algorithms

MDPI

Article

Nature-Inspired Optimization Algorithms for Text Document Clustering—A Comprehensive Analysis

Laith Abualigah [1], Amir H. Gandomi [2,*], Mohamed Abd Elaziz [3], Abdelazim G. Hussien [4], Ahmad M. Khasawneh [1], Mohammad Alshinwan [1] and Essam H. Houssein [5]

1 Faculty of Computer Sciences and Informatics, Amman Arab University, Amman 11953, Jordan; Aligah.2020@gmail.com (L.A.); a.khasawneh@aau.edu.jo (A.M.K.); mohmdsh@aau.edu.jo (M.A.)
2 Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia
3 Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt; abd_el_aziz_m@yahoo.com
4 Faculty of Science, Fayoum University, Faiyum 63514, Egypt; aga08@fayoum.edu.eg
5 Faculty of Computers and Information, Minia University, Minia 61519, Egypt; essam.halim@mu.edu.eg
* Correspondence: Gandomi@uts.edu.au

Received: 29 November 2020; Accepted: 14 December 2020; Published: 18 December 2020

Abstract: Text clustering is one of the efficient unsupervised learning techniques used to partition a huge number of text documents into a subset of clusters. In which, each cluster contains similar documents and the clusters contain dissimilar text documents. Nature-inspired optimization algorithms have been successfully used to solve various optimization problems, including text document clustering problems. In this paper, a comprehensive review is presented to show the most related nature-inspired algorithms that have been used in solving the text clustering problem. Moreover, comprehensive experiments are conducted and analyzed to show the performance of the common well-know nature-inspired optimization algorithms in solving the text document clustering problems including Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), and Bat-inspired Algorithm (BA). Seven text benchmark datasets are used to validate the performance of the tested algorithms. The results showed that the performance of the well-known nurture-inspired optimization algorithms almost the same with slight differences. For improvement purposes, new modified versions of the tested algorithms can be proposed and tested to tackle the text clustering problems.

Keywords: nature-inspired; optimization algorithms; machine learning; optimization problems; text clustering applications

1. Introduction

Since the number of digital documents is increasing excessively, an automatic organization (without human intervention/interaction) of such documents is vital and valuable [1]. Document/Text clustering can be defined as the diving process of documents pool based on their similarity to distinct subclasses, often called clusters. Text clustering can be considered a useful and essential technique [2]. It has been extensively excessively employed ineffective organization, extraction navigation, retrieval, and summarization of the massive volume of texts/documents.

We can generally classify the clustering algorithm into two categories: (1) hierarchical clustering; (2) partitional clustering. The first category (hierarchical clustering) can be divided into two subcategories: agglomerative and divisive [3,4]. In the proper subcategory (agglomerative), the cluster

used the bottom-up technique, which begin by considering every single document as a single cluster and after that tries to combine them in a large cluster [5]. In contrast, in the latter subcategory (divisive), the cluster operates using the top-down technique that initially begins with one cluster that contains all documents. After that, it split them into smaller clusters. In the second category (partitional clustering), algorithms try to classify the document to non-hierarchical disjoint clusters [6]. The most famous example of a partitional clustering algorithm is a center-based clustering algorithm. Each cluster is shown as a cluster center, such as the k-means algorithm. Its goal is to attain cluster by decreasing the sum of Euclidean distance between the cluster center, and the object [7,8].

Hierarchical methods are better than partitioning algorithms in clustering quality; however, they do not reallocate poorly classified objects. Their time complexity is quadratic of data objects number [9,10]. The partitioning clustering methods have gained a broad interest due to their advantages in handling large datasets such as time complexity [11].

Metaheuristics algorithms have been successfully used to solve versions optimization problems such as wind farm decision system [12], vehicle routing problem [13], industry applications [14], feature selection [15], parameter control [16], and social aware cognitive radio handovers [17]. Metaheuristics algorithms can be defined into two groups: (1) single-based algorithms such as hill climbing and simulated annealing. (2) Population-based algorithms such as Grey Wolf Optimizer (GWO), Grasshopper Optimization Algorithm, Harris Hawks Algorithm, and Henry gas solubility optimization [18,19]. One of the most challenging issues in metaheuristics is the balancing between exploration and exploitation. To have a good algorithm performance and reasonable outcome, the search algorithm should make trade-offs between two strategies [20,21]. In order to triumph over traditional clustering limitations, different methods and concepts have been shown recently. Different machine learning methods that is based on clustering techniques have been proposed such as Graph Theory [22], Artificial Intelligence network [23,24], statistics [25], and evolutionary algorithms [26,27]. The most popular technique is using optimization techniques with a pre-defined clustering fitness function.

Many metaheuristics algorithms have been employed in solving clustering problems such as Genetic Algorithms (GAs) [28], Particle Swarm Optimization [29], Ant Colony Optimization [30], Whale Optimization Algorithm [31,32], Lightning Search Algorithm [33], Crow Search Algorithm [34], Ant Lion Optimization [35], Moth-flame Optimization [36], Gray Wolf Optimization (GWO) hybridized with Grasshopper Optimization Algorithm (GOA) [37], Artificial Bee Colony [38,39], Salp Swarm Algorithm [40], and Arithmetic Optimization Algorithm [41]. Several surveys that reviewed the metaheuristic optimization methods in versions applications, including the big data text clustering, are found in References [42–44]. Moreover, many classification algorithms have been used also in text clustering applications including k-Nearest Neighbors [45], a set theory [46], similarity measures [47], Support Vector Machine [48], a fuzzy self-constructing [49], and ensemble scheme [50].

Nature-inspired optimization algorithms, are the optimization algorithms inspired by whence nature changes to challenging situations, have been successfully utilized to address various optimization problems, including text document clustering problems. In this paper, a comprehensive survey is given to show the most related nature-inspired algorithms that have been used to solve the text clustering problem. Moreover, comprehensive experiments are conducted and analyzed to show the performance of the common well-know nature-inspired optimization algorithms in solving the text document clustering problems, including Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), and Bat-inspired Algorithm (BA). Several evaluation measures evaluate the tested algorithms in solving the text clustering problems, including Precision, Recall, F-measure, Entropy, Purity, and Accuracy. Seven text benchmark datasets are used to validate the performance of the tested algorithms. The results showed that the performance of the well-known nurture-inspired optimization algorithms almost the same with slight differences. For improvement purposes, new modified versions of the tested algorithms can be proposed and

tested to tackle the text clustering problems. This paper can be a useful reference for future research. It can help the researcher find the published clustering methods using versions of nature-inspired optimization methods.

The main sections of this paper are prepared as follows—Section 2 gives the most related nature-inspired algorithms that have been used in solving clustering problems. In Section 3, the main ideas of the text clustering problem are shown. Evaluation criteria employed in text clustering applications are presented in Section 4. Experimental results and discussion are given in Section 5. Finally, the conclusion and possibilities for further research are given in Section 6.

2. Related Works

In this section, the most nature-inspired clustering algorithms and well-known clustering techniques are presented.

2.1. Krill Herd Algorithm (KHA)

The Krill herd algorithm is a swarm-based optimization that mimics the krill group behavior [51]. In this algorithm, the exploration and exploitation are balanced by complementing the global wide-range strength of nearby local searching. In Reference [52], a new text clustering approach introduced based on the enhancement of KHA hybrid with multiple criteria function, the improved approach called MMKHA. Furthermore, six versions of the KHA algorithm were adopted for solving the text clustering problem. MMKHA compared with other algorithms (i.e., particle swarm optimization, genetic algorithm, harmony search, and k-mean) and shows an outperformance in text document clustering.

Abualigah [53,54] proposed combinations of the harmony search (HS) method and the KHA algorithm to enhance the global search capability, namely H-KHA. The improvement includes adding the HS global search operator to the KHA to enhance the exploration process. The H-KHA approach shows superior performance in terms of high convergence rate and accurate clusters. In Reference [55], two approaches for web text document clustering are introduced based on the KHA algorithm. The first approach adopted the basic KHA, while the second approach uses the genetic operators instead of the KHA operators. The experimental result shows a superior performance comparing to the K-mean algorithm.

Abualigah [56] proposed a novel method for text document clustering, which built based on two levels: Firstly, a PSO algorithm proposed for feature selection with a new weighting system, and a particular dimension reduction method, in terms of selecting a new subset of the most informative features with less dimensional space. This subset is utilized to enhance the performance and decrease the computation time of the text clustering algorithm. The k-mean method is used to assess the efficacy of the selected subsets. Secondly, four KHA algorithms are proposed for text clustering, namely, basic KHA, hybrid KHA, modified KHA, and multi-objective hybrid KHA, each of which describes an incremental enhancement on its predecessor. The results show that the proposed method got better results compared to other methods.

2.2. Magnetic Optimization Algorithm (MOA)

In Reference [57], a novel method for text clustering is proposed based on the Magnetic optimization algorithm (MOA), called MOAC. This approach aims to select the best position of centroid particles, which is considered as the optimal cluster position. The numerical result shows an efficient accuracy and robust text clustering comparing to other algorithms such as K-means and PSO.

2.3. Particle Swarm Optimization (PSO)

The particle swarm optimization (PSO) algorithm imitates the swarm social behavior, which refers to the group of applicant solutions, and each solution is considered a particle. In References [58,59], a new method for feature selection using the PSO algorithm, namely FSPSOTC , aims to solve the

problem of feature selection in text clustering by building a new subset of informative features. These feature subset can reduce the computational time and enhance the clustering performance. The proposed method shows superior performance in feature selection comparing to the feature selection in text clustering by harmony search algorithm and genetic algorithm.

In Reference [60], a hybridization of the Spectral Clustering algorithm and PSO algorithm for text clustering, namely, SCPSO. The randomization is implemented with the initial population by considering the local and global optimization methods. The performance of the SCPSO compared with the basic PSO algorithm and K-means algorithm and show high accuracy in text clustering.

A novel approach is proposed in Reference [61] for text clustering based on the PSO algorithm and latent semantic indexing (PSO + LSI). To achieve better performance in the search space the adaptive inertia weight (AIW) is used to balance the exploration and exploitation. The proposed approach proves a high performance comparing to other algorithms such as PSO with k-means.

In Reference [62], a hybridization between binary PSO with a chaotic map, fitness based dynamic inertia weight, opposition-based learning, and mutation, is introduced for feature selection in text clustering. Further, the opposition-based algorithm used, to begin with, diversified solutions and a set of promising to accomplish the best result. Besides, the opposite position of the global best particle (gbest) generates by the opposition based learning algorithm.

A new method is proposed in Reference [63] to improve the feature selection process using the PSO algorithm. An improved PSO version was introduced to optimize the feature selection based on constant constriction factor and functional inertia weight. Then, a functional constriction factor is created based on the constant constriction factor, and combine this factor into the classic PSO method. Moreover, two improved PSO models were proposed called synchronously and asynchronously PSO. Finally, the chi-square test applies for feature selection. The proposed asynchronous approach shows a significant result in terms of the balance of various dimensions and text classification.

2.4. Social Spider Optimization (SSO)

In Reference [64], a new method based on the SSO algorithm is proposed. SSO utilizes the cooperative intelligence of spider social groups. According to spider gender, every spider tries to regenerate specific behavior and decreasing the local particles problems. This approach compared with K-means and show significant results.

In Reference [65], a novel approach based on the Social Spider Optimization (SSO) algorithm is proposed for text clustering. This method is compared with several algorithms such as Bee Colony, PSO, and Ant colony, and show improved performance. Also, two-hybrid algorithms are presented based on SSO and K-means algorithms called SSO + K-means and K-means + SSO. The SSO + K-means method shows an outperformance in clustering accuracy.

2.5. Whale Optimization Algorithm (WOA)

In Reference [66], a combination of Whale Optimization Algorithm (WOA) and Fuzzy C Means (FCM) is proposed for text clustering, called FCMWOA. WOA chose the cluster center, which supports the FCM algorithm to achieve a more reliable text clustering. The performance of this approach compared with K-mean + FCM and PSO + FCM, the FCMWOA shows an outperformance in term average of F-measure result.

In Reference [67], a modified WOA (MWOA) and FCM algorithm for text clustering. Further, an automobile insurance fraud detection system (AIFDS) is introduced, and the MWOA employs to find the optimal cluster centroids in AIFDS. In the AIFDS, the outliers were removed from the dataset by the FCM model. Then, the processed dataset is undergone by different classifiers such as Random Forest, XGBoost, LightGBM, Decision Tree, and CATBoost. The combination of MWOA, FCM, and CATBoost shows superior performance.

2.6. Ant Colony Optimization (ACO)

In order to enhance the fuzzy document clustering issues, an ant colony optimization (ACO) algorithm is proposed in Reference [68]. To extract the features of the documents and to achieve a language-independent vector representation, a specialized glossary and a thesaurus are employed, which can be used to calculate similarities among documents expressed in different languages. In order to identify membership values in a fuzzy cluster, the pheromone trails acquired in the ACO process are utilized. It was implemented to a corpus of bilingual documents in various economic and management areas to demonstrate the nature of the method.

Multi-label text categorization plays an important role in feature optimization and selection. Several features share a similar class in multi-label text categorization, and the classification process encountered a problem of selecting the relevance function for the classification. In Reference [69], feature optimization based on multi-label text categorization has been suggested using ant colony optimization. The optimization of the ant colony accumulated from document to classify the related common function. The cluster mapping classification method is employed in the classification stage. The process of feature optimization decreases data loss during the conversion of feature mapping in the classification. A typical dataset such as web page data, medical search data and RCV1 dataset was utilized for the performance evaluation of the suggested algorithm. Results demonstrated that the suggested technique is outperformed fuzzy relevance other classification techniques.

A new version of ant colony optimization (ACO) called enRiched Ant Colony Optimization (RACO) is introduced in Reference [70]. In the earlier executions, this modification attempts to consider the previous crossed edges to modify the pheromone values correctly and avoid premature convergence. Feature selection (FS) is the process of selecting similar features or ignoring irrelevant data features. RACO is also applied to the feature selection problem in order to demonstrate the efficiency of the suggested method. It could be assumed in the RACO-based feature selection (RACOFS) method that subsequent features with a higher priority are considered by the proposed method. Therefore, in another example, the algorithm is implemented with a local search procedure capability to prove that this is not the issue. The enhanced RACO method is capable of finding optimal solutions globally but is stuck in local optima. Thus, the method is integrated with a local search method in the third variation to address this problem by looking for the similarity of the optimal solution globally. Experiments were performed using two metrics, kappa statistics and classification accuracy, on several standard datasets to evaluate the efficiency of the suggested methods. The results were compared with a wide range of other swarm-based methods and different feature selection techniques. The findings show that the proposed methods have superiority over related works.

2.7. Local Search Techniques

In terms of increasing the amount of text information, it has become extremely difficult to deal with text information. Text clustering is an effective tool employed to manage a large number of text documents by classifying such text documents into clusters. In the end, it is challenging to cluster text documents with sparse, non-uniform distribution and uninformative features. The selection of the text function is a key unsupervised learning approach used to select a new subset of information text features. In Reference [71], a novel technique is suggested in based on the β-hill climbing technique for the problem of selecting text features to enhance text clustering (B-FSTC). The performance of the introduced β-hill climbing and original Hill climbing (i.e., H-FSTC) method are analyzed and compared with other techniques using k-mean text clustering. Experiments were carried out on four typical text datasets with different characteristics. Intriguingly, by generating a new subset of information text features, the suggested β-hill climbing method verifies better outcomes in comparison with the other techniques. Finally, to accomplish more accurate clusters, the β-hill climbing-based feature selection approach supports the k-mean clustering algorithm.

In Reference [72], a new local search strategy, called β-hill climbing technique is presented to tackle the text document clustering issue. The β-hill climbing technique's main invention is β. A balance

among local and global search has been implemented. Local search (exploitation) techniques are applied effectively as the k-mean to the issue of text document clustering. Experiments were carried out on five randomly taken benchmark text datasets from "Dmoz-Business" datasets with different features. The findings demonstrate that, compared to the original hill climbing technique calculated by F-measure, precision, recall, and accuracy, the suggested β-hill climbing obtained better outcomes. The results indicate that by tuning the parameter of the β-hill claiming, the suggested (β-hill climbing) achieved better results compared to other original technique.

In Reference [73], a cognitive-inspired multi-objective automated document clustering framework is suggested, which is a combination of the self-organizing map (SOM) and multi-objective differential evolution approach. In different population solutions, the variable number of cluster centers is encrypted in calculating the number of clusters from the dataset in an automated process. During evolution, these solutions undergo different genetic operations. In developing new genetic algorithm for the suggested clustering strategy, the idea of SOM is used. Two cluster validity measures, the Pakhira-Bandyopadhyay-Maulik index, and the Silhouette index, are optimized simultaneously to calculate the performance of a clustering solution. The efficiency of the introduced approach, including self- organizing map based multi-objective document clustering technique (SMODoc clust) is seen in automatic classification of some scientific papers and web-documents. The findings obtained clearly indicate that the proposed strategy is better than current strategies. The validation of the findings obtained is also shown by means of statistically relevant t-tests.

In Reference [74], a new local clustering technique, called β-hill climbing, have been proposed to address the issue of the clustering of text documents by modeling the β-hill climbing method to partition related documents into the same cluster. The key innovation in the β-hill climbing technique is the β factor. It was implemented to manage local and global search. In order to overcome the issue of text documents clustering such as k-medoid and k-mean techniques, local search techniques are effectively used. Experiments have been carried out using eight text datasets with different features. The findings showed that the suggested β-hill climbing obtained better outcomes in addressing the problem of text clustering compared to the previous hill climbing method.

2.8. Bee Colony Optimization (BCO)

One of the latest swarm intelligence (SI) methods, the conventional bee colony optimization (BCO) method, is strong at investigation while being poor at manipulation. In Reference [75], a new method, dubbed as weighted BCO (wBCO), is presented to enhance the optimization ability of BCO that enables bees to check deliberately in the solution space while considered policies to heuristically share information obtained about the food sources. For this reason, for each food source, wBCO calculates global and local weights, where the former is the level of popularity of a specific food source in the swarm and the latter is the relevance to a category classification of a food source. it implemented new policies in the recruiter choice process to ensure that uncommitted bees obey the most comparable committed ones in order to maintain population diversity. The local food source weighting and recruiter strategic objectives therefore make the method acceptable for problems of discrete optimization. The feature selection (FS) problem is formulated as a discrete optimization task to illustrate the effectiveness of wBCO and has been solved by the introduced method. The efficiency of wBCO are evaluated using standard benchmark optimization methods and datasets. Results demonstrated that wBCO outperformed state-of-the-art techniques.

In Reference [39], an improved bee colony optimization algorithm, dubbed IBCO was presented by implementing cloning and fairness principles into the BCO method and making it more effective for data clustering. These features provide BCO with good exploration and exploitation abilities to direct the search method to the proximity of high-quality solutions effectively. In general, when creating new solutions, the cloning feature enables it to take advantage of experiences learned from previous generations. In the suggested version, the problem of getting trapped in a local optimum remained open. Thus, authors hybridize it with the k-means method to take advantage of the fine-tuning power

of the commonly utilized k-means method, which showed good results in local searches, to address the lack of this swarm method in the local search. moreover, it proposed four IBCO-based hybridized methods and k-means techniques and analyze their clustering results and convergence behavior. it empirically showed that for large and high-dimensional datasets such as document clustering, the proposed hybrid methods mitigate the problem of sticking into a local solution. Particularly in comparison to k-means and other evolutionary clustering algorithms, including genetic, particle swarm optimization, ant colony, and bee-based algorithms, the experiments demonstrated that the introduced algorithms are stable and can be employed in different application.

2.9. Generic Algorithm (GA)

Text clustering method is an effective technique employed to partition a large number of text documents into sets. The size of documents impacts the text clustering by reducing its efficiency. Text documents subsequently have sparse and uninformative factors that minimize the efficiency of the method for the underlying text cluster and increase the computational time. Feature Selection is a basic unsupervised learning algorithm utilized to choose a new subset of informative text features to enhance text clustering efficiency and minimize computational time. For the feature selection problem, Reference [76] proposed a hybrid particle swarm optimization algorithm with genetic operators. The k-means clustering is employed to determine the efficiency of the subsets of features acquired. The experiments were performed out using eight main text datasets with different characteristics. The findings demonstrated that by establishing a new subset of more informative features, the introduced hybrid algorithm (H-FSPSOTC) enhanced the efficiency of the clustering algorithm. Finally, the feature selection method encouraged the clustering technique to achieve accurate clusters.

In information retrieval systems (IRs), genetic algorithms are primarily employed to improve the process of information retrieval and to improving the effectiveness of optimal information retrieval to meet the needs of users and to help them achieve what they need among the increasing amount of possible information. Improving adaptive genetic algorithms helps accurately retrieve the user's necessary information, reduces the relevant files retrieved and removes irrelevant files. In Reference [77], the way of selecting mutation likelihood and fitness function, and chose the mathematics test collection Cranfield English Corpus is discussed. This collection was performed by Cyrial Cleverdon and employed for simulation purposes at the University of Cranfield in 1960, containing 1400 documents and 225 queries. In order to compute similarity between the query and records, it also employed cosine similarity and jaccards and used two suggested adaptive fitness functions, mutation operators as well as adaptive crossover. The purpose of the process was to examine the efficiency of the findings according to precision and recall metrics. Finally, this work concluded that by using adaptive genetic algorithms, it could have many enhancements.

2.10. Harmony Search (HS)

Clustering analysis is affected by the increasing quantity of text information on Internet web pages. Text clustering is a constructive method of analysis used to divide a large amount of data into clusters. Therefore, the key issue affecting the methodology of text clustering is the inclusion of uninformative and sparse features in text documents. The feature selection (FS) is an important unsupervised method employed to remove uninformative features in effort to enhance text clustering method. Meta-heuristic techniques have recently been successfully implemented to analyze many optimization issues. In Reference [59], Harmony Search (HS) method to overcome the issue of feature selection (FSHSTC) have been proposed. The suggested approach is used to improve the text clustering (TC) technique by acquiring a new subset of informative or useful features. Experiments have been conducted using four datasets. The findings demonstrated that the designed FSHSTC increases the efficiency of the F-measure and Accuracy calculated by k-mean clustering algorithm.

The Harmony Search Algorithm (HSA) is an algorithm for optimizing swarm intelligence that has been applied to a wide variety of clustering systems, including data clustering, text clustering, fuzzy clustering, image processing, and wireless sensor networks. In Reference [42], a study of the literature on HSA and its variants, examine its strengths and limitations, and propose potential directions for research have been proposed.

In current application areas, such as the World Wide Web, clustering has become an extremely important and highly complex research field for tracking relevant and useful information. Previous works shown that the K-means method, the most widely used partitioning-based clustering technique, is more efficient for big datasets. However, a local optimal clustering can be generated by the K-means method. In Reference [78], a new document clustering algorithms based on the Harmony Search (HS) optimization method have been proposed. first, it proposed a pure HS based clustering algorithm by modeling clustering as an optimization problem, which identifies near-optimal clusters within a reasonable time. Second, harmony clustering is combined in three ways with the K-means method to reach better clustering by integrating HS's exploratory power with the K-means' refining power. In comparison to the K-means method's localized search property, the introduced algorithms conducted a globalized search throughout the solution space. In addition, by having it less dependent on input values such as randomly selected initial cluster centers, the enhanced algorithms strengthen K-means, thereby making it more robust. By modelling its population variability as a Markov chain, the behavior of the introduced method is analyzed. authors also performed experimental analysis to determine the effects of different parameters on the performance of the algorithms' clusters and convergence behavior. In the tests, on five different datasets, it applied the introduced methods along with K-means and a Genetic Algorithm (GA) based clustering algorithm. Experimental findings showed that better clusters can be identified by the proposed algorithms and the output of clusters is reasonable based on F-measure, Entropy, Purity, and Average Distance of Documents to the Centroid cluster (ADDC).

text clustering is used in several fields such as text mining, data processing, pattern recognition, image clustering. This chapter [79] demonstrated the methodology projected hybrid function section technique assisted by the harmony search principle for the method of text agglomeration. Irregular harmony memory is generated by the harmony search principle, which includes a list of candidate solutions. After creating a new range of decisions, the chapter showed the steps of the content cluster system by updating the concordance look algorithm based to enhance the operation of the content cluster, enhance the k-mean algorithm, and modify the mass and similarity of the cluster centers. The predicted content aggregation of k-means with the highlight selection approach was better than the content aggregation strategy of k-means.

One of the most effective approaches to mining and collecting information from the web is called cluster web documents. Recently, the learning and optimization methods have been oriented towards one of the most attractive patterns in clustering high-dimensional web pages. In Reference [80], a new hybrid harmony search (HS) based algorithms to cluster web documents which find a globally optimal partition into a given clusters. First, it proposed a pure harmony-based search-based clustering algorithm by modeling clustering as an optimization problem that identifies optimal clusters approximately global within a reasonable time. Second, K-means and harmony clustering are hybridized in two ways to achieve effective clustering. Experimental results indicated that when compared to similar techniques, the introduced algorithms can find good clusters and also demonstrated the effectiveness of the hybrid clustering methods.

2.11. K-Means Clustering Technique

The K-mean algorithm is a simple, robust, and rabid local search method used for text clustering. In Reference [81], an improved approach multi-objective combining the similarity and distance measure based on K-means algorithm called (MKM). The k-means algorithm is used for text clustering to evaluate the performance of the multi-objective method. Seven different datasets are used to perform

experiments. Results indicate that the proposed method shows an outperformance comparing with other measures.

Reference [82] proposed a new method to solve the text clustering issue called length feature weight (LFW). This approach enhances the text clustering by providing a fair weighting value of the most important features for each document. Further, the LFW method is proposed to improves the β-hill climbing algorithm. This approach shows great performance with three main factors max-term frequency, document frequency, and outer terms that are not included in the common weight systems.

In Reference [83], a combination of three features algorithms for text document clustering based LFW approach, the harmony search (HS), swarm optimization (PSO), and Genetic algorithm (GA), with dynamic dimension reduction and weight scheme. The important informative features are chosen by specific evaluation algorithms to classify the text documents. To reduce the number of features a novel dynamic dimension reduction (DDR) approach is proposed. The k-mean algorithm proposed to cluster the text documents depend on the features selected through dynamic reduction. The experimental on several datasets shows optimal outcomes with PSO, LFW, and DDR.

In Reference [84], analysis of three methods PSO, K-means algorithm, and hybrid algorithm of PSO and K-means for text clustering. The bag of terms used for describing the text documents, which cannot utilize the semantics. Texts are defined based on synsets matching to a word. This approach implemented in the Nepali language and the experimental evaluation is achieved using inter and intra cluster similarity.

2.12. Other Algorithms

In Reference [38], the ABC search equation is improved and integrate two local search paradigms into the standard ABC, called chaotic local search and gradient search to enhance its performance. The proposed method is called chaotic gradient artificial bee colony. Three separated benchmark text datasets, namely Reuters-21,578, Classic4, and WebKB, are evaluated the efficiency of the introduced method. The results obtained are contrasted with ABC, a recent variant of ABC, including gbest-guided ABC, a variant of the methodology suggested, called chaotic artificial bee colony, memetic ABC, and K-means standard clustering technique. The experimental results show promising outcomes in terms of solution consistency and speed of convergence.

In Reference [85], an approach using fuzzy logic techniques and self-organizing maps (SOM) is presented to handle conceptual aspects in document clusters and to minimize training time, a concept frequency formula is applied in order to calculate the degree of presence of a concept in a document. For the calculation of the polysemic degree of terms and the synonymic degree between terms, this formula is focused on new fuzzy equations. New fuzzy enhancements such as automatic topology selection, initialization of the heuristic map, a fuzzy similarity measure and an extraction method for keywords are employed in this method. In order to compare the introduced method with classic SOM approaches via Reuters selection, some experimental tests were carried out. The efficiency of the method has been evaluated in terms of F-measurement and training time. The experiments revealed that, compared to classic SOM strategies, the introduced method achieves satisfactory results with less training time.

Text analysis involves complex techniques for managing different text documents in the field of text mining. machine recruitment and pattern recognition. Computers may begin to organize a corpus document using rational text-clustering methods in some organizational frameworks of conceptual clusters. noisy, inconsequential and superfluous characteristics are included in the informative and un-informative functionality of text documents. The unsupervised selection of text features is the key method of determining a new subset of informative features for each document. There are two goals of the functional selection technique: (1) optimize the reliability of the text clustering algorithm, (2) minimize the number of uninformative traits. In Reference [37], a new technique proposed is that it achieves a mature convergence rate and needs minimal computational time and is stuck in a low dimensional space in local minima. As the input and pre-processing steps are conducted in

the document, the text data is fed. Next, by selecting the local optima from the text document and then choosing the proper global optima from the local optimum utilizing hybrid GWO-GOA, the text function selection is analyzed. In addition, the chosen optima are clustered to use of Fuzzy c-means (FCM) clustering algorithm. the proposed algorithm increases reliability and reduces the time cost. In the introduced method, eight datasets have been used and the performance is essentially predicted. The performance measures used for conducting text feature selection and text clustering are accuracy, precision, recall, F-measure, sensitivity, specificity and demonstrated better quality when comparing with numerous methods. The proposed method showed 87.6.

In several main fields of information retrieval including text mining, and natural language processing, text clustering problem (TCP) is a main method. This poses to the need for a powerful method for document clustering that can be utilized efficiently to navigate, summarize, and organize data to gather large data sets. Reference [86] provided an adaptation of the grey wolf optimizer (GWO) for TCP (TCP-GWO). Above what is possible with metaheuristic swarm-based techniques, the TCP requires a high degree of accuracy. How to break text documents based on GWO into homogeneous clusters that are sufficiently specific and usable is the key problem to be tackled. Precisely, in order to continuously optimize the distance between the clusters of documents, TCP-GWO used the average document distance to the central cluster (ADDC) as the objective feature. In order to evaluate the recall detection accuracy of the document clustering algorithm, documents of high complexity were also included in the evaluation. The extensive experiments for a test set of over a subset of 1300 documents indicated that in less than 20%, failure to properly cluster a document occurred with a classification accuracy of more than 65% for a highly complex data set. The high F-measure rate and ability to effectively cluster documents are important advances resulted from this analysis. The suggested TCP-GWO approach was compared using randomly chosen data sets to the other well-established techniques of text clustering. Intriguingly, in terms of precision, recall, and F-measure rates, TCP-GWO outperformed other state-of-the-art methods. An overview of the reported clustering algorithms is given in Table 1.

Table 1. An overview of the studied text clustering algorithms.

Name	Method	Proposed	Dataset	Measure	Year
MKM [81]	K-means algorithm	Improved approach called multi-objective combining the similarity and distance measure for text clustering based on K-means	Seven different datasets	Accuracy F-measure	2016
LFW [82]	β-hill climbing	Improve the β-hill climbing algorithm the text clustering by providing a fair weighting value of the most important features for each document	LABIC datasets	F-measure Recall Precision Accuracy	2018
DDR [83]	PSO, LFW, K-means	A combination of three features algorithms for text document clustering based LFW approach the harmony search (HS), swarm optimization (PSO), and Genetic algorithm (GA), with dynamic dimension reduction and weight scheme	LABIC datasets	Accuracy F-measure	2017
MMKHA [52]	KHA	a new text clustering approach introduced based on the enhancement of KHA hybrid with multiple criteria function	LABIC datasets	ASDC Accuracy Precision Recall F-measure Purity Entropy	2018
H-KHA [53]	Harmony-KHA	A combination of the harmony search (HS) method and the KHA algorithm proposed to enhance the global search capability	UCI dataset	ASDC Precision Recall Accuracy F-measure	2017

Table 1. *Cont.*

Name	Method	Proposed	Dataset	Measure	Year
KHA [56]	KHA	The first approach adopted the basic KHA, while the second approach uses the genetic operators instead of the KHA operators	CSTR Trec-5 Trec-6 Trec-7 Reuters 21,578	Purity Entropy	2016
MHKHA [87]	MHKHA	hybrid KH algorithm is proposed for feature selection with a new weighting system, and a particular dimension reduction method. Four KHA algorithms are proposed for text clustering	CSTR SyskillWebert tr32- TREC-5 tr12- TREC-5 tr11- TREC-5 oh15	Accuracy Purity Entropy Precision Recall F-measure	2019
MOAC [57]	MOA	This approach aims to select the best position of centroid particles, which is considered as the optimal cluster position	UCI	Accuracy Purity	2018
FSPSOTC [59]	PSO	Feature selection using the PSO algorithm, namely FSPSOTC, aims to solve the problem of feature selection in text clustering by building a new subset of informative features	LABIC	Precision Recall Accuracy F-measure Rank	2017
SCPSO [60]	PSO	The randomization is implemented with the initial population by considering the local and global optimization methods	Reuters 21,578 20 Newsgroup document TDT2	Accuracy	2019
PSO + LSI [61]	PSO LSI	Text clustering based on the PSO algorithm and latent semantic indexing (PSO + LSI)	Reuters	F-measure time	2012
PM [62]	Binary PSO	Opposition chaotic fitness mutation A hybridization between binary PSO with a chaotic map, fitness based dynamic inertia weight, opposition-based learning, and mutation, is introduced for feature selection in text clustering	Reuters 21,578 Classic4 WebKB	Precision(P) Recall(R) F-score(F)	2016
Improved PSO [63]	PSO CHI	An improved PSO version was introduced to optimize the feature selection based on constant constriction factor and functional inertia weight	N.A Paired-sample	T-test	2015
ISSO [65]	SSO K-mean	Two-hybrid algorithms are presented based on SSO and K-means algorithms called SSO+K-means and K-means+SSO	PatentCorpus5000	Cosine similarity F-Measure Accuracy	2017
SSO [64]	SSO K-mean	SSO utilizes the cooperative intelligence of spider social groups. According to spider gender, every spider tries to regenerate specific behavior and decreasing the local particles problems	Patent Corpus5000	F-measure Precision Recall	2016
FCMWOA [66]	WOA FCM	A combination of Whale Optimization Algorithm (WOA) and Fuzzy C Means (FCM) is proposed for text clustering	Webkb Re0 20Newsgroup	Precision Recall F-measure	2018
AIFDS [67]	MWOA FCM	A modified WOA (MWOA) and FCM algorithm for text clustering. Further, an automobile insurance fraud detection system (AIFDS) is introduced	Iris Wine Seed Glass *E-coli*	CMC Sensitivity Specificity Accuracy	2019

Table 1. *Cont.*

Name	Method	Proposed	Dataset	Measure	Year
CGABC [38]	Chaotic gradient artificial bee colony	Chaotic gradient artificial bee colony is proposed to enhance the performance of ABC search equation	Reuters 21,578, Classic4, and WebKB	Quality of solution and convergence speed	2014
ACO [68]	Ant colony optimization	Ant colony optimization is proposed to enhance the fuzzy document clustering issues	Accuracy	Pheromone trails	2011
MTACO [69]	Multi-label ACO	ACO is proposed to select the relevance function for the classification	webKB Yahoo RCV1	classification	2015
RACO RACOFS [70]	EnRiched Ant Colony Optimization	RACO-based feature selection enRiched Ant Colony Optimization is proposed to to modify the pheromone values correctly and avoid premature convergence. RACO-based feature selection is introduced to find optimal solutions globally.	Monk1 Monk2 Post-operative BreastCancer Glass Vowel Wine Zoo	classification accuracy	2014
β-FSTC [71]	β-hill climbing technique	β-hill climbing technique for text feature selection problem is proposed to improve the text clustering	Dmoz-Business and others	Accuracy Precision Recall F-measure	2017
β-HC [72]	β-hill climbing technique	β-hill climbing technique is proposed to solve the text document clustering problem	"Dmoz-Business" dataset	F-measure Precision Recall Accuracy	2017
IBCO [39]	An improved bee colony optimization algorithm	An improved bee colony optimization algorithm using dubbed IBCO is proposed to enhance clustering problem	Iris Wine Glass Wisconsin Breast Cancer Vowel	Classification Error Percentage (CEP) SICD	2015
SOM [85]	Optimization of SOM for document clustering	An optimization of SOM for document clustering has been presented to handle conceptual aspects in document clusters and to minimize training time	Reuters	F-measure training time	2010
H-FSPSOTC [76]	Hybrid method	Hybrid of particle swarm optimization algorithm with genetic operators have been proposed to solve the problem of feature selection	Reuters 21,578 20Newsgroups Reuters 21,578 20Newsgroups Dmoz-Business Dmoz-Science Reuters 21,578 20Newsgroups	Accuracy Precision Recall F-measure	2017
GWO–GOA [37]	Hybrid algorithm	GWO–GOA have been presented to improve meta-heuristic algorithms convergence levels	Reuters 21,578 20 newsgroups Reuters 21,578 20Newsgroups Dmoz-Business DMOZ-Science Reuters 21,578 20 Newsgroups	Accuracy Precision Recall F-measure Sensitivity Specificity	2020
TCP-GWO [86]	Improved grey wolf optimizer (GWO)	The grey wolf optimizer for TCP have been presented to address the problem of how to split text documents on the basis of GWO into homogeneous clusters	Dmoz-Business	Precision Recall F-measure rates	2018
FSHSTC [59]	Metaheuristic Harmony search (HS)	Harmony search (HS) algorithm have been proposed to solve the feature selection problem	Dmoz-Business and others	F-measure Accuracy	2016

3. Procedures of the Text Clustering Problem

Text clustering aims to provide optimal clusters that include related documents (objects). Clustering is based on partitioning many documents into a predefined number of similar groups. Each group holds many similar objects, but various groups have various objects. In this section, the text clustering problem's general procedure as an optimization problem, its formulation, mathematical notations, preprocessing steps, document representation, solution representation of clustering problem, and the fitness function are given. This can help future research in finding the general information about that problem clearly.

3.1. Problem Formulations

In this section, the text clustering problem and its information and formulations are provided as follows.

- A set of text documents (D/objects) is grouped into a established in advance number of clusters (K) [38].
- D can be given as a vector of objects $D = (d_1, d_2, d_3, \ldots\ldots, d_i, \ldots\ldots, d_n)$, d_2 gives the object number two, i presents the number of the object and n is the number of total objects provided in D Reference [88].
- Each group contains a cluster centroid, called c_k, which is represented as a vector of term weights of the words $c_k = (c_{k1}, c_{k2}, c_{k2}, \ldots\ldots, c_{kj}, \ldots\ldots, c_{kt})$.
- c_k is the k_{th} cluster centroid, c_{k2} is the value of position two (feature) in the centroid of cluster number k, and t is the number of all unique centroid terms (features) in the given object.
- The similarity or distance measures are employed to clustering each object to the closest cluster centroid [39,78,89].

3.2. Preprocessing Steps

Before creating clusters, the text needs preprocessing steps, which are as follows: (i) tokenization, (ii) stop word removal, (iii) stemming, (iv) term weighting, and (v) document representation [62]. A brief explanation of these preprocessing levels is performed, as follows:

3.2.1. Tokenization

Tokenization is the responsibility of splitting words up into pieces (word), called tokens, apparently at the same time missing some letters, such as punctuation. These tokens are typically linked to as terms/words, though it is necessary to achieve a type/token distinction [90].

3.2.2. Stop Words Removal

Stop-words are popular and important words, such as "some", "the", "that", "is", "let", "an", "our", "me", and "which", as well as other popular terms that are particularly frequently employed and small useful words in the text. List (http://www.unine.ch/Info/clef/) of stop-words includes a total of more than 500 words [62].

3.2.3. Stemming

Stemming is the process of reducing general words to their root/stem. The stem design is not identical to the morphological root method; it is normally to describe words to the identical stem, even if it is not in itself a true root. Porter (Porter stemmer. website at http://tartarus.org/martin/PorterStemmer/) stemmer is the general stemming method used in text mining [90,91].

3.3. Document Representation

Vector space model (VSM) is a powerful model that is applied to represent documents' content in an official format [92]. It was introduced in the early 1970s. Each document is intended as a vector of term weight to improve similarity calculation [93]. Equation (1) presents n documents and t terms utilizing the VSM, as follows:

$$VSM = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,(t-1)} & w_{1,t} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_{(n-1),1} & w_{(n-1),2} & \cdots & \cdots & w_{(n-1),t} \\ w_{n,1} & w_{n,2} & \cdots & w_{n,(t-1)} & w_{n,t} \end{bmatrix} \quad (1)$$

$$d_i = (w_{i,1}, w_{i,2}, w_{i,3}, ..., w_{i,j}, ..., w_{i,t}). \quad (2)$$

3.4. Solution Representation of Clustering Problem

Document clustering cab formulated as an optimization problem, which is performed based on using an optimization algorithm to be solved. Optimization algorithms use several candidate solutions to solve the clustering problem. Each solution or vector expresses the candidate solution to solve the clustering problem. Figure 1 shows the solution composition. The ith position of the solution guides to the decision of the ith document. If the amount of the given clusters is K, then each section of the solution is a state in the range $(1, \ldots\ldots, K)$. Each part meets a collection of K centroids [78]. Clearly, the number of text document clusters is normally given in advance.

In the example given in Figure 1, ten documents and four clusters are presented. Each solution designs where the documents belong. In this case, documents 1, 5, and 7 are from the same group as label 1 (i.e., cluster number one). Meantime, documents 2, and 6 belong to the same cluster as label 2 (i.e., cluster number two). Documents 3, 8, and 9 belong to the same cluster as label 3 (i.e., cluster number two). For document number 6 and 10, they belong to cluster number 4.

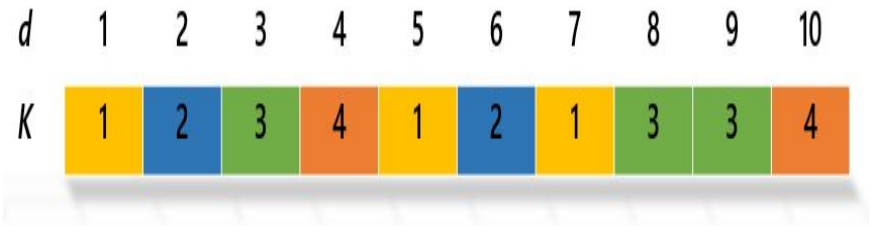

Figure 1. Solution representation of the clustering problem.

3.5. Fitness Function

The fitness value is calculated to evaluate and assess each solution based on its current positions. Each document belongs to a cluster centroids $C = (c_1, c_2, \ldots, c_k, \ldots, c_K)$, where c_k is the centroid of cluster k. The fitness function value for each candiate solution is determined by the average similarity of documents to the cluster centroid (ASDC), as given in Equation (3) [93,94].

$$ASDC = \left[\frac{\sum_{j=1}^{K} \left(\frac{\sum_{i=1}^{n} Cos(d_i, c_j)}{m_i} \right)}{K} \right], \quad (3)$$

where K is the number of given clusters in the used dataset, m_i is the number of documents that correctly should be in cluster i and $Cos(d_i, c_i)$ is the similarity value calculated between the centroid

of cluster j and the document number i. Each solution is given in a binary matrix $a_{i,j}$ of size $n * K$ to calculate the clusters centroid, as given in Equation (4) [39].

$$a_{ij} = \begin{cases} 1, & \text{if } d_i \text{ is assigned to the } j_{th} \text{ cluster} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Equation (5) is utilized to calculate the k_{th} cluster centroid, which is given as a vector $c_k = (c_{k1}, c_{k2}, c_{k3}, \ldots, c_{kj}, \ldots, c_{kt})$ [91].

$$c_{kj} = \frac{\sum_{i=1}^{n} a_{ij}(d_{ij})}{\sum_{i=1}^{n} a_{ij}}, \tag{5}$$

where a_{ij} is a matrix presents the grouped data (see Equation (4)), d_{ij} is the j_{th} feature weight of the document number i, and n is the number of all documents in the given dataset.

4. Evaluation Measures

The most common evaluation measures employed in the text clustering domain are accuracy, purity, entropy, precision, recall, and F-measure [91,95,96]. The text clustering method produces two sets of evaluation measures, namely, internal and external measures [78]. External measurements are applied to evaluate the collected clusters' accuracy (correct) based on the provided document's class labels in the dataset [97]. The following subsections define the external evaluation criteria applied in evaluating the output of the clustering algorithms.

4.1. Accuracy Measure

The accuracy test is applied to determine the correct documents selected to all groups in the provided dataset [98–100]. This measure is defined using Equation (6).

$$AC = \frac{1}{n} \sum_{i=1}^{K} n_{i,i}, \tag{6}$$

where $n_{i,i}$ is the number of all correct candidiates of class i in cluster i, n is the number of all given documents, and K is the number of all given clusters in the dataset.

4.2. Purity Measure

The purity test is applied to determine each cluster's section in a large class [38,101]. This test indicates each group to the common frequent class. An excellent value of purity is close to 1 due to the percentage of big class sizes in each group, which is calculated based on its size. Hence, the value of purity is in the interval $\left[\frac{1}{K^+}, 1\right]$. Equation (7) is used to manage the purity value of the cluster j:

$$P(c_j) = \frac{1}{n_j} \max_j n_{i,j}, \tag{7}$$

where $\max_j$ is the large class size in group j, $n_{i,j}$ is the number of all correct candidates of the class label i in cluster j, and n_j is the total number of members (documents) of cluster j. The purity test for all groups is determined using Equation (8):

$$P = \sum_{j=1}^{K} \frac{n_j}{n} P(c_j), \tag{8}$$

4.3. Entropy Measure

The entropy test measures the partitioning of class marks in each group [101,102]. This test centers on the containment of different cluster classes. A good example has 0 entropy, showing an excellent clustering solution has a low entropy case. The entropy measure of cluster j according to the quantity of each group can be defined using Equation (9):

$$E(c_j) = -\sum_i p_{i,j} \log p_{i,j}, \tag{9}$$

where $p_{i,j}$ is the probability value of class' i members that belong to group j. The entropy test for all groups is determined using Equation (10):

$$E = -\sum_{j=1}^{K} \frac{n_j}{n} E(c_j). \tag{10}$$

4.4. Precision Measure

The precision (P) test for each cluster is calculated using Equation (11) based on the assigned class label in the datasets. The precision test is the ratio of relevant documents and the total number of documents in all groups [78,95]. Precision use for class i in cluster j is defined as follows.

$$P(i,j) = \frac{n_{i,j}}{n_j}, \tag{11}$$

where $n_{i,j}$ is the number of correct candidates of the class labeled i in the group j, and n_j is the total number of objects in the group j.

4.5. Recall Measure

The recall (R) test for each cluster is determined based on the doled out course label. The recall value is the rate of relevant documents in all groups and the whole quantity of relevant objects in the dataset [62,78]. Recall analysis for class i and cluster j is defined using Equation (12).

$$R(i,j) = \frac{n_{i,j}}{n_i}, \tag{12}$$

where $n_{i,j}$ is the number of correct candidates of the class i in group j, and n_i is the number of truly members of class i as the class labels given in the main dataset.

4.6. F-Measure

The F-measure intends to evaluate clusters of the tested partition clusters at the biggest match of the class label partition clusters. This test is a common evaluation criterion in the clustering area based on the collection of precision and recalls tests [78,89,95]. The F-measure value for group j is prepared using Equation (13):

$$F(j) = \frac{2 \times P(i,j) \times R(i,j)}{P(i,j) + R(i,j)}, \tag{13}$$

where $P(i,j)$ is the precision of candidates of class i in group j, $R(i,j)$ is the recall of candidates of class i in group j. Moreover, the F-measure for all clusters is determined using Equation (14).

$$F = \frac{\sum_{j=1}^{K} F_j}{K} \tag{14}$$

5. Experiments Results and Discussion

In this section, comprehensive experiments are conducted to show the performance of the well-known nature-inspired optimization algorithms in solving the text clustering problems. All the given algorithm worked using 30 solution and 1000 iterations. The parameter settings of the used algorithms are taken from the original papers.

5.1. Document Dataset

In this section, the description of the used datasets is given. Table 2 shows the used dataset in the experiments. These datasets are freely available at The university of SAO PAULO, institute of mathematical and computer sciences-USP (http://sites.labic.icmc.usp.br/text_collections/). The numbers of documents, features, clusters, and sources are given in Table 2. Note that the datasets are selected from the comment benchmarks that have been usually used in that domain with different topics and numbers of documents, features, and clusters.

Table 2. The results of the comparative methods in terms of Accuracy measure.

Dataset	Number of			Source
	Documents	Features	Clusters	
DS1	299	1107	4	Technical Reports
DS2	333	2604	4	Web Pages
DS3	204	4252	6	TREC
DS4	313	3745	8	TREC
DS5	414	4879	9	TREC
DS6	878	4537	10	TREC
DS7	913	3100	10	MEDLINE

The datasets details are given also as follows.

- In dataset number 1 (DS1), 299 documents are given, which contains 1107 features. The documents in this dataset belong to 4 different clusters as given in Table 2.
- In dataset number 2 (DS2), 333 documents are given, which contains 2604 features. The documents in this dataset belong to 4 different clusters as given in Table 2.
- In dataset number 3 (DS3), 204 documents are given, which contains 4252 features. The documents in this dataset belong to 6 different clusters as given in Table 2.
- In dataset number 4 (DS4), 313 documents are given, which contains 3745 features. The documents in this dataset belong to 8 different clusters as given in Table 2.
- In dataset number 5 (DS5), 414 documents are given, which contains 4879 features. The documents in this dataset belong to 9 different clusters as given in Table 2.
- In dataset number 6 (DS6), 878 documents are given, which contains 5437 features. The documents in this dataset belong to 10 different clusters as given in Table 2.
- In dataset number 7 (DS7), 913 documents are given, which contains 3100 features. The documents in this dataset belong to 10 different clusters as given in Table 2.

5.2. Results and Discussion

In this section, the results of the comparative methods are given. the comparative methods includes Harmony Search (HS) Algorithm [103], Genetic Algorithm (GA) [104], Particle Swarm Optimization (PSO) Algorithm [105], Ant Colony Optimization (ACO) [106], Krill Herd Algorithm (KHA) [51], Cuckoo Search (CS) Algorithm [107], Gray Wolf Optimizer (GWO) [108], Bat-inspired Algorithm (BA) [109], and K-means technique [110].

The results are given in terms of Accuracy, Precision, Recall, F-measure, Purity, and Entropy measures. These are common external evaluation measures used in the domain of text mining, especially for evaluating the text clustering methods.

In Table 3, the results of the comparative methods using seven datasets (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) in terms of Accuracy measure. It is clear that the basic nature-inspired algorithms that have been selected in this experiment have almost the same performance in solving the given text clustering problems. For example, the GA got 0.472742 Accuracy value in dataset number 1 and 0.431106 in dataset number 7. Also, for another example, the KHA got 0.521404 Accuracy value in dataset number 1 and 0.649999 in dataset number 7. The results show differences in the outcomes according to the given values in Figure 2.

The results of the Friedman ranking test for the comparative methods (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) using the Accuracy values are given in Table 4. In this Table, the rank of each results, summation, mean ranking, and final ranking results. The GWO got the first ranking, followed by ACO, it got the second ranking, BA got the third ranking, K-means got the fourth ranking, CS got the fifth ranking, KHA got the sixth ranking, HS got the seventh ranking, PSO got the eighth ranking, and finally, GA got the ninth ranking. Moreover the summation ranking for the given algorithms using the seventh datasets are given as HS got 42, GA, got 59, PSO got 49, ACO got 18, KHA got 40, CS got 38, GWO got 17, BA got 21, and K-mean got 31. These results can show the significant ability of the nature-inspired optimization algorithms to solve the text clustering problems.

Table 3. The results of the comparative methods in terms of Accuracy measure.

Accuracy	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	0.521571	0.472742	0.498831	0.541639	0.521404	0.504682	0.622909	0.553512	0.620902
DS2	0.727778	0.670421	0.725526	0.754805	0.649999	0.677476	0.843544	0.691741	0.659609
DS3	0.426878	0.356863	0.368628	0.445097	0.425980	0.424509	0.444046	0.454167	0.412745
DS4	0.467572	0.421912	0.473804	0.559265	0.487061	0.503514	0.569808	0.546485	0.505750
DS5	0.497464	0.491048	0.488769	0.576623	0.540217	0.514372	0.483936	0.582661	0.530555
DS6	0.493679	0.463895	0.494818	0.577961	0.510308	0.508542	0.584910	0.547836	0.528644
DS7	0.437952	0.431106	0.450438	0.513801	0.453505	0.514732	0.549452	0.461172	0.530175

Table 4. The results of the Friedman ranking test for the comparative methods using the Accuracy values.

Ranking	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	5	9	8	4	6	7	1	3	2
DS2	3	7	4	2	9	6	1	5	8
DS3	4	9	8	2	5	6	3	1	7
DS4	8	9	7	2	6	5	1	3	4
DS5	6	7	8	2	3	5	9	1	4
DS6	8	9	7	2	5	6	1	3	4
DS7	8	9	7	4	6	3	1	5	2
Summation	42	59	49	18	40	38	17	21	31
Mean rank	6.00	8.42857	7.00	2.57142	5.71428	5.42857	2.42857	3.00	4.42857
Final ranking	7	9	8	2	6	5	1	3	4

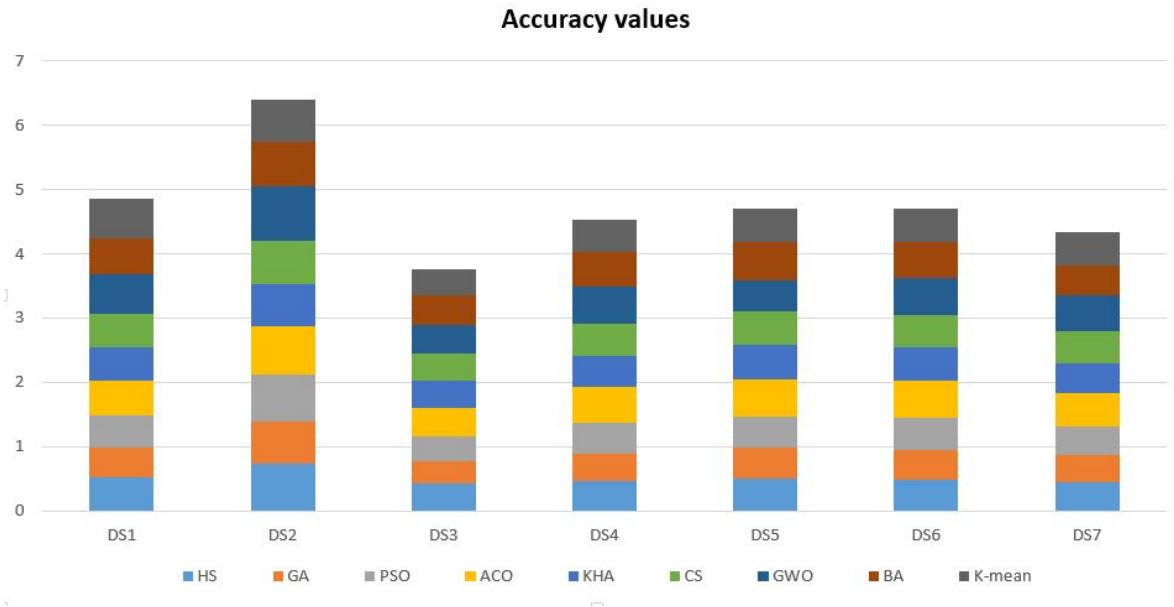

Figure 2. The Accuracy results of the comparative methods using seven datasets.

In Table 5, the results of the comparative methods using seven datasets (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) in terms of Precision measure. It is clear that the basic nature-inspired algorithms that have been selected in this experiment have almost the same performance in solving the given text clustering problems. For example, the HS got 0.695420 Precision value in dataset number 2 and 0.421941 in dataset number 7. Also, for another example, the GWO got 0.850940 Precision value in dataset number 2 and 0.401748 in dataset number 3. The results show differences in the outcomes according to the given values in Figure 3.

The results of the Friedman ranking test for the comparative methods (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) using the Precision values are given in Table 6. In this Table, the rank of each results, summation, mean ranking, and final ranking results. The GWO got the first ranking, followed by ACO, it got the second ranking, BA got the third ranking, K-means got the fourth ranking, PSO got the fifth ranking, CS got the fifth ranking, HS got the seventh ranking, KHA got the eighth ranking, and finally, GA got the ninth ranking. Moreover the summation ranking for the given algorithms using the seventh datasets are given as HS got 43, GA, got 61, PSO got 39, ACO got 20, KHA got 44, CS got 39, GWO got 13, BA got 23, and K-mean got 33. These results can show the significant ability of the nature-inspired optimization algorithms to solve the text clustering problems.

Table 5. The results of the comparative methods in terms of Precision measure.

Precision	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	0.516311	0.465179	0.506517	0.523483	0.510689	0.497316	0.626874	0.535501	0.598837
DS2	0.695420	0.660070	0.722608	0.715740	0.648346	0.667253	0.850940	0.660592	0.620807
DS3	0.394765	0.321445	0.339438	0.399019	0.386761	0.391978	0.413193	0.401748	0.369091
DS4	0.452066	0.407555	0.478154	0.533171	0.461404	0.497471	0.544943	0.499450	0.487632
DS5	0.409932	0.408273	0.444686	0.484433	0.458736	0.419713	0.417344	0.525953	0.448664
DS6	0.425038	0.408469	0.437013	0.469521	0.425535	0.418595	0.504058	0.464550	0.451067
DS7	0.421942	0.405116	0.426024	0.464437	0.418351	0.483816	0.537852	0.436314	0.494360

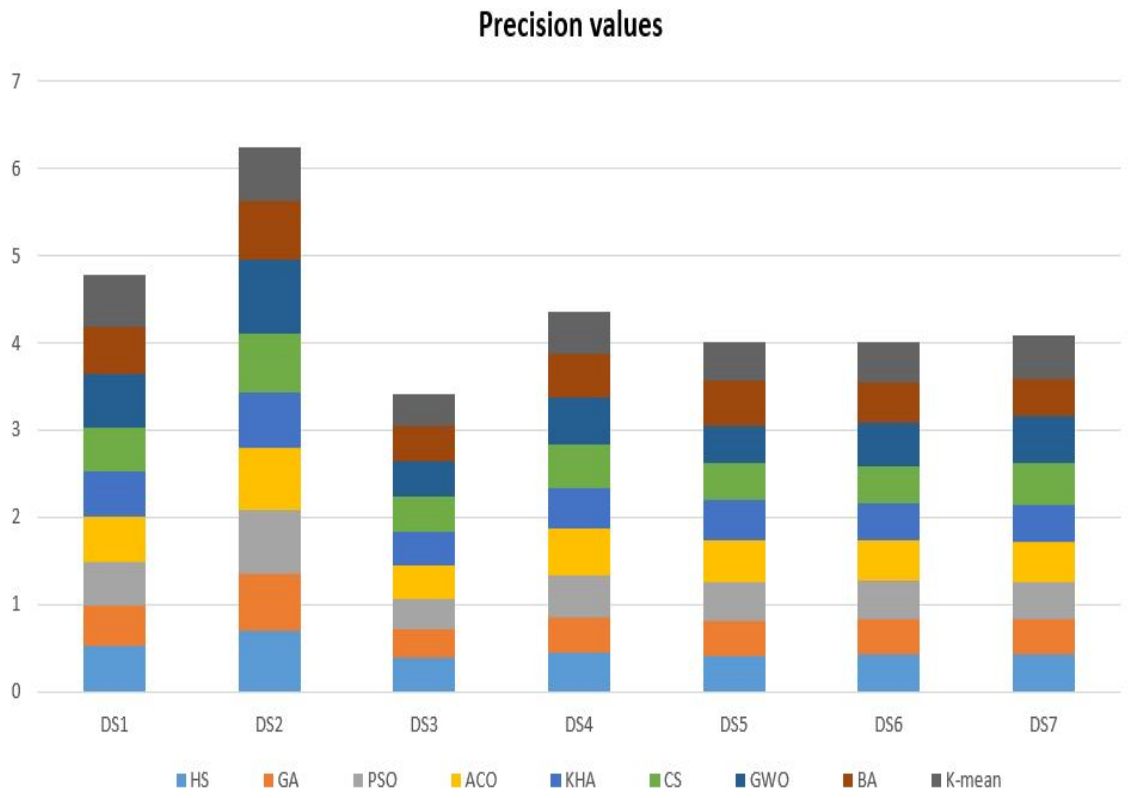

Figure 3. The Precision results of the comparative methods using seven datasets.

Table 6. The results of the Friedman ranking test for the comparative methods using the Precision values.

Ranking	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	5	9	7	4	6	8	1	3	2
DS2	4	7	2	3	8	5	1	6	9
DS3	4	9	8	3	6	5	1	2	7
DS4	8	9	6	2	7	4	1	3	5
DS5	8	9	5	2	3	6	7	1	4
DS6	7	9	5	2	6	8	1	3	4
DS7	7	9	6	4	8	3	1	5	2
Summation	43	61	39	20	44	39	13	23	33
Mean rank	6.14285	8.71428	5.57142	2.85714	6.28571	5.57142	1.85714	3.28571	4.71428
Final ranking	7	9	5	2	8	5	1	3	4

In Table 7, the results of the comparative methods using seven datasets (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) in terms of Recall measure. It is clear that the basic nature-inspired algorithms that have been selected in this experiment have almost the same performance in solving the given text clustering problems. For example, the ACO got 0.737131 Recall value in dataset number 2 and 0.409219 in dataset number 3. Also, for another example, the GA got 0.408991 Recall value in dataset number 5 and 0.428635 in dataset number 6. The results show differences in the outcomes according to the given values in Figure 4. It is clear that the performance of GWO is better than almost all the given nature-inspired optimization algorithms.

The results of the Friedman ranking test for the comparative methods (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) using the Recall values are given in Table 8. In this Table, the rank of each results, summation, mean ranking, and final ranking results. The GWO got the first ranking, followed by ACO, it got the second ranking, BA got the third ranking, K-means got the fourth ranking, HS got the fifth ranking, CS got the fifth ranking, PSO got the seventh ranking, KHA got the eighth ranking, and finally, GA got the ninth ranking. Moreover the summation ranking for the given algorithms using the seventh datasets are given as HS got 38, GA, got 59, PSO got 41, ACO got 20, KHA got 42, CS got 38, GWO got 18, BA got 26, and K-mean got 33. These results can show the significant ability of the nature-inspired optimization algorithms to solve the text clustering problems. Moreover, the differences between the obtained results are slight and there is no much variation.

Table 7. The results of the comparative methods in terms of Recall measure.

Recall	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	0.541343	0.476285	0.508354	0.536898	0.507354	0.476691	0.650435	0.561969	0.616213
DS2	0.708215	0.662504	0.722580	0.737131	0.640515	0.658443	0.820237	0.665052	0.631951
DS3	0.412106	0.346574	0.377370	0.409219	0.395009	0.414267	0.403324	0.392217	0.390691
DS4	0.454048	0.419921	0.464152	0.541043	0.464823	0.496220	0.553784	0.516053	0.489738
DS5	0.409845	0.408991	0.432807	0.467791	0.441765	0.409100	0.406372	0.522470	0.433629
DS6	0.455282	0.428635	0.458360	0.500248	0.448898	0.437551	0.537407	0.489501	0.485761
DS7	0.415760	0.404179	0.418950	0.482779	0.419679	0.484986	0.517571	0.430767	0.499522

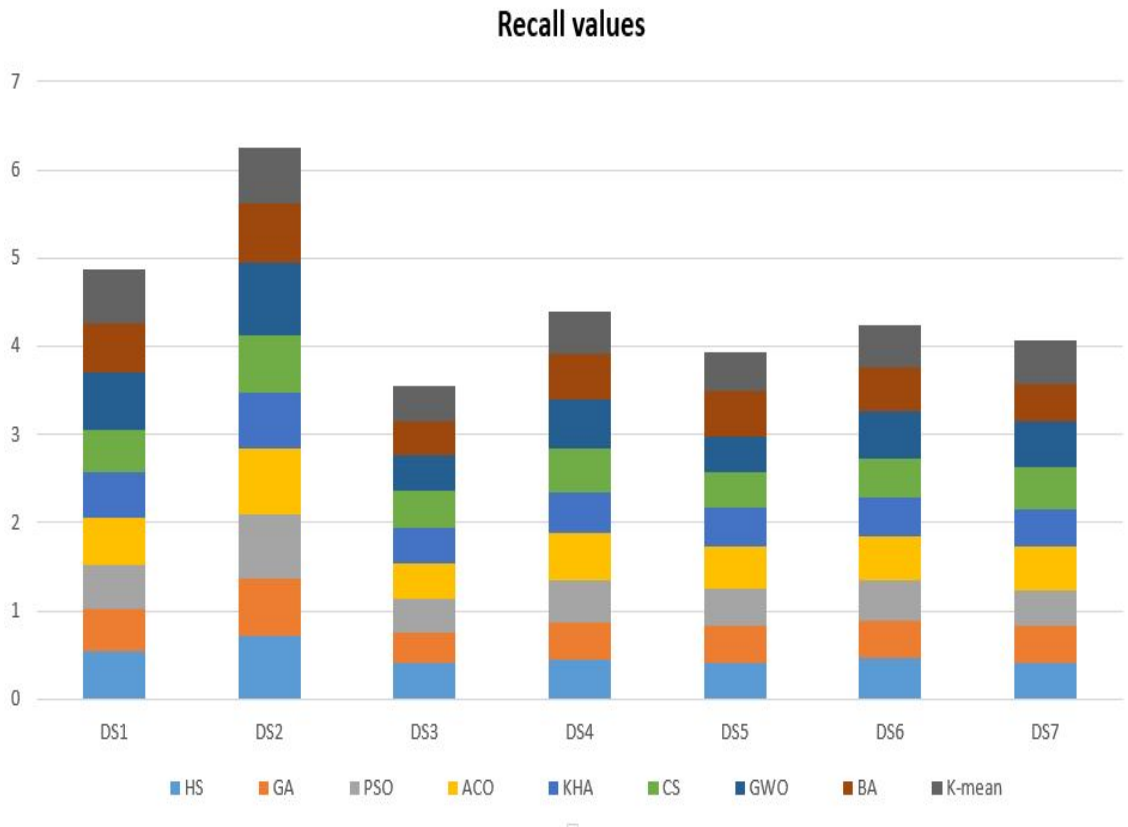

Figure 4. The Recall results of the comparative methods using seven datasets.

Table 8. The results of the Friedman ranking test for the comparative methods using the Recall values.

Ranking	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	4	9	6	5	7	8	1	3	2
DS2	4	6	3	2	8	7	1	5	9
DS3	2	9	8	3	5	1	4	6	7
DS4	8	9	7	2	6	4	1	3	5
DS5	6	8	5	2	3	7	9	1	4
DS6	6	9	5	2	7	8	1	3	4
DS7	8	9	7	4	6	3	1	5	2
Summation	38	59	41	20	42	38	18	26	33
Mean rank	5.42857	8.42857	5.85714	2.85714	6.00	5.42857	2.57142	3.71428	4.71428
Final ranking	5	9	7	2	8	5	1	3	4

In Table 9, the results of the comparative methods using seven datasets (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) in terms of F-measure measure. It is clear that the basic nature-inspired algorithms that have been selected in this experiment have almost the same performance in solving the given text clustering problems. For example, the KHA got 0.507064 F-measure value in dataset number 1 and 0.418632 in dataset number 7. Also, for another example, the GWO got 0.637902 F-measure value in dataset number 1 and 0.527244 in dataset number 7. The results show differences in the outcomes according to the given values in Figure 5. It is obvious that the performance of the used nature-inspired optimization algorithm almost the same and GWO is better than almost all the given nature-inspired optimization algorithms.

The results of the Friedman ranking test for the comparative methods (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) using the F-measure values are given in Table 10. In this Table, the rank of each results, summation, mean ranking, and final ranking results. The GWO got the first ranking, followed by ACO, it got the second ranking, BA got the third ranking, K-means got the fourth ranking, PSO got the fifth ranking, CS got the sixth ranking, HS got the seventh ranking, KHA got the eighth ranking, and finally, GA got the ninth ranking. Moreover the summation ranking for the given algorithms using the seventh datasets are given as HS got 41, GA, got 61, PSO got 39, ACO got 19, KHA got 45, CS got 39, GWO got 13, BA got 25, and K-mean got 33. These results can show the significant ability of the nature-inspired optimization algorithms to solve the text clustering problems. Moreover, the differences between the obtained results are slight and there is no much variation. The F-measure results commit with the other results obtained by the Precision and Recall measures.

Table 9. The results of the comparative methods in terms of F-measure measure.

F-Measure	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	0.526843	0.469324	0.509023	0.529366	0.507064	0.484753	0.637902	0.547258	0.606673
DS2	0.701549	0.661178	0.722382	0.725740	0.643889	0.661381	0.835154	0.666718	0.625910
DS3	0.402158	0.336810	0.356561	0.401984	0.389066	0.400461	0.410225	0.395514	0.377217
DS4	0.452630	0.412758	0.470815	0.536117	0.462653	0.501921	0.548784	0.507005	0.487956
DS5	0.409599	0.407297	0.438048	0.475219	0.449388	0.413829	0.411017	0.523684	0.440503
DS6	0.439249	0.417510	0.447198	0.484020	0.436299	0.427266	0.519341	0.475981	0.467066
DS7	0.418555	0.404324	0.422297	0.473309	0.418632	0.484100	0.527244	0.433272	0.496694

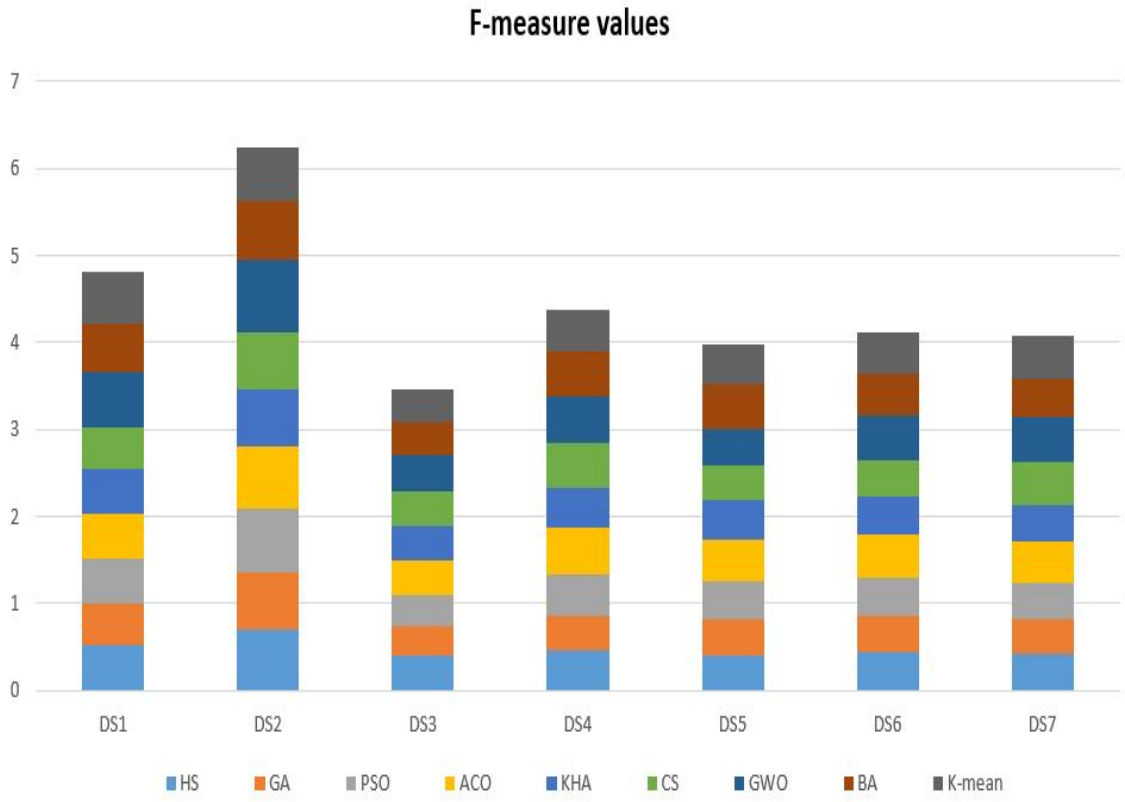

Figure 5. The F-measure results of the comparative methods using seven datasets.

Table 10. The results of the Friedman ranking test for the comparative methods using the F-measure values.

Ranking	Comparative Algorithms								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	5	9	6	4	7	8	1	3	2
DS2	4	7	3	2	8	6	1	5	9
DS3	2	9	8	3	6	4	1	5	7
DS4	8	9	6	2	7	4	1	3	5
DS5	8	9	5	2	3	6	7	1	4
DS6	6	9	5	2	7	8	1	3	4
DS7	8	9	6	4	7	3	1	5	2
Summation	41	61	39	19	45	39	13	25	33
Mean rank	5.85714	8.71429	5.57143	2.71429	6.42857	5.57143	1.85714	3.57143	4.71428
Final ranking	7	9	5	2	8	5	1	3	4

In Table 11, the results of the comparative methods using seven datasets (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) in terms of Purity measure. It is clear that the basic nature-inspired algorithms that have been selected in this experiment have almost the same performance in solving the given text clustering problems. For example, the HS got 0.636698 Purity value in dataset number 1 and 0.542493 in dataset number 7. Also, for another example, the GA got 0.578385 Purity value in dataset number 1 and 0.579877 in dataset number 7. The results show slight differences in the outcomes according to the given values in Figure 6. It is noticeable that the achievement of the used nature-inspired optimization algorithm almost identical and GWO got better results in almost all the given datasets compared to other nature-inspired optimization algorithms.

The results of the Friedman ranking test for the comparative methods (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) using the Purity values are given in Table 12. In this Table, the rank of each results, summation, mean ranking, and final ranking results. The GWO got the first ranking, followed by ACO, it got the second ranking, KHA got the third ranking, CS got the fourth ranking, K-means got the fifth ranking, BA got the sixth ranking, PSO got the seventh ranking, HS got the eighth ranking, and finally, GA got the ninth ranking. Moreover the summation ranking for the given algorithms using the seventh datasets are given as HS got 50, GA, got 53, PSO got 49, ACO got 21, KHA got 27, CS got 29, GWO got 20, BA got 35, and K-mean got 31. These results can reveal the significant ability of the nature-inspired optimization algorithms to address the text clustering problems. Besides, the differences between the achieved results are small, and there is no much difference. The Purity results commit with the other results obtained by the Accuracy, F-measure, Precision, and Recall measures.

Table 11. The results of the comparative methods in terms of Purity measure.

Purity	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	0.636698	0.578385	0.632088	0.687157	0.753064	0.713239	0.781025	0.701243	0.778676
DS2	0.743519	0.721757	0.806309	0.910619	0.798183	0.789050	0.989379	0.776432	0.754982
DS3	0.626638	0.572274	0.629077	0.642075	0.667246	0.714183	0.693792	0.622121	0.632117
DS4	0.630758	0.580626	0.557168	0.788665	0.708666	0.658809	0.718479	0.648318	0.688434
DS5	0.718831	0.750700	0.661226	0.831996	0.661298	0.664520	0.648362	0.733289	0.683852
DS6	0.633277	0.557320	0.644802	0.688866	0.729106	0.674527	0.678664	0.680488	0.656595
DS7	0.542493	0.579877	0.564461	0.644539	0.583298	0.662628	0.750014	0.658338	0.663414

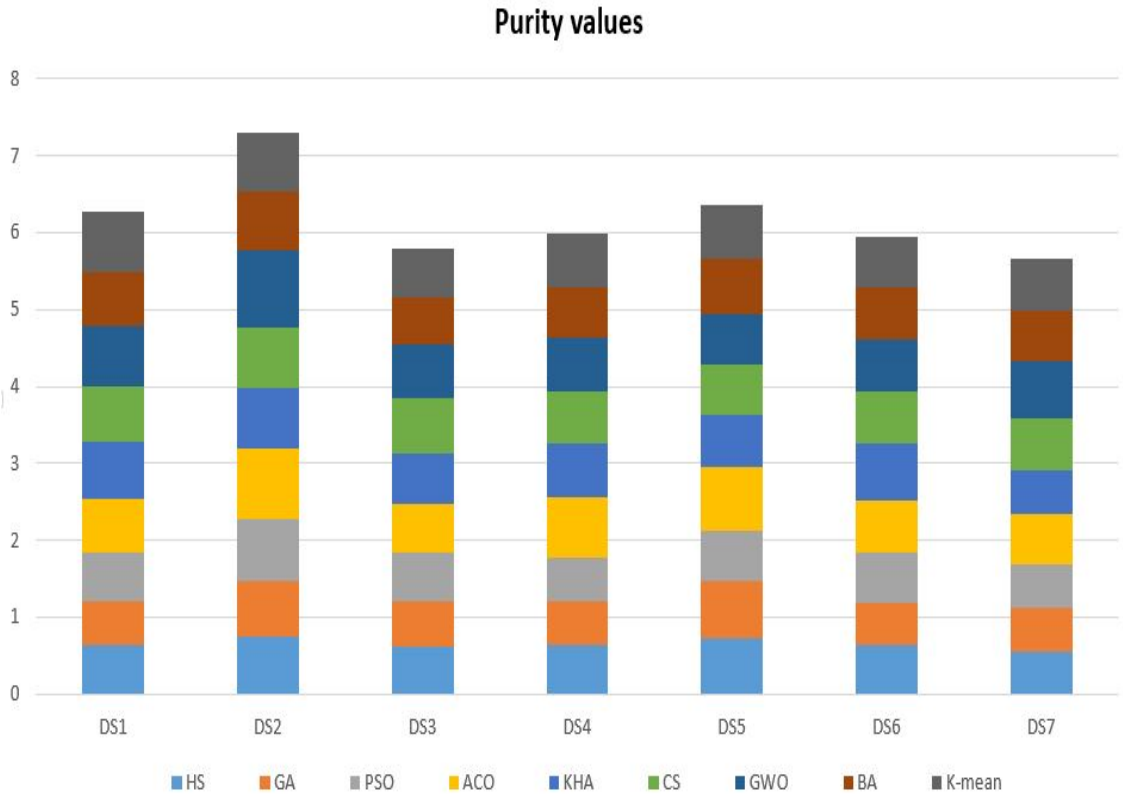

Figure 6. The Purity results of the comparative methods using seven datasets.

Table 12. The results of the Friedman ranking test for the comparative methods using the Purity values.

Ranking	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	7	9	8	6	3	4	1	5	2
DS2	8	9	3	2	4	5	1	6	7
DS3	7	9	6	4	3	1	2	8	5
DS4	7	8	9	1	3	5	2	6	4
DS5	4	2	8	1	7	6	9	3	5
DS6	8	9	7	2	1	5	4	3	6
DS7	9	7	8	5	6	3	1	4	2
Summation	50	53	49	21	27	29	20	35	31
Mean rank	7.14285	7.57142	7.00	3.00	3.85714	4.14285	2.85714	5.00	4.42857
Final ranking	8	9	7	2	3	4	1	6	5

In Table 12, the results of the comparative methods using seven datasets (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) in terms of Entropy measure. It is obvious that the basic nature-inspired algorithms that have been selected in this experiment have almost the same performance in solving the given text clustering problems. For example, the ACO got 0.388168 Entropy value in dataset number 1 and 0.580499 in dataset number 7. Also, for another example, the KHA got 0.320995 Entropy value in dataset number 1 and 0.625289 in dataset number 7. The results show slight differences in the outcomes according to the given values in Figure 7. It is obvious that the achievement of the used nature-inspired optimization algorithm almost identical and ACO got better results in almost all the given datasets compared to other nature-inspired optimization algorithms.

The results of the Friedman ranking test for the comparative methods (i.e., Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), Bat-inspired Algorithm (BA), and K-means technique.) using the Entropy values are given in Table 13. In Table 14, the rank of each results, summation, mean ranking, and final ranking results. The ACO got the first ranking, followed by KHA, it got the second ranking, GWO got the second ranking, K-means got the fourth ranking, BA got the fifth ranking, CS got the sixth ranking, HS got the seventh ranking, GA got the eighth ranking, and finally, PSO got the ninth ranking. Moreover the summation ranking for the given algorithms using the seventh datasets are given as HS got 44, GA, got 48, PSO got 53, ACO got 15, KHA got 27, CS got 36, GWO got 27, BA got 34, and K-mean got 31. These results can reveal the significant ability of the nature-inspired optimization algorithms to address the text clustering problems. Besides, the differences between the achieved results are small, and there is no much difference. The Entropy results commit with the other results obtained by the Accuracy, F-measure, Precision, Purity, and Recall measures.

Table 13. The results of the comparative methods in terms of Entropy measure.

Entropy	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	0.448296	0.465955	0.445373	0.388168	0.320995	0.335901	0.375879	0.421111	0.333152
DS2	0.352708	0.464191	0.388575	0.324644	0.346803	0.346968	0.239784	0.411863	0.395838
DS3	0.412584	0.423777	0.408261	0.357795	0.414517	0.419521	0.398005	0.391949	0.333981
DS4	0.591471	0.575612	0.736625	0.546981	0.583180	0.580376	0.616758	0.619459	0.621091
DS5	0.498653	0.447260	0.509601	0.440492	0.456882	0.491119	0.514431	0.404461	0.466960
DS6	0.437201	0.453595	0.501121	0.349836	0.379381	0.407821	0.371808	0.416385	0.387643
DS7	0.626637	0.641534	0.662544	0.580499	0.625289	0.634617	0.510802	0.609279	0.624934

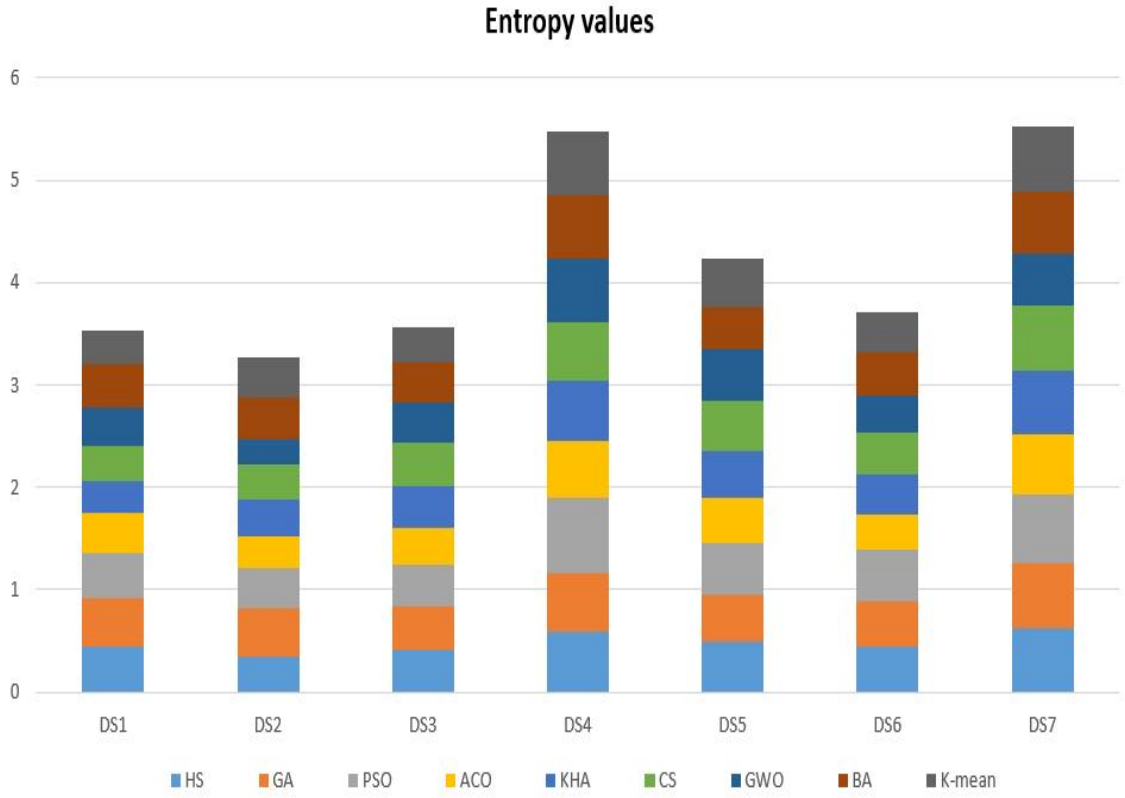

Figure 7. The Entropy results of the comparative methods using seven datasets.

Table 14. The results of the Friedman ranking test for the comparative methods using the Entropy values.

Ranking	**Comparative Algorithms**								
Dataset	**HS**	**GA**	**PSO**	**ACO**	**KHA**	**CS**	**GWO**	**BA**	**K-Mean**
DS1	8	9	7	5	1	3	4	6	2
DS2	5	9	6	2	3	4	1	8	7
DS3	6	9	5	2	7	8	4	3	1
DS4	5	2	9	1	4	3	6	7	8
DS5	7	3	8	2	4	6	9	1	5
DS6	7	8	9	1	3	5	2	6	4
DS7	6	8	9	2	5	7	1	3	4
Summation	44	48	53	15	27	36	27	34	31
Mean rank	6.28571	6.85714	7.57142	2.14285	3.85714	5.14285	3.85714	4.85714	4.42857
Final ranking	7	8	9	1	2	6	2	5	4

We concluded that the performance of the nature-inspired algorithms is better than the K-means clustering technique and the tested optimization algorithms got almost the same performance on the tested seven datasets. The basic algorithms can get better results when modifying or hybridizing it with other algorithms components and other local search methods.

6. Conclusions and Future Works

Text clustering is one of the efficient unsupervised learning techniques used to partition a huge number of text documents into a subset of clusters. In which, each cluster contains similar documents and the clusters contain dissimilar text documents. Nature-inspired optimization algorithms have been successfully used to solve various optimization problems, including text document clustering problems. Nature-inspired optimization algorithms demonstrated their achievement in solving different kinds of

text clustering problems. However, local optima can be trapped because of its focus on exploration (i.e., global search) instead of exploitation (i.e., local search). This effect may be improved over time, as to how well the sets of rules governing various search algorithms work are better understood. There are two main problems in the text clustering application: the initial cluster centroids and the number of clusters. Parameter tuning will also play a critical role in future studies since the parameters' values and settings govern the algorithm's overall performance. From this discussion, we see that nature-inspired optimization algorithms are robustly feasible for continuing use in machine learning domains.

In this paper, a comprehensive review is presented to show the most related nature-inspired algorithms that have been used in solving the text clustering problem. This paper summarizes the most common papers published in the literature until the end of the year 2020. Most of the gathered papers describe the optimization methods that have been used in text clustering applications. Several variants of algorithms, including standard, basic, modified, hybrid methods, and others, are studied. Moreover, comprehensive experiments are conducted and analyzed to show the performance of the common well-know nature-inspired optimization algorithms in solving the text document clustering problems including Harmony Search (HS) Algorithm, Genetic Algorithm (GA), Particle Swarm Optimization (PSO) Algorithm, Ant Colony Optimization (ACO), Krill Herd Algorithm (KHA), Cuckoo Search (CS) Algorithm, Gray Wolf Optimizer (GWO), and Bat-inspired Algorithm (BA). Seven text benchmark datasets are used to validate the performance of the tested algorithms. The results showed that the performance of the well-known nurture-inspired optimization algorithms almost the same with slight differences. Moreover, according to the accuracy measure, GWO is the best and GA is the worst. Also, the GWO is the best, and GA is the worst according to the F-measure. For improvement purposes, new modified versions of the tested algorithms can be proposed and tested to tackle the text clustering problems.

Finally, we recommend new future investigation directions on text clustering-based nature-inspired optimization algorithms. The most vital features of these algorithms (i.e., GA, GWO, KHA, PSO, and HSA) might be blended for better overall performance in solving the text clustering problems. New hybrid and modified algorithms can be recommended to tackle the text clustering problems. Moreover, numerous new nature-inspired optimization algorithms have been introduced recently, which can be employed to solve the clustering problems. These algorithms are Slime Mould Algorithm, Lightning Search Algorithm, Moth-flame Optimization Algorithm, Marine Predators Algorithm, Equilibrium Optimizer, Sine Cosine Algorithm, Salp Swarm Algorithm, Group Search Optimizer, Harris Hawks Optimization, Multi-verse Optimizer Algorithm, Ant Lion Optimizer, Henry Gas Solubility Optimization, and others. Some limitations that have been recognized in the text clustering domain are given as following.

- The behavior of the selected clustering algorithm.
- The number of clusters.
- The initial clusters centroids.
- The selected features from the given documents for applying the clustering process.
- The dimension size of the given text documents
- The weighting score of the used features

Author Contributions: L.A.: Conceptualization, supervision, methodology, formal analysis, resources, data curation, writing—original draft preparation. A.H.G.: Conceptualization, supervision, writing—review and editing, project administration, funding acquisition. M.A.E.: Conceptualization, writing—review and editing, supervision. A.G.H.: Conceptualization, writing—review and editing, supervision. A.M.K.: Conceptualization, writing—review and editing. M.A.: Conceptualization, writing—review and editing. E.H.H.: Conceptualization, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Punitha, S.; Punithavalli, M. Performance evaluation of semantic based and ontology based text document clustering techniques. *Procedia Eng.* **2012**, *30*, 100–106. [CrossRef]
2. Lu, Q.; Conrad, J.G.; Al-Kofahi, K.; Keenan, W. Legal document clustering with built-in topic segmentation. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Scotland, UK, 24–28 October 2011; pp. 383–392.
3. Karypis, G.; Han, E.H.; Kumar, V. Chameleon: Hierarchical clustering using dynamic modeling. *Computer* **1999**, *32*, 68–75. [CrossRef]
4. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An efficient data clustering method for very large databases. *ACM Sigmod Rec.* **1996**, *25*, 103–114. [CrossRef]
5. Xu, S.; Zhang, J. A parallel hybrid web document clustering algorithm and its performance study. *J. Supercomput.* **2004**, *30*, 117–131. [CrossRef]
6. Bradley, P.S.; Fayyad, U.; Reina, C. Scaling EM (Expectation-Maximization) Clustering to Large Databases. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.2850&rep=rep1&type=pdf (accessed on 28 November 2020).
7. Han, J.; Pei, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Elsevier: Amsterdam, The Netherlands, 2011.
8. Tan, P.N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; Pearson Education India: London, UK, 2016.
9. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: A review. *ACM Comput. Surv. (CSUR)* **1999**, *31*, 264–323. [CrossRef]
10. Karypis, M.S.G.; Kumar, V.; Steinbach, M. A Comparison of Document Clustering Techniques; TextMining Workshop at KDD2000 (May 2000). Available online: http://www.stat.cmu.edu/~rnugent/PCMI2016/papers/DocClusterComparison.pdf (accessed on 28 November 2020).
11. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, 21 June–18 July 1965.
12. Zhao, X.; Wang, C.; Su, J.; Wang, J. Research and application based on the swarm intelligence algorithm and artificial intelligence for wind farm decision system. *Renew. Energy* **2019**, *134*, 681–697. [CrossRef]
13. Pasha, J.; Dulebenets, M.A.; Kavoosi, M.; Abioye, O.F.; Wang, H.; Guo, W. An Optimization Model and Solution Algorithms for the Vehicle Routing Problem with a "Factory-in-a-Box". *IEEE Access* **2020**, *8*, 134743–134763. [CrossRef]
14. Slowik, A.; Kwasnicka, H. Nature inspired methods and their industry applications—Swarm intelligence algorithms. *IEEE Trans. Ind. Inform.* **2017**, *14*, 1004–1015. [CrossRef]
15. Brezočnik, L.; Fister, I.; Podgorelec, V. Swarm intelligence algorithms for feature selection: A review. *Appl. Sci.* **2018**, *8*, 1521. [CrossRef]
16. Dulebenets, M.A.; Kavoosi, M.; Abioye, O.; Pasha, J. A self-adaptive evolutionary algorithm for the berth scheduling problem: Towards efficient parameter control. *Algorithms* **2018**, *11*, 100. [CrossRef]
17. Anandakumar, H.; Umamaheswari, K. A bio-inspired swarm intelligence technique for social aware cognitive radio handovers. *Comput. Electr. Eng.* **2018**, *71*, 925–937. [CrossRef]
18. Hussien, A.G.; Houssein, E.H.; Hassanien, A.E. A binary whale optimization algorithm with hyperbolic tangent fitness function for feature selection. In Proceedings of the 2017 IEEE Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 5–7 December 2017; pp. 166–172.
19. Abualigah, L.; Shehab, M.; Alshinwan, M.; Mirjalili, S.; Abd Elaziz, M. Ant Lion Optimizer: A Comprehensive Survey of Its Variants and Applications. In *Archives of Computational Methods in Engineering*; Springer: Berlin/Heidelberg, Germany, 2020.
20. Hussien, A.G.; Hassanien, A.E.; Houssein, E.H.; Bhattacharyya, S.; Amin, M. S-shaped binary whale optimization algorithm for feature selection. In *Recent Trends in Signal and Image Processing*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 79–87.
21. Abualigah, L.; Shehab, M.; Alshinwan, M.; Alabool, H. Salp swarm algorithm: A comprehensive survey. *Neural Comput. Appl.* **2019**, *32*, 11195–11215. [CrossRef]
22. Zahn, C.T. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.* **1971**, *100*, 68–86. [CrossRef]

23. Mao, J.; Jain, A.K. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans. Neural Netw.* **1995**, *6*, 296–317.
24. Liao, S.H.; Wen, C.H. Artificial neural networks classification and clustering of methodologies and applications–literature analysis from 1995 to 2005. *Expert Syst. Appl.* **2007**, *32*, 1–11. [CrossRef]
25. Forgy, E.W. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics* **1965**, *21*, 768–769.
26. Paterlini, S.; Minerva, T. Evolutionary approaches for cluster analysis. In *Soft Computing Applications*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 165–176.
27. Sarkar, M.; Yegnanarayana, B.; Khemani, D. A clustering algorithm using an evolutionary programming-based approach. *Pattern Recognit. Lett.* **1997**, *18*, 975–986. [CrossRef]
28. Cole, R.M. *Clustering with Genetic Algorithms*; University of Western Australia: Crawley, Australia, 1998.
29. Cui, X.; Potok, T.E.; Palathingal, P. Document clustering using particle swarm optimization. In Proceedings of the 2005 IEEE Swarm Intelligence Symposium, SIS 2005, Pasadena, CA, USA, 8–10 June 2005; pp. 185–191.
30. Runkler, T.A. Ant colony optimization of clustering models. *Int. J. Intell. Syst.* **2005**, *20*, 1233–1251. [CrossRef]
31. Hussien, A.G.; Hassanien, A.E.; Houssein, E.H.; Amin, M.; Azar, A.T. New binary whale optimization algorithm for discrete optimization problems. *Eng. Optim.* **2020**, *52*, 945–959. [CrossRef]
32. Hussien, A.G.; Oliva, D.; Houssein, E.H.; Juan, A.A.; Yu, X. Binary Whale Optimization Algorithm for Dimensionality Reduction. *Mathematics* **2020**, *8*, 1821. [CrossRef]
33. Abualigah, L.; Abd Elaziz, M.; Hussien, A.G.; Alsalibi, B.; Jalali, S.M.J.; Gandomi, A.H. Lightning Search Algorithm: A Comprehensive Survey. Available online: https://link.springer.com/article/10.1007/s10489-020-01947-2 (accessed on 28 November 2020).
34. Hussien, A.G.; Amin, M.; Wang, M.; Liang, G.; Alsanad, A.; Gumaei, A.; Chen, H. Crow search algorithm: Theory, recent advances, and applications. *IEEE Access* **2020**, *8*, 173548–173565. [CrossRef]
35. Assiri, A.S.; Hussien, A.G.; Amin, M. Ant Lion Optimization: Variants, hybrids, and applications. *IEEE Access* **2020**, *8*, 77746–77764. [CrossRef]
36. Hussien, A.G.; Amin, M.; Abd El Aziz, M. A comprehensive review of moth-flame optimisation: Variants, hybrids, and applications. *J. Exp. Theor. Artif. Intell.* **2020**, *32*, 705–725. [CrossRef]
37. Purushothaman, R.; Rajagopalan, S.; Dhandapani, G. Hybridizing Gray Wolf Optimization (GWO) with Grasshopper Optimization Algorithm (GOA) for text feature selection and clustering. *Appl. Soft Comput.* **2020**, *96*, 106651. [CrossRef]
38. Bharti, K.K.; Singh, P.K. Chaotic gradient artificial bee colony for text clustering. *Soft Comput.* **2016**, *20*, 1113–1126. [CrossRef]
39. Forsati, R.; Keikha, A.; Shamsfard, M. An improved bee colony optimization algorithm with an application to document clustering. *Neurocomputing* **2015**, *159*, 9–26. [CrossRef]
40. Hussien, A.G.; Hassanien, A.E.; Houssein, E.H. Swarming behaviour of salps algorithm for predicting chemical compound activities. In Proceedings of the 2017 IEEE Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 5–7 December 2017; pp. 315–320.
41. Abualigah, L.; Diabat, A.; Mirjalili, S.; Abd Elaziz, M.; Gandomi, A.H. The Arithmetic Optimization Algorithm. *Comput. Methods Appl. Mech. Eng.* **2020**, in press.
42. Abualigah, L.; Diabat, A.; Geem, Z.W. A Comprehensive Survey of the Harmony Search Algorithm in Clustering Applications. *Appl. Sci.* **2020**, *10*, 3827. [CrossRef]
43. Abualigah, L. Group Search Optimizer: A Nature-Inspired Meta-Heuristic Optimization Algorithm with Its Results, Variants, and Applications. Available online: https://link.springer.com/article/10.1007/s00521-020-05107-y (accessed on 29 November 2020).
44. Abualigah, L. Multi-Verse Optimizer Algorithm: A Comprehensive Survey of Its Results, Variants, and Applications. Available online: https://link.springer.com/article/10.1007/s00521-020-04839-1 (accessed on 29 November 2020).
45. Rossi, R.G.; Marcacini, R.M.; Rezende, S.O. Benchmarking Text Collections for Classification and Clustering Tasks. Available online: http://repositorio.icmc.usp.br/bitstream/handle/RIICMC/6641/Relat%C3%B3rios%20T%C3%A9cnicas_395_2013.pdf?sequence=1 (accessed on 29 November 2020).
46. Amer, A.A.; Abdalla, H.I. A set theory based similarity measure for text clustering and classification. *J. Big Data* **2020**, *7*, 1–43. [CrossRef]

47. Nalawade, R.; Samal, A.; Avhad, K. Improved similarity measure for text classification and clustering. *Int. Res. J. Eng. Technol. (IRJET)* **2016**, *3*, 214–219.
48. Yang, X.; Zhang, G.; Lu, J.; Ma, J. A kernel fuzzy c-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises. *IEEE Trans. Fuzzy Syst.* **2010**, *19*, 105–115. [CrossRef]
49. Jiang, J.Y.; Liou, R.J.; Lee, S.J. A fuzzy self-constructing feature clustering algorithm for text classification. *IEEE Trans. Knowl. Data Eng.* **2010**, *23*, 335–349. [CrossRef]
50. Onan, A. Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes* **2017**, *46*, 330–348. [CrossRef]
51. Gandomi, A.H.; Alavi, A.H. Krill herd: A new bio-inspired optimization algorithm. *Commun. Nonlinear Sci. Numer. Simul.* **2012**, *17*, 4831–4845. [CrossRef]
52. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S. Hybrid clustering analysis using improved krill herd algorithm. *Appl. Intell.* **2018**, *48*, 4047–4071. [CrossRef]
53. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S.; Gandomi, A.H. A novel hybridization strategy for krill herd algorithm applied to clustering techniques. *Appl. Soft Comput.* **2017**, *60*, 423–435. [CrossRef]
54. Abualigah, L.M.; Khader, A.T.; AlBetar, M.A.; Hanandeh, E.S. A new hybridization strategy for krill herd algorithm and harmony search algorithm applied to improve the data clustering. In Proceedings of the 1st EAI International Conference on Computer Science and Engineering, Penang, Malaysia, 11–12 November 2016.
55. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A.; Awadallah, M.A. A krill herd algorithm for efficient text documents clustering. In Proceedings of the 2016 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Batu Feringghi, Malaysia, 30–31 May 2016; pp. 67–72.
56. Abualigah, L.M.Q. *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering*; Springer: Berlin/Heidelberg, Germany, 2019.
57. Kushwaha, N.; Pant, M.; Kant, S.; Jain, V.K. Magnetic optimization algorithm for data clustering. *Pattern Recognit. Lett.* **2018**, *115*, 59–65. [CrossRef]
58. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S. A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *J. Comput. Sci.* **2018**, *25*, 456–466. [CrossRef]
59. Abualigah, L.M.; Khader, A.T.; AI-Betar, M.A. Unsupervised feature selection technique based on harmony search. In Proceedings of the 2016 IEEE 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 13–14 July 2016.
60. Janani, R.; Vijayarani, S. Text document clustering using spectral clustering algorithm with particle swarm optimization. *Expert Syst. Appl.* **2019**, *134*, 192–200. [CrossRef]
61. Hasanzadeh, E.; Rokny, H.A.; Poyan, M. Text clustering on latent semantic indexing with particle swarm optimization (PSO) algorithm. *Int. J. Phys. Sci.* **2012**, *7*, 16–120. [CrossRef]
62. Bharti, K.K.; Singh, P.K. Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering. *Appl. Soft Comput.* **2016**, *43*, 20–34. [CrossRef]
63. Lu, Y.; Liang, M.; Ye, Z.; Cao, L. Improved particle swarm optimization algorithm and its application in text feature selection. *Appl. Soft Comput.* **2015**, *35*, 629–636. [CrossRef]
64. Chandran, T.R.; Reddy, A.; Janet, B. A social spider optimization approach for clustering text documents. In Proceedings of the 2016 IEEE 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, India, 27–28 February 2016; pp. 22–26.
65. Chandran, T.R.; Reddy, A.; Janet, B. Text clustering quality improvement using a hybrid social spider optimization. *Int. J. Appl. Eng. Res.* **2017**, *12*, 995–1008.
66. Gopal, J.; Brunda, S. Text Clustering Algorithm Using Fuzzy Whale Optimization Algorithm. *Int. J. Intell. Eng. Syst.* **2019**, *12*. [CrossRef]
67. Majhi, S.K. Fuzzy clustering algorithm based on modified whale optimization algorithm for automobile insurance fraud detection. *Evol. Intell.* **2019**, *2*, 1–12. [CrossRef]
68. Cobo, A.; Rocha, R. Document management with ant colony optimization metaheuristic: A fuzzy text clustering approach using pheromone trails. In *Soft Computing in Industrial Applications*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 261–270.

69. Nema, P.; Sharma, V. Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique. In Proceedings of the 2015 IEEE International Conference on Computers, Communications, and Systems (ICCCS), Kanyakumari, India, 2–3 November 2015; pp. 1–5.
70. Forsati, R.; Moayedikia, A.; Jensen, R.; Shamsfard, M.; Meybodi, M.R. Enriched ant colony optimization and its application in feature selection. *Neurocomputing* **2014**, *142*, 354–371. [CrossRef]
71. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A.; Alyasseri, Z.A.A.; Alomari, O.A.; Hanandeh, E.S. Feature selection with β-hill climbing search for text clustering application. In Proceedings of the 2017 IEEE Palestinian International Conference on Information and Communication Technology (PICICT), Gaza City, Palestine , 8–9 May 2017; pp. 22–27.
72. Abualigah, L.M.; Sawaie, A.M.; Khader, A.T.; Rashaideh, H.; Al-Betar, M.A.; Shehab, M. β-hill climbing technique for the text document clustering. In *New Trends in Information Technology (NTIT)–2017*; University of Jordan: Amman, Jordan, 2017; p. 60.
73. Saini, N.; Saha, S.; Bhattacharyya, P. Automatic scientific document clustering using self-organized multi-objective differential evolution. *Cogn. Comput.* **2019**, *11*, 271–293. [CrossRef]
74. Abualigah, L.M.; Hanandeh, E.S.; Khader, A.T.; Otair, M.A.; Shandilya, S.K. An Improved B-hill Climbing Optimization Technique for Solving the Text Documents Clustering Problem. *Curr. Med. Imaging* **2020**, *16*, 296–306. [CrossRef]
75. Moayedikia, A.; Jensen, R.; Wiil, U.K.; Forsati, R. Weighted bee colony algorithm for discrete optimization problems with application to feature selection. *Eng. Appl. Artif. Intell.* **2015**, *44*, 153–167. [CrossRef]
76. Abualigah, L.M.; Khader, A.T. Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. *J. Supercomput.* **2017**, *73*, 4773–4795. [CrossRef]
77. Abualigah, L.M.Q.; Hanandeh, E.S. Applying genetic algorithms to information retrieval using vector space model. *Int. J. Comput. Sci. Eng. Appl.* **2015**, *5*, 19.
78. Forsati, R.; Mahdavi, M.; Shamsfard, M.; Meybodi, M.R. Efficient stochastic algorithms for document clustering. *Inf. Sci.* **2013**, *220*, 269–291. [CrossRef]
79. Mohammad Abualigah, L.; Al-diabat, M.; Al Shinwan, M.; Dhou, K.; Alsalibi, B.; Said Hanandeh, E.; Shehab, M. Hybrid Harmony Search Algorithm to Solve the Feature Selection for Data Mining Applications. *Recent Adv. Hybrid Metaheuristics Data Clust.* **2020**, 19–37. [CrossRef]
80. Mahdavi, M.; Chehreghani, M.H.; Abolhassani, H.; Forsati, R. Novel meta-heuristic algorithms for clustering web documents. *Appl. Math. Comput.* **2008**, *201*, 441–451. [CrossRef]
81. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A. Multi-objectives-based text clustering technique using K-mean algorithm. In Proceedings of the 2016 IEEE 7th International Conference on Computer Science and Information Technology (CSIT), Amman, Jordan, 13–14 July 2016; pp. 1–6.
82. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S. A novel weighting scheme applied to improve the text document clustering techniques. In *Innovative Computing, Optimization and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 305–320.
83. Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A.; Alomari, O.A. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. *Expert Syst. Appl.* **2017**, *84*, 24–36. [CrossRef]
84. Sarkar, S.; Roy, A.; Purkayastha, B. A comparative analysis of particle swarm optimization and K-means algorithm for text clustering using Nepali Wordnet. *Int. J. Nat. Lang. Comput. (IJNLC)* **2014**, *3*. [CrossRef]
85. Romero, F.P.; Peralta, A.; Soto, A.; Olivas, J.A.; Serrano-Guerrero, J. Fuzzy optimized self-organizing maps and their application to document clustering. *Soft Comput.* **2010**, *14*, 857–867. [CrossRef]
86. Rashaideh, H.; Sawaie, A.; Al-Betar, M.A.; Abualigah, L.M.; Al-Laham, M.M.; Ra'ed, M.; Braik, M. A grey wolf optimizer for text document clustering. *J. Intell. Syst.* **2018**, *29*, 814–830. [CrossRef]
87. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S. A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis. *Eng. Appl. Artif. Intell.* **2018**, *73*, 111–125. [CrossRef]
88. Prabha, K.A.; Visalakshi, N.K. Improved Particle Swarm Optimization Based K-Means Clustering. In Proceedings of the 2014 International Conference on IEEE Intelligent Computing Applications (ICICA), Coimbatore, India, 6–7 March 2014; pp. 59–63.
89. Basu, T.; Murthy, C. A similarity assessment technique for effective grouping of documents. *Inf. Sci.* **2015**, *311*, 149–162. [CrossRef]

90. Zhong, N.; Li, Y.; Wu, S.T. Effective pattern discovery for text mining. *Knowl. Data Eng. IEEE Trans.* **2012**, *24*, 30–44. [CrossRef]
91. Bharti, K.K.; Singh, P.K. Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Syst. Appl.* **2015**, *42*, 3105–3114. [CrossRef]
92. Salton, G.; Wong, A.; Yang, C.S. A vector space model for automatic indexing. *Commun. ACM* **1975**, *18*, 613–620. [CrossRef]
93. De Vries, C.M. Document Clustering Algorithms, Representations and Evaluation for Information Retrieval. Ph.D. Thesis, Queensland University of Technology, Brisbane City, Australia, 2014.
94. Forsati, R.; Mahdavi, M. Web text mining using harmony search. In *Recent Advances in Harmony Search Algorithm*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 51–64.
95. Kaur, S.P.; Madan, N. Document Clustering Using Firefly Algorithm. *Artif. Intell. Syst. Mach. Learn.* **2016**, *8*, 182–185.
96. Kumar, L.; Bharti, K.K. A novel hybrid BPSO–SCA approach for feature selection. *Nat. Comput.* **2019**, 1–23. [CrossRef]
97. Mahdavi, M.; Abolhassani, H. Harmony K-means algorithm for document clustering. *Data Min. Knowl. Discov.* **2009**, *18*, 370–391. [CrossRef]
98. Del Buono, N.; Pio, G. Non-negative Matrix Tri-Factorization for co-clustering: An analysis of the block matrix. *Inf. Sci.* **2015**, *301*, 13–26. [CrossRef]
99. Inbarani, H.H.; Bagyamathi, M.; Azar, A.T. A novel hybrid feature selection method based on rough set and improved harmony search. *Neural Comput. Appl.* **2015**, *26*, 1859–1880. [CrossRef]
100. Bharti, K.K.; Singh, P.K. A three-stage unsupervised dimension reduction method for text clustering. *J. Comput. Sci.* **2014**, *5*, 156–169. [CrossRef]
101. Chen, L.; Liu, M.; Wu, C.; Xu, A. A Novel Clustering Algorithm and Its Incremental Version for Large-Scale Text Collection. *Inf. Technol. Control.* **2016**, *45*, 136–147. [CrossRef]
102. Singh, V.K.; Tiwari, N.; Garg, S. Document clustering using k-means, heuristic k-means and fuzzy c-means. In Proceedings of the Computational Intelligence and Communication Networks (CICN), 2011 International Conference on IEEE, Gwalior, India, 7–9 October 2011; pp. 297–301.
103. Geem, Z.W.; Kim, J.H.; Loganathan, G.V. A new heuristic optimization algorithm: Harmony search. *Simulation* **2001**, *76*, 60–68. [CrossRef]
104. Whitley, D. A genetic algorithm tutorial. *Stat. Comput.* **1994**, *4*, 65–85. [CrossRef]
105. Zhou, C.; Gao, H.; Gao, L.; Zhang, W.G. Particle Swarm Optimization (PSO) Algorithm. *Appl. Res. Comput.* **2003**, *12*, 7–11.
106. Dorigo, M.; Birattari, M.; Stutzle, T. Ant colony optimization. *IEEE Comput. Intell. Mag.* **2006**, *1*, 28–39. [CrossRef]
107. Gandomi, A.H.; Yang, X.S.; Alavi, A.H. Cuckoo search algorithm: A metaheuristic approach to solve structural optimization problems. *Eng. Comput.* **2013**, *29*, 17–35. [CrossRef]
108. Abualigah, L.; Shehab, M.; Alshinwan, M.; Alabool, H.; Abuaddous, H.Y.; Khasawneh, A.M.; Al Diabat, M. Ts-gwo: Iot tasks scheduling in cloud computing using grey wolf optimizer. In *Swarm Intelligence for Cloud Computing*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2020; pp. 127–152.
109. Alomari, O.A.; Khader, A.T.; Al-Betar, M.A.; Abualigah, L.M. Gene selection for cancer classification by combining minimum redundancy maximum relevancy and bat-inspired algorithm. *Int. J. Data Min. Bioinform.* **2017**, *19*, 32–51. [CrossRef]
110. Al-Sai, Z.A.; Abualigah, L.M. Big data and E-government: A review. In Proceedings of the 2017 IEEE 8th International Conference on Information Technology (ICIT), Amman, Jordan, 17–18 May 2017; pp. 580–587.

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).