PSO-KNN for Arabic text classification. The authors of [67] employed Polynomial Neural Networks to classify Arabic text data after performing features selection using CHI Square. The proposed method achieved 0.94% in precision measurement. In [68], a comparative analysis was carried out on two feature selection techniques (i.e. CHI Square and Information Gain) combined with a number of classifiers, including KNN, Naive Bayes multinomial, Naive Bayesian method, and decision tree. The results revealed that the combination between CHI Square and Information Gain with classifiers provided good results except for KNN. In [69], the authors proposed an efficient hybrid feature selection approach based on a couple of filters including F-measure, Odd Ratio (OR), Class Discriminating Measure (CDM), GSS, IG, and TF-IDF of training text features (FM) and enhanced Genetic Algorithm (EGA). In EGA, the crossover operator was applied on the chromosome (feature subset) derived from term and document frequencies, while in mutation operators, two factors were considered, which are feature importance and the classification performance of the original parents. In this study, NB is used for classification and three datasets collected from a well-known Arabic news website (including Akhbar Al-Khaleej, Al-waten, and Al- Jazeerah). The results showed that the performance of the EGA outperformed GA and also six well-known filters (i.e., OR, CDM, GSS, IG, TF-IDF, and FM). In [70], the authors combined chi-square and Artificial Bee Colony (CHI-ABC) as feature selection techniques to classify Arabic text data. In this study, NB was used to perform classification. The proposed method was experimented on BBC Arabic dataset, and the results demonstrated that CHI-ABC outperformed CHI and ABC when running individually. A hybrid filter-wrapper feature selection for Arabic text classification was proposed in [71]. The proposed method embedded IG to perform a filtering approach and then passed the top-ranked features to the wrapper approach guided by a modified version of the Sine Cosine Algorithm. The proposed method was experimented on three new Hadith datasets. The results showed that the proposed method provided a good compromise of classification accuracy and the total reduced features.

In [72], the authors proposed Neural Networks (NN) for Arabic text categorization utilizing self-organization Maps (SOM) and Learning Vector Quantization (LVQ). A satisfactory results were produced on a small size datasets. Similarly, the authors in [73] confirm the superiority of NN over SVM after reducing feature space. Recently, deep-learning based approaches used for Arabic text classification, which have yielded outstanding results. In [15], the authors implemented nine deep learning models for Arabic text classification and they used word2vec embedding models to boost the classification performance. the results showed that all deep learning models yielded very promising classification accuracy, moreover, the utilize of word embedding enhanced the overall classification performance.

III. RESEARCH BACKGROUND

This section presents some of the widely used optimization algorithms, which is Grey Wolf Optimizer.

A. GREY WOLF OPTIMIZER (GWO)

The GWO is a well-known swarm-based metaheuristic algorithm that simulates the life cycle and hunting mechanism of grey wolves in nature. The GWO was introduced and proposed by Mirjalili in 2014 [74], who also brought up its mathematical expression.

The grey wolves' pack divides into four hierarchical levels of wolves: alpha (α), beta (β), delta (δ), and omega (ω) wolves which are dispersed on the bases of their levels of domination, with being at the highest and lowest levels of the wolves pack as presented in Figure 1.

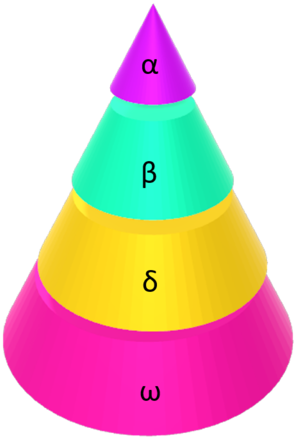


Figure 1: Grey Wolves hierarchy

In this pack of grey wolves, the α wolf is the wisest. It is highly efficient in managing the pack as well as in taking decision regarding the control of the pack and appropriate hunting style. It is also excellent at selecting a habitat. The α wolf is succeeded by the β wolves in hierarchy of domination. Normally, β wolves follows the α wolf wherever it is, supporting the α wolf in management and pack control.

The third domination level of the hierarchy is made up of the δ wolves. Wolves in this level are responsible for providing help, support and at the same time being guard to the members of the territory that are weak and old. The remaining wolves in the pack are the ω wolves. This stratification is based on the lifestyle of the wolves in the pack, their points of transaction management, hunting strategies, and their overall day-to-day activities. The main advantage of this hierarchy is that it assists in the leadership of the wolves during prey hunting. As soon as a prey is found, then α wolf orders the encircling of this prey by members of the pack while it leads the β and δ wolves in attacking the prey.

B. GREY WOLF OPTIMIZER ALGORITHM

Two main elements are based on the inspiration of GWO (hierarchy of grey wolves and their domination levels). Individual wolves serve as candidate solutions to various optimization issues that are being encountered. In the first three hierarchies (i.e., α, β, and δ levels), there is only single solution contained in each. While δ level contains a good solution, β holds a better. However, α level has the best solution. The remaining solutions are held in level ω. It should be noted that the members of ω level are obliged to help the members of the α, β, and δ levels to encircle and hunt the prey by means of the following formulation.

$$z = |b \times X_{p,e} - X_e|, \tag{1}$$

$$X_{e+1} = X_{p,e} - a \times z, \tag{2}$$

$$a_i = 2 \times \mu \times d_1 - \mu, \tag{3}$$

$$b_i = 2 \times d_2, \tag{4}$$

$$\mu = 2 - e \times \frac{2}{I}, \tag{5}$$

where $X_{p,e}$ represents for the location of the prey at e^{th} iteration, $X_{(e)}$ denotes the location at e^{th} and $(e+1)^{th}$ iterations, respectively. d_1 and d_2 stands for two arbitrary values, a and b are two coefficient vectors, and I represents the total iterations. a and b mainly aim to optimize balance between exploitation and exploration, and escape from the local optima. Through randomly altering the value of b, GWO it is capable of staying way from stagnation in local optima, as well as exploiting and exploring a given search space when $|a| < 1$ and $|a| > 1$, correspondingly. It is necessary to update the solutions in ω level after every iteration based on the solutions in α, β, and δ levels through applying the following formula:

$$z_\alpha = |b_1 \times X_\alpha - X|, \tag{6}$$

$$z_\beta = |b_2 \times X_\beta - X|, \tag{7}$$

$$z_\delta = |b_3 \times X_\delta - X|, \tag{8}$$

$$X_1 = X_\alpha - a_1 \times z_\alpha, \tag{9}$$

$$X_2 = X_\beta - a_2 \times z_\beta, \tag{10}$$

$$X_3 = X_\delta - a_3 \times z_\delta, \tag{11}$$

$$X_{e+1} = \frac{X_1 + X_2 + X_3}{3}, \tag{12}$$

IV. THE PROPOSED TEXT CLASSIFICATION APPROACH

In this section, an intelligent hybrid feature selection approach is proposed for the classification of Arabic texts. In the proposed method, PCA is used as a filtering approach, and its main task is to search over the term/feature search space of all extracted features from the raw Arabic text datasets and find the best subset of relevant and informative features. To further seek informative features and better classification accuracy, GWO is coupled with Logistic Regression (LR) classifier, where GWO is utilized to optimize the feature subsets produced by PCA and then passes the candidate feature subsets to assess them by carrying out LR on those selected features by measuring their classification accuracy. The proposed system consists of several stages, including preprocessing, feature selection, and classification, as depicted in Figure 2. The stages are elaborated in detail in the subsequent sections.

A. PREPROCESSING

The raw Arabic data needs to be converted into an appropriate format that automatic text analysis can process, and because there are various ways of reporting text in Arabic language, the Arabic text data documents were fed into preprocessing task according to the following steps:

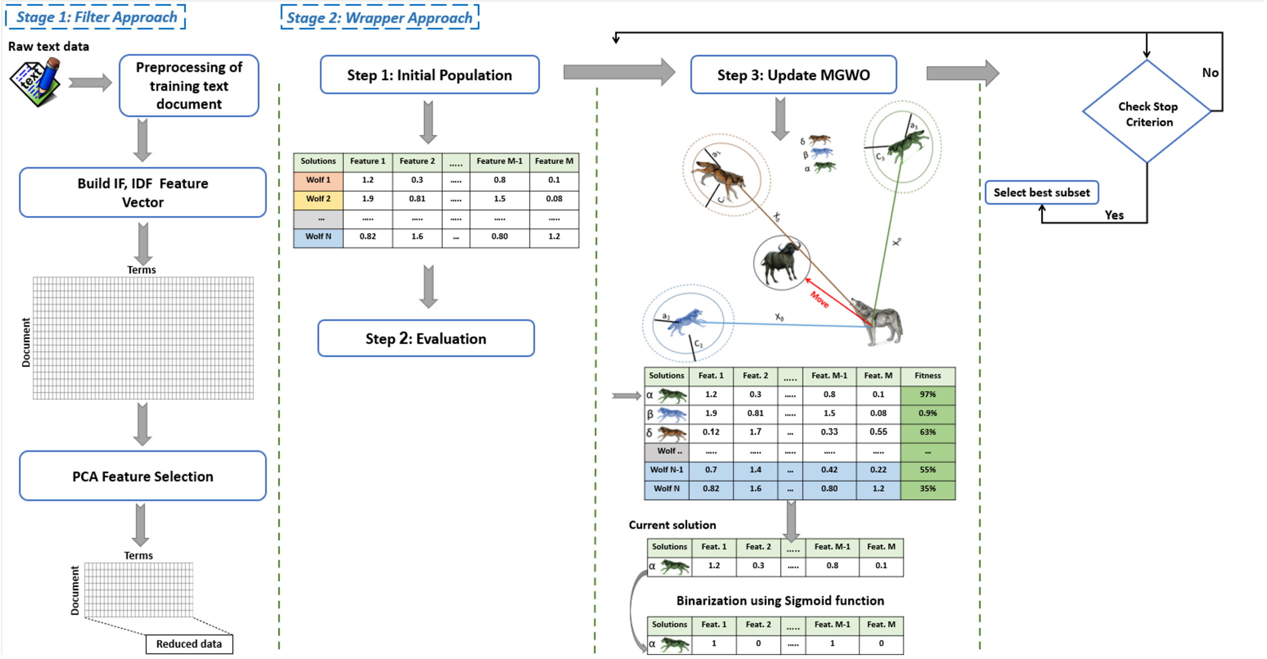
The raw Arabic textual data needs to be converted into an appropriate format that automatic text analysis can process. There are various ways of representing text in Arabic language, the Arabic documents preprocessing stage includes:

- Use UTF-8 encoding.
- Eliminate non-Arabic letters, digits, punctuation marks, and diacritics.
- Drop stop words that appear in the raw text like prepositions and pronouns.
- The Words with frequency less than five times are ignored.
- Vector Space Model is adopted in this stage to formulate the Arabic text data into a proper representation, and TFIDF (term frequency inverse document frequency) is employed for weighting the terms.

TFIDF is a popular scheme used for weighting the terms in the field of text classification. TFIDF has been proven to be a practical statistical approach for assigning weight for the terms [75]. The TF scheme, in practice, stands for the feature/term weight F_i in the feature space, which is computed by counting the number of times the F_i found in text document d_j [76].

Document frequency is calculated at the level of corpora, where the feature F_i is assigned a weight based on the number of the document text in the corpus that contains F_i at least once. The inverse document frequency that associates with the feature F_i can be computed as shown in the equation [76]:

$$IDF(t_i) = \log \frac{D}{DF(F_i)} \tag{13}$$



Solutions	Feature 1	Feature 2	.....	Feature M-1	Feature M
Wolf 1	1.2	0.3	.....	0.8	0.1
Wolf 2	1.9	0.81	.....	1.5	0.08
...	.....	.....	.....	.....	.....
Wolf N	0.82	1.6	...	0.80	1.2

Solutions	Feat. 1	Feat. 2	.....	Feat. M-1	Feat. M	Fitness
α	1.2	0.3	.....	0.8	0.1	97%
β	1.9	0.81	.....	1.5	0.08	0.9%
δ	0.12	1.7	...	0.33	0.55	63%
Wolf ..	.....	.....	.....	.....	.....	...
Wolf N-1	0.7	1.4	...	0.42	0.22	55%
Wolf N	0.82	1.6	...	0.80	1.2	35%

Current solution

Solutions	Feat. 1	Feat. 2	.....	Feat. M-1	Feat. M
α	1.2	0.3	.....	0.8	0.1

Binarization using Sigmoid function

Solutions	Feat. 1	Feat. 2	.....	Feat. M-1	Feat. M
α	1	0	.....	1	0

Figure 2: Flowchart of PCA-GWO method

where D stands for the number of text documents. The weight of the term F_i in the document d_j by means of TFIDF is estimated as follow:

$$TFIDF(t_i, d_j) = TF(t_i, d_j) \times IDF(t_i) \tag{14}$$

B. FEATURE SELECTION

In this stage, the feature selection process is carried out to the Arabic news datasets to reduce the dimension of the data and also to find a set of informative features that may be a better representative for the Arabic datasets instead of using all features. In this research, a filter feature selection approach called PCA is initially applied to the dataset to produce a strong and relevant subset of features, and thereafter, a pruning process is applied to this subset to seek further informative and discriminative features using a wrapper approach that guided by GWO algorithm. PCA and the proposed method PCA-GWO are thoroughly discussed in the subsequent sections.

1) Principal Component Analysis (PCA)

The Arabic text data is highly dimensional data that deteriorates the classification performance of machine learning algorithms. Therefore, in this work, PCA is employed to produce a subset of the most relevant features. To mathematically formulate dimension reduction, the number of features is denoted as N and the features are denoted as vector x. The features in the raw data is represented as $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, and the output from applying PCA in the raw data is $\mathbf{y} = (y_1, y_2, \ldots, y_D)$, where D is less than N.

PCA is a linear transformation model, which changes several feasible correlated variables to a few number of uncorrelated variables referred to as principal components (PCs) [77]. The PCs are represented form the linear combinations of the original variables measured by the degree of contribution they make to provision of explanation on the variance in a given orthogonal dimension. The principal components are ordered based on the variability for which they stand. Larger variance is found in the lower-order PCs, whereas, the higher-order PCs have lower variance. The chosen feature selection module involves the elimination of higher-order PCs while keeping the lower-order PCs. The authors of [77] suggested that PCA determines the correlation and dependencies that exist in the extracted features (i.e, x_1, x_2,..x_N) through the development of a covariance matrix U of the dimension N×N in which N denotes the number of extracted features, as shown below:

$$U = \begin{pmatrix} Cov(x_1, x_2) & \ldots & Cov(x_1, x_N) \\ \vdots & \ddots & \vdots \\ Cov(x_N, x_1) & \cdots & Cov(x_N, x_N) \end{pmatrix} \tag{15}$$

Eigenvectors $(v_1, v_2, \ldots, v_N)$ and corresponding eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_N)$ are values derived from the covariance matrix U to identify the PCs. Eigenvalues are arranged from the top to bottom fashion, and the k eigenvectors that are similar to the K largest eigenvalues are selected for constructing a reduced matrix U_K, in which K represents the number of dimensions of the new feature subspace ($K \leqslant N$). A project matrix W is created from the chosen K eigenvectors through the multiplication of the transpose of the reduced matrix by the original set of extracted features X, so as the newly formulated PCs replace the original data axis. This is expressed as:

$$W = U_K^T X \tag{16}$$

2) PCA-GWO implementation process

In this section, a hybrid feature selection method PCA-GWO is proposed and thoroughly illustrated. The proposed method utilizes PCA filter approach as pruning process to the raw features in the Arabic textual data. The output from the PCA filtering process is a feature subset, where it is further optimized by GWO to produce a discriminative and informative feature. The PCA-GWO process is divided into four basic steps, which will be explained in details below. The PCA-GWO flow chart is shown in Figure 2 and pseudo coded in Algorithm 1.

Algorithm 1 Pseudo code of the proposed GWO

```
Step1: Initialization.
PCA-feature subset= ${y_1, y_2, ..., y_D}$
Initialize MGWO parameters (n, $Max_e$).
while ($e \leq Max_e$) do
    for each solution ($j$) do
        Step2: Evaluation.
        Compute the fitness of the solution
        $X_\alpha$ = the fittest solution
        $X_\beta$ = the second-best solution
        $X_\delta$ = the third-best solution
    end for
    Step3: Update MGWO population
    for each single solution ($j$) do
        Update $d_1, d_2$
        The variable $a_1$ value is updated using (Eq. 3)
        The variable $b_1$ value is updated using (Eq. 4)
        Calculate $X_1$ (Eqs. 6, 9)
        Update $d_1, d_2$
        The variable $a_2$ value is updated using (Eq. 3)
        The variable $b_2$ value is updated using (Eq. 4)
        Calculate $X_2$ (Eqs. 7, 10)
        Update $d_1, d_2$
        The variable $a_3$ value is updated using (Eq. 3)
        The variable $b_3$ value is updated using (Eq. 4)
        Calculate $X_3$ (Eqs. 8, 11)
        Produce a new solution $X_{(e+1)}$ (Eq. 12)
        Transfer $X_{(e+1)}$ to binary vector using sigmoid function (Eq. 19)
    end for
    Step 4: Check the stop criterion
    if The maximum number of the iterations is not met then
        $e = e + 1$
    end if
end while
Return $X_\alpha$
```

Step1: Initialization.

This step involves the initialization of the number of iterations. Here, each wolf represents a standalone solution for a feature selection problem in which every solution serves as a binary vector of size D as expressed in Equation (17). This implies that the solution's decision variables accepts either 0 or 1, and this is a termed position in GWO.

The launching of GWO searching processes can be achieved through the generation of n wolves to serve as random binary vectors.

$$GWOP = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_D^1 \\ x_1^2 & x_2^2 & \cdots & x_D^2 \\ \vdots & \vdots & \cdots & \vdots \\ x_1^n & x_2^n & \cdots & x_D^n \end{bmatrix} \tag{17}$$

subject to:

$$x_i^j \in \{0, 1\}$$

where x_i^j refers to the jth decision variable of solution (wolf) xi.

Step2: Evaluation.

In this step, wolves are assessed based on their position vectors, where each position in the vector is either 1 or 0. The positions that have the value of 1 indicate that these features/terms form the new reduced data. Later, the reduced data is divided into training and testing, where the LR classifier is learned from the training data and assessed in the testing data. The LR model is evaluated using a classification accuracy metric. The objective function utilized to evaluate the classification performance of each grey wolf position vector is formulated below:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

where Acc denotes the objective function (accuracy rate), and TP, TN, FP, and FN represent the true positive, true negative, false positive, and false negative, respectively. The top three fitness values are X_α, X_β, and X_δ, respectively. This solutions hierarchy is inspired by the social hierarchy of wolves, as explained in Algorithm 2.

Algorithm 2 Social hierarchy component of the proposed GWO

```
while ($e < Max_e$) do
    for each solution ($j$) do
        Calculate the fitness of the solution
        $X_\alpha$ = the fittest solution
        $X_\beta$ = the second-best solution
        $X_\delta$ = the third-best solution
    end for
    $e = e + 1$
end while
Return $X_\alpha, X_\beta, X_\delta$;
```

Step3: Update GWO population

Here, GWO involves three main operators including seeking for prey (exploration), attacking prey (exploitation), encircling prey, and hunting mechanism. This is applied during the

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3222516

navigation of the search space of the feature selection problem and updating the GWO population. The equations from 1 to 12 as explained earlier in Section III-A, are used to update the solutions in GWO population. The mechanism for this is operated through the assessment of the distance between each solution in the population and social hierarchy-based solutions (i.e., X_α, X_β, and X_δ). The current wolf/solution gets its position or decision variables updated based on X_α, X_β, and X_δ. This results in the generation of a new solution (X_1) that is based on X_α using Eqs. 6, 3, 4, and 9. The steps are taken again in order to derive two new solutions, X_2 and X_3, in which X_2 is obtained on the basis of X_β by applying Eqs. 7, 3, 4, and 10, and X_3 is generated on the basis of X_δ by applying Eqs. 8, 3, 4, and 11. At last, the solutions X_1, X_2, and X_3 are aggregated by use of the mean to obtain a new solution $X(e+1)$. However, it is necessary to note that positions of $X(e+1)$ have continuous values, they are converted to binary vector through the use of Eqs. 19 and 20.

$$\sigma(X(e+1)) = \frac{1}{1+e^{-X(e+1)}} \tag{19}$$

$$X(e+1) = \begin{cases} 1 & \text{if } sigmoid(X(e+1)) > U(0,1) \\ 0 & Otherwise \end{cases} \tag{20}$$

Where U(0,1) is a uniform random number between 0 and 1. Furthermore, the new solution $X(e+1)$ is generated at every iteration and can be assessed by use of fitness function which mainly relies on classification accuracy.

Step4: Check the stop criterion

The step 3 is an iterative process that tends to achieve a better search around the best solution. This iterative process is controlled by the stopping condition (which commonly sets the maximum number of iterations). Once the stopping condition is met, the best solution that carried the distinctive features/terms for Arabic text classification problem is produced.

C. CLASSIFICATION

The most well-known machine learning algorithms which are widely applied in pattern recognition (in particular text classification field) including DT, LR, SVM, Ada boost, and RF, are discussed in detail and implemented in this research.

1) decision Tree (DT)

The decision tree is a machine learning methodology which is widely recognized for automation of the induction of classification trees with respect to training data [78]. A typical decision tree training algorithm comprises of two phases. The tree growing phase is the first one. This phase involves the building of tree through greedy splitting of respective tree nodes. The second phase involves removal of overfitted tree branches as the branches of the tree are capable of overfitting the training data [79]. C4.5 is a univariate decision tree algorithm. Only one of the attributes of instances at a given node can be adopted for decision making purposes. Details of C4.5 are obtained from Fuhr and Buckley [80].

2) Support vector machine (SVM)

In the SVM approach, linear kernel is involved, it is known to possess a significantly high performance in terms of text categorization due to its linear separable nature [81]. The important merits of this classifiers include high generalization ability, success in resolving the problem of overfitting and global optimization capabilities [82]. Moreover, this classifier possesses a satisfactory performance in the large-scale feature space, it also has the ability of managing any distributional dataset [82]. It is however, not suitable for managing massive dataset, as it needs feature scale to operate adequately. The task of training and tuning classifier tend to be exhaustive and memory intensive [82].

3) Logistic regression (LR)

Logistic regression is a well-known classifier. It has a simple coding procedure and it is highly reliable [57], [83]. The logistic regression is a classifier candidate that is effective in carrying out polarity classification tasks. It relies on Sigmoid function in generating a report related to the probabilities of the predicted labels. The maximum likelihood estimation adopts the use of the gradient descent algorithm in maximizing the likelihood of accurately classifying a arbitrary set of input features. The prediction of multi-class problem can be done through formulating the problem into a polarity classification (one-versus-the-rest). Otherwise, loss function (i.e., cross-entropy) can be used to get a solution.

4) Ada boost (AB)

Ada boost is a type of machine learning algorithm that was introduced by Yao Froud and Robert Shaper [58]. It is a meta- algorithm that is useful in enhancing performance as well as troubleshooting unbalanced categories together with similar algorithms. Classification of each step is advantageous to wrongly set samples in previous steps. It does not tolerate data that is useless and noisy; however, its operation is simpler than others classifiers. With numerous iterations, the performance of Ada boost is enhanced. In each round, a weak class is added and weights are displayed according to sample importance. Weights of wrongly classified sample increase with the increase of number of cycles, while in case the number of the samples which are correctly classified decreased, the new class focus on examples that are not easily learnt.

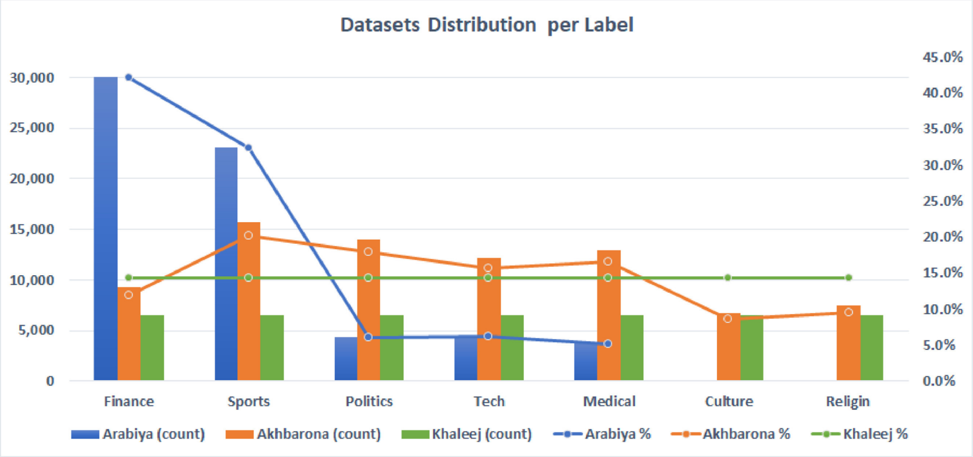


Figure 3: Distribution of categories in the SANAD datasets.

5) Random forest

Random forest is a supervised machine learning algorithm. It is among the strong machine learning algorithms that is important in classification and regression problems. The random forest is a set of decision trees that are developed by randomly selected training data by the random forest classifier. The final class of the test object is chosen through combination of votes from variant decision trees. The operation of this model is more accurate since it involves the combination of several decision trees; it yields more reliable results and less noise. A major demerit of this algorithm includes complex nature, prolonged duration of training, slow speed and its less effectiveness in terms of real-time predictions due to the high number of trees [55].

D. ARABIC DATASETS

Three different datasets constructed by [16] are used in this study. These data sets were collected by means of web scraping (Python Selenium, Requests and BeautifulSoup or PowerShell), obtained from three well-known news websites (alarabiya.net, alkhaleej.ae and akhbarona.com). The datasets are grouped in one corpus referred to as SANAD. These datasets excluding al-arabiya.net (which lacks Culture or Religion categories) have all the categories [Tech, Sports, Religion, Politics, Medical, Finance and Culture]. It does not have to do with dialect because datasets are collected from news sites, since all the articles were produced in modern standard Arabic. Moreover, samples of the selected features extracted from Arabic text data is presented in Table 1.

1) alarabiya.net

For this dataset, respective articles in the primary domain and sub-domains (i.e Aswaq and Ahadath) were carefully examined. The articles were then categorized into seven classes out of which two possess inadequate data (Iran Culture News) as compared to the others. We merged the “Iran News” and the “Politics” categories so as to develop an effective dataset. Consequently, when the “Culture category was dropped, the categories were reduced to five categories. Articles that were compiled are up-to-date till early 2018. These five categories of datasets are described in Figure. 3:

2) Alkhaleej.ae

About 1.2M (4GB) articles were collected through the examination of articles of this website for ten years (i.e., 2008-2018). It has been realized that the categorization of this website is somewhat not complete and ambiguous in several aspects. Therefore, there was the need for a manual categorization of certain amount of the article into the seven categories mentioned earlier, giving rise to a total of over 46000 articles. Moreover, it has resulted in the need to classify some articles as categories in which they don’t belong to certain category, that made the data sets not as reliable as the other two datasets (i.e, alarabiya.net and Akhbarona.com). In Figure. 3, the seven categories involved in this dataset, which are in balanced distribution, is illustrated. Manual categorization of Khaleej dataset involves the selection of the tags gotten from the website so as to categorize them into a category from the seven categories. For instance, articles of the tags ‘Technology’, ‘Digital Life’,‘Computer Internet’, and ‘GITEX’ are classified as a generic category termed ‘Technology’.

3) Akhbarona.com

All required categories of articles were collected. It happened that one of those categories (Religion) had 50% of the features possessed by other categories. As such, the remaining 50% was sourced from a newspaper website of relevant interest (Alanba.com).

This dataset’s seven categories are distributed and plotted in Figure 3.

Table 1: Samples of selected features.

Akhbarona		Arabiya		Khaleej	
Feature	Meaning	Feature	Meaning	Feature	Meaning
الكواليس	Backastages	مسلسل	Series	رواية	Novel
مهرجان	Festival	الجمهور	Audience	الإعلام	Media
الموارد	Resources	البنك	Bank	الإستثمار	Investment
المقاولات	Constructions	الاقتصاد	Economy	التمويل	Finance
الخزينة	Treasury	الأسواق	Markets	المساهمين	Shareholders
الدماغ	Brain	فيروسات	Viruses	الطبية	Medical
السعادة	Happiness	الجينات	Genes	المرضى	Patients
الوحدة	Loneliness	الطبيب	Doctor	العلاج	Treatment
البرلمان	Parliament	الأدوية	Drugs	الأطفال	Children
التصريحات	Statements	الحدود	Borders	المستشفيات	Hospitals
الحكومات	Governments	الكأس	Cup	الجيش	Army
الصلاة	Prayer	مباراة	Match	الاجتماع	Meeting
الخيرية	Charity	النادي	Club	الشرطة	Police
الإخلاص	Devotion	اللاعب	Player	الدين	Religion
كرة القدم	Football	التعادل	Tie	المهاجم	Striker
الفريق	Team	المدرب	Coach	المدافع	Defender
الملعب	Stadium	الأهداف	Goals	الشبكات	Networks
الإصدارات	Versions	الأنظمة	Systems	تكنولوجيا	Technology
هاتف ذكي	Smart phone	النصوص	Texts	اتصالات	Telecommunication
الشاشة	Screen	مايكروسوفت	Microsoft	الإنترنت	Internet

V. EXPERIMENTAL SETUP AND RESULTS

A. EXPERIMENTAL SETUP

In this study, all algorithms were implemented using Python and RapidMiner software. To assess the performance of the utilized algorithms, two metrics were used which are classification accuracy and the number of reduced features. Moreover, to validate the performance and effectiveness of these algorithms, three benchmark Arabic datasets were employed from [16] as described in Table [16]. The three datasets were compiled from Arabic news portals, including alarabiya.net, alkhaleej.ae and akhbarona.com. The number of articles for Alarabiya, Alkhaleej and Akhbarona datasets is 1207, 1408, and 1404, respectively. Alkhaleej and Akhbarona datasets have seven categories while the Alarabiya dataset has six categories. Each category has 200 articles. In the following experiments, the datasets are divided into 90:10 ratio as training and test set. In Alarabiya dataset, the number of samples in training and testing datasets are 1086 and 121, respectively. In Alkhaleej dataset, the number of samples in training and testing datasets are 1267 and 141, respectively. In Akhbarona dataset, the number of samples in training and testing datasets are 1263 and 141, respectively. Three experiment settings were carried out in the current study. In the first experiment we implemented and compared five different classifiers on the three datasets in order to pick the best classifier for the following two experiments. In the second experiment, to confirm the selection of the best classifier, PCA feature selection was combined with each classifier, then we compared the accuracy of these classifiers by using PCA feature selection. In the third experiment, the results of the proposed hybrid approach using filter and wrapper feature selection (PCA-GWO) were compared against three popular optimization algorithms (that were adapted as feature selection approaches), LR, and LR with PCA feature selection.

B. EXPERIMENTAL RESULTS OF CLASSIFIERS ONLY

The first experiment was conducted without using features selection by applying only DT, RF, SVM-RBF, LR, or AB on the full features set. Each one of the mentioned classifiers was applied on the full features set without using any reduction method such as PCA filter feature selection or wrapper method using an optimization algorithm. Table 2 shows that the best classification accuracy results were achieved using the LR classifier over the three datasets. Therefore, this confirms the superiority of the LR classifier over all other used classifiers. However, we noticed from the results that there is a possibility of further improvement by applying feature selection methods. Thus, in the next two experiments, our task is to apply PCA and wrapper feature selection to select the most relevant features and enhance the classification performance.

Table 2: Classification accuracy of DT, RF, SVM-RBF, LR, and AB on all datasets

	Datasets		
Method	Khaleej	Akhbarona	Arabiya
DT	74.47	72.34	73.05
RF	85.11	85.82	90.07
SVM-RBF	**92.91**	92.20	**92.91**
LR	**92.91**	**92.91**	**92.91**
AB	56.03	62.41	65.96

C. EXPERIMENTAL RESULTS OF CLASSIFIERS WITH PCA.

In this experiment, PCA feature selection was applied with all classifiers (i.e, DT, RF, SVM-RBF, LR, and AB). PCA is a type of feature subset algorithm, where it produces multi-feature subsets and chooses the best feature subset that mostly represents the entire dataset. The results of integration of the PCA with all classifiers are presented in Table 3 and Figure 4. The classification accuracy achieved on Khaleej dataset using PCA with all classifiers (with 1010 feature is better than using only classifiers without feature selection, except for AB classifier. It should be noted that PCA feature selection results were achieved with less than 10% from the total number of original features in the Khaleej dataset. In the Akhbarona dataset, PCA with less than 12% from the total number of original features managed to archive higher classification accuracy when integrated with DT and SVM-RBF. Furthermore, PCA-LR and LR have similar classification accuracy. On the other hand, the baseline classifiers (i.e, DT and AB) have better classification accuracy than involving PCA feature selection in their classification task. On the Arabiya dataset, PCA feature selection reduced the dimensionality n of the feature space by less than 13%. In this new dimension-reduction data, the classification accuracy of PCA with all classifiers is higher than using all baseline classifiers without feature selection. The results confirms the significance of applying PCA as a feature selection technique with most of the classifiers for all datasets. However, to obtain a more accurate automatic Arabic text classification system, PCA is hybridized with a wrapper approach guided by GWO to seek further accurate and robust features.

D. EXPERIMENTAL RESULTS OF PCA-GWO AND OTHER APPROACHES.

In this experiment, the effectiveness of the proposed method is validated by comparing it with several well-known optimization algorithms, including Bat-inspired Algorithm (BAT) [84], Firefly Algorithm (FFA) [85], Particle Swarm Optimization (PSO) [86], White Shark Algorithm (WSO) [87], Marine Predators Algorithm (MPA) [88], and Slime Mould Algorithm (SMA) [25]. The BAT, PSO, FFA, WSO, MPA, and SMA algorithms are used as wrapper feature selection approaches, and LR is used to perform classification. In Table 4, the results of the proposed method and other optimization feature selection approaches are summarized in terms average classificatiof accuracy along with standard deviation values that are expressed in the form (average ± standard deviation). The best results are highlighted in bold font. It can be observed that PCA-GWO is managed to yield the best classification results for two out of three datasets (i.e., Khaleej and Akhbarona). On the other hand, for Arabiya dataset, the best result was achieved by PCA-PSO. In respect to the average number of selected features, PCA-SMA identifies the lowest number of features for all datasets; however, it doesn't achieve the highest classification accuracy.

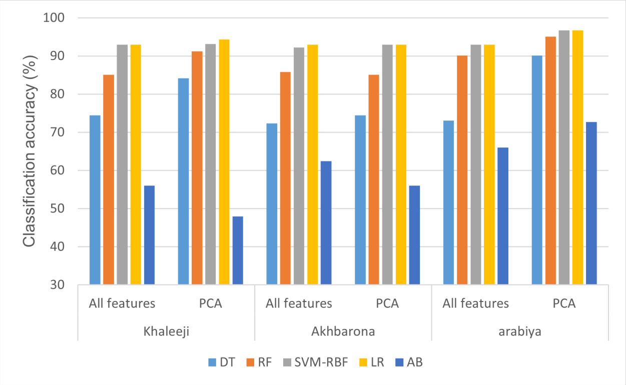


Figure 4: Comparison of classification accuracy of DT, RF, SVM-RBF, LR, and AB with and without feature selection on all datasets

Furthermore, The best classification accuracy results and number of selected features are reported in Table 5. It can be inferred that PCA-GWO yields the best results for Khaleej and Akhbarona datasets. For Arabiya dataset, the best result was achieved by PCA-MPA. In respect to the number of selected features, PCA-SMA achieved the best result, where they successfully identify the fewest number of features on all datasets. However, the classification accuracy obtained by PCA-GWO is higher than PCA-SMA on all datasets (i.e., Khaleej and Akhbarona).

To further validate the results yielded by PCA-GWO and other optimization algorithms, Wilcoxon signed-rank statistical test [89] is used in this study to demonstrate if there is statistically significant difference between these algorithms. In Table 6, Z-value stands for standardized test statistics, and P-value stands for the statistical significance ($P-Value < 0.05$). A $P-Value < 0.05$ implies that there is statistical significant difference between the compared algorithms; otherwise, there is no statistical significant difference. From Table 6, it can be inferred that PCA-GWO obtained statistical significant results in most

Table 3: Classification accuracy results of DT, RF, SVM-RBF, LR, and AB with and without feature selection on all datasets

	Datasets					
Method	Khaleej		Akhbarona		Arabiya	
	Full dataset (10040)	PCA Selected features (1010)	Full dataset (8793)	PCA selected features (1036)	Full dataset (7242)	PCA selected features (932)
DT	74.47	**84.11**	72.34	**74.47**	73.05	**90.08**
RF	85.11	**91.21**	**85.82**	85.11	90.07	**95.04**
SVM-RBF	92.91	**93.12**	92.20	**92.91**	92.91	**96.69**
LR	92.91	**94.33**	**92.91**	**92.91**	92.91	**96.69**
AB	**56.03**	47.94	**62.41**	56.03	65.96	**72.73**

of the datasets when compared with other algorithms.

Additionally, the execution time of the proposed method PCA-GWO is compared with LR (without feature selection) and PCA-LR, as shown in Figure 5. The results demonstrate that PCA-GWO has the minimum computational time. The proposed method PCA-GWO managed to effectively increase the classification accuracy while reducing the computation time. In summary, PCA-GWO provides superior and competitive results when compared to other feature selection approaches. This fruitful result is owed to robust searching operators in GWO represented by searching for prey (exploration), encircling prey, attacking prey (exploitation), and hunting mechanism, which resulted in searching the feature space of Arabic textual data effectively.

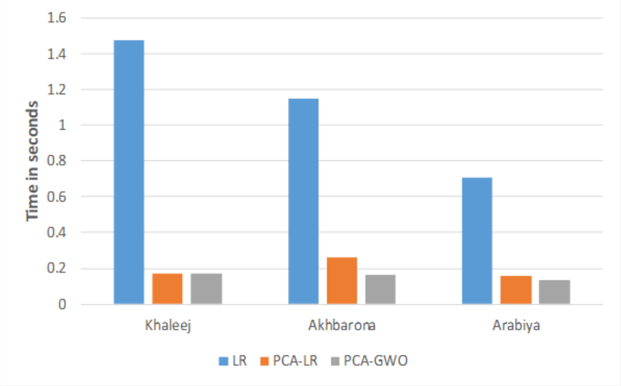


Figure 5: Comparison of computational time between LR, PCA-LR, and PCA-GWO on all datasets

E. EXPERIMENTAL RESULTS OF PCA-GWO AND OTHER DEEP LEARNING MODELS

As for deep learning (DL) models, we compare the proposed method with nine DL models that were proven to produce top results [15]. Table 7 shows the accuracy results for the 9 DL models versus the GWO-LR on the three experimented datsets. The results confirm that the GWO-LR outperforms the DL models on two datasets with scores 94.34 and 98.35 for Akhbarona and Arabiya, respectively. For the Khaleej dataset, GWO-LR produced slightly less accuracy score (96.86%) when compared to CGRU DL model (96.86%). Therefore, the performance of the our proposed method is at least comparable if not better. However, top performing DL models on the datasets are different. Table 8 provides further analysis between DL models and GWO-LR. It is clear that our proposed method is favoured over DL models with respect to size of dataset used (less than 1% of the number of samples in the original dataset), number of features (10% of the original set of features), and accuracy.

VI. CONCLUSIONS AN FURTHER DIRECTIONS

In this paper, we presents a hybrid filter-wrapper feature selection method for categorizing Arabic documents that combines PCA (filter approach) and GWO (wrapper approach). PCA is used to determine a robust feature subset that is more representative of the Arabic textual data when compared to using all features. GWO is optimized for the PCA feature subset to further select informative features. The LR classifieris used to perform classification for each feature subset produced by GWO. Three Arabic datasets Alkhaleej, Akhbarona, and Arabiya are experimented with to test the performance of the proposed PCA-GWO. The results obtained by PCA-GWO are superior to results produced by baseline classifiers with and without PCA feature selection method. We also compared GWO with other optimization feature selection algorithms. Namely, PSO, FFA, and BAT. As PCA-GWO confirmed its superiority as a feature selection method for the Arabic text classification task. However, similar to the most metaheuristic algorithms, GWO suffers from premature convergence and falling in local optima. As future work, GWO can be further enhanced, by empowering the wrapper approach via different strategies like i) hybridized GWO with other local-based approaches; ii) modifying its optimization framework by adding extra efficient and robust optimization search operators to provide more accurate results for Arabic text classification task.

References

[1] Laila Khreisat. A machine learning approach for arabic text classification using n-gram frequency statistics. Journal of Informetrics, 3(1):72–77, 2009.

[2] Fabrizio Sebastiani. Text categorization. In Encyclopedia of Database Technologies and Applications, pages 683–687. IGI Global, 2005.

[3] Selma Ayşe Özel. A web page classification system based on a genetic algorithm using tagged-terms as features. Expert Systems with Applications, 38(4):3407–3415, 2011.

[4] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4):1093–1113, 2014.

[5] Ali Bou Nassif, Abdollah Masoud Darya, and Ashraf Elnagar. Empirical evaluation of shallow and deep learning classifiers for arabic sentiment analysis. Transactions on Asian and Low-Resource Language Information Processing, 21(1):1–25, 2021.

[6] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature

Table 4: Comparison between PCA-GWO and other optimization feature selection approaches based on average classification accuracy

	Datasets					
Method	Khaleej		Akhbarona		Arabiya	
	Accuracy	#Features	Accuracy	#Features	Accuracy	#Features
PCA-BAT	93.01±1.20	599.7±53.81	85.49±1.78	592.5±59.92	94.87±2.74	543±54.95
PCA-PSO	93.15±3.70	662.6±16.73	89.53±0.68	683.55±24.78	**98.05**±0.40	620.65±28.15
PCA-FFA	91.66±0.60	572.65±20.09	87.41±1.05	581.1±26.47	97.06±0.78	529.05±12.96
PCA-WSO	92.84±0.39	641.42± 22.35	88.33±0.43	659.65±30.16	94.13± 0.65	595.00± 26.93
PCA-MPA	94.82 ±0.98	360.26±107.71	92.14±1.43	382.25±98.93	96.67±1.18	297.15±54.64
PCA-SMA	94.60±01.32	**15.79**±13.58	90.28±1.37	**108.80**±157.74	93.33±02.24	**27.65**±28.08
PCA-GWO	**95.24**±0.40	810.15±14.45	**93.54**±0.39	798.75±35.23	97.93±0.42	724.9±29.71

Table 5: Comparison between PCA-GWO and other optimization feature selection approaches based on the best classification accuracy

	Datasets					
Method	Khaleej		Akhbarona		Arabiya	
	Accuracy	#Features	Accuracy	#Features	Accuracy	#Features
LR	92.91	10040	92.92	8793	92.91	7242
PCA-LR	94.33	1010	92.91	1036	96.69	932
PCA-BAT	95.04	601	87.94	625	97.52	548
PCA-PSO	95.74	682	90.78	706	98.35	704
PCA-FFA	92.91	606	89.36	575	98.35	543
PCA-WSO	94.00	636	89.36	684	95.04	583
PCA-MPA	96.42	448	94.28	408	**99.16**	273
PCA-SMA	96.42	**20**	92.14	**11**	97.5	**37**
PCA-GWO	**96.45**	801	**94.33**	833	98.35	740

Table 6: Wilcoxon signed-rank test of PCA-GWO and other metaheuristic algorithms

	Datasets					
Method	Khaleej		Akhbarona		Arabiya	
	P-Value	Significance	P-Value	Significance	P-Value	Significance
PCA-BAT	0.0002	Yes	0.00008	Yes	0.00014	Yes
PCA-PSO	0.00528	Yes	0.00008	Yes	0.4654	No
PCA-FFA	0.00008	Yes	0.00008	Yes	0.00084	Yes
PCA-WSO	.00026	Yes	0.00008	Yes	0.01108	Yes
PCA-MPA	0.04036	Yes	0.0002	Yes	0.00068	Yes
PCA-SMA	0.00008	Yes	< .05	Yes	0.00008	Yes

Table 7: Results of comparison between PCA-GWO and other deep learning models

Model	Khaleej	Akhbarona	Arabiya
BIGRU	96.46	92.23	97.41
BILSTM	95.05	90.14	96.43
CGRU	**96.86**	94.00	97.19
CLSTM	96.59	92.66	96.97
CNN	96.33	92.72	95.62
GRU	96.04	89.56	96.76
HANGRU	96.66	92.95	96.00
HANLSTM	96.55	92.21	96.38
LSTM	94.09	90.29	96.54
GWO-LR	96.45	**94.33**	**98.35**

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3222516

Table 8: Summarization results comparison between PCA-GWO and other deep learning techniques in terms of dataset size, no. of features, and performance

criteria	DL techniques	GWO-LR
Dataset size	KH:45500 AB:46900 AR:18500	KH:1408 AB:1404 AR:1207
No. of features	All features (100%)	Subset (less than 15%)
Performance	KH: higher	AR: higher KH: comparable AB: higher

selection techniques in bioinformatics. bioinformatics, 23(19):2507–2517, 2007.

[7] Osama Ahmad Alomari, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, and Zaid Abdi Alkareem Alyasseri. A hybrid filter-wrapper gene selection method for cancer classification. In 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS), pages 113–118. IEEE, 2018.

[8] Osama Ahmad Alomari, Sharif Naser Makhadmeh, Mohammed Azmi Al-Betar, Zaid Abdi Alkareem Alyasseri, Iyad Abu Doush, Ammar Kamal Abasi, Mohammed A Awadallah, and Raed Abu Zitar. Gene selection for microarray data classification based on gray wolf optimizer enhanced with triz-inspired operators. Knowledge-Based Systems, 223:107034, 2021.

[9] Chunxia Zhang, Xindong Wu, Zhendong Niu, and Wei Ding. Authorship identification from unstructured texts. Knowledge-Based Systems, 66:99–111, 2014.

[10] Ashraf Elnagar, Sane M Yagi, Ali Bou Nassif, Ismail Shahin, and Said A Salloum. Systematic literature review of dialectal arabic: identification and detection. IEEE Access, 9:31010–31042, 2021.

[11] Thiago S Guzella and Walmir M Caminhas. A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7):10206–10222, 2009.

[12] Ismaila Idris and Ali Selamat. Improved email spam detection model with negative selection algorithm and particle swarm optimization. Applied Soft Computing, 22:11–27, 2014.

[13] Jianping Zeng and Shiyong Zhang. Variable space hidden markov model for topic detection and analysis. Knowledge-Based Systems, 20(7):607–613, 2007.

[14] TM Eldos. Arabic text data mining: a root-based hierarchical indexing model. International Journal of Modelling and Simulation, 23(3):158–166, 2003.

[15] Ashraf Elnagar, Ridhwan Al-Debsi, and Omar Einea. Arabic text classification using deep learning models. Information Processing & Management, 57(1):102121, 2020.

[16] Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. Sanad: Single-label arabic news articles dataset for automatic text categorization. Data in brief, 25:104076, 2019.

[17] Mahdieh Labani, Parham Moradi, and Mahdi Jalili. A multi-objective genetic algorithm for text feature selection using the relative discriminative criterion. Expert Systems with Applications, 149:113276, 2020.

[18] Hozayfa El Rifai, Leen Al Qadi, and Ashraf Elnagar. Arabic text classification: the need for multi-labeling systems. Neural Computing and Applications, 34(2):1135–1159, 2022.

[19] Hamouda K Chantar and David W Corne. Feature subset selection for arabic document categorization using bpso-knn. In 2011 Third World Congress on Nature and Biologically Inspired Computing, pages 546–551. IEEE, 2011.

[20] Yannis Haralambous, Yassir Elidrissi, and Philippe Lenca. Arabic language text classification using dependency syntax-based feature selection. arXiv preprint arXiv:1410.4863, 2014.

[21] Mohammed A Awadallah, Mohammed Azmi Al-Betar, Abdelaziz I Hammouri, and Osama Ahmad Alomari. Binary jaya algorithm with adaptive mutation for feature selection. Arabian Journal for Science and Engineering, 45(12):10875–10890, 2020.

[22] Ahmed M Anter, Mohamed Abd Elaziz, and Zhiguo Zhang. Real-time epileptic seizure recognition using bayesian genetic whale optimizer and adaptive machine learning. Future Generation Computer Systems, 127:426–434, 2022.

[23] Ahmed M Anter, Hany S Elnashar, and Zhiguo Zhang. Qmvo-scdl: A new regression model for fmri pain decoding using quantum-behaved sparse dictionary learning. Knowledge-Based Systems, 252:109323, 2022.

[24] Ahmed M Anter, Gan Huang, Linling Li, Li Zhang, Zhen Liang, and Zhiguo Zhang. A new type of fuzzy-rule-based system with chaotic swarm intelligence for multiclassification of pain perception from fmri. IEEE Transactions on Fuzzy Systems, 28(6):1096–1109, 2020.

[25] Ahmed M Anter, Diego Oliva, Anuradha Thakare, and Zhiguo Zhang. Afcm-lsma: New intelligent model based on lévy slime mould algorithm and adaptive fuzzy c-means for identification of covid-19 infection from chest x-ray images. Advanced Engineering Informatics, 49:101317, 2021.

[26] Ahmed M Anter and Mumtaz Ali. Feature selection strategy based on hybrid crow search optimization algorithm integrated with chaos theory and fuzzy c-means algorithm for medical diagnosis problems. Soft Computing, 24(3):1565–1584, 2020.

[27] Ahmed M Anter and Zhiguo Zhang. E-health parkinson disease diagnosis in smart home based on hybrid intelligence optimization model. In International Conference on Advanced Intelligent Systems and Informatics, pages 156–165. Springer, 2019.

[28] Ahmed M Anter, Yasmine S Moemen, Ashraf Darwish, and Aboul Ella Hassanien. Multi-target qsar modelling of chemo-genomic data analysis based on extreme learning machine. Knowledge-Based Systems, 188:104977, 2020.

[29] Mohamed Abu ElSoud and Ahmed M Anter. Computational intelligence optimization algorithm based on meta-heuristic social-spider: case study on ct liver tumor diagnosis. International Journal of Advanced Computer Science and Applications, 7(4), 2016.

[30] Chao-Ton Su and Jyh-Hwa Hsu. An extended chi2 algorithm for discretization of real value attributes. IEEE transactions on knowledge and data engineering, 17(3):437–441, 2005.

[31] Osama Ahmad Alomari, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, and Mohammed A Awadallah. A novel gene selection method using modified mrmr and hybrid bat-inspired algorithm with β-hill climbing. Applied Intelligence, 48(11):4429–4447, 2018.

[32] Newton Spolaôr, Everton Alvares Cherman, Maria Carolina Monard, and Huei Diana Lee. Relieff for multi-label feature selection. In 2013 Brazilian Conference on Intelligent Systems, pages 6–11. IEEE, 2013.

[33] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02):185–205, 2005.

[34] Harun Uğuz. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24(7):1024–1032, 2011.

[35] Anil Jain and Douglas Zongker. Feature selection: Evaluation, application, and small sample performance. IEEE transactions on pattern analysis and machine intelligence, 19(2):153–158, 1997.

[36] K Thirumoorthy and K Muneeswaran. Feature selection using hybrid poor and rich optimization algorithm for text classification. Pattern Recognition Letters, 147:63–70, 2021.

[37] Wojdan BinSaeedan and Salwa Alramlawi. Cs-bpso: Hybrid feature selection based on chi-square and binary pso algorithm for arabic email authorship analysis. Knowledge-Based Systems, page 107224, 2021.

[38] Yuhao Qing, Wenyi Liu, Liuyan Feng, and Wanjia Gao. Improved transformer net for hyperspectral image classification. Remote Sensing, 13(11):2216, 2021.

[39] Zaid Abdi Alkareem Alyasseri, Mohammed Azmi Al-Betar, Iyad Abu Doush, Mohammed A Awadallah, Ammar Kamal Abasi, Sharif Naser Makhadmeh, Osama Ahmad Alomari, Karrar Hameed Abdulkareem, Afzan Adam, Robertas Damasevicius, et al. Review on covid-19 diagnosis models based on machine learning and deep learning approaches. Expert systems, page e12759, 2021.

[40] Zaid Abdi Alkareem Alyasseri, Ahmad Tajudin Khader, Mohammed Azmi Al-Betar, Joao P Papa, Osama Ahmad Alomari, and Sharif Naser Makhadmeh. Classification of eeg mental tasks using multi-objective flower pollination algorithm for person identification. International Journal of Integrated Engineering, 10(7), 2018.

[41] Mohammed Azmi Al-Betar, Osama Ahmad Alomari, and Saeid M Abu-Romman. A triz-inspired bat algorithm for gene selection in cancer classification. Genomics, 112(1):114–126, 2020.

[42] Said Bahassine, Abdellah Madani, Mohammed Al-Sarem, and Mohamed Kissi. Feature selection using an improved chi-square for arabic text classification. Journal of King Saud University-Computer and Information Sciences, 32(2):225–231, 2020.

[43] Souad Larabi Marie-Sainte and Nada Alalyani. Firefly algorithm based feature selection for arabic text classification. Journal of King Saud University-Computer and Information Sciences, 32(3):320–328, 2020.

[44] Xun Li and Kwai Man Luk. The grey wolf optimizer and its applications in electromagnetics. IEEE Transactions on Antennas and Propagation, 2019.

[45] Xin Ma, Xie Mei, Wenqing Wu, Xinxing Wu, and Bo Zeng. A novel fractional time delayed grey model with grey wolf optimizer and its applications in forecasting the natural gas and coal consumption in chongqing china. Energy, 178:487–507, 2019.

[46] Mohammed Azmi Al-Betar, Mohammed A Awadallah, and Monzer M Krishan. A non-convex economic load dispatch problem with valve loading effect using a hybrid grey wolf optimizer. Neural Computing and Applications, pages 1–28, 2019.

[47] Qasem Al-Tashi, Helmi Md Rais, Said Jadid Abdulkadir, Seyedali Mirjalili, and Hitham Alhussian. A review of grey wolf optimizer-based feature selection methods for classification. In Evolutionary Machine Learning Techniques, pages 273–286. Springer, 2020.

[48] Xiaohu Yan, Yongjun Zhang, Dejun Zhang, and Neng Hou. Multimodal image registration using histogram of oriented gradient distance and data-driven grey wolf optimizer. Neurocomputing, 2020.

[49] Sharif Naser Makhadmeh, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, Syibrah Naim, Zaid Abdi Alkareem Alyasseri, and Ammar Kamal Abasi. A min-conflict algorithm for power scheduling problem in a smart home using battery. In Proceedings of the 11th national technical seminar on unmanned system technology 2019, pages 489–501. Springer, 2021.

[50] Sharif Naser Makhadmeh, Ammar Kamal Abasi, and Mohammed Azmi Al-Betar. Hybrid multi-verse optimizer with grey wolf optimizer for power scheduling problem in smart home using iot. The Journal of Supercomputing, pages 1–36, 2022.

[51] Zaid Abdi Alkareem Alyasseri, Osama Ahmad Alomari, Mohammed Azmi Al-Betar, Mohammed A Awadallah, Karrar Hameed Abdulkareem, Mazin Abed Mohammed, Seifedine Kadry, V Rajinikanth, and Seungmin Rho. Eeg channel selection using multiobjective cuckoo search for person identification as protection system in healthcare applications. Computational Intelligence and Neuroscience, 2022, 2022.

[52] Xiaoqiang Zhao, Shaoya Ren, Heng Quan, and Qiang Gao. Routing protocol for heterogeneous wireless sensor networks based on a modified grey wolf optimizer. Sensors, 20(3):820, 2020.

[53] Qusay M Alzubi, Mohammed Anbar, Zakaria NM Alqattan, Mohammed Azmi Al-Betar, and Rosni Abdullah. Intrusion detection system based on a modified binary grey wolf optimisation. Neural Computing and Applications, pages 1–13, 2019.

[54] Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6):275–285, 2004.

[55] S Santhosh Baboo and M Amirthapriya. Comparison of machine learning techniques on twitter emotions classification. SN Computer Science, 3(1):1–8, 2022.

[56] Shunjie Han, Cao Qubo, and Han Meng. Parameter selection in svm with rbf kernel function. In World Automation Congress 2012, pages 1–4. IEEE, 2012.

[57] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1):3–14, 2002.

[58] Rehane Hafezi Fard and Soodeh Hosseini. Machine learning algorithms for prediction of energy consumption and iot modeling in complex networks. Microprocessors and Microsystems, page 104423, 2021.

[59] S Al-Harbi, A Almuhareb, A Al-Thubaity, MS Khorsheed, and A Al-Rajeh. Automatic arabic text classification. 2008.

[60] Rehab M Duwairi. Machine learning for arabic text categorization. Journal of the American society for information science and technology, 57(8):1005–1010, 2006.

[61] Fouzi Harrag, Eyas El-Qawasmeh, and Pit Pichappan. Improving arabic text categorization using decision trees. In 2009 First International Conference on Networked Digital Technologies, pages 110–115. IEEE, 2009.

[62] Riyad Al-Shalabi, Ghassan Kanaan, and M Gharaibeh. Arabic text categorization using knn algorithm. In Proceedings of The 4th International Multiconference on Computer Science and Information Technology, volume 4, pages 5–7, 2006.

[63] Mohammad S Khorsheed and Abdulmohsen O Al-Thubaity. Comparative evaluation of text classification techniques using a large diverse arabic dataset. Language resources and evaluation, 47(2):513–538, 2013.

[64] Mostafa M Syiam, Zaki T Fayed, and Mena B Habib. An intelligent system for arabic text categorization. International Journal of Intelligent Computing and Information Sciences, 6(1):1–19, 2006.

[65] Abdelwadood Moh'd Mesleh and Ghassan Kanaan. Support vector machine text classification system: Using ant colony optimization based feature subset selection. In 2008 International Conference on Computer Engineering & Systems, pages 143–148. IEEE, 2008.

[66] Bilal M Zahran and Ghassan Kanaan. Text feature selection using particle swarm optimization algorithm 1. 2009.

[67] Mayy M Al-Tahrawi and Sumaya N Al-Khatib. Arabic text classification using polynomial networks. Journal of King Saud University-Computer and Information Sciences, 27(4):437–449, 2015.

[68] Ghazi Raho, Riyad Al-Shalabi, Ghassan Kanaan, and Asmaa Nassar. Different classification algorithms based on arabic text classification: feature selection comparative study. International Journal of Advanced Computer Science and Applications Ijacsa, 6(2):23–28, 2015.

[69] Abdullah Saeed Ghareb, Azuraliza Abu Bakar, and Abdul Razak Hamdan. Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Systems with Applications, 49:31–47, 2016.

[70] Musab Hijazi, Akram Zeki, and Amelia Ismail. Arabic text classification using hybrid feature selection method using chi-square binary artificial bee colony algorithm. Computer Science, 16(1):213–228, 2021.

[71] Mohammad Tubishat, Salinah Ja'afar, Norisma Idris, Mohammed Azmi Al-Betar, Mohammed Alswaitti, Hazim Jarrah, Maizatul Akmar Ismail, and Mardian Shah Omar. Improved sine cosine algorithm with simulated annealing and singer chaotic map for hadith classification. Neural Computing and Applications, pages 1–22, 2021.

[72] Muhammad Fahad Umer and M Sikander Hayat Khiyal. Classification of textual documents using learning vector quantization. Information Technology Journal, 6(1):154–159, 2007.

[73] Fouzi Harrag, Eyas El-Qawasmah, and Abdul Malik S Al-Salman. Stemming as a feature reduction technique for arabic text categorization. In 2011 10th international symposium on programming and systems, pages 128–133. IEEE, 2011.

[74] Seyedali Mirjalili, Seyed Mohammad Mirjalili, and Andrew Lewis. Grey wolf optimizer. Advances in engineering software, 69:46–61, 2014.

[75] Ian H Witten, E Frank, MA Hall, and CJ Pal. The weka workbench. online appendix for "data mining: practical machine learning tools and techniques". In Morgan Kaufmann. 2016.

[76] Catarina Silva and Bernardete Ribeiro. The importance of stop word removal on recall values in text categorization. In Proceedings of the International Joint Conference on Neural Networks, 2003., volume 3, pages 1661–1666. IEEE, 2003.

[77] Arnaz Malhi and Robert X Gao. Pca-based feature selection scheme for machine defect classification. IEEE transactions on instrumentation and measurement, 53(6):1517–1525, 2004.

[78] J. Ross Quinlan. Induction of decision trees. Machine learning, 1(1):81–106, 1986.
[79] Fred J Damerau, Tong Zhang, Sholom M Weiss, and Nitin Indurkhya. Text categorization for a comprehensive time-dependent benchmark. Information processing & management, 40(2):209–221, 2004.
[80] HongFang Zhou, JiaWei Zhang, YueQing Zhou, XiaoJie Guo, and YiMing Ma. A feature selection algorithm of decision tree based on feature weight. Expert Systems with Applications, 164:113842, 2021.
[81] Wen Zhang, Taketoshi Yoshida, and Xijin Tang. Text classification based on multi-word with support vector machine. Knowledge-Based Systems, 21(8):879–886, 2008.
[82] Hariom Tatsat, Sahil Puri, and Brad Lookabaugh. Machine learning and data science blueprints for finance, 2021.
[83] Mohammed Lataifeh, Ashraf Elnagar, Ismail Shahin, and Ali Bou Nassif. Arabic audio clips: Identification and discrimination of authentic cantillations from imitations. Neurocomputing, 418:162–177, 2020.
[84] Xin-She Yang. A new metaheuristic bat-inspired algorithm. In Nature inspired cooperative strategies for optimization (NICSO 2010), pages 65–74. Springer, 2010.
[85] Xin-She Yang. Nature-inspired metaheuristic algorithms. Luniver press, 2010.
[86] Sharif Naser Makhadmeh, Ahamad Tajudin Khader, Mohammed Azmi Al-Betar, Syibrah Naim, Zaid Abdi Alkareem Alyasseri, and Ammar Kamal Abasi. Particle swarm optimization algorithm for power scheduling problem using smart battery. In 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT), pages 672–677. IEEE, 2019.
[87] Malik Braik, Abdelaziz Hammouri, Jaffar Atwan, Mohammed Azmi Al-Betar, and Mohammed A Awadallah. White shark optimizer: A novel bio-inspired meta-heuristic algorithm for global optimization problems. Knowledge-Based Systems, 243:108457, 2022.
[88] Afshin Faramarzi, Mohammad Heidarinejad, Seyedali Mirjalili, and Amir H Gandomi. Marine predators algorithm: A nature-inspired metaheuristic. Expert systems with applications, 152:113377, 2020.
[89] Denise Rey and Markus Neuhäuser. Wilcoxon-signed-rank test. In International encyclopedia of statistical science, pages 1658–1659. Springer, 2011.