Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

Multi-objective of wind-driven optimization as feature selection and clustering to enhance text clustering

Mehdi G. Duaimi[a], Qusay Bsoul[b*] and Abbas F. J. AL-Gburi[c]

[a]*Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq*
[b]*Cybersecurity and Cloud Computing Department, Applied Science Private University, Amman, Jordan*
[c]*Iraq general commission for custom, Iraqi Ministry of Finance, Baghdad, Iraq*

C H R O N I C L E

Article history:
Received: November 2, 2023
Received in revised format: November 25, 2023
Accepted: January 18, 2024
Available online: January 18, 2024

Keywords:
Text Clustering
Multi-Objectives
Wind Driven Optimization
K-Means
Unsupervised Feature Selection
Meta-heuristics optimization

A B S T R A C T

Text Clustering consists of grouping objects of similar categories. The initial centroids influence operation of the system with the potential to become trapped in local optima. The second issue pertains to the impact of a huge number of features on the determination of optimal initial centroids. The problem of dimensionality may be reduced by feature selection. Therefore, Wind Driven Optimization (WDO) was employed as Feature Selection to reduce the unimportant words from the text. In addition, the current study has integrated a novel clustering optimization technique called the WDO (Wasp Swarm Optimization) to effectively determine the most suitable initial centroids. The result showed the new meta-heuristic which is WDO was employed as the multi-objective first time as unsupervised Feature Selection (WDOFS) and the second time as a Clustering algorithm (WDOC). For example, the WDOC outperformed Harmony Search and Particle Swarm in terms of F-measurement by 93.3%; in contrast, text clustering's performance improves 0.9% because of using suggested clustering on the proposed feature selection. With WDOFS more than 50 percent of features have been removed from the other examination of features. The best result got the multi-objectives with F-measurement 98.3%.

1. Introduction

Clustering (Mehra et al., 2020) is a method of grouping related items together. This method primarily focuses on two main forms of clustering, namely hierarchical clustering and partitional clustering. While hierarchical clustering methods are generally regarded as more effective, they do not always ensure the reassignment of documents that are prone to initial missclassification during the clustering procedure (Jain et al.,1999) because of quadratic time complexity of the number of data items. As a result of their comparatively low computational cost, partitioning cluster approaches are more advantageous (Wu et al., 2018). Meanwhile, K- means algorithm is utilized as a fundamental aspect of partitioning clustering (Bsoul et al., 2013). This approach is primarily employed for partitional clustering, demonstrating a linear time complexity (Wu et al., 2018). It is also said that one of the primary goals of Hartigan's k-means method (Hartigan, 1981) is to represent each of the k groups by using the average document weights given to that group. This weighted average is referred to as the centroid of that group. On the contrary, the k-means algorithm exhibits reduced sensitivity to initialization and requires prior knowledge of the number of clusters. The selection of initial centroids is a crucial factor in determining the performance of the algorithm since it has been observed that the algorithm tends to get trapped in local optima solutions (Selim & Ismail, 1984).The main meta-heuristic used before for this work are Particle Swarm (Cagnina et al., 2014) and Harmony Search (Devi et al., 2015), where these algorithms are called global search optimization and they have weaknesses related to trapping into local optima

* Corresponding author.
E-mail address: abu4212@yahoo.com (Q. Bsoul)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2024.1.014

and exploration is poor when trying to solve complex multi-problems, global search optimization in convergence rate is slowly, and trapped in a local optimum (Zhe et al., 2011).

The affliction of dimensionality may be alleviated by feature selection. It reduces the problem's dimensionality by reducing unneeded and redundant features, which in turn enhances learning performance. The high dimensionality of textual data poses a recurring challenge in the field of text clustering (Mafarja & Mirjalili, 2017). The utilization of Unsupervised Feature Selection (UFS) is necessary to reduce the dimensions of textual data and select an optimal subset of features with high quality. However, this can be negative to the overall method's performance. Hence, this study focuses on investigating novel approaches to address the challenges of feature selection (FS) and dimensionality reduction, particularly in the domain of clustering textual data. In order to accomplish this task, researchers employ FS techniques. These techniques serve three main purposes: (a) enhancing performance through metrics such as f-measurement, (b) employing data visualization and simplification to facilitate model selection, and (c) reducing dimensionality by eliminating noise and irrelevant features (Mafarja & Mirjalili, 2017).

This study is structured objectively: The second part covers feature selection and clustering. Sections 3.1 and 3.2 use the multi-objective WDO approach for feature selection and clustering, respectively. The fourth section presents feature selection and cluster algorithm results using two internal assessment methods. Finally, the overview and next work recommendations are presented.

2. Related Work

The performance of cluster algorithms, including text clustering, is significantly influenced by the selection and distribution of characteristics (Abualigah et al., 2016, July; Bsoul et al., 2014). In contrast, the latter approach tends to prioritize the identification and analysis of local minima. In the majority of instances, the results are favorably received. This is particularly true if the first feature selections are a significant distance from one another. As a result, it can generally identify the most important class or category in each dataset. In addition, the cluster algorithms' core processing and extraction techniques' quality are both impacted by the features selection initialization procedure. Therefore, the quality of the product relies heavily on the beginning points (Mafarja & Mirjalili, 2017). For example, the cluster method may be unable to recognize the qualities of the primary categories in specific data if they are near or comparable. If the feature selection algorithm is left unregulated, this failure may also occur. A strong initial feature selection and improved performance in refining features to discover the best feature choices will be achieved (Abualigah et al., 2017, February). However, only few works incorporate optimization techniques for the purpose of unsupervised feature selection. Tabakhi et al. (2014) studied the ant colony optimization approach to UFS for classifiers. Their findings indicated that this method exhibited superior efficiency and success rates compared to alternative feature selection techniques previously employed by researchers. Bharti and Singh (2014) further postulated that a genetic algorithm using Mean Absolute Difference (MAD) as a measure of feature similarity might improve text clustering efficiency, which they tested against k-means clustering without feature selection. To evaluate this hypothesis, they conducted a comparative analysis between the aforementioned genetic algorithm and k-means clustering without feature selection. Abualigah et al. (2017, February) used harmony search as a method for FS. They conducted a comparison between their proposed approach and k-means clustering, utilizing MAD as a metric to measure the similarity between features. The authors asserted that their proposed method using FS had improved the performance of text clustering.

Previously in (Abualigah et al., 2018) an optimization technique was employed for feature selection. Specifically, particle swarm optimization (PSO) was utilized to address the feature selection problem. While MAD in (Bharti & Singh, 2014) was serving as the objective for measuring feature similarity. The proposed feature selection method outperformed alternative approaches, including Genetic Algorithm (GA) and Harmony Search (HS) algorithm, in terms of feature selection efficacy. Abualigah and Khader (2017) employed PSO and GA as alternative approaches to address the feature selection problem. Their work demonstrated that the proposed feature selection method outperforms previous approaches, as assessed using MAD. Empirical evaluations have been conducted on diverse datasets to assess the performance of clustering algorithms. Some researchers have focused on identifying the optimal cluster, while others have aimed to ascertain the appropriate number of clusters. To enhance Agglomerative Hierarchical Clustering, Dai et al. (2010, July) used the relevance of a story's first paragraph to improve the clustering algorithm. When the phrase appeared in the title, the title's weight was increased accordingly. The results demonstrated that the suggested strategy is successful in grouping financial news documents. The clustering of topics or events, however, has been studied by several scholars such that Bação et al. (2005) who utilized a self-organizing map for the purpose of clustering internal ribosome entry sites (IRES), and subsEq. uently conducted a comparison with the K-means algorithm. It was found that Self-Organizing Maps (SOM) exhibited superior performance compared to k-means in terms of performance. Subsequent work by Bouras and Tsogkas (2010, May) utilized several clustering methodologies in their research, encompassing k-means, ordinary k-means, k-medians, and k-means++ (Jo, 2009, July). Accordingly, it has been determined that the k-means algorithm demonstrates superior performance in both the internal evaluation of the clustering index function and in real-world user experimentation. The clustering subjects or occurrences in the news to compare the efficacy of k-means, single-pass algorithms, and other algorithms in the context of theme grouping has garnered significant attention from scholars in the academic community. A study conducted by Jo (2009, July) found that K-means clustering outperformed single-pass clustering. Further Dai et al. (2010, October) introduced a novel two-layer text

clustering methodology aimed at detecting retroactive news events. This approach leverages the Affinity Propagation (AP) clustering algorithm for its implementation. The clusters developed by researchers underwent a subsequent feature selection process. The ultimate news events were generated utilizing a conventional Agglomerative Hierarchical Clustering (AHC) algorithm. Researchers employed traditional hierarchical clustering and conventional k-Means for comparison. The results demonstrated that the proposed strategy exhibited the highest level of accuracy in which AP clustering, k-means clustering, and AHC emerged subsequently. The suggested technique and k-means algorithm demonstrated superior performance in terms of recall, while the AHC and AP clustering methods exhibited slightly lower results.

Velmurugan and Santhanam (2011) investigated of three cluster techniques applied to a geographic map data set. The k-means algorithm demonstrated favorable performance when applied to small datasets, while the k-medoids algorithm exhibited strong performance when dealing with large datasets. Additionally, the fuzzy c-means algorithm yielded results that were comparable to those obtained from both the k-means and k-medoids algorithms. Dueck and Frey (2007, October) suggested that Affinity Propagation (AP) has the potential to build multiple clusters automatically. However, based on the F-measure evaluation, it can be concluded that k-means outperforms AP in terms of recall. Qasim et al (2013) used 65 texts to compare four groups. The findings also showed that AP exhibited the highest performance in clustering, followed by Spectral, Hierarchical, and k-means algorithms. One of the primary limitations associated with the PSO algorithm is its inefficiency in addressing intricate multi-objective problems. Specifically, the PSO algorithm exhibits a slow convergence rate and tends to become trapped in a local optimum (Zhe et al., 2011). The Harmony Search (HS) has weaknesses in the control parameters and a slow level of convergence. The weaknesses observed in the previous algorithm necessitates the introduction of a meta-heuristic optimization, which is referred to as wind-driven optimization in this study. The Wind Driven Optimization (WDO) was employed in the selection of the most optimal initial centroids for the purpose of text clustering. One of the primary drawbacks of HS and PSO algorithms is their slow convergence rate, which often leads to encountering local optima rather than the desired global optimum (Yang et al., 2009; Labani et al., 2018).

Another limitation of the HS and PSO algorithms is their reliance on a larger number of control parameters compared to the Whale Optimization Algorithm (WOA) (Bayraktar et al., 2010, July). The only exceptions to the general control parameters, such as population size and maximum iteration number, are the control parameters limits. In contrast, the HS and PSO algorithms are characterized by the presence of seven control parameters (Cagnina et al., 2014; Devi et al., 2015). A detailed explanation of the problem is presented in part 3 of this study. Bayraktar et al. (2010, July) developed a novel meta-heuristic algorithm, known as the WDO optimization algorithm, to address certain limitations in existing approaches. Nevertheless, the effectiveness of WDO has been comparatively lower in comparison to text clustering. However, a notable drawback associated with the WDO algorithm is its susceptibility to early convergence, particularly when tackling complex global optimization problems characterized by large feature spaces. We propose two novel multi-objective algorithms, namely unsupervised feature selection and cluster-based global search optimization, to address the limitations of a single method in effectively handling global optimization problems.

3. The Proposed Text Clustering

In this study, we employ the Weighted Dominance Operator (WDO) as a methodology to devise novel approaches for the task of multi-objective unsupervised feature selection and clustering. The WDO algorithm encompasses several key procedures, namely population initialization, multi-objective fitness evaluation, selection, air parcel velocity calculation, solution position update, objectives fitness evaluation, and termination criteria. We have developed a novel wind-driven optimization approach with multiple objectives. The initial step involves modifying the solution space and the encoding of chromosomes, followed by the initialization of the solution. Next, we will discuss the techniques employed in generating novel solutions. Subsequently, an optimal solution will be selected and the structures will be advocated for wind-driven optimization based on outlined objectives. The multi-objective optimization of wind-driven advances typically involves the following phases.

Stage 1: Parameter arrangement and arrangement introduction
Stage 2: Multi-objective wellbeing estimation calculation
Stage 3: The generation of novel arrangement alternatives by considering the velocity and spatial position of the air parcel
Stage 4: The optimization solutions of unsupervised feature selection and clustering
Stage 5: Evaluation of the final condition. In the event that the circle is not complete, proceed to Stage 2. Stage 6 is typically executed at the conclusion of the cycle
Stage 6: The outcome of the preceding arrangement and serves as the final stage in this computational approach.

3.1. The Proposed Wind Driven as Unsupervised Feature Selection

Feature selection reduces the curse of dimensionality due to high-dimensional data. Eliminating redundant characteristics reduces the problem's dimensionality, improving learning performance. The challenges of feature subset selection is a frEq. uently encountered an issue in data mining (Mafarja & Mirjalili, 2017). The term “TC” refers to a particular context in which a problem persists in its ineffectiveness within the Text Classification. This issue necessitates the implementation of

unsupervised feature selection techniques, as the text contained in the documents exhibits a high level of dimensionality. Hence, the process of feature selection becomes imperative to effectively reduce the dimensionality of textual data and identify a suitable subset of high-quality features that can significantly influence performance outcomes. Therefore, novel methods were often discussed by researchers to address feature selection and the challenges posed by the curse of dimensionality, particularly in the field of TC. For example, a feature can be employed to enhance performance, such as accuracy, aid in data visualization and simplification for model selection, or decrease the number of features to mitigate noise and unnecessary features (Mafarja & Mirjalili, 2017). Feature selection and data distribution are highly influential upon algorithmic performance (Abualigah et al., 2016, July) and TC processes (e.g. extraction) (Romeo et al., 2019). The latter indicates a preference for generating local minima, rather than global minima, as a result. No matter how widely apart an individual's initial traits may be, an algorithm is typically capable of identifying a given data set's major category or class. Furthermore, the efficacy of extraction techniques and the primary TC process are both impacted by the initial feature selection, leading to the improvement of outcome quality (Mafarja & Mirjalili, 2017). For example, it is possible for TC to fail in accurately ranking documents based on certain selected features when those features are closely or identically related. The process of feature selection is delegated into two main components: filter and wrapper methods. The filter method disregards feature dependencies (Mafarja & Mirjalili, 2017; Abualigah et al., 2016, July; Zhang et al., 2019), while the wrapper method, through the utilization of optimization for feature selection, is able to establish a strong initial feature selection and ultimately achieve superior performance in the process of refining the features (Zhang et al., 2019).

The algorithms that were recommended employed the vector-space model to generate textual representations. Therefore, each term represents an individual component within the multi-dimensional term spaces. Additionally, each document $d_i = (w_{i1}, w_{i2} \ldots w_{in})$ is regarded as a vector within the term space, consisting of n distinct terms. Furthermore, the vector of features represented each potential solution for document detection. Hence, the problem of unsupervised feature selection was characterized as an optimization problem with the primary objective of identifying the most optimal features, rather than utilizing all available features. The objective task chosen for this study is to evaluate the quality of unsupervised feature selection. To achieve this, the unsupervised feature selection method of Wind-Driven was employed to optimize the objective.

This approach can effectively assess the performance of the TC objective through explicit testing. This, in turn, enhances our comprehension of how the TC performs on specific data types, thereby enabling the application of unsupervised feature selection objectives. Additionally, this study elucidated another advantage of this approach, specifically the possibility of concurrently evaluating multiple objectives (Bayraktar et al., 2010, July). In addition, when employing a global objective optimization meta-heuristic for UFS, it is imperative to consider various design options. The main considerations include the selection of the objective function and the representation of the problem. Both factors have significant effects on the optimization performance and quality of TC. Hence, it could be argued that the issue, characterized by the absence of supervision in the selection process, could be resolved through the optimization of tasks that primarily involve determining the value of a local optimum derived from a set of outcomes, treated as a fixed attribute. In this context, it can be argued that the identification of the quality of unsupervised features derived from the objective function and the identification of the Air Parcel value as a means of feature selection may encounter the aforementioned challenge of observing local optima and determining the most appropriate value. The utilization of specific features for the selection of the objective function can provide a comprehensive understanding of the algorithm's performance by leveraging various available features. Task-specific unsupervised feature selection enables the utilization of objectives by exclusively selecting exceptional features through the use of specific types of data.

Fig. 1 depicted the proposed technique using a variety of representations to encode the entirety of feature collection F, denoted by m representing the feature number. The aforementioned label is assigned to each of these vectors, irrespective of the presence or absence of the aforementioned characteristics which, as depicted in Fig. 1, serves as an example of a solution representation. The selected characteristics consist of numbers 1, 2, 5, 9, and 12, while the remaining features, namely 3, 4, 6, 7, 8, 10, and 11, are excluded in favor of the aforementioned 12.

F 1	F 2	F 3	F 4	F 5	F 6	F 7	F 8	F 9	F 10	F 11	F 12
1	1	0	0	1	0	0	0	1	0	0	1

Fig. 1. Some features represented by 0 or 1 [F= Feature]

3.1.1. Development of Initial features

The problem pertained to the variable values of the chosen features, which exhibited an array-like structure. The term “wind” was used to describe the horizontal movement of air. Temperature changes between regions can be ascribed to fluctuations in solar energy to reach the earth surface (Bayraktar et al., 2010, July). Regions with high temperatures are characterized by an abundance of warm air, whereas regions with lower temperatures exhibit a scarcity of cold air. The density and pressure of atmospheric air exhibit variability due to fluctuations in temperature that is due to horizontal pressure gradients at the higher-

pressure regions (Bayraktar et al., 2010, July). In the domain of Genetic Algorithms (GA), the problem variables' values are organized in arrays referred to as "chromosomes", while in Particle Swarm Optimization (PSO), these values are encapsulated within arrays known as "particles". The array utilized for features is referred to as "Air Parcel" in the presumed methodology. The air parcel in the N-dimensional space of problems referral system is represented by a 1 × N array, which can be defined as follows:

$$\text{Feature of Air Parcel} = [X_1, X_2, X_3 \ldots X_N] \tag{1}$$

At the onset of the analysis, an eligible representative is generated for a matrix of dimensions $N_{opt} \times N_{features}$, which corresponds to the air parcel matrix. This matrix can be interpreted as a collection of features, analogous to raindrops. The matrix X is generated randomly and presented as follows, with the columns representing the number of design variables and the rows representing the number of unsupervised feature selections.

$$\text{Air Parcels of features} = \begin{bmatrix} AirParcel_1 \\ AirParcel_2 \\ AirParcel_3 \\ \vdots \\ AirParcel_{opt} \end{bmatrix} \tag{2}$$

$$\begin{bmatrix} x_1^1 x_2^1 x_3^1 & \cdots & x_{N\text{features}}^1 \\ \vdots & \ddots & \vdots \\ x_1^{opt} x_2^{opt} x_3^{opt} & \cdots & x_{N\text{features}}^{opt} \end{bmatrix}$$

Rows and columns, conversely, denote the quantity of characteristics and design variables. Where opt the number of initial features (*AirParcels*) and $N_{features}$ denotes number of design variables. Leading to the opt and cost of each AirParcels after that is computed by assessing the following cost function (Cost):

$$Cost_i = f(x_1^i, x_2^i, \ldots x_{Nfeatures}^i) \quad i=1, 2, 3, \ldots, opt. \tag{3}$$

3.1.2. Cost of Solution

As aforementioned, each row in N_{pop} correspond to the number of features in the document. Let $f = (f_1, f_2 \ldots f_k)$ represent the set of features corresponding to a row in N_{pop}. The objective function of each row in N_{pop} is to confirm the Mean Absolute Difference (MAD) adopted from the study of Bharti and Singh (2014). MAD is correlated to determine the most relevant features within text classification, which aims to provide a score for every feature that reflects its relevancy. Such score is computed by measuring the difference between the sample and the mean values. The equation can be represented as follows:

$$MAD_i = \frac{1}{n} \sum_{j=1}^{n} \left| X_{ij} - \bar{X}_i \right| \tag{4}$$

where X_{ij} is the face with the document j and X_i is the mean of the feature i, as expressed in the equation as follows:

$$X_i = \left(\frac{1}{n}\right) \sum_{j=1}^{n} X_{ij} \tag{5}$$

The difference in horizontal air pressure "opt" induces the movement of air from regions of higher pressure to lower pressure, seeking equilibrium at the lowest pressure points.

3.1.3. The process of wind-driven as unsupervised feature selection

The air parcels are classified based on their pressure values in descending order as stated in Eq. (3) (Bayraktar et al., 2010, July), such as follows:

$$u_{new} = (1-\alpha)u_{cur} - gx_{cur} + \left(RT \left|\frac{1}{i} - 1\right| (x_{opt} - x_{cur})\right) + \left(\frac{cu_{cur}^{otherdim}}{i}\right) \tag{6}$$

As further clarification, i stand for the ranking among all air parcels. The most efficient resolution solution, which exhibits the least compression, is assigned the top rank (rank 1) within this framework and is determined at the $\boldsymbol{x}_{opt}$ point. Eq. (6) influences the final form of speed used in WDO. The positions of air parcels can be rationalized by utilizing the Eq. uation as follows:

$$x_{new} = x_{cur} + (u_{new}\Delta t) \tag{7}$$

On the contrary, $\boldsymbol{x}_{cur}$ represents the current location of the air parcel within the designated search area, $\boldsymbol{x}_{new}$ denotes the new position for the subsEq. uent iteration, with a time step of $\Delta t = 1$ being executed. By using Eq. (6) and Eq. (7), the speed and position of each are changed. These changes occur iteratively as the parcel progresses towards the optimal pressure point. Consequently, the optimal solution is attained at the conclusion of the final iteration and the velocity of the adjusted parcel is constrained to the maximum iteration value to avoid air parcel from excessively consuming large steps in the search space. If the velocity value in any dimension surpasses the specified maximum, it is reduced accordingly.

$$u^*_{new} = \begin{cases} u_{max} \;\; if \;\; u_{new} > u_{max} \\ -u_{max} \;\; if \;\; u_{new} < -u_{max} \end{cases} \tag{8}$$

After all updates, the evaluation of parcel pressures will be conducted in the new location until the maximum number of iterations is attained. Ultimately, the optimal objective functions of the MAD area will be documented at the conclusion of the final cycle, resulting from the optimization process. These optimal objective functions represent the most favorable candidate solution for the selected features, providing further insights into the deep wind-driven optimization algorithm (Bayraktar et al., 2010, July).

Set user parameters of the WDO: *Popt, α, RT, C, g, and Maximum_ Iteration.*

Generate randomly initial population of feature selection using Eqs. (1-2)
Randomize the velocity and the position of air parcels
Set the fitness function using Eq. (5)
Evaluate the population and find the minimum pressure value for air parcels ($P_0 = min(f(x_j))$)
Set global $\boldsymbol{P} = P_0$
for all *i* **do**
for all *j* **do**
Update the velocity by Eq. (6)
Choose random dimension
Choose velocity based on random dimension
Check velocity by Eq. (8)
Update air parcel positions by Eq. (7)
Call CMAES
Return the new set of WDO coefficients for *i+1*
Evaluate the new soluation and update *P(i;j)* using Eq. (5)
if *P(i;j)* < *P0* then
global*P* = *P(i;j)*
end if
end for
Rank air parcels and find the local optimum (*xlocal*)
end for
Rank air parcels and find the global optimum (*xglobal*)
Return the values of *xglobal* and *fmin = min(f(xbest))*
End of the algorithm

Fig. 2. pseudo-code of WDO as feature selection

3.1.4. Stop Criteria

The stop criteria for the WDOFS algorithm consist of two conditions: either the fitness average remains unchanged within a predetermined value Ɛ= dmax following a specific number of iterations, or the maximum number of generations is achieved.

3.2. Text Clustering Evaluation

The clustering process (Mehra et al., 2020) typically involves the grouping of objects based on their similarities. Specifically, the k-means is important in partitioning during the clustering process (Anitha & Patil, 2022). This approach is commonly employed for partition-based clustering, with a linear time complexity (Gbadoubissa et al., 2020). The centroid of a cluster is determined by calculating the mean of its nearest objects. However, it is highly sensitive to its initialization hence rEq. uires prior knowledge of the number of clusters in the original dataset.

The selection of initial centroids has a crucial role in determining the effectiveness of the clustering algorithm and may lead to the algorithm becoming stuck in a suboptimal solution (Wan et al., 2018). In order to mitigate reliance on a particular dataset and initialization, it is advantageous to acquire well-performing initial clustering centroids, enhance the precision of these centroids, and ascertain the ideal clustering centres (Sahmoudi & Lachkar, 2017). The application of nature-inspired

metaheuristic algorithms is a prevalent in computer science, data mining, industry, agriculture, forecasting, medicine and biology, scheduling, economy, and engineering (Brahimi et al., 2021; Manoj & Elias, 2012). Therefore, this study used a text optimisation cluster. However, this investigation only showed the inadequacies of the harmony search algorithm with k-means clustering, as provided by Forsati et al (2013), using their optimum parameter settings. Thus, the harmony search algorithm's drawbacks compel the search for an alternative optimisation method. Harmony Search (HS) has three essential control parameters: pitch adjustment rate (phmcr), pitch adjustment range (p par), and bandwidth (BW) (Akay & Karaboga, 2009; Yang et al., 2009). Wind-Driven optimization was proposed as a clustering strategy to mitigate the limitations associated with harmony search.

3.2.1 The Proposed Wind Driven Optimization Clustering

In this section, we describe how to employ the new meta-heuristic called wind-driven clustering. The vector-space model shows each phrase as a matrix space dimension. Thus, every document $d_i = (w_{i1}, w_{i2} \ldots w_{in})$ is a term space vector with n terms. Clustering solutions include the vector of centroids. Thus, clustering is an optimisation task to find the optimal cluster centroids instead of optimal partitions. An objective function selected the clustering quality, and Wind-Driven optimization Clustering was used to improve the clustering algorithm's performance on specific data types and enabled task-specific clustering objectives. A prior study has identified an additional advantage of utilizing the approach, specifically the ability to simultaneously consider different objectives (Handl & Knowles, 2007). Clustering using a general-purpose optimisation metaheuristic requires careful design choice. The objective function and issue representation affect clustering optimization quality and performance, hence the major option represents them.

Fig. 3 shows how the algorithm encodes the whole document set partition using specified representations. Each document has an m-length vector and a P-number. This vector uniquely identifies documents. Each solution vector has an integer value from 1 to *K*, and *K* is the number of clusters. Each correct assignment of *K* non-empty clusters includes the correspondence between *K* centroids, ensuring legality. Permutation sets of size m from 1 to *K* confine the algorithm's search space. Every document can be assigned to a search space cluster, ensuring no cluster is empty. NP-hard despite *K*=2. After that, one way to describe this permutation is to see each Wind-Driven Optimisation Cluster (WDOC) row as a vector of integers with m locations. This representation maps the ith location to the document's cluster. Fig. 3 presents solutions. In this case, the cluster assigned label 3 produced four documents {*2, 3, 7, and 8*} originating from the cluster were assigned to label *3*, while label 2 produced three documents {*4, 6, 9*}, and so on.

D 1	D 2	D 3	D 4	D 5	D 6	D 7	D 8	D 9	D 10	D 11	D 12
5	3	3	2	1	4	3	3	2	5	2	5

Fig. 3. Documents represented by their number of groups from 1 to 5 [D=Document]

3.2.2. Generate the Initial Clusters

The values pertaining to the problem variables typically exhibit a tendency to congregate into an array. In terms of PSO (Cagnina et al., 2014) and GA (Devi et al., 2015) lexicons, such array is referred to as 'Particle Position' and 'Chromosome', respectively. Hence, the label is termed as 'Air Parcels cluster' and defines a single cluster. In the context of an N-*dimensional* cluster problem, an Air Parcel can be formally characterized as a one-dimensional array with dimensions of $1 \times N$. The following array is defined below:

$$\text{Air Parcel } cluster = [X_1, X_2, X_3 \ldots X_N cluster] \tag{9}$$

The clustering algorithm commences by executing the generation of the candidate genera, which represents a matrix of raindrops denoted as $Nopt \times Ncluster$. Therefore, the matrix *X*, which is generated arbitrarily, can be represented in a format where the columns represent the number of design variables, and the rows indicate the number of clusters.

$$Raindrops\ of\ cluster = \begin{bmatrix} Raindrop_1 \\ Raindrop_2 \\ Raindrop_3 \\ \vdots \\ Raindrop_{opt} \end{bmatrix} \begin{bmatrix} x_1^1 x_2^1 x_3^1 & \cdots & x_N^1 \\ \vdots & \ddots & \vdots \\ x_1^{opt} x_2^{opt} x_3^{opt} & \cdots & x_N^{opt} \end{bmatrix} \tag{10}$$

Floating-point numbers, also known as real values, can be utilized to represent the values of each decision variable ($X_1, X_2, X_3 \ldots X_N$). In this context, *Nopt* denotes the number of Air Parcels (initial cluster) and *Nclusters* denotes the number of design

variables. The initial step involves the establishment of *Nopt Air Parcels*, followed by the evaluation of the cost of a raindrop through the analysis of the cost function (Cost) depicted as follows:

$$\text{Cost}_i = f(x_1^i,\ x_2^i, \ldots x_{Nvar}^i)\ \ i=1, 2, 3,\ldots, \text{Nopt.} \tag{11}$$

3.2.3. Cost of Solutions

The calculation of each solution in *Npop* corresponds to a document cluster, where each element in the solution represents the cluster number as C = (c_1, c_2…$c_k is$). The *C* represents the set of *K* centroids corresponding to a row in *Nopt*. The centroid of the kth cluster is c_k= *(*c_{k1}*…,* which can be computed as follows:

$$c_{kj} = \frac{\sum_{i=1}^{m} aki\ dij}{\sum_{I=1}^{n} aki} \tag{12}$$

The objective is to validate cluster centroids and optimises similarity within each cluster by minimising distances inside the cluster and between clusters by minimising distances between them. The row concerns the average document distance to the cluster centroid and its fitness value. This information is subsequently linked to a viable solution. The condition is commonly known as Attention Deficit Disorder with Hyperactivity (ADDC).

$$Costi = [\sum_{i=1}^{UB} \quad]/UB \tag{13}$$

The cosine similarity is represented as the *D (.,.)*, m_i denotes the number of documents in cluster *i* (e.g., (ni=$\sum_{j=1}^{n}$ aij), *dij* is the j^{th} document of cluster *i*, and *UB* is the number of clusters. In the event that the locally optimized vector results in a superior cost value compared to the solutions in *Not*, the newly generated solution can be substituted with a row in *Nopt*.

3.2.4. The process of wind-driven clustering

The air parcels are classified based on their pressure values in descending order as Eq. (13), such as follows:

$$u_{new} = (1-\alpha)u_{cur} - gx_{cur} + \left(RT\ \left|\frac{1}{i} - 1\right|\left(x_{opt} - x_{cur}\right)\right) + \left(\frac{cu_{cur}^{otherdim}}{i}\right) \tag{14}$$

To clarify, *i* represent the ranking among all air parcels. The optimal resolution solution has the lowest compression ranked 1 in this scheme and is determined at the $\boldsymbol{x}_{opt}$ point. Eq. (14) influences the final form of speed update used in WDO. The positions of an air parcel can be rationalized through the utilization of the subsequent equation:

$$x_{new} = x_{cur} + (u_{new}\Delta t) \tag{15}$$

On the other hand, $\boldsymbol{x}_{cur}$ means the current position of the air parcel in the search area, $\boldsymbol{x}_{new}$ is the new position for the next iteration and the time step Δt = 1 is performed. By using Eq. (14) and Eq. (15), the speed and position of each entity are altered. The characteristics of the parcel undergo modifications during each iteration as it transitions towards the pressure point that yields the most favorable outcome. Hence, it is advisable to position the optimal solution at the conclusion of the final iteration to prevent the modified air parcel from taking excessive steps and disregarding specific search regions. This is achieved by reducing the speed of the modified air parcel to the maximum iteration value. The velocity diminishes when the velocity value in each dimension attains the specified maximum.

$$u_{new}^{*} = \begin{cases} u_{max} & if\ \ u_{new} > u_{max} \\ -u_{max} & if\ \ u_{new} < -u_{max} \end{cases} \tag{16}$$

After all updates, an evaluation of the parcel pressures in the new location will be conducted until the maximum number of iterations is attained. The optimal objective functions of the Adaptive Differential Dynamic Cluster (ADDC) are documented at the conclusion of the final iterations due to the optimization process, making it the most favorable candidate solution for the initial cluster in order to obtain further insights into the deep wind-driven optimization algorithm (Bayraktar et al., 2010, July).

Set user parameters of the WDO: *Popt, α, RT, C, g, and Maximum_ Iteration.*

Generate randomly initial population of initial centroids using Eq. (9-10)
Randomize the velocity and the position of air parcels
Set the fitness function using Eq. (13)
Evaluate the population and find the minimum pressure value for air parcels ($P_0 = min(f(x_j))$)
Set global $\boldsymbol{P} = P_0$

```
for all i do
 for all j do
      Update the velocity by Eq.  (14)
      Choose random dimension
      Choose velocity based on random dimension
      Check velocity by Eq.  (16)
      Update air parcel positions by Eq.  (15)
      Call CMAES
      Return the new set of WDO coefficients for i+1
      Evaluate the new soluation and update P(i;j) using Eq.  (3)
            if P(i;j) < P0 then
                  globalP = P(i;j)
            end if
  end for
      Rank air parcels and find the local optimum (xlocal)
 end for
 Rank air parcels and find the global optimum (xglobal)
Return the values of xglobal and fmin = min(f(xbest))
 End of the algorithm
```

Fig. 4. pseudo-code of WDO as clustering

3.3.5. Stop Criteria

The stop criteria for the WDOC algorithm are twofold: either the fitness average remains unchanged within a predetermined value Ɛ= dmax after a certain number of iterations, or the maximum number of generations is achieved.

4. Experimental Settings

4.1. Parameter Setting Of Wind Driven Optimization As Unsupervised Feature Selection And As Clustering

There are five parameters in wind-driven optimization, and the main setting of tuning parameters in the experiments as in Table 1.

Table 1
The parameter settings of WDO and SA

Parameter	Meaning	Value
NP	The number of *opting* solutions	100
G	gravitational constant	0.2
RT	coefficient	3
Gmax	The number of iterations	500
C	Coriolis effect	0.4
A	constants alpha	0.4

4.2. Performance Measurement and Datasets

This study evaluated the external condition using the universal F-measure (Larsen & Aone 1999, August; Jardine & van Rijsbergen 1971). The ideal cluster has a higher F-measure, which includes information retrieval accuracy and recall. It is anticipated that every class will possess a distinct assortment of essential documents, and each class will maintain its obligatory compilation of documents. The F-measure is a metric that ranges from 0 to 1, with higher values indicating superior clustering performance. In this study, a comprehensive evaluation of algorithm performance was conducted by utilizing four distinct and autonomous datasets. This approach ensured a rigorous and unbiased comparison and assessment of the algorithms. During the text mining procedure, the primary dataset, referred to as classic 3, was utilized as a standard for comparison. The dataset comprises 3892 documents that are classified into three distinct categories. Specifically, there are 1399 documents pertaining to aviation systems (CRAN), 1033 documents related to medical conditions (MED), and 1460 documents focused on information retrieval (CISI) (Common IR Test Collection, 2010).

The second dataset consisted of 1,445 CNN news documents that were selected from the TDT2 and TDT3 corpora and subsEq. uently incorporated. The i-Event experiment incorporates data from the TDT2 and TDT3 corpora as a replication of the dataset (Mohd et al.,2012). Typically, a sufficient number of selected documents is necessary for experimentation and cluster generation on the user interface. Furthermore, the selection of these sources was influenced by the concise nature of CNN's articles and the relevance of their events. The dataset used in this investigation, known as the 20 newsgroups data (Lang, 2008), comprises a total of 10,000 messages. These messages were gathered from 10 distinct Usenet newsgroups, with each newsgroup containing 1000 messages. This dataset is the third one employed in the current study. After undergoing pre-

processing, the dataset contained a total of 3831 documents. ConsEq. uently, a data set comprising 20 newsgroups was employed to evaluate the efficacy of algorithms in handling large-scale datasets. One additional dataset that has been widely employed in previous scholarly investigations is Reuters-21578 (Lewis, 1997), which serves as a test collection for text classification. Nevertheless, there exist several constraints associated with the process of data collection in this context. A considerable number of documents are assigned to multiple classes, however, a majority of the documents lack annotations for class labels. Furthermore, the data distribution within this dataset exhibits uniformity across various categories. Certain classes, such as “earn” and “acquisition”, exhibit a substantial volume of documents, whereas others, such as “reserve” and “veg-oil”, possess a comparatively limited number of documents. Therefore, it was determined that a dataset comprising 8 primary categories and 1100 documents per category would be employed in this study to address these limitations.

Table 2
Summary description of document set

Document	Source	#of document	#of cluster
DS1	Classic 3	3892	3
DS2	TDT2 and TDT3 of TREC 2001	1445	53
DS3	20 NEWSGROUP	3831	10
DS4	routers	4195	8

4.3. Results of the WDOFS with WDOC algorithm

As presented in Table 3, the summary consists of cluster performance and the number of features obtained using the Bag-of-Words (BOW) approach and the k-means algorithm. The article cites the works of (Abualigah et al., 2018; Cagnina et al., 2014; Devi et al., 2015; Forsati et al., 2013), as well as our proposed novel approaches, WDOUFS and WDOC, for text clustering optimization. The results indicate that the highest f-measure attained using the k-means clustering algorithm was 86.9% in DS1, while the lowest f-measure was observed to be 58.2% in DS3. The PSOC optimization yielded the highest f-measure of 89.1% for dataset DS1, while the lowest f-measure of 60.6% was attained with PSOC for DS3. The highest f-measure achieved by the HSC was 90.9% for the dataset DS1, while the lowest f-measure obtained with the HSC was 61.4% in DS3. The WDOC algorithm demonstrated the highest f-measure of 93.3% for the dataset DS1, while the lowest f-measure of 64.9% was observed in DS3 when employing the WDOC algorithm. The KHCluster algorithm demonstrated the highest f-measure of 93.5% for the dataset DS1, while the lowest f-measure of 65.9% was obtained with KHCluster for DS3.

The WDOFS algorithm, when combined with k-means, yielded the highest f-measure of 95.8% for dataset DS1. Conversely, the lowest f-measure of 66.1% was obtained using WDOFS with k-means for dataset DS3. WDOFS successfully reduced the number of features in DS1 by 7096 out of 13310, in DS2 by 4731 out of 6737, in DS3 by 10346 out of 27211, and in DS4 by 5651 out of 12152. The WDOFS method, in conjunction with WDOC, demonstrated the highest f-measure of 98.3% for dataset DS1, while the KHCluster method achieved the lowest f-measure of 68.6% for DS3. WDOFS successfully reduced the number of features in DS1 by 7096 out of 13310, in DS2 by 4731 out of 6737, in DS3 by 10346 out of 27211, and in DS4 by 5651 out of 12152. The highest performing text clustering approach was observed with the proposed Weighted Distance-based Optimal Feature Selection (WDOFS) combined with Weighted Distance-based Optimal Clustering (WDOC). Following this, the WDOFS combined with the k-means algorithm exhibited a relatively strong performance. The KHCluster algorithm ranked third in terms of performance, while the proposed WDOC, Hierarchical Spectral Clustering (HSC), Particle Swarm Optimization Clustering (PSOC) algorithms demonstrated moderate performance. The k-means algorithm displayed the lowest performance among the evaluated methods. The use of UFS and k-means algorithms has been shown to improve the performance of text clustering. In this experiment, the features selected by the WDOFS algorithm in conjunction with k-means and WDOC were utilized to ensure unbias comparison between the hybrid k-means and multi-objective WDOC approaches, as both methods employed the same number of features.

Table 3
The f-measurement of the proposed hybrid multi-objective clustering

Comparison	***DS1***		***DS2***		***DS3***		***DS4***	
	features	*f - measure*	*features*	*f - measure*	*features*	*f - measure*	*features*	*f - measure*
k-means	13310	0.869	6737	0.804	27211	0.582	12152	0.636
PSOC	13310	0.891	6737	0.841	27211	0.606	12152	0.688
HSC	13310	0.909	6737	0.837	27211	0.614	12152	0.683
WDOC	13310	0.933	6737	0.85	27211	0.649	12152	0.704
KHCluster	13310	0.935	6737	0.853	27211	0.659	12152	0.712
WDOFS with k-means	**7096**	0.958	**4731**	0.859	**10346**	0.661	**5651**	0.726
WDOFS with WDOC	**7096**	**0.983**	**4731**	**0.875**	**10346**	**0.686**	**5651**	**0.744**

****best result: underline and bold**

The proposed hybridise WDO multi-objective technique for unsupervised feature selection and clustering was further evaluated statistically. This investigation will evaluate text clustering's ideal performance and whether this multi-objective technique for unsupervised feature selection and clustering yields meaningful differences. Table 4 presents the rankings of the proposed multi-objectives of unsupervised feature selection, as well as clustering and other algorithms, based on the Friedman criterion. The ranking is determined by the criterion, where a lower value corresponds to a higher rank.

Table 4's last two rows show Friedman and Iman-Davenport p-values. Table 10 shows that our strategy, which uses WDO for unsupervised feature selection and grouping, had the lowest value and ranked top. Whilst Hybridize WDO as unsupervised feature selection with k-mean, KHCluster, WDOCe, HSC, PSOC, and K-means is in the 2nd, 3rd, 4th, 5th, 6th, and 7th ranks, respectively.

Table 4
The ranking of the proposed algorithms using the Friedman test

Algorithms	Ranking
K-means	10.18
PSOC	10.15
HSC	10.01
WDOC	9.39
KHCluster	9.3
WDOFS with k-mean	9.09
WDOC and WDOFS	**8.53**
Friedman test (p-value)	0.00
man-Davenport (p-value)	0.00

***The best is in bold font**

5. Conclusion

This study aimed to enhance existing text clustering algorithms, namely PSO, HS global search, and local search methods like K-means, which have been utilized in prior research. One of the primary limitations associated with the PSO algorithm is its poor performance in addressing complex multi-objective problems. Specifically, it exhibits a slow convergence rate in global search optimization and often becomes trapped in local optima (Zhe et al., 2011). The weakness associated with HS pertains to deficiencies in the control parameters and a relatively sluggish rate of convergence (Yang et al., 2009; Labani et al., 2018). An additional limitation of the HS and PSO algorithms is their reliance on a larger number of control parameters compared to the WOA (Bayraktar et al., 2010, July). The HS, PSO algorithm is characterized by the presence of seven control parameters (Cagnina et al., 2014; Devi et al., 2015). Therefore, we proposed the utilization of the Weighted Differential Evolutionary Optimization (WDO) algorithm as a multi-objective approach. The decision to utilise the WDO algorithm is driven by its ability to effectively overcome the drawbacks of sluggish convergence rates and the propensity to be locked in local optima.

The k-means algorithm's main drawbacks include poor accuracy and slow execution. K-means is susceptible to local optima and responsive to early proposals (Bsoul et al., 2013). One issue that arises in the context of large-scale features is the potential for data mis-clustering (Mafarja & Mirjalili, 2017). The foundation of our proposed multi-objective global search optimization approach is the utilization of wind-driven optimization for unsupervised feature selection and cluster formation. In the present study, two internal fitness functions were implemented, which were derived from Mean Absolute Deviation (MAD) and the Average Distance to Desired Class (ADDC) metrics.

In this study, we have conducted a comparative analysis with the k-means algorithm as introduced by Abualigah et al. (2018), the PSOC algorithm as presented by Cagnina et al (2014), the HSC algorithm as introduced by Devi et al (2015), and the KHCluster algorithm as described by Forsati et al (2010). These findings indicate that the hybridization of WDOFS-WDOC for text clustering enhances efficiency and surpasses the performance of other evaluated methods. Nevertheless, there are still unresolved research inquiries in this paper. Our organization provides opportunities for further research and development in the field of text clustering, specifically focusing on the utilization of combined WDOC (Weighted Document) with local search as a clustering technique. It is also advisable to consider assessing certain novel internal evaluation methods that are more intricate than the MAD and ADDC techniques employed in this study.

References

Abualigah, L. M., & Khader, A. T. (2017). Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. *The Journal of Supercomputing, 73*, 4773-4795.

Abualigah, L. M., Khader, A. T., & Al-Betar, M. A. (2016, July). Unsupervised feature selection technique based on genetic algorithm for improving the text clustering. *In 2016 7th international conference on computer science and information technology (CSIT)* (pp. 1-6). IEEE.

Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *Journal of Computational Science, 25*, 456-466.

Abualigah, L. M., Khader, A. T., AlBetar, M. A., & Hanandeh, E. S. (2017, February). A new hybridization strategy for krill herd algorithm and harmony search algorithm applied to improve the data clustering. *In First EAI international conference on computer science and engineering* (pp. 54-63).

Akay, B., & Karaboga, D. (2009). Parameter tuning for the artificial bee colony algorithm. In Computational Collective Intelligence. *Semantic Web, Social Networks and Multiagent Systems: First International Conference, ICCCI 2009, Wrocław, Poland, October 5-7, 2009. Proceedings 1 (pp. 608-619). Springer, Berlin.*

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences, 34*(5), 1785-1792.

Bação, F., Lobo, V., & Painho, M. (2005). Self-organizing maps as substitutes for k-means clustering. *In Computational Science–ICCS 2005: 5th International Conference, Atlanta*, GA, USA, May 22-25, 2005, Proceedings, Part III 5 (pp. 476-483). Springer Berlin Heidelberg.

Bayraktar, Z., Komurcu, M., & Werner, D. H. (2010, July). Wind Driven Optimization (WDO): A novel nature-inspired optimization algorithm and its application to electromagnetics. *In 2010 IEEE antennas and propagation society international symposium (pp. 1-4). IEEE.*

Bharti, K. K., & Singh, P. K. (2014). A three-stage unsupervised dimension reduction method for text clustering. *Journal of Computational Science, 5*(2), 156-169.

Bouras, C., & Tsogkas, V. (2010, May). Assigning web news to clusters. *In 2010 Fifth International Conference on Internet and Web Applications and Services* (pp. 1-6). IEEE.

Brahimi, B., Touahria, M., & Tari, A. (2021). Improving sentiment analysis in Arabic: A combined approach. *Journal of King Saud University-Computer and Information Sciences, 33*(10), 1242-1250.

Bsoul, Q., Al-Shamari, E., Mohd, M., & Atwan, J. (2014). Distance measures and stemming impact on arabic document clustering. *In Information Retrieval Technology: 10th Asia Information Retrieval Societies Conference*, AIRS 2014, Kuching, Malaysia, December 3-5, 2014. Proceedings 10 (pp. 327-339). Springer International Publishing.

Bsoul, Q., Salim, J., & Zakaria, L. Q. (2013). An intelligent document clustering approach to detect crime patterns. *Procedia Technology, 11*, 1181-1187.

Cagnina, L., Errecalde, M., Ingaramo, D., & Rosso, P. (2014). An efficient particle swarm optimization approach to cluster short texts. *Information Sciences, 265*, 36-49.

Common IR Test Collection (2010). http://web.eecs.utk.edu/research/lsi/corpa.html

Dai, X. Y., Chen, Q. C., Wang, X. L., & Xu, J. (2010, July). Online topic detection and tracking of financial news based on hierarchical clustering. *In 2010 International Conference on Machine Learning and Cybernetics* (Vol. 6, pp. 3341-3346). IEEE.

Dai, X., He, Y., & Sun, Y. (2010, October). A two-layer text clustering approach for retrospective news event detection. *In 2010 International Conference on Artificial Intelligence and Computational Intelligence* (Vol. 1, pp. 364-368). IEEE.

Devi, S. S., Shanmugam, A., & Prabha, E. D. (2015). A proficient method for text clustering using harmony search method. *International Journal of Scientific Research in Science, Engineering and Technology, 1,* 145-150.

Dueck, D., & Frey, B. J. (2007, October). Non-metric affinity propagation for unsupervised image categorization. In 2007 IEEE 11th international conference on computer vision (pp. 1-8). IEEE.

Forsati, R., Mahdavi, M., Shamsfard, M., & Meybodi, M. R. (2013). Efficient stochastic algorithms for document clustering. *Information Sciences, 220,* 269-291.

Gbadoubissa, J. E. Z., Ari, A. A. A., & Gueroui, A. M. (2020). Efficient k-means based clustering scheme for mobile networks cell sites management. *Journal of King Saud University-Computer and Information Sciences, 32*(9), 1063-1070.

Handl, J., & Knowles, J. (2007). An evolutionary approach to multiobjective clustering. *IEEE transactions on Evolutionary Computation, 11*(1), 56-76.

Hartigan, J. A. (1981). Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association, 76*(374), 388-394.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31*(3), 264-323.

Jardine, N., & van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information storage and retrieval, 7*(5), 217-240.

Jo, T. (2009, July). Clustering news groups using inverted index based NTSO. *In 2009 First International Conference on Networked Digital Technologies* (pp. 1-7). IEEE.

Labani, M., Moradi, P., Ahmadizar, F., & Jalili, M. (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence, 70*, 25-37.

Lang, K. (2008). The 20 news groups data set. http://people. csail. mit. edu/jrennie/20Newsgroups/.

Larsen, B., & Aone, C. (1999, August). Fast and effective text mining using linear-time document clustering. *In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 16-22).

Lewis, D. D. (1997). Test Collections: Reuters-21578. URL: http://www. daviddlewis. com/resources/testcollections/reuters21578.

Manoj, V. J., & Elias, E. (2012). Artificial bee colony algorithm for the design of multiplier-less nonuniform filter bank transmultiplexer. *Information Sciences, 192*, 193-203.

Mafarja, M. M., & Mirjalili, S. (2017). Hybrid whale optimization algorithm with simulated annealing for feature selection. *Neurocomputing, 260*, 302-312.

Mehra, P. S., Doja, M. N., & Alam, B. (2020). Fuzzy based enhanced cluster head selection (FBECS) for WSN. *Journal of King Saud University-Science, 32*(1), 390-401.

Mohd, M., Crestani, F., & Ruthven, I. (2012). Evaluation of an interactive topic detection and tracking interface. *Journal of information science, 38*(4), 383-398.

Qasim, I., Jeong, J. W., Heu, J. U., & Lee, D. H. (2013). Concept map construction from text documents using affinity propagation. *Journal of Information Science, 39*(6), 719-736.

Romeo, S., Da San Martino, G., Belinkov, Y., Barrón-Cedeño, A., Eldesouki, M., Darwish, K., ... & Moschitti, A. (2019). Language processing and learning models for community question answering in arabic. *Information Processing & Management, 56*(2), 274-290.

Sahmoudi, I., & Lachkar, A. (2017). Formal concept analysis for Arabic web search results clustering. *Journal of King Saud University-Computer and Information Sciences, 29*(2), 196-203.

Selim, S. Z., & Ismail, M. A. (1984). K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, *1*, 81-87.

Tabakhi, S., Moradi, P., & Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence, 32*, 112-123.

Velmurugan, T., & Santhanam, T. (2011). A survey of partition based clustering algorithms in data mining: An experimental approach. *Information Technology Journal, 10*(3), 478-484.

Wan, Y., Zhong, Y., & Ma, A. (2018). Fully automatic spectral–spatial fuzzy clustering using an adaptive multiobjective memetic algorithm for multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing, 57*(4), 2324-2340.

Wu, J., Dong, M., Ota, K., Li, J., & Guan, Z. (2018). Big data analysis-based secure cluster management for optimized control plane in software-defined networks. *IEEE Transactions on Network and Service Management, 15*(1), 27-38.

Yang, F., Sun, T., & Zhang, C. (2009). An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization. *Expert Systems with Applications, 36*(6), 9847-9852.

Zhang, Y., Li, H. G., Wang, Q., & Peng, C. (2019). A filter-based bare-bone particle swarm optimization algorithm for unsupervised feature selection. *Applied Intelligence, 49,* 2889-2898.

Zhe, G., Dan, L., Baoyu, A., Yangxi, O., Wei, C., Xinxin, N., & Yang, X. (2011). An analysis of ant colony clustering methods: Models, algorithms and applications. *International Journal of Advancements in Computing Technology, 3*(11), 112-121.