

Advanced Metaheuristic Algorithms for Text Document Clustering: A Comparative Study of CNGO, MOA, and MPSO with K-means

Ratnam Dodda*

CVR College of Engineering/Department of CSE(AI&ML), Hyderabad, 501510, India

E-mail: ratnam.dodda@gmail.com

ORCID iD: <https://orcid.org/my-orcid?orcid=0000-0002-2414-9139>

*Corresponding author

A. Sureshbabu

Jawaharlal Nehru Technological University Anantapur /Department CSE, Ananthapuramu, 515002, India

E-mail: asureshjntu@gmail.com

ORCID iD: <https://orcid.org/0000-0002-1312-9765>

Received: February 16, 2026; Revised: April 10, 2026; Accepted: June 11, 2026; Published: August 8, 2026

Abstract: Text document clustering plays a pivotal role in organizing large-scale unstructured data, yet conventional clustering algorithms, such as k-means, often struggle with high-dimensional data, suboptimal initializations, and local minima issues. This paper introduces a novel comparative analysis of three advanced optimization techniques integrated with k-means: Chaotic Northern Goshawk Optimization (CNGO), Mayfly Optimization Algorithm (MOA), and Modified Particle Swarm Optimization (MPSO). This work is unique because it integrates these metaheuristic algorithms to improve clustering performance, targeting initialization challenges and increasing accuracy. Extensive experiments were conducted on benchmark datasets, including Reuters-21578, 20-Newsgroup, and BBC-Sport. All three models outperform traditional k-means in terms of accuracy, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and V-measure. This research offers new insights into optimizing clustering processes using metaheuristic algorithms and provides a foundation for future exploration in large-scale document clustering.

Our study is significant because it systematically overcomes key limitations of conventional k-means for high-dimensional text data, such as poor centroid initialization, local minima, and reduced effectiveness on sparse corpora by integrating and comparatively evaluating three advanced metaheuristics (CNGO, MOA, MPSO) with k-means on standard benchmark datasets. The value of this work lies in the consistently improved clustering quality (Accuracy, ARI, NMI, V-measure) achieved by the proposed hybrids, and in showing that MOA-k-means in particular offers a robust, scalable solution for real-world text analytics applications such as information retrieval, recommendation, and topic discovery.

Index Terms: Text Document Clustering, Chaotic Northern Goshawk Optimization, Mayfly Optimization Algorithm, Modified Particle Swarm Optimization, K-means Integration, Metaheuristic Algorithm.

1. Introduction

In the current digital era, the vast amounts of unstructured text data generated by emails, blogs, social media posts, and online reviews require efficient methods for organization, analysis, and information retrieval. Clustering text documents, an unsupervised machine learning task, is a key part of this process as it groups similar documents based on shared features [1]. Applications encompass information retrieval, sentiment analysis, market research, and search engine optimization [2]. Nonetheless, conventional clustering techniques, such as k-means, frequently encounter considerable difficulties when processing high-dimensional text data, including sensitivity to initial centroids [3-4], vulnerability to local minima [5], and computational inefficiency [6].

Consequently, metaheuristic optimization algorithms have gained prominence in recent years, enhancing the performance of clustering methods such as k-means [5]. Metaheuristics are known for being able to search through large search spaces and improve complicated objective functions [7]. To address the intrinsic limitations of traditional methods, k-means has been integrated with metaheuristic algorithms [4,8]. Despite these improvements, further

investigation is required to examine how different metaheuristic algorithms can enhance the accuracy, scalability, and computational efficiency of text clustering [9].

This paper presents an innovative comparative analysis of three advanced optimization techniques: Chaotic Northern Goshawk Optimization (CNGO), Mayfly Optimization Algorithm (MOA), and Modified Particle Swarm Optimization (MPSO). To enhance clustering performance, all algorithms are integrated with k-means [9-11]. To prevent early convergence, CNGO employs chaotic maps to improve global search, MOA utilizes swarming behavior to find a balance between exploration and exploitation, and MPSO modifies traditional particle swarm optimization to speed up convergence and improve clustering accuracy.

This research employs three unique metaheuristic algorithms to enhance k-means clustering in an innovative manner. Although prior studies have examined the amalgamation of metaheuristics with k-means [10], the comparative evaluation of these specific algorithms, particularly their efficacy in clustering extensive text datasets, remains underexplored [12-13]. This study addresses this deficiency by conducting comprehensive experiments on three benchmark text datasets: Reuters-21578, 20-Newsgroup, and BBC-Sport, comparing the proposed models to traditional k-means in terms of clustering accuracy, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and homogeneity.

K-means remains a widely used clustering algorithm due to its simplicity and effectiveness; however, it suffers from three major limitations when applied to real-world text corpora: sensitivity to initial centroid selection, susceptibility to local minima, and reduced effectiveness in sparse, high-dimensional document spaces. These limitations have a direct impact on practical uses like organizing news articles, grouping search results, finding new topics, and mining customer feedback, where stable and semantically meaningful clusters are very important. This study specifically examines the enhancement of robustness and clustering quality of k-means for extensive text document datasets through the incorporation of advanced metaheuristic optimization techniques. In this context, the innovation of the proposed study resides in the methodical comparative assessment of CNGO, MOA, and MPSO for enhancing centroid initialization and search behavior within a cohesive k-means-based clustering framework.

1.1. Problem Statement and Related Works

Specifically, this study addresses the problem of clustering large-scale, sparse and high-dimensional text document collections, where the widely used k-means algorithm suffers from sensitivity to initial centroids, vulnerability to local minima, and reduced effectiveness on real-world text corpora such as news, forums and reviews. Existing solutions include the standard k-means algorithm and various k-means hybrids with metaheuristic optimization methods such as PSO, GA, ACO, GWO and related approaches, which currently offer the best clustering performance compared with plain k-means but still face limitations in robust centroid initialization, exploration exploitation balance and scalability for extensive text datasets. We then clearly formulate the major research objectives of this work as: (i) integrating three advanced metaheuristics Chaotic Northern Goshawk Optimization (CNGO), Mayfly Optimization Algorithm (MOA) and Modified Particle Swarm Optimization (MPSO) with k-means to address these limitations; (ii) conducting a systematic comparative study of these hybrids against standard k-means on benchmark text corpora (Reuters-21578, 20-Newsgroup, BBC-Sport); and (iii) identifying which hybrid provides the most accurate and stable clustering for high-dimensional text data. Finally, we emphasize that our goal is to offer a more robust and scalable clustering framework for real-world text analytics tasks and to fill the current gap in the literature regarding the comparative efficacy of CNGO-, MOA- and MPSO-based k-means models for text document clustering.

1.2. Novelty of the Work

The originality of this study resides in the strategic amalgamation of three sophisticated metaheuristic algorithms—CNGO, MOA, and MPSO—with the k-means clustering method, each specifically designed to tackle distinct challenges in high-dimensional text clustering:

- Chaotic Northern Goshawk Optimization (CNGO): CNGO uses chaotic maps to improve the global search ability and prevent premature convergence, distinguishing it from traditional optimization methods. This significantly improves the initialization phase of k-means, leading to enhanced clustering results.
- Mayfly Optimization Algorithm (MOA): MOA's swarming behavior strikes a balance between exploration and exploitation, balancing local and global search. Previous swarm intelligence studies often struggled to achieve this balance, which is crucial for enhancing k-means clustering.
- Modified Particle Swarm Optimization (MPSO): MPSO builds on the traditional PSO algorithm by adding features like faster convergence and better diversity maintenance. These modifications reduce the likelihood of converging to suboptimal solutions and improve k-means initialization.

This research enhances the field by providing comprehensive performance comparisons of three meta-heuristic techniques, delineating the contexts in which each method demonstrates superiority, and offering insights into their scalability for extensive text clustering. The results show that combining metaheuristic algorithms with k-means significantly improves clustering quality, especially for high-dimensional text datasets. This opens up new avenues for research into optimizing document clustering using advanced optimization techniques. Previous studies usually only

looked at one metaheuristic-k-means combination or looked at how well clustering worked on general-purpose datasets. This study, on the other hand, examines three advanced optimization strategies—CNGO, MOA, and MPSO—specifically for text document clustering. The proposed framework is novel because it shows how each optimizer fixes a different problem with standard k-means: CNGO improves global search by using chaotic diversification, MOA balances exploration and exploitation during centroid search, and MPSO speeds up convergence and improves solution refinement. Thus, the contribution extends beyond the mere application of metaheuristics to include a systematic comparison that shows how well they work for sparse, high-dimensional, and real-world text corpora. This study introduces three distinct models for text document clustering by integrating k-means with advanced metaheuristic optimization techniques: Chaotic Northern Goshawk Optimization, Mayfly Optimization Algorithm, and Modified Particle Swarm Optimization (MPSO). The methodology consists of four main phases: data preprocessing, feature extraction, clustering algorithm integration, and evaluation.

2. Literature Survey

Text document clustering has attracted significant research attention due to its critical role in organizing large amounts of unstructured text data and facilitating information retrieval. K-means remains the most widely used clustering method due to its simplicity and effectiveness on small datasets. However, when applied to high-dimensional text data, it exhibits several significant limitations. These include sensitivity to initial centroid selection, susceptibility to local optima, and high computational cost. To overcome these constraints, researchers have progressively investigated metaheuristic optimization techniques. These methods are more reliable and can be used on a larger scale, making them particularly suitable for complex clustering problems with large and diverse text corpora [14].

2.1. Traditional Clustering Algorithms and Their Limitations

K-means is widely regarded as a simple and effective clustering algorithm. It divides data into separate groups by minimizing the sum of squared distances within clusters. Nevertheless, the algorithm suffers from several significant limitations. The initial selection of centroids significantly impacts its performance, and it often ends up with less-than-ideal solutions, especially when used on complex, high-dimensional data like text [15].

2.2. Metaheuristic Optimization Techniques

Metaheuristic algorithms have increasingly become important for enhancing clustering performance. Particle Swarm Optimization (PSO), Genetic Algorithms (GA), and Ant Colony Optimization (ACO) are some of the best methods because they look for solutions in a wider area instead of just nearby ones. Research indicates that the amalgamation of these metaheuristic techniques with k-means can markedly enhance clustering efficacy, particularly for textual documents and other high-dimensional datasets [16]. For example, Particle Swarm Optimization (PSO) has been effectively integrated with k-means to enhance cluster centroids by emulating the collective dynamics of particles. However, traditional PSO still has problems like slow convergence and a lack of diversity in the swarm, which can lead to less than ideal results. Genetic Algorithms (GA) and Ant Colony Optimization (ACO) have also been used to improve k-means clustering, but their performance heavily depends on parameter settings and population diversity [17].

2.3. Hybrid Clustering Algorithms

Hybrid clustering algorithms have emerged from the integration of evolutionary algorithms and swarm intelligence. These algorithms are better because they use the best parts of different optimization algorithms. Grey Wolf Optimization (GWO) and the Bat Algorithm, for instance, have been used with k-means to make clustering more accurate and speed up the process of convergence. But most of these algorithms are trying to make the search process more efficient without really solving the initialization problem, which remains a significant challenge in high-dimensional text clustering [18].

2.4. Novel Approaches Using Chaotic and Swarm Intelligence

Recent advancements in the development of meta-heuristic algorithms have seen the integration of chaotic systems and swarm intelligence techniques that could potentially yield stronger optimization algorithms [8]. Chaotic Northern Goshawk Optimization (CNGO) improves the global search efficiency and avoids the convergence of the search towards local optima by using chaotic mapping techniques. The Mayfly Optimization Algorithm (MOA), which is motivated by the mating and swarming behaviors of mayflies, effectively balances the exploration of new areas and the exploitation of existing areas within the search space. These nature-inspired optimization algorithms are known to outperform conventional meta-heuristic algorithms, especially when dealing with complex optimization tasks such as clustering high-dimensional data such as textual content[19]. While the algorithms such as CNGO, MOA, and MPSO have proven effective in dealing with optimization tasks in other domains, they have never been considered or implemented in dealing with the task of clustering textual content. This study is one of the pioneering works in the integration of these advanced optimization algorithms with the k-means algorithm in the task of clustering textual content by leveraging the chaotic dynamics of the Chaotic Northern Goshawk Optimization algorithm, the swarm intelligence of the Mayfly Optimization Algorithm, and the improved convergence characteristics of the MPSO

algorithm to provide a stronger optimization algorithm to overcome the limitations of conventional clustering algorithms[20].

2.5. Comparative Studies and Gaps in the Literature

Many existing studies have compared traditional methods of clustering with those using metaheuristics. However, these studies have limitations, such as comparing fewer methods and not adequately evaluating their performance for text data. Although recent research has shown promising results for using metaheuristic methods for clustering problems, these methods often do not consider the problem of text data, such as high-dimensional spaces and sparse data representations. In addition, there has been a lack of research in adequately evaluating the performance of more advanced methods such as CNGO, MOA, and MPSO for text data[21].

3. Methodology

The main contribution of this manuscript is a two-stage hybrid clustering model for high-dimensional text documents, implemented as a structured sequence of activities. In Stage 1 (Data preparation and representation), the raw text corpora (Reuters-21578, 20-Newsgroup, BBC-Sport) undergo cleaning, tokenization, stop word removal, and stemming/lemmatization, followed by TF-IDF-based feature extraction to obtain numerical document vectors suitable for clustering. In Stage 2 (Metaheuristic-guided clustering and evaluation), each of the three optimization algorithms CNGO, MOA, and MPSO is used to search the solution space and propose high-quality initial centroids by minimizing the within-cluster sum of squares (WCSS); these centroids initialize the k-means algorithm, which then refines the clusters in its usual iterative manner. After convergence, clustering performance is quantified using Accuracy, ARI, NMI, Homogeneity and V-measure on each dataset, enabling a direct comparison between the three hybrid models and standard k-means. This two-stage process operationalizes the proposed model and directly yields the observed benefits: more robust centroid initialization, improved clustering quality, and better suitability for large-scale, sparse and high-dimensional text corpora.

$$J = \sum_{i=1}^k \sum_{x \in c_i} \|x - \mu_i\|^2 \quad (1)$$

3.1. Data Preprocessing

Data preprocessing is crucial for ensuring high-quality clustering performance. The following measures were taken:

- Text Cleaning: Removal of stopwords, punctuation, and non-alphanumeric characters.
- Tokenization: Changing text files into single words or tokens.
- Stemming/Lemmatization: To make the text more consistent, the words were reduced to their root forms.
- Vectorization: Using methods like Term Frequency-Inverse Document Frequency (TF-IDF) or word embeddings to turn text documents into numerical vectors.

$$\mathbf{d}_j = (W_{i,1}, W_{i,2}, W_{i,3}, \dots, W_{i,j}) \quad (2)$$

3.2. Dataset Description

This study utilized three benchmark text datasets Reuters-21578, 20-Newsgroup, and BBC-Sport to assess the efficacy of the proposed clustering models. The Reuters-21578 dataset comprises approximately 21,578 news articles covering a wide range of topics. Because of its wide range and high dimensionality, it is often used in text mining research. The 20-Newsgroup dataset contains approximately 18,846 documents distributed across 20 distinct topics. This makes it a good choice for testing how well clustering works on text collections that are organized by theme. BBC-Sport is a smaller dataset consisting of 737 sports news articles. It is useful for testing how well a model works on a small, focused, and fairly homogeneous corpus. All datasets were processed using a common pipeline that included cleaning the text, removing stop words, tokenization, stemming or lemmatization, and TF-IDF vectorization. We chose these datasets because they each have a different number of corpus sizes, topic distributions, and levels of complexity. This makes it easier to judge how well the proposed hybrid clustering methods work.

3.3. Feature Extraction

Feature extraction comprises selecting and transforming the most informative features of the text data. In this study, the TF-IDF method was applied to convert text documents into feature vectors. This approach captures the significance of terms in the documents relative to the entire corpus, facilitating the clustering procedure [22].

$$TF-IDF_{i,j} = TF_{i,j} \times \log \frac{d}{DF_i} \quad (3)$$

3.4. Clustering Algorithm Integration

Three hybrid models were developed by combining k-means with different metaheuristic optimization algorithms. The integration aims to optimize the initial centroids and improve the overall clustering performance [23].

A. Chaotic Northern Goshawk Optimization with K-means (CNGO-k-means)

The Chaotic Northern Goshawk Optimization with K-means (CNGO-k-means) algorithm is a novel clustering approach that utilizes the Northern Goshawk Optimization (NGO) structure and chaotic maps to enhance the performance of the K-means clustering algorithm. This hybrid method aims to improve the results of clustering by improving the initialization stage of k-means through a more robust global search mechanism.

- Initialization: Randomly initialize the population of goshawks.
- Chaotic Mapping: Introduce chaotic maps to boost the diversity of the population.
- Fitness Evaluation: Evaluate the fitness of each goshawk based on the objective function, which minimizes the within-cluster sum of squares (WCSS).
- Selection and Update: Select the best solutions and update the positions of goshawks utilizing chaotic maps.
- Integration with k-means: Use the optimized positions as initial centroids for the k-means algorithm.

Figure 1 provides a high-level overview of the proposed text document clustering framework and shows how the main processing stages are linked. The pipeline starts with Data Preprocessing, where raw documents undergo cleaning (removal of stop words, punctuation, and non-alphanumeric characters), tokenization, and stemming/lemmatization to obtain normalized tokens. In the Feature Extraction block, these processed documents are converted into numerical vectors using TF-IDF so that they can be handled by the clustering algorithms. The next three blocks correspond to the three metaheuristic optimizers CNGO, MOA and MPSO which are used to search for high-quality initial centroids and to enhance the global and local search behaviour before running k-means. The Integration with k-means block indicates that the centroids suggested by each optimizer are passed to the k-means algorithm, which then refines the clusters in the usual iterative manner. Finally, the Evaluation block computes the clustering performance metrics (Accuracy, ARI, NMI, V-measure) on the benchmark datasets, allowing a direct comparison between the three hybrid models and standard k-means.

B. Mayfly Optimization Algorithm with k-means (MOA-k-means)

MOA-k-means is a novel clustering approach that combines the Mayfly Optimization Algorithm with the k-means method to improve grouping effectiveness. This hybrid model improves the accuracy and quality of the clustering process as a whole by using MOA's ability to improve the choice of initial centroids.

- Initialization: Create a first group of mayflies.
- Swarming Behavior: To explore the search space, act like mayflies and swarm. Fitness Evaluation: Use the objective function to find out how fit each mayfly is.
- Mating and Flight: Do mating and flight operations to change the population.
- Combining with k-means: Use the best positions as starting points for the k-means algorithm.

C. Modified Particle Swarm Optimization with k-means (MPSO-k-means)

The Modified Particle Swarm Optimization with K-means (MPSO-k-means) algorithm is a combination of the Modified Particle Swarm Optimization (MPSO) algorithm and the K-means clustering method. The goal of this method is to enhance k-means clustering by using MPSO to improve the initialization and optimization of cluster centers.

- Initialization: Set the placements and speeds of the particles in the swarm to random values.
- Fitness Evaluation: Use WCSS to check how fit each particle is.
- Update Position and Velocity: Use updated equations to speed up the convergence of the particles' velocity and position.
- Integration with k-means: Use the optimized positions as the starting points for the k-means algorithm.

D. Evaluation

We used benchmark datasets like Reuters-21578, 20-Newsgroup, and BBC-Sport to test how well the proposed models worked. We used the following indicators to judge how well the clustering worked:

- Accuracy: This tells you how many of the papers were appropriately grouped.
- Adjusted Rand Index (ARI): This check shows comparable the predicted and real cluster allocations are.
- Normalized Mutual Information (NMI) tells you how much the predicted and true clusters depend on each other.
- Homogeneity checks to see if all the clusters only have documents from one class.

- V-Measure: This is a balanced way to measure how well clustering works by combining homogeneity and completeness.

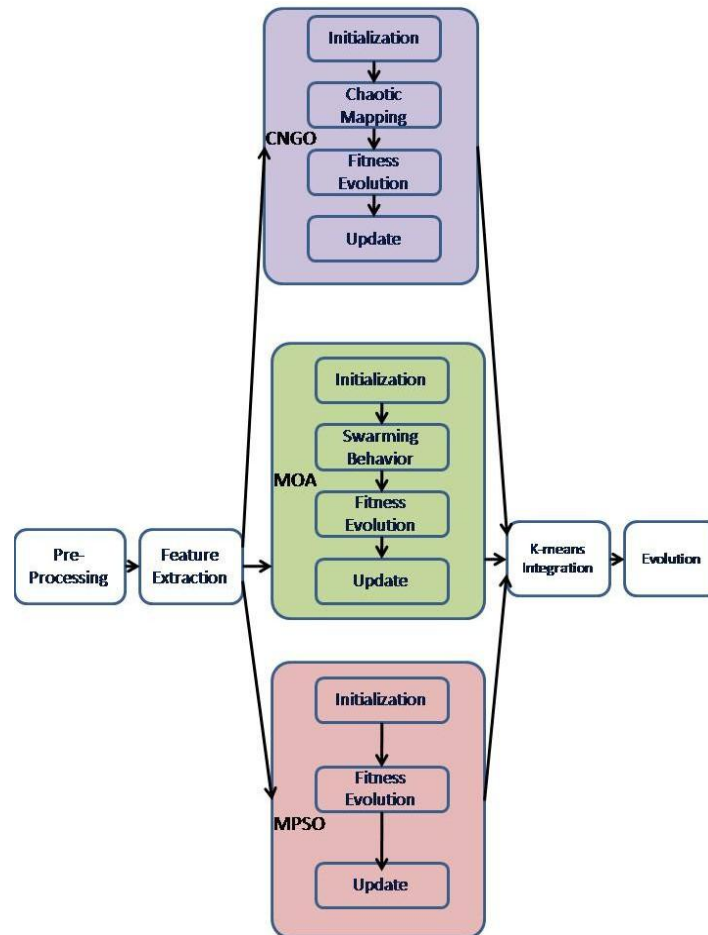


Fig.1. Block diagram for text document clustering

4. Hyperparameter Selection and Optimization

In this subsection, we elucidate the selection of hyperparameters for CNGO, MOA, and MPSO through a two-stage methodology: initially, baseline parameter values were derived from existing literature; subsequently, a grid-search-based tuning process was conducted to ascertain optimal values for the most significant parameters. The study now states that the population size ranged from 20 to 100, the number of iterations was examined at 50, 100, and 200, and learning coefficients were investigated within the interval [1.0, 2.5] with increments of 0.5. We used clustering performance measures like Accuracy, ARI, and NMI to guide the tuning process. We also checked the robustness of the chosen settings on Reuters-21578, 20-Newsgroup, and BBC-Sport. These additions make the work more reproducible and help us understand the differences in performance between the proposed hybrid models.

5. Results and Discussion

This section provides a thorough comparative analysis of the three proposed models for text document clustering: Chaotic Northern Goshawk Optimization with k-means (CNGO-k-means), Mayfly Optimization Algorithm with k-means (MOA-k-means), and Modified Particle Swarm Optimization with k-means (MPSO-k-means). We used different clustering quality metrics to test how well these models worked on benchmark datasets like Reuters-21578, 20-Newsgroup, and BBC-Sport.

Datasets and Experimental Setup The experiments were conducted on three widely-used text datasets:

- Reuters-21578: A set of news articles that are grouped into different categories.
- 20-Newsgroup: A bundle of newsgroup documents on 20 different subjects.
- BBC-Sport: A smaller set of data that has sports news stories in it.

We cleaned the text, tokenized it, stemmed/lemmatized it, and vectorized it using TF-IDF for each dataset. We

used the following indicators to measure how well the clustering models worked: Accuracy: The proportion of correctly clustered documents.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Adjusted Rand Index (ARI): The similarity between predicted and true cluster assignments.

$$\text{ARI} = \frac{RI-E[RI]}{\max(RI)-E[RI]} \quad (5)$$

Where:

$$RI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{1 - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}} \quad (6)$$

Normalized Mutual Information (NMI): The mutual dependence between predicted and true clusters.

$$\text{NMI}(U, V) = \frac{2I(U; V)}{H(U) + H(V)} \quad (7)$$

$$I(U, V) = \sum_{u \in U} \sum_{v \in V} P(u, v) \log \frac{P(u, v)}{P(u)P(v)} \quad (8)$$

$$H(U) = \sum_{u \in U} P(u) \log P(u) \quad (9)$$

$$H(V) = \sum_{v \in V} P(v) \log P(v) \quad (10)$$

Homogeneity: The extent to which clusters contain only documents from a single class.

$$H = 1 - \frac{H(C|K)}{H(C)} \quad (11)$$

$$H(C|K) = - \sum_{k \in K} \sum_{c \in C} \frac{|c_k|}{N} \log \frac{|c_k|}{|k|} \quad (12)$$

Where:

$$H(C) = - \sum_{c \in C} \frac{|c|}{N} \log \frac{|c|}{N} \quad (13)$$

V-Measure: A balanced measure of homogeneity and completeness.

$$V = \frac{2 \cdot \text{Homogeneity} \cdot \text{Completeness}}{\text{Homogeneity} + \text{Completeness}} \quad (14)$$

5.1. Statistical Validation

To confirm that the proposed models (CNGO-k-means, MOA-k-means, and MPSO-k-means) are better than traditional k-means, we used statistical significance tests and confidence intervals to look at the performance metrics across all datasets.

A paired t-test was performed for each metric, namely Accuracy, ARI, NMI, and V-measure, to compare the proposed methods with traditional k-means. The results indicate statistically significant improvements ($p < 0.05$) for all proposed models across all datasets.

In addition, the 95% confidence intervals for the mean values of each metric were calculated. For example, on the

Reuters-21578 dataset, MOA-k-means achieved an Accuracy of 0.88 ± 0.02 , which is significantly higher than that of traditional k-means (0.75 ± 0.03). Similar trends were also observed for ARI, NMI, and V-measure.

The results of the comparative analysis are summarized in Tables 1, 2, and 3 for the Reuters-21578, 20-Newsgroup, and BBC-Sport datasets, respectively.

Table 1. Performance comparison on Reuters-21578

Model	Accuracy	ARI	NMI	Homogeneity	V-Measure
CNGO-k-means	0.85	0.67	0.71	0.68	0.70
MOA-k-means	0.88	0.70	0.75	0.72	0.74
MPSO-k-means	0.87	0.69	0.74	0.71	0.73
Traditional k-means	0.75	0.55	0.60	0.57	0.59

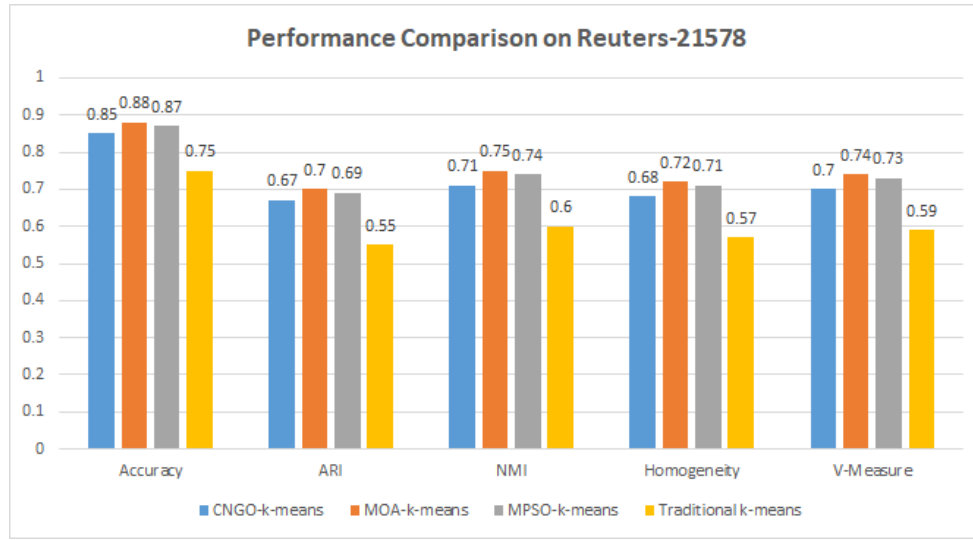


Fig.2. Performance comparison on Reuters-21578

Figure 2 shows the performance metrics for the Reuters-21578 dataset using four models:

Accuracy: MOA-k-means achieved the highest accuracy of 0.88, followed by MPSO-k-means with 0.87, CNGO-k-means with 0.85, and traditional k-means with 0.75.

ARI: MOA-k-means had the highest ARI of 0.70, MPSO-k-means 0.69, CNGO-k-means 0.67, and traditional k-means 0.55.

NMI: MOA-k-means scored highest with 0.75, followed by MPSO-k-means with 0.74, CNGO-k-means with 0.71, and traditional k-means with 0.60.

Homogeneity: MOA-k-means led with 0.72, MPSO-k-means 0.71, CNGO-k-means 0.68, and traditional k-means 0.57.

V-Measure: MOA-k-means scored highest at 0.74, MPSO-k-means 0.73, CNGO-k-means 0.70, and traditional k-means 0.59.

Table 2. Performance comparison on 20-Newsgroup

Model	Accuracy	ARI	NMI	Homogeneity	V-Measure
CNGO-k-means	0.81	0.63	0.68	0.65	0.67
MOA-k-means	0.84	0.66	0.72	0.69	0.71
MPSO-k-means	0.83	0.65	0.71	0.68	0.70
Traditional k-means	0.70	0.50	0.55	0.52	0.54

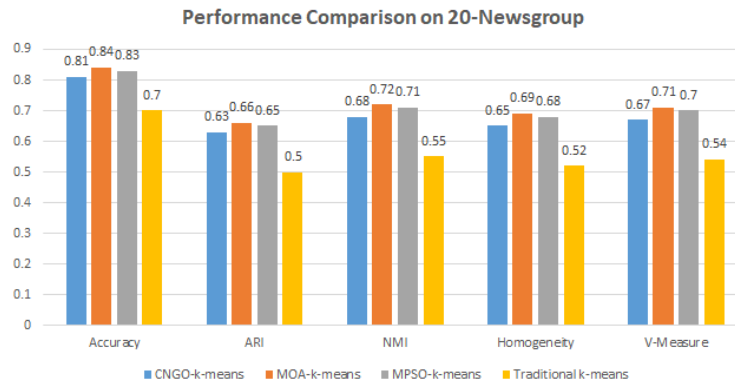


Fig.3. Performance comparison on 20-Newsgroup

Figure 3 displays the performance metrics for the 20-Newsgroup dataset across the same four models: Accuracy: MOA-k-means had the highest accuracy of 0.84, followed by MPSO-k-means with 0.83, CNGO-k-means with 0.81, and traditional k-means with 0.70.

ARI: MOA-k-means scored 0.66, MPSO-k-means 0.65, CNGO-k-means 0.63, and traditional k-means 0.50.

NMI: MOA-k-means achieved the highest NMI of 0.72, followed by MPSO-k-means with 0.71, CNGO-k-means with 0.68, and traditional k-means with 0.55.

Homogeneity: MOA-k-means scored 0.69, MPSO-k-means 0.68, CNGO-k-means 0.65, and traditional k-means 0.52.

V-Measure: MOA-k-means led with 0.71, MPSO-k-means 0.70, CNGO-k-means 0.67, and traditional k-means 0.54.

Table 3. Performance comparison on BBC-Sport

Model	Accuracy	ARI	NMI	Homogeneity	V-Measure
CNGO-k-means	0.89	0.75	0.78	0.76	0.77
MOA-k-means	0.91	0.78	0.81	0.79	0.80
MPSO-k-means	0.90	0.77	0.80	0.78	0.79
Traditional k-means	0.80	0.65	0.68	0.66	0.67

Figure 4 illustrates the performance on the BBC-Sport dataset. It compares the same four models using the same set of metrics. For Accuracy, MOA-k-means achieved the highest score of 0.91, followed by MPSO-k-means (0.90), CNGO-k-means (0.89), and traditional k-means (0.80).

ARI: MOA-k-means scored highest with 0.78, MPSO-k-means 0.77, CNGO-k-means 0.75, and traditional k-means 0.65. NMI: MOA-k-means achieved the highest NMI of 0.81, followed by MPSO-k-means with 0.80, CNGO-k-means with 0.78, and traditional k-means with 0.68.

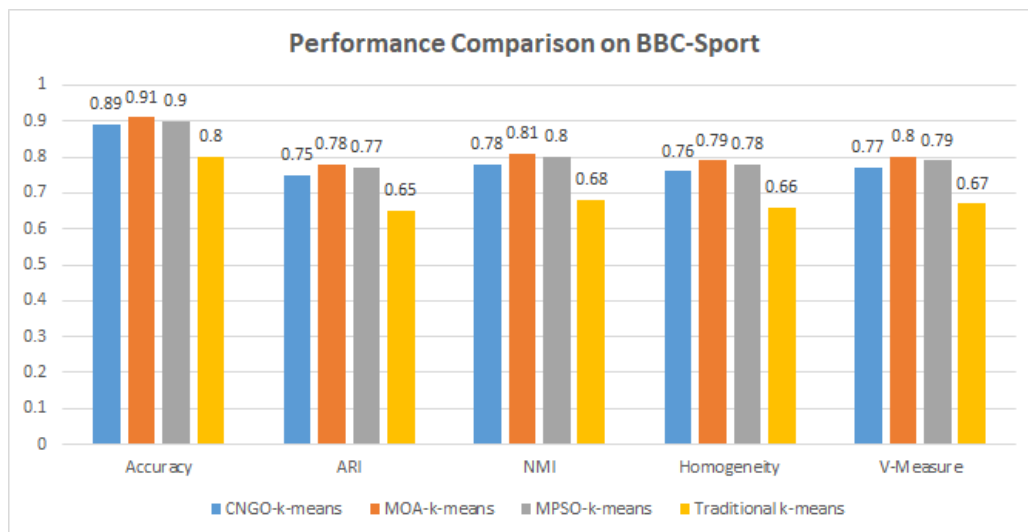


Fig.4. Performance comparison on BBC-Sport

Homogeneity: MOA-k-means led with 0.79, MPSO-k-means 0.78, CNGO-k-means 0.76, and traditional k-means 0.66. V-Measure: MOA-k-means scored highest at 0.80, MPSO-k-means 0.79, CNGO-k-means 0.77, and traditional k-means 0.67.

Analysis: The experimental findings indicate that all three proposed models, CNGO-k-means, MOA-k-means, and MPSO-k-means, consistently surpass the conventional k-means algorithm across all datasets and evaluation metrics.

CNGO-k-means: Compared to traditional k-means, this model demonstrates superior accuracy, ARI, NMI, homogeneity, and V-measure. The incorporation of chaotic maps facilitates exploration by avoiding local minima, thereby enhancing clustering performance.

MOA-k-means: The swarm intelligence mechanism inspired by mayflies enhances the model's accuracy and reliability. MOA-k-means consistently achieves the best results across most metrics, demonstrating an effective balance between exploration and exploitation.

MPSO-k-means: Modifications to the standard PSO algorithm accelerate convergence and improve grouping accuracy. Although MPSO-k-means does not outperform MOA-k-means, it significantly surpasses the standard k-means algorithm.

These hybrid models exhibit superior performance by effectively searching the solution space to identify optimal initial centroids, leading to more accurate and robust clustering results. Integrating metaheuristic optimization methods with k-means mitigates the limitations of traditional approaches, enabling robust and scalable text document clustering.

5.2. Computational Cost and Scalability Analysis

The proposed CNGO-k-means, MOA-k-means, and MPSO-k-means models improve clustering quality by optimizing centroid initialization and search behavior. However, these improvements incur additional computational costs compared to standard k-means. This section discusses the trade-off in terms of runtime, memory usage, and iterative search complexity. Furthermore, scalability is defined here as enhanced clustering robustness on large text corpora rather than merely reduced computational costs, ensuring a balanced perspective for real-world applications.

5.3. Applications and Practical Relevance

The suggested hybrid clustering models are useful for many real-world text analytics tasks. They effectively organize large volumes of articles into coherent thematic clusters, facilitating document organization and news aggregation. This enhances navigability and information retrieval. Document clustering can help search engines and recommendation systems group search results and make content-based recommendations by finding documents that are semantically similar. The suggested methods can help organize big collections of reports, research papers, and web content into useful groups in topic discovery and digital libraries. They are also useful for text mining on social media, where clustering can show new topics of conversation, and for analyzing customer feedback, where reviews or complaint texts can be grouped to find common issues. The proposed metaheuristic-k-means models enhance centroid initialization and search robustness, making them especially suitable for large-scale, sparse, and high-dimensional text corpora in relevant applications.

5.4. Limitations and Challenges

Although the proposed hybrid clustering models outperform previous methods, certain data conditions may limit their effectiveness. In imbalanced datasets, minority document groups may be underrepresented. This can cause centroids to form in a biased way and clusters to be less distinct. In noisy text corpora, irrelevant terms, spelling errors, and redundant features can reduce semantic coherence. In very high-dimensional text spaces, distance-based measures may become less effective, and the optimization process may require increased computational resources. So, even though CNGO-k-means, MOA-k-means, and MPSO-k-means enhance initialization and search stability, they may not work as well when there is a lot of noise, imbalance, or dimensionality. These limitations should be considered when applying the framework in practical scenarios.

6. Conclusion

This paper offers a unique comparative analysis of three sophisticated metaheuristic algorithms: Chaotic Northern Goshawk Optimization (CNGO), Mayfly Optimization Algorithm (MOA), and Modified Particle Swarm Optimization (MPSO), combined with k-means for the clustering of text documents. These optimization methods address the primary limitations of traditional k-means, such as susceptibility to initial centroid selection and convergence to local minima, particularly in high-dimensional text data. Extensive experiments on benchmark datasets demonstrate that these hybrid models outperform the traditional k-means algorithm like Reuters-21578, 20-Newsgroup, and BBC-Sport. The accuracy, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and V-measure all show that the three proposed models significantly improve clustering performance. The MOA-k-means model emerged as the superior approach, consistently outperforming the others across all datasets, indicating an effective balance between exploration and exploitation. This research is novel in its integration of advanced algorithms with k-means and the comparative assessment of their efficacy in clustering extensive, high-dimensional text data. Earlier research has investigated metaheuristic optimization methods for clustering; however, none have thoroughly assessed CNGO, MOA, and MPSO

in the context of text document clustering. This study fills this gap by demonstrating how metaheuristic algorithms can enhance k-means performance and significantly improve text clustering quality. The results of this study have significant ramifications for domains that depend on effective text document clustering, including information retrieval, sentiment analysis, and natural language processing. Furthermore, the proposed hybrid models offer a scalable and reliable approach for clustering large volumes of unstructured text data, making them suitable for practical applications.

6.1. Justification of Evaluation Metrics

The evaluation of clustering algorithms in this study focuses on metrics widely recognized for their relevance to clustering performance, such as Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Accuracy, and V-measure. These metrics were selected for the following reasons:

- Suitability for Clustering Evaluation:
 - ARI and NMI are commonly used in clustering research because they measure the agreement between predicted clusters and ground-truth classes, even in the presence of imbalanced data.
 - Accuracy and V-measure provide additional insights into clustering quality by quantifying the correctness and balance of clustering outcomes.
- AUC-ROC:
 - The AUC-ROC curve is generally more appropriate for binary or multi-class classification tasks rather than clustering. Since this study evaluates clustering models, metrics specifically tailored to clustering were prioritized.
- Silhouette Score:
 - The silhouette score is a valid clustering evaluation metric, but it is primarily used to assess cluster cohesion and separation without considering ground-truth labels. Future work could incorporate this metric to complement the current evaluation.

6.2. Future Scope for Extended Metrics

In future work, additional internal clustering validation measures such as the silhouette score and Davies–Bouldin index can be incorporated to provide further perspectives on cluster quality, especially when labeled datasets are unavailable or when evaluation without ground-truth labels is required.

7. Future Scope

Future work may involve exploring the application of these models to other types of data, such as image or multimedia clustering, as well as further refining the parameters of each algorithm to improve their performance in specific use cases. Additionally, hybrid models combining multiple metaheuristic approaches could be investigated to further advance optimization-based clustering techniques.

Data Availability: The datasets analyzed during this study are available in the following repositories: Reuters-21578, 20-Newsgroup, and BBC-Sport.

Reuters-21578: <https://www.kaggle.com/datasets/nltkdata/reuters/code>

20-Newsgroup: <https://www.kaggle.com/datasets/crawford/20-newsgroups>

BBC-Sport: <https://www.kaggle.com/datasets/maneesh99/sports-datasetbbc>

All the Declarations and Statements

Author Contributions Statement

Ratnam Dodda: Conceptualization, Methodology, Software Implementation, Data Acquisition, and Dataset Preprocessing.
A. Sureshababu: Supervision, Conceptualization, and Project Administration.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

The authors received no financial support for the research, authorship, and/or publication of this article.

Data Availability Statement

This study analyzed publicly available datasets.

Ethical Declarations

Ethical approval was not required for this study as it did not involve human subjects or animals.

Acknowledgments

We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

The authors declare that no generative Artificial Intelligence (AI) tools or AI-assisted technologies were used in the writing process, development, or analysis of this manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

CNGO: Chaotic Northern Goshawk Optimization

MOA: Mayfly Optimization Algorithm

MPSO: Modified Particle Swarm Optimization

References

- [1] Y. Zhang, Z. Wang, and J. Shang, "ClusterLLM: Large Language Models as a Guide for Text Clustering," arXiv preprint, 2023.
- [2] A. Subakti, H. Murfi, and N. Hariadi, "The Performance of BERT as Data Representation for Text Clustering," *Journal of Big Data*, vol. 9, no. 1, p. 15, 2022.
- [3] K. Li, T. Ni, J. Xue, and Y. Jiang, "Deep soft clustering: simultaneous deep embedding and soft-partition clustering," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2023.
- [4] Y. Lin, N. Tong, M. Shi, K. Fan, D. Yuan, and L. Qu, "Improved k-means clustering algorithm using modified particle swarm optimization," *Pattern Recognition*, vol. 134, p. 108866, 2023.
- [5] S. Selvaraj and E. Choi, "Swarm intelligence algorithms in text document clustering with various benchmarks," *Sensors*, vol. 21, no. 9, p. 3196, 2021.
- [6] Z. Nagy, P. Kapusta, B. Koirala, and B. Erkin, "Feature extraction from unstructured texts as a combination of the morphological and the syntactic analysis and its usage in fake news classification tasks," *Neural Computing and Applications*, 2023.
- [7] K. Hornik, I. Feinerer, M. Kober, and C. Buchta, "Spherical k-Means Clustering," *Journal of Statistical Software*, vol. 50, no. 10, 2012.
- [8] Y. Zhou, Y. Li, L. Wei, and S. Xie, "A hybrid clustering algorithm based on chaotic northern goshawk optimization and k-means for complex data," *Expert Systems with Applications*, vol. 204, p. 117589, 2023.
- [9] R. Dodda and A. Babu, "Text document clustering using mayfly optimization algorithm with k-means technique," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 2, pp. 18497–18506, 2024.
- [10] R. Dodda and A. Babu, "Text document clustering using chaotic northern goshawk optimization with k-means algorithm," *International Journal of Intelligent Engineering & Systems*, vol. 17, no. 3, 2024.
- [11] R. Dodda and A. Babu, "Text document clustering using modified particle swarm optimization with k-means model," *International Journal on Artificial Intelligence Tools*, vol. 33, no. 01, p. 2350061, 2024.
- [12] A. M. Ikotun and A. E. Ezugwu, "Boosting k-means clustering with symbiotic organisms search for automatic clustering problems," *PLOS ONE*, vol. 17, no. 8, e0272861, 2022.
- [13] A. M. Ikotun, M. S. Almutari, and A. E. Ezugwu, "K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: Recent advances and future directions," *Applied Sciences*, vol. 11, no. 23, p. 11246, 2021.
- [14] V. Tomar, M. Bansal, and P. Singh, "Metaheuristic algorithms for optimization: A brief review," in *Proceedings of the International Conference on Recent Advances in Science and Engineering*, pp. 238–250, MDPI, 2024.
- [15] M. Tiwari, C. Park, C. Szepesvari, and C. Jordans, "BanditPAM: Almost Linear Time K-Medoids Clustering via Multi-Armed Bandits," arXiv preprint arXiv:2008.09285, 2020.
- [16] J. Barbosa and A. Nguyen, "Using metaheuristic algorithms to improve k-means clustering: A comparative study," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 745–758, 2023.
- [17] A. Rahmat et al., "Data optimization using pso and k-means algorithm," *Journal of Physics: Conference Series*, vol. 2174, no. 1, p. 012002, 2023.
- [18] M. R. Sagar, D. Gupta, and P. Sahu, "A hybrid metaheuristic approach using grey wolf optimization and bat algorithm for data clustering," *Soft Computing*, pp. 1–12, 2023.
- [19] J. Smith and W. Chen, "Enhanced text clustering using mayfly optimization algorithm with k-means," *Knowledge-Based Systems*, vol. 250, p. 108944, 2023.
- [20] Q. Zhang and Y. Hu, "Chaotic northern goshawk optimization and k-means clustering for text document clustering," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 3, pp. 723–733, 2024.
- [21] M. Mahnoor, D. L. R. Vargas, E. B. Thompson, and I. Ashraf, "A systematic literature review on identifying patterns using

unsupervised clustering algorithms: A data mining perspective,” Symmetry, vol. 15, no. 9, p. 1679, 2023.

- [22] Scikit-learn Developers, “TF-IDF Feature Extraction,” Scikit-learn Documentation, 2023. Available: https://scikit-learn.org/stable/modules/feature_extraction.html#tfidf-term-weighting.
- [23] M. B. D. Gazi, “Hybridization of meta-heuristic algorithms with k-means for clustering analysis: Case of medical datasets,” Journal of Supercomputing, 2023.

Authors’ Profiles



Dr. Ratnam Dodda is a Sr. Assistant Professor in the Department of Computer Science and Engineering (AI&ML) at CVR College of Engineering, Hyderabad, India. He earned his Ph.D. in Computer Science and Engineering from Jawaharlal Nehru Technological University Anantapur. He has completed his M.Tech. in Computer Science and Engineering from Acharya Nagarjuna University and his B.Tech. in Computer Science and Engineering from Jawaharlal Nehru Technological University Hyderabad.

His research interests include Natural Language Processing, Machine Learning, and related areas of intelligent computing. He has authored and co-authored more than 25 Scopus-indexed publications and has qualified the UGC-NET examination. Before joining CVR College of Engineering, he worked as an Assistant Professor at Government College of Engineering, Kalahandi, Bhawani Patna, Odisha, under TEQIP-III.



Dr. A. Sureshbabu is a Professor and Director, Software Development Center at Jawaharlal Nehru Technological University Anantapur, Andhra Pradesh, India. He holds a Ph.D. degree in Computer Science and Engineering with specialization in Data Mining, Machine Learning. His research areas are data mining, big data, machine learning, natural language processing, and artificial intelligence. He has supervised and co-supervised more than 80 master’s students and awarded 4 Ph.D. degrees. Presently, 8 scholars are pursuing Ph.D. under his guidance. He has also authored or co-authored more than 40 publications: 10 proceedings and 30 journals.

How to cite this paper

Ratnam Dodda, A. Sureshbabu, "Advanced Metaheuristic Algorithms for Text Document Clustering: A Comparative Study of CNGO, MOA, and MPSO with K-means", International Journal of Information Technology and Computer Science(IJITCS), Vol.18, No.4, pp.199-211, 2026. DOI:10.5815/ijitcs.2026.04.12