

Unsupervised Feature Selection Using the Atomic Orbital Search Algorithm for Information Retrieval

Sattam Abdallah Alyusuf^{1*}, and Mohd Zakree Ahmad Nazri²

¹Faculty of Information Science & Technology, University Kebangsaan Malaysia, Bangi, Selangor, Malaysia. alyusefsattam@gmail.com, <https://orcid.org/0009-0009-0477-2140>

²Faculty of Information Science & Technology, University Kebangsaan Malaysia, Bangi, Selangor, Malaysia. zakree@ukm.edu.my, <https://orcid.org/0000-0003-2267-4965>

Received: September 27, 2025; Revised: November 18, 2025; Accepted: December 23, 2025; Published: March 31, 2026

Abstract

Classical information retrieval methods face increasing difficulty in handling large-scale, high-dimensional datasets due to the rapid growth of digital content. As feature dimensionality increases, traditional retrieval techniques suffer from high computational complexity, increased noise sensitivity, and reduced retrieval efficiency. This study introduces a new method based on the principles of quantum mechanics for unsupervised feature selection (UFS) known as Adaptive Optical Search for Unsupervised Feature Selection (AOSUFS). This is aimed at exploring high-dimensional data for information retrieval in the absence of labeled data. The new approach is based on a multi-layer search space and a criterion using the mean absolute difference to obtain the optimal feature subsets. AOSUFS is evaluated using the Reuters dataset comprising 12,152 bag-of-words features and is compared with several optimisation algorithms, including Genetic Algorithm, Harmony Search, Particle Swarm Optimisation, Simulated Annealing, and Krill Herd. The results of the experiments show that AOSUFS cuts the dimensionality by 51.4%, leaving only 5,904 features in the feature space. The proposed method achieves the highest mean average precision of 0.251. This is 9 percent higher than the baseline that does not use feature selection. The Mean Average Recall drops to 0.1384. This shows a 73 percent drop. Krill Herd got second place with a MAP of 0.2499. The unfiltered Harmony Search variant got the lowest score. This work presents the first application of adaptive optical search to unsupervised information retrieval, demonstrating improved retrieval effectiveness, reduced computational requirements, and efficient dimensionality reduction for large, sparse datasets.

Keywords: Unsupervised Feature Selection, Atomic Orbital Search, Mechanistic Optimization, Text Mining, Information Retrieval, Dimensionality Reduction.

1 Introduction

The proliferation of digital media and the improvement in the techniques of data collection have led to the creation of extremely high-dimensional text data sets. Traditional information retrieval techniques, designed for systems with few variables, will have difficulty when these systems are subjected to high-dimensional data (Patel et al., 2023). With an increasing number of variables, it is noticed that the computational cost increases, the noise in the system also accumulates, and so does the likelihood that

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 1 (March-2026), pp. 209-228 DOI: 10.58346/JOWUA.2026.11.013

*Corresponding author: Faculty of Information Science & Technology, University Kebangsaan Malaysia, Bangi, Selangor, Malaysia.

retrieval will be less accurate. Feature selection is thus of great importance in information retrieval systems, helping to decrease redundancy, increase the precision of results and enhance computational efficiency (Alyasiri et al., 2022). While supervised feature selection has been shown to produce excellent results, it does, however, require data to be labelled. This can be expensive and, in some situations, is unavailable. The UFS method offers a possible alternative to the other algorithms but many of the metaheuristic methods such as Harmony Search, Genetic Algorithm, Krill herd, Particle swarm optimisation and Ant colony optimisation all suffer from the problem of slow convergence and are very sensitive to the parameters used in the algorithm (Kang et al., 2023). In reality, many of these machine learning algorithms were created for general optimisation tasks, and they have not been adapted to suit the problems of information retrieval, due to the challenges of IR, including very sparse data, high levels of noise, and the lack of labels for the classes. In high-dimensional text retrieval contexts, UFS needs a more adaptable, robust, and domain-aware optimisation strategy. To address this gap, we present the Atomic Orbital Search for Unsupervised Feature Selection (AOUSFS), the first application of the quantum-mechanics-inspired Atomic Orbitals Search algorithm to Information Retrieval. In contrast to traditional evolutionary algorithms, a quantum-inspired evolutionary strategy models the process of search by quantum mechanical properties, including electron density distributions, energy levels, and transitions based on photons (Azizi, 2021). The ability to change movement through different phases of the search process allows for exploration and exploitation to be adjusted on the fly. This approach uses the Mean Absolute Difference (MAD) as a unique fitness function suited for information retrieval, allowing for the evaluation of the relevance of features even when there is no labelled data.

Rationale for Information Retrieval as an Ideal Domain for AOS

A crucial aspect of information retrieval systems is the selection of the most relevant items from a cluster of ideal texts. This is for three main reasons: (i) high-dimensional data - information retrieval systems are often dealing with tens of thousands of features or terms, as seen with the Reuters dataset, 20 Newsgroups, or large web crawls (Raza & Ding, 2022). In many cases, the features selected for a classification or regression task are not only unnecessary but also superfluous. This type of search strategy can efficiently explore these sparse data spaces without having to consider all possible combinations (Theng & Bhoyar, 2024), scenarios with a scarcity of labels - numerous information retrieval tasks, notably in unsupervised document retrieval or clustering, and the absence of labelled training data (Alyasiri et al., 2022; Iqbal et al., 2020). Unlike conventional feature selection methods, which depend on supervised signals, AOSUFS uses unsupervised signals to determine the value of the features. This is because the signals required for supervised assessment may either be too costly to obtain or not be available at all. Another issue in feature selection is the trade-off between retrieval performance and feature reduction (Iqbal et al., 2020). The adaptive exploitation-exploration trade-off in AOSUFS makes it possible to reach a high accuracy while reducing the number of attributes that are used a benefit that other search methods can't achieve (Tiwari & Chaturvedi, 2022).

Major Problem Gap Addressed by AOSUFS

Metaheuristic-based Uniform Fragmentation Search (UFS) has been studied in the context of Information Retrieval (Nassef et al., 2023). Current approaches in feature subset selection either ignore the specific difficulties encountered in information retrieval or apply optimisation techniques without any modification for text retrieval. This gap is addressed by AOSUFS through its discrete high-dimensional IR feature spaces optimisation of the search mechanism for AOS, incorporation of a relevant objective function to information retrieval which is label free (mean average difference), use of adaptive photon driven multi-layered search dynamics to manage quality and diversity of feature subsets,

and by providing superior performance in retrieval quality (MAP and MAR) and computational efficiency which is important in real time, large scale IR systems. In addition to employing a robust, quantum-inspired optimisation algorithm for the information retrieval domain, AOSUFS also adjusts its search mechanism and its evaluation criteria so as to satisfy the key demands of unsupervised feature selection within information retrieval from text. Various high-dimensional data reduction projects have employed unsupervised feature selection techniques. These include Genetic Algorithm (Abualigah et al., 2016), Bat Algorithm (Agarwal & Kumar, 2022), and Ant Colony Optimisation (Yilmaz Eroglu & Akcan, 2024). These algorithms exhibit improved performance when solving the USF problem. They too have to address significant challenges before implementation can be successful. Key challenges in evolutionary computation include the parameter sensitivity of the algorithm, the exploration/exploitation trade-off, and convergence speed. The efficiency and effectiveness of the optimisation process are significantly reduced by these limitations. Research indicates that atomic orbital search algorithms possess superior convergence speed as well as computational efficiency when compared to metaheuristic techniques. They are also better at balancing local and global searches (Abd Elaziz et al., 2022). Although AOS has many benefits, it has yet to be used extensively within the field of Information Retrieval. This research's principal contributions are as follows:

1. To propose the first adaptation of the unsupervised feature selection (UFS) using the AOS algorithm (AOSUFS), inspired by quantum mechanics in information retrieval (IR) systems.
2. To show that the AOSUFS method can effectively navigate the high-dimensional feature space that characterizes the textual data in IR systems.
3. To experimentally evaluate the new model using the benchmark IR datasets DS1 from REUTERS and unsupervised metrics such as MAP and MAR.

The subsequent sections of this paper are arranged as follows: Section 2 summarises the relevant literature. This includes the review of the related studies on AOS techniques, specifically unsupervised feature selection approaches based on AOS algorithms. Section 3 discusses the proposed approaches. Sections 4 and 5 cover the experimental settings and findings of this study, respectively. Lastly, section 6 offers the conclusion of this research.

2 Literature Review

The more recent advancements in the field of information retrieval have led to the documents, their filtering, categorization, and routing, all being looked at as very large ranking tasks. This ranking is based on how relevant the documents are in relation to a user's query (Sharma & Panda, 2023; Aguilar et al., 2020). In established paradigms of binary, multi-class, and multi-label classification, current information retrieval systems are encountering difficulties due to the nature of high-dimensional feature spaces, which are extremely sparse and where there is a lack of labelled data. This is especially true in retrieval operations without supervision. Typically, information retrieval systems based on text are comprised of four stages, including learning and ranking, feature selection, representation, and preprocessing (HS & Shenoy, 2020). The reduction of the dimensionality of the feature space is crucial in order to reduce the cost of the computational work, decrease the noise, and improve the effectiveness of the retrieval of the information. Recent work in text representation and feature extraction from the data not supervised is considered here, focusing on the search for metaheuristics with their drawbacks in large databases.

Representation Methods for a Textual Document

Computers can handle numerical data much more efficiently than text. Therefore, we need to transform unstructured text data into a numerical format before a machine learning algorithm can be applied to it. One of the reasons why the traditional n-gram model, which involves the bag-of-words with TF-IDF or term frequency weighting, is popular is that it is a very simple and efficient technique. These representations often yield sparse and high-dimensional feature spaces, where retrieval efficiency is decreased and noise amplified. To overcome the semantic limitations of single-word representations, several algorithms are now available to represent words by a set of real numbers (vectors) that encode word semantics and capture semantic relationships between words (Khomsah et al., 2022; Ravichandran et al., 2023). Current contextual embedding models that are more advanced do, however, have the result of generating large sets of features when they are combined at the document level. Both traditional and neural network-based text representations require appropriate feature selection processes to manage the dimensionality of the data while retaining relevant information, particularly in unsupervised information retrieval applications.

Feature Selection Methods for Document Classification

In dimensionality reduction, an essential step is featuring subset selection, where the objective is to select those attributes of a pattern that convey the most information while filtering out those that are either irrelevant or redundant (Theng & Bhoyar, 2024). Data mining techniques can be broadly classified into three categories, namely, embedded methods, filter methods, and wrappers (Biernacki, 2025). The wrapper approach is often more efficient computationally, but it ignores the interactions between the features of the data. It works by using a pre-existing learning algorithm and wrapping it within a feature selection method. The algorithms that are most commonly used include decision trees, support vector machines, and k-nearest neighbours. The embedded approach, on the other hand, includes Bayesian regularization of the neural networks, neural networks themselves, and regularised logistic regression. These methods work by learning the weights of the original learning algorithm and the features that are selected. However, they are computationally expensive. In high-dimensional spaces where the data can be text, conventional feature selection strategies often find it difficult to be robust and to scale. Because of their capability to strike a balance between exploration and exploitation and conduct a global search, metaheuristic optimisation methods have been increasingly used. Metaheuristics have been proven to be highly effective in the domain of feature selection, especially in the areas of text classification, the detection of network intrusions, bio-medical applications, and IoT systems (Kaur et al., 2023). They have some advantages, yet a particular class of data pre-processing methods based on metaheuristics is restricted in its application by a few drawbacks. They are highly susceptible to parameters being tuned correctly in high-dimensional data spaces and are also prone to the issue of converging on the optimal solution too early (Das et al., 2026). In addition, many of the assessment activities are performed in supervised circumstances, thus reducing the relevance of these assessments to retrieval tasks involving information without a human supervisor. However, commonly used evaluation measures fail to capture information retrieval-specific properties like sparsity, ranking sensitivity, and the precision-recall trade-off (Cisternas-Caneo et al., 2025). Current approaches to feature selection are not tailored to the needs of unsupervised learning in information retrieval.

Unsupervised Feature Selection and Emerging Directions

With large-scale text mining, information retrieval, and IoT applications, where labelled data are not available, there has been a growing interest in unsupervised feature selection. These approaches exploit the intrinsic data structure to find the informative features, which has led to the development of both

wrapper and embedded methods. A key reason for the growing popularity of unsupervised wrapper learning procedures based on metaheuristics is their ability to conduct a global search (Mohmmadzadeh & Gharehchopogh, 2021; Du et al., 2020; Baysal et al., 2021). Studies have shown that modified versions of Symbiotic Organism Search have performed better in terms of exploration and robustness for large datasets. Recent structure-aware, unsupervised methods that make use of feature correlation, manifold learning and spectral clustering to preserve the inherent data geometry are proposed in parallel. Generally, their large parameter lists and computational complexities restrict their use in large collections of documents. Several unsupervised feature selection techniques in high-dimensional environments, unfortunately, suffer from loss of diversity and premature convergence. At the moment, information retrieval systems often fall short of optimally finding the information we require because the algorithms we use are too complex. In order to overcome these difficulties researchers have started looking at simulation techniques that are based on the principles of physics.

This paper describes the atomic orbital search as a meta-heuristic that uses a physical model, representing optimisation as transitions from one energy state to another. The process is based upon quantum mechanics' atomic orbital interactions (Azizi, 2021). A method with a layered architecture facilitates search convergence while allowing for extensive exploration of the solution space. Atomic Orbital search has been used to solve global optimisation and selected feature discovery problems with considerable success, but its value in unsupervised feature discovery in the field of information retrieval remains largely unexplored (Azizi et al., 2022). The proposed atomic orbital search-based unsupervised feature selection algorithm is driven by the requirement to overcome the sparsity, scalability, and ranking effectiveness in very large data retrieval systems.

3 The Proposed AOSUFS Optimization Algorithm

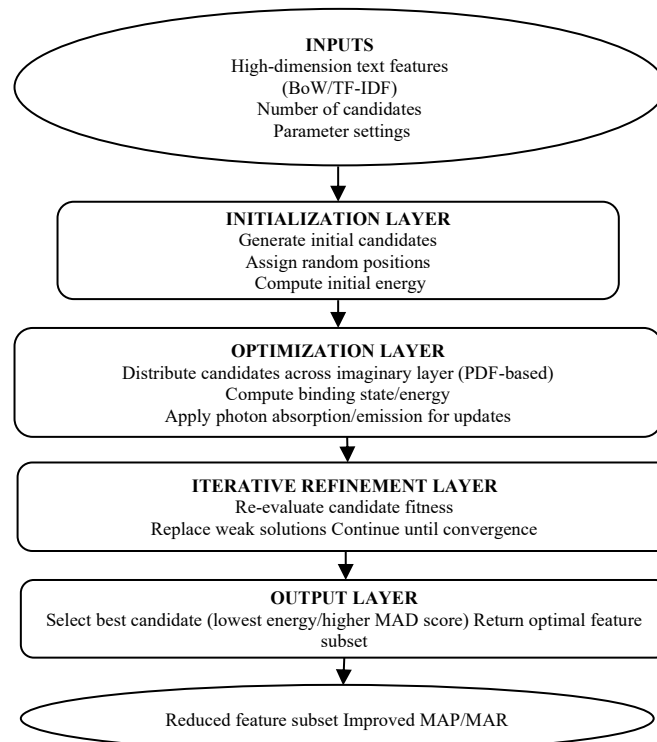


Figure 1: The proposed Atomic Orbital Search (AOS) algorithm-based unsupervised feature selection system

This section presents the proposed AOSUFS algorithm, which is derived from the AOS optimisation framework introduced in (Azizi, 2021). The algorithm models how atoms work in quantum physics. The algorithm treats each possible solution as an electron moving between energy levels. In this case, lower-energy states correspond to more important feature sets.

Most metaheuristic algorithms pick groups of solutions at random. AOSUFS uses organized solution candidates. AOSUFS moves forward step by step to improve solutions. This helps AOSUFS search spaces with high-dimensional feature spaces, which often occur in IR systems. As illustrated in figure 1, the proposed framework consists of four main stages: initialization, layered optimization, iterative refinement, and output. The objective is to identify an optimal subset of features that minimises an energy-based fitness function while preserving discriminative information.

Initialization Layer

The process starts by generating an initial population of candidate solutions (X). Each member of the population is a possible feature set configuration for the document retrieval task. The candidates are embedded within a d -dimensional space corresponding to the original set of features.

Solution Candidate Representation

To define a set of initial solution candidates, let n denote the number of solution candidates (atoms), and let d denote the number of features (dimensions in the search space). The population of solution candidates is represented as a matrix (Eq. 1).

$$\mathbf{X} = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^d \\ \vdots & \vdots & & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^d \end{bmatrix}, \begin{cases} i = 1, 2, \dots, n \\ j = 1, 2, \dots, d \end{cases} \quad (1)$$

Where n is the number of candidates and d is the number of features; x_i^j denotes the value of the j -th feature in the i -th candidate.

Each row of the matrix represents the features of a particular solution under consideration, and each column represents a particular dimension of the features of the system. The initial position of each candidate is generated using uniform random initialization (Eq. 2).

$$x_i^j(0) = x_{min}^j + r \times (x_{max}^j - x_{min}^j) \quad (2)$$

Where $r \in [0,1]$, $x_{min}^j = 0$, and $x_{max}^j = 1$.

Fitness Evaluation

Once the solution candidates are initialized, each candidate is evaluated using an energy-based fitness function. In the proposed AOSUFS framework, the fitness value represents the binding energy of an electron in the atomic analogy, where lower energy corresponds to a more informative and stable feature subset.

Let E_i denote the energy value associated with the i -th solution candidate. The fitness values of all candidates are aggregated into an energy vector, as defined in Eq. (3):

$$\mathbf{E} = [E_1, E_2, \dots, E_n]^T \quad (3)$$

Where n is the total number of solution candidates (population size), E_i is the fitness (energy) of the i -th candidate, and $(\cdot)^T$ denotes the transpose operator.

Each fitness value E_i is computed based on the quality of the feature subset encoded by the candidate solution \mathbf{X}_i . In this study, the energy (fitness) is defined using the Mean Absolute Deviation (MAD) criterion, as expressed in Eq. (4):

$$E_i = MAD(X_i) \quad (4)$$

Where $MAD(\mathbf{X}_i)$ measures the dispersion of the selected feature values in the candidate \mathbf{X}_i . In the AOSUFS framework, lower MAD values correspond to lower energy states, indicating more consistent and less noisy feature subsets for retrieval.

Binding State and Energy Initialization

In order to simulate the stability of atomic nuclei within the AOSUFS model, each candidate solution is associated with a binding state (BS) and a binding energy (BE). The binding of candidate solutions to the optimisation level is controlled by these quantities, dictating what information should be kept and what should be discarded during the optimisation process. A smaller atomic binding energy signifies a more stable configuration of electrons in the atom. Upon startup, candidate solutions are grouped into conceptual optimization levels. In one embodiment, for a given layer k , the state of binding is defined as the average location of all the solution candidates that have been assigned to that layer, while the binding energy is defined as the average fitness of those candidates.

Let the number of solution candidates at the k -th layer be denoted by p . The binding state of the layer k is computed as shown in Eq. (5):

$$BS^k = \frac{1}{p} \sum_{i=1}^p \mathbf{X}_i^k \quad (5)$$

Where \mathbf{X}_i^k represents the position vector (feature subset) of the i -th candidate in layer k , and p is the total number of candidates assigned to that layer. The binding state BS^k reflects the central tendency of candidate solutions within the layer.

Similarly, the binding energy of the layer k is defined as the average fitness (energy) of all candidates in that layer, as expressed in Eq. (6):

$$BE^k = \frac{1}{p} \sum_{i=1}^p E_i^k \quad (6)$$

Where E_i^k denotes the fitness (energy) value of the i -th candidate in layer k , computed using the fitness function defined in Eq. (4). The binding energy BE^k represents the average stability of solutions within the layer.

At each iteration, the values of BS^k and BE^k are used to guide transitions between layers, regulate exploration and exploitation, and determine whether candidate solutions should be retained, updated, or replaced. This mechanism enables AOSUFS to progressively converge toward an optimal feature subset while maintaining diversity across the search space.

Layered Optimization Core (AOS Engine)

The core of the proposed AOSUFS is a layered optimisation strategy which includes several fictitious layers, known as the LE^k layer, each LE^k representing a refinement phase within the optimization process. In the search for solutions, a layered approach is employed to mimic the atomic strata found in

quantum mechanics, thereby promoting exploration of potential answers while maintaining genetic variety. In each layer, the population is a subset of the total population selected using a probabilistic method. The distribution of the layers is controlled by a probability density function, which prevents early convergence by encouraging exploration at different refinement levels. Formally, the k -th optimization layer is defined as shown in Eq. (7):

$$LE^k = \{X_i \in X | P_k(X_i) \geq \tau_k\} \quad (7)$$

Where $\mathbf{X} \in \mathbb{R}^{n \times d}$ denotes the matrix of all candidate solutions, n is the population size, and d is the number of features. The set LE^k contains all candidates assigned to the k -th layer, $P_k(\mathbf{X}_i)$ represents the probability that the candidate \mathbf{X}_i belongs to the layer k , and τ_k is a predefined threshold that controls layer membership. The total number of layers is denoted by K .

Due to the high-dimensional and overlapping nature of feature spaces in information retrieval, spatial relationships between candidates may vary across layers. Consequently, upper layers tend to focus on exploitation within promising regions of the search space, while lower layers emphasize broader exploration.

Layer-wise Evaluation

Within each optimisation level, candidate solutions are assessed in the vicinity using their bound variable state BS^k and their bound energy BE^k . In order to regulate updates and maintain a population's diversity, a threshold value, PR which is probabilistic, is used. The candidates are updated at random with a probability of P . These rankings are adjusted based on how much the respective candidates have progressed. If the candidate has energy which is less than or equal to the global minimum binding energy, BE^* , a global update is then carried out. Otherwise, the layer-specific update is applied. Those candidates not fulfilling the revised selection criteria remain in their existing positions. The selective feature subset updating procedure assists the algorithm in striking an optimal balance between feature subset exploration and exploitation. This process allows the search process to converge towards an effective subset of features in the context of information retrieval from high dimensional spaces while retaining the feature subset diversity.

Output Layer

Following each stage of layer-by-layer processing, all the candidate solutions are updated so that the global binding pattern and binding energy are maintained. The algorithm can now learn as it iterates, adjusting itself to emerging trends in the features that are relevant. The algorithm's iterations continue until certain termination conditions are fulfilled. The termination criterion used here is that either the maximum number of generations is reached or the fitness of the whole population does not improve by ε or less since the last improvement. The candidate with the lowest remaining binding energy is the system's solution. This candidate is the optimal subset of features identified by the AOSUFS algorithm.

Algorithm 1 presents the complete workflow of the proposed Atomic Orbital Search-based Unsupervised Feature Selection (AOSUFS) algorithm. The algorithm aims to find the best subset of features by treating the feature selection process as a quantum-inspired energy minimization problem in a search space with many dimensions. The algorithm starts with the initialization stage, where it randomly creates a group of possible solutions in the d -dimensional feature space. Each candidate encodes a possible group of features. The fitness evaluation stage then uses the MAD-based fitness function to find the energy of each candidate. The function measures the uniformity of surface characteristics and the variance of the random component.

After the initial stage and fitness calculation, the algorithm proceeds to the layered optimization stage. In this stage, candidate solutions are assigned to multiple conceptual layers in a probabilistic way. Local and global updates are performed in each layer based on the binding state and binding energy. This results in a mechanism that dynamically changes the balance between exploration and exploitation and prevents the algorithm from converging prematurely. Finally, the algorithm proceeds to the output layer, where convergence is assessed, and the optimal solution is selected.

Algorithm 1: Pseudocode of the Proposed AOSUFS Algorithm

Input:

n : the number of possible solutions

d : the number of features

Max_{iter} : the most times you can repeat something

Output:

X^* : the best solution candidate with the lowest binding energy

Steps:

- 1: Use Eq. (2) to start solution candidates $\mathbf{X} \in \mathbb{R}^{n \times d}$
 - 2: Use Eq. (3) to find the fitness values E_i .
 - 3: Use Eqs. to find the initial binding state and binding energy. (4)– (5)
 - 4: Find the best candidate in the world with the least amount of energy.
 - 5: while iteration $< Max_{iter}$ do:
 - 6: Make up an imaginary layer LE^k
 - 7: Use the PDF (Eq. (6)) to put candidates into layers.
 - 8: For each layer, compute BS^k and BE^k
 - 9: Update candidate positions using the layer-wise update rules
 - 10: Update global binding state and binding energy
 - 11: Update the global best candidate
 - 12: end while
 - 13: Return X^*
-

Fitness Function Definition and Convergence Criteria

Fitness Function Based on Mean Absolute Deviation

In the proposed AOSUFS methodology, a solution is evaluated based on a MAD-based fitness function. The Mutual Aspects Development (MAD) method seeks to reduce the variability of the selected variables by quantifying the variability in the feature values (Hu et al., 2025). Low MAD values correspond to lower energy states in this analogy and indicate the presence of a good subset of features.

For a given candidate solution \mathbf{X}_i , the fitness function is defined using the MAD criterion, as shown in Eq. (8):

$$MAD(X_i) = \frac{1}{n_i} \sum_{j=1}^d |x_{ij} - \bar{X}_i| \quad (8)$$

Where n_i denotes the number of selected features in the i -th candidate solution, d is the total number of features in the original feature space, and x_{ij} represents the value of the j -th feature in candidate \mathbf{X}_i . The term \bar{x}_i represents the mean value of the selected features in \mathbf{X}_i , computed as defined in Eq. (9):

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^d x_{ij} \quad (9)$$

This formulation enables AOSUFS to favor feature subsets with low internal variability, thereby reducing redundancy and noise while preserving discriminative information. The use of MAD is particularly suitable for sparse TF-IDF representations, as it is less sensitive to extreme feature values and better captures the underlying distribution of relevant features in information retrieval tasks.

Parameter Settings and Convergence Criteria

The key parameters controlling the AOSUFS algorithm are population size, maximum number of iterations allowed, and a threshold of convergence. In order to achieve the balance between computational efficiency and exploration of the solution space, MaxIter is set to 100. In such evolutionary algorithms, the process stops when the fit of the population either stops improving or a maximum number of generations has been reached. The conditions for terminating the process guarantee a stable solution and minimize unnecessary computation.

4 Experimental Setup

This section describes the datasets, experimental configuration, baseline algorithms, evaluation metrics, and implementation details used to assess the performance of the proposed AOSUFS framework.

Dataset Description

This research utilizes the Reuters-21578 document collection, a standard resource in text mining and information retrieval studies. This corpus comprises approximately 10,788 English text files drawn from a diverse array of sources across the Internet and contains 12,152 unique word types following initial preprocessing. This collection, a news archive of Reuters, is highly sparse, with nearly all its vectors being zero. This makes it very useful in the evaluation of methods for selecting features automatically in the area of information retrieval, where the number of features of a document is large. Documents were all searched separately within the query-by-document methodology.

Feature Representation

All relevant texts were encoded in a numerical format suitable for the process of retrieval and optimisation by the BoW model weighted by the TF-IDF system. Because of its adaptability, this model was selected for the benchmark. It is widely used in many information retrieval studies, and it is also highly scalable. In the case of the resultant document-term matrix, it forms a high-dimensional, sparse feature space. This acted as input to the unsupervised feature selection techniques, including the proposed method and other algorithms of comparison. In order to guarantee that the comparison that was made was fair and unbiased, the feature vector construction method used was the same throughout.

Parameter Initialization and Experimental Configuration

To ensure reproducibility and fair comparison, all optimization algorithms were executed using explicitly defined and consistent parameter settings. For the proposed AOSUFS framework, the initial population of solution candidates was generated using a uniform random distribution over the interval $[0, 1]$, where each candidate represents a potential feature subset in the d -dimensional search space. The population size was fixed for all experiments, and the maximum number of iterations was set to 100 to balance computational efficiency and convergence stability.

The layered optimization structure of AOSUFS was initialized with a fixed number of optimization layers, and candidate solutions were probabilistically assigned to layers using the defined probability density function. The convergence criterion was met when either the maximum number of iterations was reached or when there was no significant improvement in fitness between iterations.

For all baseline algorithms, the parameter values were chosen based on settings that are often recommended in the literature to ensure that the convergence is stable and that the computational budgets are similar. To keep the experiment fair, all methods used the same stopping criteria and population sizes.

Baseline Algorithms and Comparative Methods

A comparative study has been conducted to evaluate the performance of the proposed Atomic Orbital Search-based Optimisation algorithm for Unsupervised Feature Selection, against other well-established global and local search heuristics used for unsupervised feature selection. In addition to GA and PSO, several other metaheuristics have been proposed for global optimisation, including Harmony Search in both its original and its upgraded form (Abualigah & Dulaimi, 2021; Wang et al., 2023; Ahmed et al., 2020), simulated annealing (Venkateswaran et al., 2022), the Krill Herd algorithm (Abualigah et al., 2024), the Salp Swarm algorithm (Abuain, 2024; Zivkovic et al., 2022), and its improvement. These algorithms are commonly used in Information Retrieval and Text Mining literature, which is one of the motivations for choosing them. The other motivation is that all of them have long-standing traditions in high-dimensional feature selection and optimisation problems. In addition to the optimisation techniques mentioned, experiments were conducted without feature selection to assess the effect of reducing dimensionality on retrieval performance. For comparison purposes, a baseline model has been developed, which includes all the features of the original data. This can be used to assess the extent to which the different feature selection methods perform. In order to guarantee the fairness of the results obtained in the experiment, the same feature representation, evaluation methods, stopping criteria, and data configuration were utilised for all algorithms.

Evaluation Metrics

The algorithms were evaluated using two well-known measures from the field of information retrieval. These are the mean average precision and the mean average recall. These metrics jointly measure ranking quality and retrieval performance using a single score that summarizes the performance of the system. Mean Average Precision measures how well relevant documents are ranked near the top of the retrieval list. It is defined as the mean of the average precision scores computed over all queries, as shown in Eq. (10):

$$MAP = \frac{1}{Q} \sum_{q=1}^Q \frac{1}{R_q} \sum_{k=1}^N P_q(k) \cdot rel_q(k) \quad (10)$$

Here the number of relevant documents for a query q is given by R_q , where q is the query number. Precision at rank k for query q is given by, $P_q(k)$ and $rel_q(k)$ is the binary indicator which is 1 if the document at the rank k is relevant to the query q , otherwise it is 0. The total number of queries is given by Q .

Mean Average Recall evaluates the system's ability to retrieve relevant documents while minimizing the inclusion of non-relevant items. This is determined using Eq. (11):

$$MAR = \frac{1}{Q} \sum_{q=1}^Q \frac{1}{R_q} \sum_{k=1}^N rel_q(k) \quad (11)$$

Lower MAR values mean that the system is more selective, which means that it finds fewer documents that aren't relevant. This is especially important in systems that need to find information with high accuracy. The Cosine Similarity and Euclidean Distance are used as ranking measures to calculate both MAP and MAR. This enables the comparison between retrieval performance across different similarity metrics.

Implementation Details

All experiments were conducted within a unified and reproducible environment using Python 3. Numerical computations and matrix operations were performed using the libraries NumPy and SciPy, data pre-processing and result management were handled with the library Pandas. To enable different methods to be evaluated fairly and consistently, scikit-learn was used for TF-IDF weighting, feature extraction using the Bag-of-Words model, and similarity calculations. The experiments were carried out on a computer running on an Intel Core i7 processor along with 16 gigabytes of RAM. Data for experimental results and comparative visualisation were plotted using matplotlib.

5 Experimental Results

This section analyzes the retrieval performance of the proposed AOSUFS framework in comparison with competing optimization-based feature selection methods.

Retrieval Effectiveness Under Cosine Similarity

The results from applying the cosine similarity technique are presented in table 1 and they indicate that using unsupervised feature selection can enhance retrieval effectiveness when compared to a method without feature extraction. In comparison to other methods, AOSUFS provides the best results, yielding a MAP of 0.2510, which signifies that the ranking of AOSUFS is superior, thanks to its ability to locate relevant documents at the beginning of a list. Other approaches, AOSUFS still outperformed, although marginally, other algorithms such as GA, ISSA and KH in terms of MAP score. In comparison, variants of Harmony Search (HS-1 and HS-2), Particle Swarm Optimisation and Simulated Annealing give lower MAP values, highlighting weaker ranking performance. In terms of selectivity, the AOSUFS achieved the lowest mean average recall with 0.1384, signifying its effectiveness in excluding non-relevant documents while maintaining relevant ones. In comparison with the evolutionary methods, K-HS and ISSA obtained similar results, whereas the PSO, Simulated Annealing, and Harmony Search algorithms all produced high MAR values, which implies that these methods had over-emphasised certain features. While the best balance between precision and recall is achieved by AOSUFS, it is still not as high as with AOSUFS.

Table 1: Result of nine optimization algorithms as unsupervised feature selection with cosine as similarity index ranking measure

Feature Selection Method	MAP	MAR
Without Feature Selection	0.2305	0.5125
KH	0.2499	0.1394
AOSUFS	0.2510	0.1384
HS-2	0.1445	0.2977
HS-1	0.1542	0.4001
PSO	0.1563	0.4737
SA	0.1643	0.4820
GA	0.2486	0.1441
SSA	0.2428	0.1400
ISSA	0.2500	0.1400

Feature Reduction Behavior and its Impact

The graph in figure 2 shows how the feature selection methods reduce the number of attributes. The baseline model keeps all of the attributes, whereas attribute selection and feature selection techniques decrease the number of attributes. Although HS-1 and HS-2 give the largest reduction in the number of features, their poor performance in classifying instances in the test set indicates that pruning removes informative attributes along with noise. In comparison AOSUFS achieved a modest but useful reduction in dimensionality, keeping roughly 5904 features left, akin to the results for ISSA and SSA. The proposed indexing method of AOSUFS efficiently reduces storage requirements without significantly degrading retrieval performance by eliminating redundant information. Feature selection can improve the efficiency of neural networks by reducing the input dimensionality while maintaining performance.

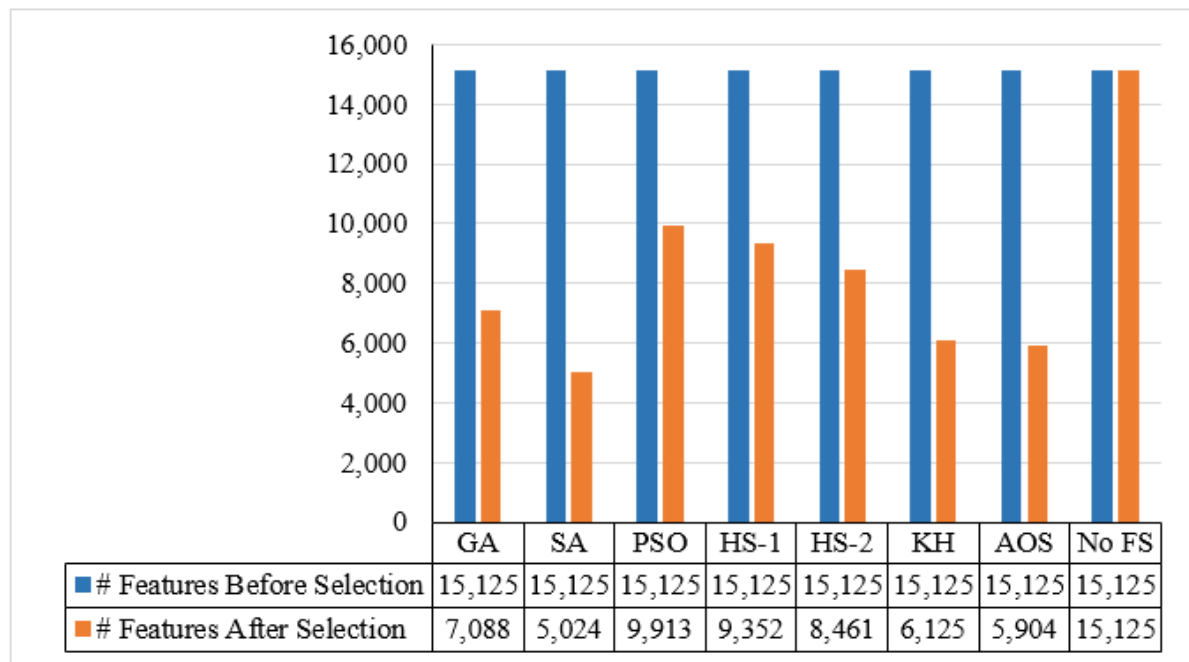


Figure 2: Comparison between 9 algorithms and the AOSUFS proposed unsupervised feature selection based on the number of features

Performance Under Euclidean Distance

The retrieval performance of all feature selection methods under the Euclidean Distance ranking measure is reported in table 2. Following the cosine similarity analysis, the suggested AOSUFS framework has the highest mean average precision (MAP = 0.2533) thus highlighting its strong ranking sensitivity compared to other similarity measures. In regards to selectivity, AOSUFS has one of the lowest mean average recalls (MAR = 0.1380), with a close resemblance to Krill Herd (KH), which achieves the lowest MAR; a fact that demonstrates the ability of AOSUFS to reduce the retrieval of irrelevant documents and maintain those that are relevant, despite the use of a distance based ranking criterion. The rival strategies of HS-1, HS-2, PSO, and SA show significantly different MAP indicators and higher MAR values, which speaks of a poor quality of ranking and the presence of unimportant elements. Figures 3 and 4 also outline the relative performances of MAP and MAR with Euclidean distance showing that AOSUFS overpowers performance and perturbations in other approaches are not significant. Thereby, the findings support the idea that the suggested AOS model ensures retrieval efficiency and selectivity to different measures of similarity or distance.

Table 2: Optimization algorithms and AOSUFS proposed as unsupervised feature selection with Euclidean distance index ranking measure

Feature Selection Method	MAP	MAR
Without Feature Selection	0.2340	0.5010
KH	0.2499	0.1388
AOS	0.2533	0.1380
HS-2	0.1450	0.2910
HS-1	0.1509	0.4102
PSO	0.1571	0.4710
SA	0.1655	0.4804
GA	0.2499	0.1499
SSA	0.2416	0.1403
ISSA	0.2500	0.1392

Comparative Analysis Across Similarity Measures

Figures 3 and 4 present a comparative study of the Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) of MAR and MAP under both Euclidean Distance and Cosine Similarity. When evaluating all these algorithms, Euclidean distance was seen to result in lower mean average recall values, which in turn gives a higher degree of selectivity. The AOSUFS approach shows considerable robustness as it performs consistently well in terms of MAP and also keeps the MRR very low, regardless of the measure used. Mostly, the Cosine metric results in slightly higher mean average precision for the methods, though this is achieved at the expense of an increased mean average recall. Compared to Euclidean Distance, a number of algorithms such as AOSUFS and KH are more geared towards finding a balance between the accurate ranking of relevant documents and the selectivity for top-ranking methods.

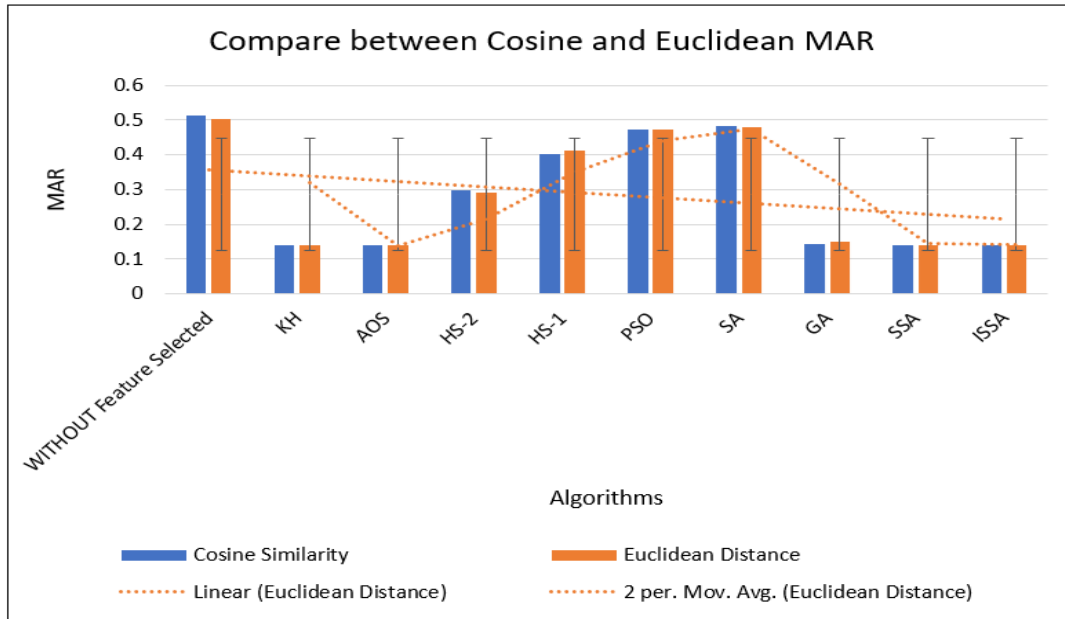


Figure 3: Comparison of MAR using cosine similarity and Euclidean distance for all unsupervised feature selection methods, including the proposed AOSUFS

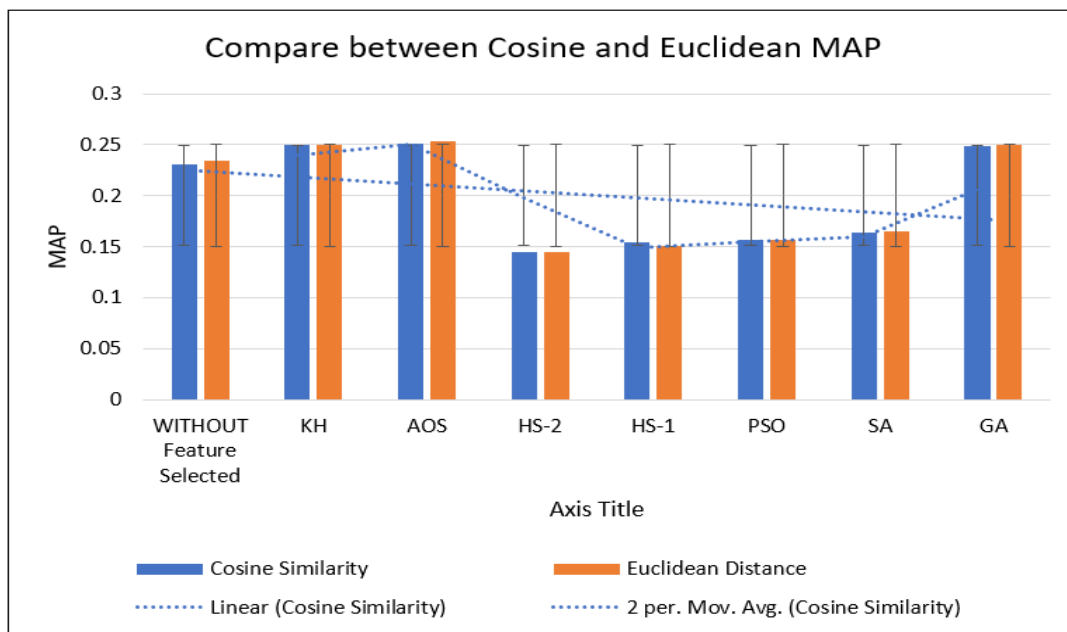


Figure 4: Comparison of MAP using cosine similarity and Euclidean distance for all unsupervised feature selection methods, including the proposed AOSUFS

Ablation Study of AOSUFS

In order to gain a deeper understanding of the AOSUFS system's various components, experiments were performed to remove key elements of the system and observe any resulting performance changes. The evaluation was carried out by employing both Euclidean Distance and Cosine Similarity measures and the resulting information is presented in table 3 (Aguilar et al., 2020). In particular, four ablated variants were examined: the model without layered optimization (AOSUFS), the model without the assignment

of layers based on the probability density function (AOSUFS), the model without energy-driven update rules (AOSUFS), and AOSUFS which uses an alternative fitness function instead of mean absolute difference. Each component of AOSUFS as shown in table 3 had a significant impact on the retrieval performance. Optimisation is greatly diminished through the removal of this layered retrieval method leading to a noticeable drop in the mean average precision and an associated increase in the mean average recall using either measure of similarity. Our layered search approach, which includes a top-level search of clusters of prototypes, followed by a search within the clusters, has been shown to efficiently reduce the complexity of the problem. This results in the avoidance of premature convergence and the maintenance of diversity within the population. When we turn off the use of the PDFs to select which layer to use, performance drops even further; this indicates that using the probability distributions helps the network to balance exploration and exploitation effectively. This is because candidate solutions prematurely converge to a suboptimal solution within the search space before a thorough exploration has taken place. The removal of the energy driven update equation has the most significant impact on performance. In comparison to the full model, this model performs the worst in terms of Mean Average Precision and the best in terms of Mean Average Recall, as indicated in table 3. This observation supports the atomic-orbital analogy of AOSUFS theory, where the transitions between the different energy levels drive the convergence of the system towards the feature set with the most information. Using an alternative to the mean absolute deviation, namely the variance, to compute the fitness function results in a lower quality. MAD appears to perform more effectively in sparse TF-IDF vector space, particularly because it is less influenced by the extreme vector elements and better captures the distribution of vectors pertinent to the information retrieval.

Table 3: Ablation study results for AOSUFS under cosine similarity and Euclidean distance

AOSUFS Variant	MAP (Cosine)	MAR (Cosine)	MAP (Euclidean)	MAR (Euclidean)
Without Layered Optimization	0.2458	0.1462	0.2471	0.1455
Without PDF-Based Assignment	0.2439	0.1497	0.2453	0.1479
Without Energy-Guided Update	0.2364	0.1688	0.2381	0.1665
Fitness without MAD (Variance)	0.2417	0.1524	0.2430	0.1508
Full AOSUFS	0.2510	0.1384	0.2533	0.1380

6 Discussion

The experimental results reported in Section 5 demonstrate that the proposed AOSUFS framework consistently outperforms competing optimization-based methods across different similarity measures. From table 1 and table 2, AOSUFS ranks highest in MAP yet keeps MAR among the lowest, showing strong alignment between precision and restraint. When using Cosine Similarity (Table 1), AOSUFS stands out with a top MAP value of 0.2510 alongside the lowest MAR at 0.1384, showing strong early document relevance plus minimal noise. Its pattern holds true even in Euclidean Distance (Table 2), where results mirror the first case - MAP at 0.2533 paired with MAR at 0.1380. Because these outcomes match so closely regardless of similarity method, it becomes clear the approach works well under variation. On the flip side, approaches like PSO, SA, and some Harmony Search versions show much higher MAR numbers along with lower MAP scores - these point to weak filtering, meaning irrelevant traits keep getting pulled in. The original setup without feature trimming appears in both lists with the peak MAR value, underlining how crowded, noisy high-dimensional data drags down search results.

The influence of feature reduction on retrieval effectiveness is further illustrated in figure 2. Algorithms such as HS-1 and HS-2, though they reduce the dimensionality significantly, have low values

of MAP, which means that they retain informative features together with noise. AOSUFS is more balanced in its reduction approach, and it will maintain around 5,904 features, which will not discard discriminatory information and reduce redundancy. This is the reason as to why AOSUFS attains better values of MAP and MAR than those that either over-prune (e.g., SA) or retains too many features (e.g., PSO). These findings are further proved by the comparative performance of MAR and MAP with Cosine Similarity and Euclidean Distance as indicated in Figures 3 and 4 respectively. Despite the fact that Euclidean Distance tends to provide lower value of MAR over algorithms, AOSUFS is the only algorithm that has high MAP and low MAR under both measures. It means that the AOSUFS search mechanism latticing and energy-guided more attractively depends on the similarity metric used and is more resilient to changes in ranking criteria. On the whole, the findings validate that successful unsupervised feature selection in information retrieval, is not about dimensionality reduction maximization but classification of task relevant features. The suggested AOSUFS framework has been able to balance the noise reduction and the information preservation effectively and thus creating uniform changes in retrieval preciseness and selectivity in various evaluation circumstances.

7 Conclusion

This study examined unsupervised feature selection for high-dimensional information retrieval using a quantum mechanics-inspired optimisation strategy. The traditional unsupervised approaches are often faced with a trade-off between dimensionality loss and recall performance in sparse lexical domains leading to unstable ranking results and less discriminatory ability. As a reaction, the Atomic Orbital Search for Unsupervised Feature Selection (AOSUFS) was introduced as the first adaptation of the Atomic Orbital Search algorithm to feature selection on information search without the use of supervision. It combines an energy fitness criterion that is based on a median absolute deviation and a multi-layered optimization structure. Consistent and statistically measurable improvements in retrieval performance are empirically validated by performing empirical validation on the Reuters-21578 benchmark corpus. AOSUFS reduced 12,152 candidate attributes to 5,904, which constitute 51.4% reduction, and retained strong discriminative power. When tested with cosine similarity, the algorithm gave a mean average precision of 0.2510, which is 9% higher than that of the no feature selection, and the mean average recall changed by 0.5125 to 0.1384 indicating 73% decrease in the recall of irrelevant documents. Similar behaviour was observed under Euclidean Distance, where AOSUFS attained the highest MAP of 0.2533 and one of the lowest MAR values at 0.1380, demonstrating stability across different ranking measures. The ablation study provides the quantitative evidence of the contribution of each algorithmic element. The elimination of the layered optimisation mechanism resulted in a significant reduction in mean average precision which was accompanied with a rise in mean average recall; whereas even the removal of the probability density-based layer assignment further deteriorated the retrieval performance. The most significant drop in the performance was realized in the absence of energy guided update rules thus highlighting the importance of energy-based transitions in leading convergence. The use of the variance instead of maximum absolute deviation fitness function produced lower retrieval effectiveness which supported the appropriateness of MAD when using sparse TF-IDF feature space. The suggested AOSUFS framework offers an efficient unsupervised method to feature selection on large-scale information retrieval system to enhance ranking accuracy, selectiveness, and a decrease in cost of computation with a regulated dimensionality reduction (Sharma & Panda, 2023). Future directions will focus on making the analysis be extended to other benchmark databases, add semantic and contextual data and explore adaptive parameter measures to strengthen the robustness and generalisation of the evaluation in different retrieval context.

Acknowledgement

The authors would like to appreciate all colleagues and research peers who provided valuable feedback during the planning and refinement stages of the study. We also acknowledge the management of Universiti Kebangsaan Malaysia (UKM), Bangi, 43600 Selangor, Malaysia that provided support and an enabling environment for this research.

References

- [1] Abd Elaziz, M., Ouadfel, S., Abd El-Latif, A. A., & Ali Ibrahim, R. (2022). Feature selection based on modified bio-inspired atomic orbital search using arithmetic optimization and opposite-based learning. *Cognitive Computation*, 14(6), 2274-2295. <https://doi.org/10.1007/s12559-022-10022-6>
- [2] Abuain, W. A. (2024). Improved Salp Swarm Algorithm for Text Document Clustering. *Journal of Theoretical and Applied Information Technology*, 102(14), 5396-5407
- [3] 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. <https://doi.org/10.1109/CSIT.2016.7549453>
- [4] Abualigah, L., & Dulaimi, A. J. (2021). A novel feature selection method for data mining tasks using hybrid sine cosine algorithm and genetic algorithm. *Cluster Computing*, 24(3), 2161-2176. <https://doi.org/10.1007/s10586-021-03254-y>
- [5] Abualigah, L., Al-Zyod, M., Ikotun, A. M., Shehab, M., Otair, M., Ezugwu, A. E., ... & El-kenawy, E. S. M. (2024). A review of krill herd algorithm: optimization and its applications. *Metaheuristic Optimization Algorithms*, 231-239. <https://doi.org/10.1016/B978-0-443-13925-3.00017-0>
- [6] Agarwal, T., & Kumar, V. (2022). A systematic review on bat algorithm: Theoretical foundation, variants, and applications. *Archives of Computational Methods in Engineering*, 29(5), 2707-2736. <https://doi.org/10.1007/s11831-021-09673-9>
- [7] Aguilar, J., Salazar, C., Velasco, H., Monsalve-Pulido, J., & Montoya, E. (2020). Comparison and evaluation of different methods for the feature extraction from educational contents. *Computation*, 8(2), 30. <https://doi.org/10.3390/computation8020030>
- [8] Ahmed, S., Ghosh, K. K., Singh, P. K., Geem, Z. W., & Sarkar, R. (2020). Hybrid of Harmony Search Algorithm and Ring Theory-Based Evolutionary Algorithm for Feature Selection. *IEEE Access*, 8, 102629-102645. <https://doi.org/10.1109/ACCESS.2020.2999093>
- [9] Alyasiri, O. M., Cheah, Y. N., Abasi, A. K., & Al-Janabi, O. M. (2022). Wrapper and hybrid feature selection methods using metaheuristic algorithms for English text classification: A systematic review. *IEEE Access*, 10, 39833-39852. <https://doi.org/10.1109/ACCESS.2022.3165814>
- [10] Azizi, M. (2021). Atomic orbital search: A novel metaheuristic algorithm. *Applied Mathematical Modelling*, 93, 657-683. <https://doi.org/10.1016/j.apm.2020.12.021>
- [11] Azizi, M., Talatahari, S., Khodadadi, N., & Sareh, P. (2022). Multiobjective atomic orbital search (MOAOS) for global and engineering design optimization. *IEEE Access*, 10, 67727-67746. <https://doi.org/10.1109/ACCESS.2022.3186696>
- [12] Baysal, Y. A., Ketenci, S., Altas, I. H., & Kayikcioglu, T. (2021). Multi-objective symbiotic organism search algorithm for optimal feature selection in brain computer interfaces. *Expert Systems with Applications*, 165, 113907. <https://doi.org/10.1016/j.eswa.2020.113907>
- [13] Biernacki, A. (2025). Evaluating Filter, Wrapper, and Embedded Feature Selection Approaches for Encrypted Video Traffic Classification. *Electronics*, 14(18), 3587. <https://doi.org/10.3390/electronics14183587>

- [14] Cisternas-Caneo, F., Barrera-Garcia, J., Crawford, B., Soto, R., Sánchez, M. G., Gomez-Pulido, J. M., & Garces-Jimenez, A. (2025). How Is the Objective Function of the Feature Selection Problem Formulated? In *International Conference on Optimization and Learning* (pp. 3-13). Springer, Cham. https://doi.org/10.1007/978-3-031-77941-1_1
- [15] Das, A., Mollick, S., Mondal, H., Bala, T. J., Nag, A., & Guho, A. (2026). Meta-heuristic Algorithms for High-Dimensional Feature Selection: A Systematic Review of Methodologies, Applications, and Emerging Challenges with Future Research Directions. *Feature Fusion for Next-Generation AI*, 51-61. https://doi.org/10.1007/978-3-031-94386-7_5
- [16] Du, Z. G., Pan, J. S., Chu, S. C., & Chiu, Y. J. (2020). Improved binary symbiotic organism search algorithm with transfer functions for feature selection. *IEEE Access*, 8, 225730-225744. <https://doi.org/10.1109/ACCESS.2020.3045043>
- [17] HS, C., & Shenoy, M. K. (2020). Advanced text documents information retrieval system for search services. *Cogent Engineering*, 7(1), 1856467. <https://doi.org/10.1080/23311916.2020.1856467>
- [18] Hu, C., Ren, D., Li, H., Zhang, S., & Pan, G. (2025). Improving the robustness of circular curve fitting with equality constraints using the median absolute deviation method. *Survey Review*, 57(404), 498-508. <https://doi.org/10.1080/00396265.2025.2480919>
- [19] Iqbal, M., Abid, M. M., Khalid, M. N., & Manzoor, A. (2020). Review of feature selection methods for text classification. *International Journal of Advanced Computer Research*, 10(49), 138-152. <http://dx.doi.org/10.19101/IJACR.2020.1048037>
- [20] Kang, Y., Peng, L., Guo, J., Lu, Y., Yang, Y., Fan, B., & Pu, B. (2023). A fast hybrid feature selection method based on dynamic clustering and improved particle swarm optimization for high-dimensional health care data. *IEEE Transactions on Consumer Electronics*, 70(1), 2447-2459. <https://doi.org/10.1109/TCE.2023.3334373>
- [21] Kaur, S., Kumar, Y., Koul, A., & Kumar Kamboj, S. (2023). A Systematic Review on Metaheuristic Optimization Techniques for Feature Selections in Disease Diagnosis: Open Issues and Challenges *Archives of Computational Methods in Engineering*, 30(3), 1863-1895. <https://doi.org/10.1007/s11831-022-09853-1>
- [22] Khomsah, S., Ramadhani, R. D., & Wijaya, S. (2022). The accuracy comparison between Word2Vec and FastText on sentiment analysis of hotel reviews. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 6(3), 352-358.
- [23] Mohmmadzadeh, H., & Gharehchopogh, F. S. (2021). An efficient binary chaotic symbiotic organisms search algorithm approaches for feature selection problems. *The Journal of Supercomputing*, 77(8), 9102-9144. <https://doi.org/10.1007/s11227-021-03626-6>
- [24] Nassef, A. M., Abdelkareem, M. A., Maghrabie, H. M., & Baroutaji, A. (2023). Metaheuristic-based algorithms for optimizing fractional-order controllers a recent, systematic, and comprehensive review. *Fractal and Fractional*, 7(7), 553. <https://doi.org/10.3390/fractalfract7070553>
- [25] Patel, V., Hiran, D., & Dangarwala, K. (2023, April). Recent trends of information retrieval system: Review based on ir models and applications. In *International Conference on Recent Trends in Machine Learning, IOT, Smart Cities & Applications*, 873, (pp. 619-629). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-9442-7_51
- [26] Ravichandran, V., Sadhu, S., Convey, D., Guerrier, S., Chomal, S., Dupre, A. M., ... & Mankodiya, K. (2023). iTex gloves: design and in-home evaluation of an e-textile glove system for tele-assessment of Parkinson's disease. *Sensors*, 23(6), 2877. <https://doi.org/10.3390/s23062877>
- [27] Raza, S., & Ding, C. (2022). News recommender system: a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review*, 55(1), 749-800. <https://doi.org/10.1007/s10462-021-10043-x>

- [28] Sharma, S., & Panda, S. P. (2023). Efficient information retrieval model: overcoming challenges in search engines-an overview. *Indonesian Journal of Electrical Engineering and Computer Science*, 32(2), 925-932. <https://doi.org/10.11591/ijeecs.v32.i2.pp925-932>
- [29] Theng, D., & Bhoyar, K. K. (2024). Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems*, 66(3), 1575-1637. <https://doi.org/10.1007/s10115-023-02010-5>
- [30] Tiwari, A., & Chaturvedi, A. (2022). A hybrid feature selection approach based on information theory and dynamic butterfly optimization algorithm for data classification. *Expert Systems with Applications*, 196, 116621. <https://doi.org/10.1016/j.eswa.2022.116621>
- [31] Venkateswaran, C., Ramachandran, M., Ramu, K., Prasanth, V., & Mathivanan, G. (2022). Application of simulated annealing in various field. *Materials and its Characterization*, 1(1), 01-08. <https://doi.org/10.46632/mc/1/1/1>
- [32] Wang, J., Wang, X., Li, X., & Yi, J. (2023). A hybrid particle swarm optimization algorithm with dynamic adjustment of inertia weight based on a new feature selection method to optimize SVM parameters. *Entropy*, 25(3), 531. <https://doi.org/10.3390/e25030531>
- [33] Yilmaz Eroglu, D., & Akcan, U. (2024). An adapted ant colony optimization for feature selection. *Applied Artificial Intelligence*, 38(1), 2335098. <https://doi.org/10.1080/08839514.2024.2335098>
- [34] Zivkovic, M., Stoean, C., Chhabra, A., Budimirovic, N., Petrovic, A., & Bacanin, N. (2022). Novel improved salp swarm algorithm: An application for feature selection. *Sensors*, 22(5), 1711. <https://doi.org/10.3390/s22051711>

Authors Biography



Sattam Abdallah Alyusuf is a Lecturer in Information Science and Technology at Dar Al Uloom University, Riyadh, Saudi Arabia. His academic work at the university involves teaching students and conducting research about information systems, data and information management, and new information technology developments. He dedicates his time to teaching and learning activities about information science and technology while conducting research that focuses on modern data and information systems and digital technology challenges.



Mohd Zakree Ahmad Nazri holds MSc and PhD degrees from University Teknologi Malaysia in Intelligent Decision Support and Machine Learning. His research involves the development of nature- and biologically inspired clustering algorithms, including Auto-Immune Systems and Artificial Neural Networks. He teaches Business Intelligence and Analytics and Decision Support Systems, and his current research focuses on unstructured data processing, particularly for Malay and mixed-language (rojak) text.