Int. J. Advanced Networking and Applications
Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

An Efficient Document Clustering Method Using Improved Bacterial Colony Optimization

S Dhanalakshmi
Department of Computer Science
Thiruvalluvar Government Arts College, Rasipuram, Tamilnadu, India,
Email: dhanalakshmisubramaniam79@gmail.com
S Sathiyabama
Department of Computer Science
Government Arts and Science College, Kangeyam, India
Email: drsathyaabama@gmail.com
D Ayyamuthukumar
Department of Artificial Intelligence and Data Science
Muthayammal College of Engineering, Rasipuram, India
E-Mail: ukdkumar@gmail.com

----------ABSTRACT----------

This paper presents an efficient document clustering method based on an improved Bacterial Colony Optimization (BCO) algorithm. The proposed approach integrates a centroid opposition- based learning (COBL) strategy for effective population initialization and incorporates Quantum Computing-inspired position updates to enhance the search capability of BCO. To strengthen clustering performance, multiple feature extraction techniques—Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and GloVe—are utilized in a hybrid manner to capture both statistical and semantic document representations. The framework is rigorously evaluated on six benchmark datasets: Reuters-21578, Classic4, WebKB, SMS, BBC, and DMOZ, and its performance is compared against traditional and evolutionary clustering. Experimental results demonstrate that the proposed COL-QBCO framework significantly outperforms the baseline methods in terms of clustering accuracy, normalized mutual information (NMI), F-measure, and convergence rate, highlighting its robustness and effectiveness in handling diverse and high-dimensional text datasets.

Keywords: Text document clustering; Bacterial colony optimization; Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and GloVe

---

Date of Submission: October 15, 2025 Date of Acceptance: December 10, 2025

I. INTRODUCTION

The sudden increase in the amount of written information in the last decades is due to the active development of digital information that has created the world of numerous textual materials created using digital technologies of online sources, social networks, digital libraries, news websites, and business databases. Sorting and deriving significant patterns in such large amounts of text has become a critical information access, knowledge discovery and natural language processing challenge [1]. One method that has become central in this respect is document clustering in which documents sharing a common content are automatically clustered together into a cluster without labeling [2]. In contrast to supervised learning techniques, clustering is unsupervised and especially effective in the context of unstructured data of large size, thus it is of great importance in any search engine, recommendation system, text categorization, and archiving of data in digital form [3]. K-Means and other traditional clustering methods have been extensively applied because they are simple and efficient. Nevertheless, K-Means has a number of drawbacks that are present in its nature such as sensitivity to initialization, local optimum convergence, and poor performance with high-dimensional sparse document vectors [4, 5]. In overcoming these some metaheuristic optimization algorithms, Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Bacterial Colony Optimization (BCO) have been applied in clustering. BCO has demonstrated potential among them because of biologically inspired chemotaxis, reproduction and elimination-dispersal practices that provide analogies to bacterial groups in their search of nutrient rich areas. Nevertheless, standard BCO has its drawbacks, including random initialization resulting in low convergence rates, low exploration capability of high-dimensional feature space, and sensitivity to being stuck in local minima [6].

To overcome such shortcomings, this paper presents a better-sourced BCO scheme of document clustering, where Centroid Opposition-Based Learning (COBL) will be used to do an effective initializing step, and Quantum-inspired position updating will be utilized to optimize the search process. COBL improves the diversity and quality of the original bacterial population by producing not only candidate solutions, but also their complementary solutions as far as the boundaries of the feature space are concerned. This is so that the search is

Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

commenced at promising and varied areas, and the chances of premature convergence are minimized. Conversely, quantum-inspired updates mean the inclusion of probabilistic movement strategies based on quantum mechanics to allow bacteria to escape local optima and conduct global exploration and local refinement in a more efficient way. Collectively, these mechanisms can offer a balanced optimization approach, and make BCO more efficient and robust in task clustering.

Along with enhanced optimization process, various document feature extraction algorithms, such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec and GloVes, are incorporated into a hybrid representation scheme in this study. BoW and TF-IDF do statistical significance of words, whereas Word2Vec and GloVe offer semantic representations, which are contextual data about words. With these representations combined, the proposed framework will make sure that syntactic and semantic representations of documents are taken into account thereby improving the quality of the resulting clusters. The main focus of the study is to create an effective and stable document clustering system through refining BCO with COBL and Quantum-inspired updates and combining several feature extraction techniques to achieve better clustering.

The proposed approach is strictly tested on six benchmark collections (Reuters-21578, Classic4, WebKB, SMS, BBC, and DMOZ). These datasets have a wide variety of types of documents, such as news articles, scholarly text, short messages, and web content, hence guaranteeing a thorough validation of the approach. They are compared to a number of base clustering algorithms, such as K-Means, GA, PSO, IPSO, GQPSO and regular BCO. Measures of the performance are taken in terms of the popular evaluation metrics of performance, which include accuracy, precision, recall, and F-measure, and convergence rate. The key findings of this study may be concluded as follows:

- A new methodological approach of integrating BoW, TF-IDF, Word2Vec, and GloVe into a unified feature space, so that both syntactic and semantic aspects of the process are reflected in the clustering process.
- Application of COBL to create high-quality and diverse initial population of bacteria bypassing standard BCO constraints on random initialization.
- Introduction of quantum-inspired probabilistic updates to the chemotaxis phase of BCO, which makes the algorithm more flexible and able to find the optimal tradeoff between exploration and exploitation.
- Improvement of a better BCO-based clustering framework (QCOBL-BCO) that can reach better conversion velocity and clustering results.
- Strict validation on six benchmark datasets and comparative analysis with the state-of-the-art algorithms on clustering, which showed substantial improvement.

The rest of the paper is structured in such a way that section 2 talks about the related works; section 3 talks about research methods such as BCO, COBL, quantum methods and proposed IBCO. The experimental results and discussions are presented in section 4. Section 5 talks about the end of the research work and its improvements in the future.

II. RELATED WORKS

The past few years have been characterized by an explosion in the number of studies aimed at utilizing swarm intelligence and metaheuristic algorithms in the process of text document clustering (TDC) and text document summarization. A number of hybrid approaches have been suggested to ameliorate clustering accuracy, robustness and scalability in situations where large amounts of unstructured text data in high dimensions are involved. The current clustering document techniques rely to a large extent on shallow textual features and their traditional evolutionary algorithms, thereby limiting their capability of semantic relationship mapping and resulting in poor clustering in high dimensional text spaces. In addition, classical swarm-based methods have a low population diversity and tend to converge early making them impractical when dealing with large scale and intricate data sets.

In order to address these shortcomings, the suggested COL-QBCO framework combines hybrid representations of features (BoW, TF-IDF, Word2Vec, and GloVe) together to simultaneously capture both statistical and semantic features of documents. Diversification of population by adding Centroid Opposition-Based Learning (COBL) can improve convergence speed, yet quantum-inspired updates of positions can help to improve global search power and avoid stalling. The high performance can be further proven by large-scale test over a range of benchmark datasets with various performance indices and the fact that the proposed approach is relatively robust and generalizable.

Table 1 : Related works for document clustering problem

Author & Year	Method	Dataset(s)	Highlights	Performance
Dodda et al. (2024) [7]	MPSO + K-Means	Reuters, 20-Newsgroups, BBC-Sport	TF-IDF + Modified PSO	Accuracy: **0.85–0.86**
Sanchez-Gomez et al. (2025) [1]	MOSIDOC	ODP-239	Optimized compactness & DB-index	Up to **86.81% improvement**
Abuain et al. (2024) [8]	LBSSA (Link-Based Salp Swarm)	Reuters, WebKB	Neighborhood-based exploitation	Higher accuracy & faster convergence
Hassan et al. (2025) [9]	K-Means++ + PSO	Reuters, 20-Newsgroups	Feature extraction + smart initialization	Purity up to **0.95**

Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

Al-Betar et al. (2023) [10]	BBSSA (Bare-Bones SSA)	Reuters, 20-Newsgroups, WebKB	Gaussian-based greedy exploitation	Outperformed SSA & others
Dodda et al. (2023) [11]	CNGO + K-Means	Reuters, 20-Newsgroups, BBC-Sport	Chaotic NGO improves convergence	Accuracy: **96–98%**
Kumar et al. (2024) [12]	SWI	WebKB, Reuters	Tackled scalability & local optima	Improved clustering efficiency
Pitchandi et al. (2023) [13]	AJWJSC–COA Hybrid	Reuters, 20-Newsgroups	Hybrid clustering	Better performance vs. K-Means, MFO
Larabi-Marie-Sainte et al. (2023) [14]	GWO + SOM	Arabic Text Dataset	Replaced K-Means with SOM	Higher accuracy & efficiency
Nachaoui et al. (2024) [15]	HPSO (Adaptive + Chaotic)	Reuters, WebKB	Feature selection optimization	Improved precision & clustering quality
Abdulsahib et al. (2024) [16]	DGBPSO + DBSCAN	Reuters	Automatic parameter optimization	Higher accuracy & F-measure
Azarshab et al. (2024) [17]	TLBO-GWO Hybrid	Reuters	Feature selection for redundancy reduction	Enhanced clustering performance
Debnath et al. (2023) [18]	Cat Swarm Optimization (CSO)	DUC datasets	Extractive summarization	25% ROUGE-1 improvement
Aote et al. (2023) [19]	BPSO + IGA Hybrid	Hindi multi-doc	Sentence selection optimization	Outperformed 6 state-of-the-art methods
Das et al. (2024) [20]	Binary GWO (BGWO)	DUC datasets	Binary encoding for summarization	Better precision & recall
Sheikh et al. (2024) [21]	BMOGWO (Multi-objective GWO)	CNN/DailyMail, DUC	Multi-objective extractive summarization	Significant ROUGE improvements

III. RESEARCH METHODS

The following subsection discuses about the research method which are used in the research works in details.

III.1 BCO

BCO is a population-based method planned by Niu et al. [22]. BCO is a population-based technique that will be implemented by Niu et al. (2012) [22] and have been utilized in numerous real-world applications [22-29]. To address the above issue, a new bacterial algorithm named BCO was created with SI attributes in order to accelerate the optimization process. BCO stages comprise chemotaxis and communication, elimination and reproduction, migration, and the others four stages. The whole process of BCO exploits the chemotaxis and communication stage. The bacteria respond to the population statistics by altering their swimming and tumbling behaviors. A special chemotaxis and communication strategy is used to refresh the location of the microorganisms. Bacteria may undergo two types of chemotaxis in their lives, swimming and tumbling. A stochastic direction is involved in the process of actual swimming when tumbling. The joint actions of the best searching director in Tumbling and the turbulent director affect the new locations and new search direction of every bacteria. These effects are outlined as follows:

$$Position_i(T) = Position_i(T-1) + C(i) * [f_i.(G_{best} - Position_i(T-1)) + (1-f_i) * (P_{best_i} - Position_i(T-1)) + turb_i] \quad (1)$$

Bacteria do not however have a director of turbulence; he will direct them to optimum state which is as follows:

$$Position_i(T) = Position_i(T-1) + C(i) * [f_i.(G_{best} - Position_i(T-1)) + (1-f_i) * (P_{best_i} - Position_i(T-1))] \quad (2)$$

$$C(i) = C_{min} + \left(\frac{Iter_{max} - Iter_j}{Iter_{max}}\right)^n (C_{max} - C_{min}) \quad (3)$$

$turb_i$ - direction variance of turbulence $f_i \in \{0,1\}$, $C(i)$ - chemotaxis step-size. P_{best} and G_{best} denotes the personal and global best value, respectively. n is a linearly decreasing way of the chemotaxis step. $Iter_{max}$ and $Iter_j$ - the maximum number of iterations and current iteration respectively. During the stage of eradication and reproduction, the unhealthy bacterium will be substituted by the one of high energy and will reproduce them to create the latest people. The fact that it is of high energy demonstrates that the bacterium is efficient in seeking resources. When the conditions are satisfied, the bacterium is able to move within a defined range of search space during the

Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

migration phase. The bacteria usually move towards the latest nutrients in a given probability during the migration stage.

III.2 Quantum Computing

A quantum computing refers to a field of computing that utilizes quantum physics in computing. In quantum computation, quantum bits (qubits) are used rather than binary digits (bits) of information as in classical computation. Superposition The qubits may be more than one state concurrently. Such an aspect enables quantum computers to accomplish some tasks much faster compared to traditional computers [23]. The state of a qubit can be represented as a two-dimensional complex vector in a complex two-dimensional space known as a Bloch sphere. A qubit system contains two situations, $|0\rangle$ and $|1\rangle$. As stated earlier, a qubit may exist in a superposition of both $|0\rangle$ and $|1\rangle$ state, which is denoted by $|\Psi\rangle$ as follows,

$$|\Psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle \tag{4}$$

where the probability amplitudes of positions 0 and 1 are denoted by the complex numbers α and β, respectively. A qubit state is measured by use of an observation procedure. Once the process of observation has been completed, the state of a qubit can be determined as being 0 or 1. The likelihood of the qubit being in the 0 and 1 states respectively, are designated by the modulus $|\alpha|^2$ and $|\beta|^2$, and both α and β should meet the normalization necessity.

$$|\alpha|^2 + |\beta|^2 = 1 \tag{5}$$

Therefore, a qubit can be stated as a pair of complex integers using the method $q = \left[\frac{\alpha}{\beta}\right]$. Similarly, a multi-qubit scheme with a n number of qubits can be shown as

$$Q = [q_1, q_2, \dots . q_n] = \left[\frac{\alpha_1}{\beta_1}\middle|\frac{\alpha_2}{\beta_2}\middle|\frac{\dots}{\dots}\middle|\frac{\alpha_n}{\beta_n}\middle|\right] \tag{6}$$

The state of a qubit may be updated through quantum gates, e.g., the NOT-gate, CNOT-gate, Hadamard-gate, etc. In the case of a qubit q_i, the quantum angle may be written as follows:

$$\theta_i = arctan\frac{\beta_i}{\alpha_i} \tag{7}$$

III.3 Centroid opposition-based learning (COBL)

COBC is an extremely popular scheme of OBL strategy that demonstrated impressive success in DE algorithm in all the competitions of the same sort. The algorithm takes into account the whole population when it is calculating the centroid opposite points in the metaheuristic algorithm. Let $X = (X_1 \dots X_N)$ be N points in a D-dimensional search space. Carrying unit mass in the search space. Then the centroid of the body may be defined in the following way:

$$M = \frac{x_1 + x_2 + x_3 + \dots + x_N}{N} \tag{8}$$

The centroid value in the j^{th} dimension of the data set can be computed as follows:

$$M_j = \frac{1}{N}\sum_{i=1}^{N} x_i, j \tag{9}$$

With M the centroid of a discrete uniform body, the opposite-point $X_i^{\check{}}$ of a point X_i of the body is determined by using the formula:

$$X_i^{\check{}} = 2 \times M - X_i \tag{10}$$

Optimization problems may also be solved using centroid approach, which normally performs well as compared to min-max method. The boundary, which is estimated and determined by the generated sample points is computed as $[X_{min} \dots X_{max}]$.

III.4 Proposed improved BCO (IBCO)

The traditional BCO algorithm has demonstrated potential in solving clustering problems due to its population-based nature and adaptive mechanisms of chemotaxis and communication, reproduction, and elimination-dispersal. However, standard BCO often suffers from two major limitations: (1) poor initialization of the population, which may lead to slow convergence and premature trapping in local optima, and (2) limited exploration ability during position updates, which can reduce its effectiveness when applied to high-dimensional and sparse datasets such as text corpora. To overcome these limitations, the present work integrates COBL for population initialization and Quantum Computing-inspired position updates into the BCO framework, making it more suitable for document clustering tasks. COBL enhances the initialization phase of BCO by generating not only candidate solutions but also their corresponding opposite solutions in the feature space. The principle is grounded in opposition-based learning, where evaluating both a solution and its opposite increases the probability of starting closer to the global optimum. When adapted to clustering, COBL generates candidate bacterial populations around the centroids of the data distribution and simultaneously considers their opposites with respect to the feature boundaries. This process increases the diversity of the initial bacterial population and ensures that the search begins from a broad and promising region of the solution space.

Consequently, the algorithm minimizes the risk of bias in initializing and it enhances the convergence speed. To further improve the exploration and exploitation steps in the optimization process, the chemotaxis step of BCO is boosted by the Quantum-inspired position updating mechanisms. The benefit of using COBL and quantum-inspired updates together, as opposed to using them separately, is theoretical as they are complementary. COBL guarantees that the original bacterial colony is spread to varied and promising areas of search space, which serves to decrease the reliance on arbitrary launching. In the meantime, the quantum-inspired update scheme provides persistent exploration in the optimization process, which prevents the solution process and improves the quality of solutions. These improvements are realized when combined to provide the BCO algorithm with the ability to dynamically search the hybrid feature space comprising of Bag-of-Words, TF-IDF, Word2vec and GloVe representations to ensure that both semantic and statistical associations between documents are reflected. All in all, the suggested enhanced BCO framework theoretically ensures better clustering performance due to more informed start up strategy and a more powerful

Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

search mechanism. Through improved starting points given by COBL and using quantum inspired updates to do efficient explorations of the trajectory, the algorithm converges faster, has increased robustness and accuracy, than the traditional clustering methods and standard BCO.

The process of optimization starts with the initiation of the bacterial population. This is instead of using randomization during initialization, COBL is proposed to generate solutions and their complementary solutions around dataset centers. The mechanism ensures that the starting population is diverse, and the algorithm has a higher exploration ability, which would lower the chance of the poor convergence. The locations of the bacteria, assumed to be the possible centers of the clusters, are also updated as part of an iterative optimization process based on a Quantum Computing-inspired process. In contrast to the classical deterministic dynamics, the quantum-based model uses the probabilistic superposition principles to investigate a number of possible states at the same time. This enables the algorithm to leave the local optima and have a better balance between exploration and exploitation. This leads to a stronger and more efficient clustering process particularly when large and sparse sets of the documents are involved.

The better BCO goes through the evolutionary process of chemotaxis, reproduction and elimination- dispersal. Bacteria are attracted to promising cluster centroids, more stable solutions are replicated and weaker ones are substituted by new ones. The fitness is aimed at the minimum distance between the intra cluster and the maximum distance between the inter cluster so that documents in the same cluster are as close as possible whereas those in different clusters are different. Lastly, the best positions that the enhanced BCO obtained are the best cluster centres.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The effectiveness of the suggested clustering algorithm or optimization technique in comparison to current approaches is shown by experimental results. They demonstrating the method's effectiveness on actual datasets of varying complexity. The following subsections are discussed about the preprocessing methods, results analysis and discussions.

IV.1 Preprocessing

Preprocessing is the foundation of document clustering, ensuring that the textual data is clean, meaningful, and computationally efficient for clustering algorithms. *Tokenization* is the first step in document preprocessing where the raw text is broken down into smaller units that is called tokens and typically it would be words (or subwords). It is what transforms the unstructured textual information into a pattern of meaningful elements that can be calculated in a manner of its computerization. With the assistance of tokenization, stop-word and stemming processing may be carried out because it offers the foundation of the subsequent steps of processing to remove features in an effective manner to be grouped. Stop-word removal is used to eliminate commonly used words that do not carry much meaning such as the, is, and, in, and others. Such words are universal in papers but they do not facilitate the distinction of content meaning hence, their removal aids in dimensionality reduction and also complexity of computation without information loss.

Stemming and lemmatizing are techniques of reducing words to their roots. Stemming relies on computational rule-based truncation which is costly but it is able to produce non-dictionary forms. Rather, Lemmatization refers to the use of linguistic knowledge and vocabulary to restore the canonical version. These techniques help to reduce inflectional variation of words and the semantically related terms are projected into the feature space in a more stable manner that improves the results of clustering. *Lowercasing standardizes* all text by converting words into lowercase form. This avoids treating terms like *"Computer"* and *"computer"* as different entities, thereby reducing redundancy in the vocabulary. Noise removal involves eliminating punctuation, numbers, special characters, URLs, or other irrelevant symbols that do not contribute to semantic meaning. This step ensures that the processed text is clean, consistent, and suitable for feature extraction methods such as TF-IDF, Word2Vec, or GloVe.

IV.2 Feature extraction

- **Word2Vec:** One of the most popular embedding-based techniques for learning continuous vector representations of words from massive text corpora is Word2Vec, which was first presented by Mikolov et al. The Continuous Bag of Words (CBOW), which predicts a word based on its surrounding context, and the Skip-gram model, which predicts surrounding words from a target word, are its two primary architectures. Word2Vec embeddings are frequently combined to create document vectors for document clustering.

- **Glove:** Global Vectors for Word Representation (GloVe), was created by Stanford researchers. GloVe learns word embeddings from global co-occurrence statistics of words across the entire corpus, whereas Word2Vec depends on word prediction based on local context windows. This means that it captures both local and global relationships, embedding words into a space where their geometric distances reflect semantic similarity. GloVe embeddings are usually aggregated using weighted, pooling, or averaging techniques to create dense vector representations of documents for document clustering. To put related documents together, these vectors are subsequently fed into clustering algorithms.

- **FastText :** By adding subword information to embeddings, Facebook AI Research's FastText model expands on the Word2Vec model. Instead of learning vectors only for entire words, FastText represents each word as a bag of character n-grams and learns embeddings for these subword units. Because of this, it works especially well with morphologically complex languages, uncommon words, and terms that are not in the dictionary. Similar to Word2Vec and GloVe, FastText embeddings

Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

are aggregated for document clustering, frequently by pooling or averaging, to produce document-level vectors. When clustering real-world text corpora, which frequently contain misspellings or domain-specific terms, the incorporation of subword information enables FastText to generate meaningful embeddings even for words not seen during training. Because of this, FastText typically outperforms Word2Vec and GloVe in clustering scenarios that involve noisy data or languages with intricate word structures.

IV.3 Datasets details

The investigations are based on six distinct standard datasets [24] Table 1 will be addressed in detail below.

- ***Reuters- 21578:*** This is a compilation of news articles published by Reuters Newswire in the year 1987. The collection consists of 21,578 documents, distributed unevenly across 135 distinct theme groupings [25].
- ***WebKB:*** The webpages contained inside the WebKB database were collected by the World-Wide Knowledge Base (WebKb) project, which is affiliated with the CMU text learning group. These webpages were obtained from the 4 Universities Data Set Homepage. They were subsequently subjected to manual classification, resulting in their categorization into seven distinct classes: student, faculty, staff, department, course, project, and other [25].
- ***Classic4:*** The classic datasets consist of 7,095 records, which are categorized into four distinct groups: CACM, CRAN, CISI, and MED. For our experimental inquiry, a random selection of 2,000 papers was made, with 500 documents chosen from each group[26].
- ***BBC:*** The dataset known as BBC3 comprises a collection of 2225 textual documents sourced from the BBC news website. These documents have been categorized into five distinct domains, including business, entertainment, politics, sport, and technology [27, 28].
- ***SMS:*** The SMS Spam Collection refers to a curated collection of SMS messages that have been specifically tagged to research SMS spam. The dataset comprises a collection of 5,574 SMS messages written in English, which have been appropriately labeled as either ham or spam [25, 29].
- ***DMOZ:*** The DMOZ directory encompasses a dataset that comprises classifications of URLs. The dataset comprises over 100,000 URL descriptions. In this study, we examined four distinct categories of documents [30].

IV.4 Performance metrics

For text document clustering, measurements of cluster validity are crucial to assessing the caliber and efficacy of the clustering outcomes. They allow evaluating whether the clusters that are created are relevant, well-organized, and helpful for the intended usage. The quality of clustering is measured using the extrinsic and intrinsic approaches. Intrinsic methods are employed in the absence of the ground truth (cluster label) for the dataset, whereas extrinsic methods are utilized in the latter case. The performance of the suggested clustering algorithms is assessed using four common extrinsic assessment metrics: accuracy, precision, recall, and F-Score.

- Accuracy: The fraction of documents correctly classified into the ground truth cluster label is known as the clustering accuracy

$$Accuracy = \frac{n_{correct}}{n} \quad (11)$$

where $n_{correct}$ is the number of documents that, according to the actual ground truth of the cluster label, are appropriately placed in the related cluster.

- Precision is the proportion of documents that are accurately assigned to the ground truth cluster label.

Table 1: Dataset's Description

S. No	Datasets	Categories	#Documents	Features
		Acq	1596	
		Crude	253	
		Earn	2840	
		Grain	41	
	Reuters	interest	190	7118
		money-fx	206	
		ship	108	
		trade	251	
		Total	**5485**	
		Project	504	
		Course	930	
	WebKB	Faculty	1124	7178
		Student	1641	
		Total	**4199**	
		CACM	500	
		CRAN	500	
	Classic4	CISI	500	8830
		MED	500	

Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

$$Precision = \frac{1}{K}\sum_{k=1}^{K} \frac{nk_{correct}}{|C_k|} \quad (12)$$

where $|C_k|$ is the total number of documents placed in cluster C_k, and $nk_{correct}$ is the number of documents correctly placed in cluster C_k.

- The recall is the ratio of the number of documents that are truly available in the ground truth cluster label C_k to the number of documents that are accurately placed in cluster C_k.

$$Recall = \frac{nk_{correct}}{|GTC_k|} \quad (13)$$

Where the number of documents that are truly present in the ground truth cluster label C_k is denoted by $|GTC_k|$.

- The weighted harmonic mean of the precision and recall is known as the F-Score as follows,

$$F - Measure = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \quad (14)$$

IV.5 Results analysis and discussions

The comparative evaluation across six benchmark datasets—Reuters, WebKB, Classic4, BBC, SMS, and DMOZ—clearly demonstrates the superiority of the proposed Improved Bacterial Colony Optimization (IBCO) approach over existing methods. In every dataset, IBCO consistently achieved higher values of accuracy, precision, recall, and F-measure compared to traditional optimization and clustering techniques such as BCO, GQPSO, IPSO, PSO, GA, and K-Means. For instance, in the Reuters dataset (Table 3 and Figures 1–2), IBCO attained a maximum accuracy of 97.95% using FastText, which is about 1.26% higher than BCO and nearly 9.1% higher than GA. This improvement is particularly noteworthy given the complexity of Reuters, where documents often share overlapping vocabulary across multiple thematic categories.

Similarly, in the WebKB dataset (Table 2 and Figures 3–4), IBCO achieved 97.71% accuracy, representing an improvement of 2.21% over BCO and a significant margin of over 12% compared to GA. This highlights the ability of IBCO to effectively cluster sparse and distributed web documents. The Classic4 dataset (Table 4 and Figures 5–6) further reinforces this trend, where IBCO recorded 84.55% accuracy, outperforming BCO by 3.24% and surpassing K-Means by an impressive 22.76%. These findings demonstrate IBCO's effectiveness in clustering heterogeneous academic texts with varying term distributions.

On the BBC dataset (Table 5 and Figures 7–8), IBCO delivered 96.17% accuracy, which is 1.96% higher than BCO and 11.06% higher than GA. Importantly, it also maintained the highest recall value of 98.87%, ensuring minimal misclassification of relevant documents. In the SMS dataset (Table 6 and Figures 9–10), where short and noisy text presents unique challenges, IBCO achieved 95.04% accuracy, improving upon BCO by 1.87% and GA by 8.17%, demonstrating adaptability for short-text clustering tasks. Finally, in the DMOZ dataset (Table 7 and Figures 11–12), IBCO achieved 92.27% accuracy, which is 2.64% higher than BCO and 5.81% higher than GA, showcasing robustness in handling hierarchical and multi-level document structures.

The implications of these findings are significant. First, the consistent performance across all datasets confirms that metaheuristic optimization techniques, when enhanced through IBCO, can substantially improve clustering performance in diverse domains. The improvements across datasets demonstrate the algorithm's generalizability and practicality for real-world applications such as digital libraries, academic indexing, spam detection, and news categorization. Moreover, the results indicate that FastText-based embeddings consistently yielded the highest performance with IBCO, proving the importance of subword-level semantic representation in document clustering.

In contrast, TF-IDF showed competitive performance in structured datasets like Reuters but was less effective for shorter or hierarchical texts. Another important finding is observed in the convergence analyses (Figures 2, 4, 6, 8, 10, and 12), which show that IBCO not only enhances accuracy but also achieves faster and more stable convergence compared to other algorithms. This characteristic makes IBCO highly suitable for large-scale and dynamic text mining tasks, where both efficiency and accuracy are critical. Overall, the percentage improvements ranging from 2–4% over BCO and 8–12% over GA and K-Means across different datasets confirm that IBCO is not merely an incremental enhancement but a substantial advancement in optimization-driven document clustering. These improvements, as highlighted in Tables 2–7, not only demonstrate strong experimental validity but also carry theoretical and practical significance for future developments in document clustering and large-scale text analytics.

V. Conclusions

The results of the experiment which were conducted on six benchmark datasets viz.- Reuters, WebKB, Classic4, BBC, SMS and DMOZ are unquestionable and affirm the effectiveness of the proposed Improved IBCO method of document clustering. As IBCO was integrated with the other models of document representation, such as Word2Vec, TF-IDF and FastText, the methodology proved to be superior to the classical clustering and optimization models as they were BCO, GQPSO, IPSO, PSO, GA and K-Means. The results showed improvements in accuracy of 2 to 4 percent on average over BCO and 8 to 12 percent on average over GA and K-Means with FastText-based embeddings being the most prone to yield the best results. In addition to the superiority of clustering, the convergence analysis also showed that IBCO is steadier and converges more rapidly, thereby, making it a highly appropriate method of text mining on a large and complex scale of text mining.

.

Int. J. Advanced Networking and Applications
Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

Table 3: Comparative results analysis for the Reuters dataset

Methods	Accuracy			Precision			Recall			F-Measure		
	Word2Vec	TF-IDF	FastText	Word2Vec	TF-IDF	FastText	Word2Vec	TF-IDF	FastText	Word2Vec	TF-IDF	FastText
IBCO	96.89	97.27	97.95	96.44	96.93	97.48	97.31	97.60	98.41	96.87	97.26	97.94
BCO	95.79	96	96.69	92.28	92.51	93.44	99.24	99.45	99.93	95.63	95.85	96.57
GQPSO	93.75	94.16	95.34	90.08	90.71	91.76	97.22	97.44	98.84	93.51	93.95	95.17
IPSO	93.21	94.05	95.27	90.93	91.49	92.28	95.27	96.43	98.15	93.05	93.90	95.13
PSO	88.86	89.62	93.11	86.61	87.47	90.79	90.70	91.40	95.21	88.61	89.39	92.95
GA	86.74	88.15	91.09	84.21	85.77	88.91	88.70	90.05	92.96	86.40	87.86	90.89
K-Means	83.57	85.03	87.21	81.87	82.79	84.62	84.75	86.67	89.25	83.28	84.68	86.87

Table 2: Comparative results analysis for the WebKB dataset

Methods	Accuracy			Precision			Recall			F-Measure		
	Word2Vec	TF-IDF	FastText	Word2Vec	TF-IDF	FastText	Word2Vec	TF-IDF	FastText	Word2Vec	TF-IDF	FastText
IBCO	96.64	96.96	**97.71**	96.28	96.93	97.43	96.97	96.99	97.98	96.62	96.96	97.71
BCO	94.43	95.11	**95.60**	90.23	90.94	91.28	98.51	99.21	99.92	94.19	94.89	95.40
GQPSO	91.46	92.32	**94.02**	87.93	88.41	89.76	94.61	95.91	98.13	91.15	92	93.76
IPSO	88.69	90.19	**91.05**	85.11	86.61	87.49	91.66	93.28	94.20	88.27	89.82	90.72
PSO	84.34	87.1	**90.19**	79.89	82.41	86.61	87.69	90.93	93.28	83.61	86.46	89.82
GA	81.17	83.05	**85.01**	77.93	79.31	80.73	83.33	85.71	88.28	80.54	82.39	84.34
K-Means	78.97	80.02	**81.22**	76.93	77.60	78.49	80.19	81.55	83.03	78.53	79.52	80.70

Table 3: Comparative results analysis for the Classic4 dataset

Methods	Accuracy			Precision			Recall			F-Measure		
	Word2Vec	TF-IDF	FastText	Word2Vec	TF-IDF	FastText	Word2Vec	TF-IDF	FastText	Word2Vec	TF-IDF	FastText
IBCO	82.40	83.38	**84.55**	81.24	82.29	83.49	83.17	84.12	85.29	82.19	83.19	84.38
BCO	79.36	80.1	**81.31**	76.94	77.86	79.34	80.86	81.51	82.59	78.85	79.64	80.93
GQPSO	75.49	76.02	**77.02**	72.87	73.43	74.12	76.90	77.45	78.69	74.83	75.38	76.33
IPSO	73.71	74.23	**76.20**	71.28	71.98	73.43	74.91	75.38	77.74	73.05	73.64	75.52
PSO	70.03	72.21	**74.22**	69.62	70.98	72.14	70.20	72.76	75.27	69.91	71.86	73.67
GA	65.10	66.04	**69.64**	64.26	65.29	69.29	65.35	66.28	69.78	64.80	65.78	69.53
K-Means	57.11	59.81	**61.79**	54.94	57.93	59.81	57.43	60.19	62.28	56.15	59.04	61.02

Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

Table 4: Comparative results analysis for the BBC dataset

Methods	Accuracy			Precision			Recall			F-Measure		
	Word2Vec	TF- IDF	FastText	Word2Vec	TF- IDF	FastText	Word2Vec	TF- IDF	FastText	Word2Vec	TF- IDF	FastText
IBCO	94.53	95.28	**96.17**	91.78	92.74	93.41	97.12	97.71	98.87	94.38	95.16	96.06
BCO	92.75	93.03	**94.21**	90.27	90.78	91.48	94.99	95.05	96.75	92.57	92.87	94.04
GQPSO	89.18	90.38	**91.09**	82.12	85.78	86.24	95.62	94.47	95.50	88.36	89.91	90.63
IPSO	87.11	88.10	**89.18**	79.81	81.23	82.12	93.46	94.18	95.62	86.10	87.22	88.36
PSO	84.62	85.28	**87.26**	82.62	80.63	82.78	86.06	88.90	90.93	84.30	84.56	86.66
GA	81.7	82.85	**85.11**	79.46	81.24	83.43	83.18	83.94	86.33	81.28	82.57	84.85
K-Means	81.3	82.28	**84.11**	79.16	80.61	82.29	82.69	83.39	85.40	80.89	81.98	83.81

Table 5: Comparative results analysis for the SMS dataset

Methods	Accuracy			Precision			Recall			F-Measure		
	Word2Vec	TF- IDF	FastText	Word2Vec	TF- IDF	FastText	Word2Vec	TF- IDF	FastText	Word2Vec	TF- IDF	FastText
IBCO	93.57	94.29	**95.04**	90.93	91.77	92.43	96.01	96.65	97.53	93.40	94.14	94.91
BCO	91.38	92.14	**93.17**	89.48	89.81	90.62	93.01	94.20	95.50	91.21	91.96	92.99
GQPSO	88.78	89.25	**90.10**	85.62	86.49	86.94	91.40	91.54	92.81	88.41	88.94	89.77
IPSO	88.78	90.09	**92.97**	86.76	88.41	89.26	90.42	91.48	96.42	88.55	89.92	92.70
PSO	86.63	88.12	**90.53**	85.29	86.77	88.26	87.64	89.18	92.47	86.45	87.96	90.31
GA	85.59	85.99	**86.87**	83.27	83.52	83.94	87.32	87.86	89.16	85.24	85.64	86.47
K-Means	81.22	83.04	**85.19**	79.96	81.29	82.91	82.03	84.24	86.88	80.98	82.74	84.84

Table 6: Comparative results analysis for the DMOZ dataset

Methods	Accuracy			Precision			Recall			F-Measure		
	Word2Vec	TF- IDF	FastText	Word2Vec	TF- IDF	FastText	Word2Vec	TF- IDF	FastText	Word2Vec	TF- IDF	FastText
IBCO	90.63	91.41	**92.27**	89.48	89.99	90.78	91.59	92.63	93.56	90.52	91.29	92.15
BCO	87.11	88.27	**89.63**	85.77	86.77	87.99	88.14	89.46	90.98	86.94	88.09	89.46
GQPSO	86.60	87.68	**88.11**	84.93	85.44	86.77	87.87	89.45	89.17	86.37	87.40	87.95
IPSO	85.9	86.70	**87.18**	83.79	84.46	84.93	87.48	88.42	88.94	85.59	86.39	86.89
PSO	85.71	86.14	**86.71**	83.42	83.79	84.49	87.42	87.92	88.42	85.37	85.80	86.41
GA	84.64	85.22	**86.46**	82.11	82.94	83.92	86.49	86.91	88.40	84.24	84.87	86.10
K-Means	84.18	85.06	**86.13**	81.70	82.69	83.76	85.96	86.81	87.92	83.78	84.70	85.79

Int. J. Advanced Networking and Applications
Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

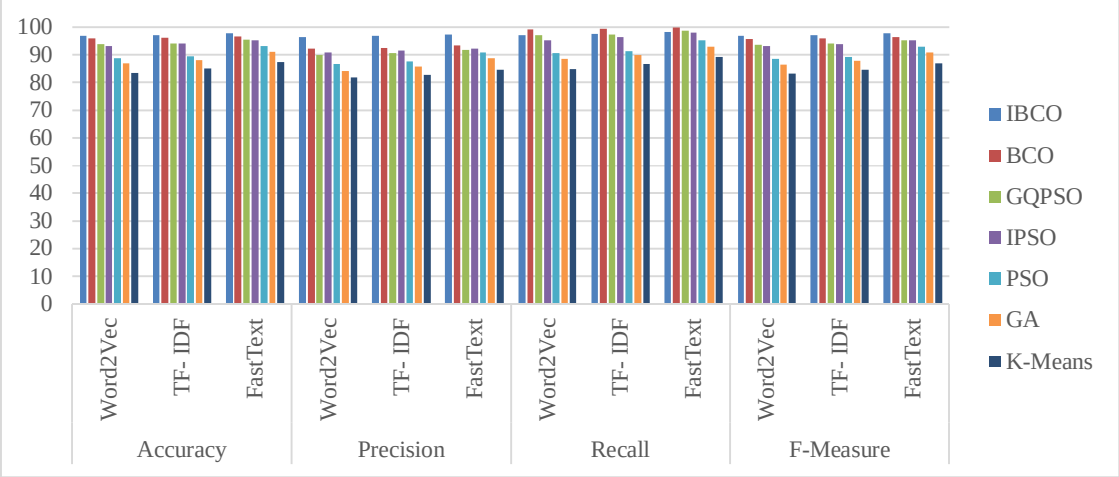

Figure 1: Comparative results analysis for the Reuters dataset

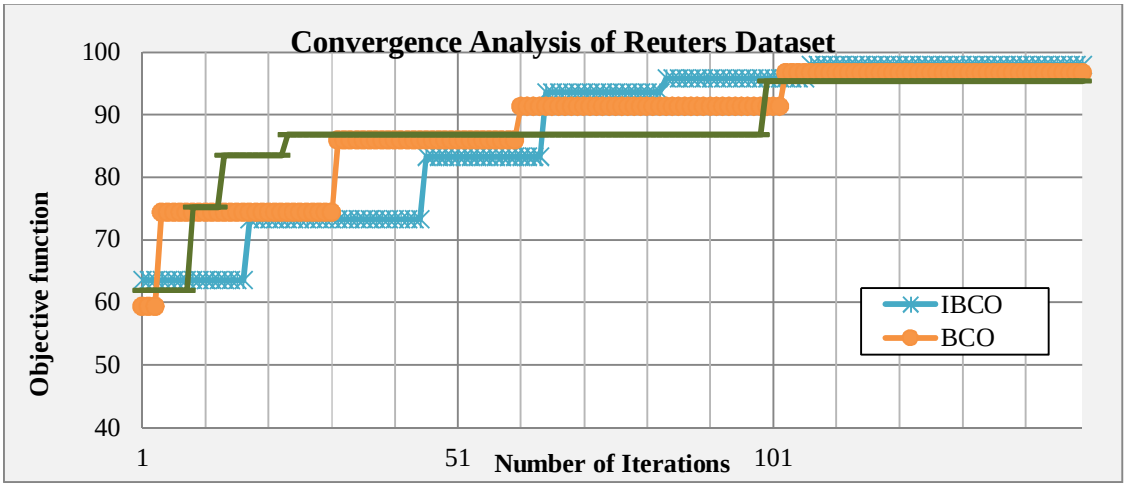

Figure 2: Convergence Analysis of Reuters Dataset

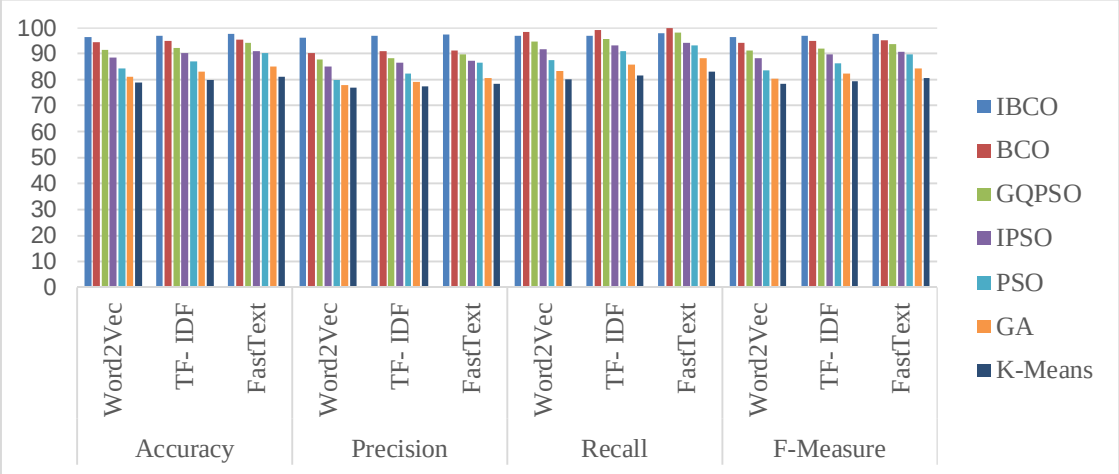

Figure 3: Comparative results analysis for the WebKB dataset

Int. J. Advanced Networking and Applications
Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

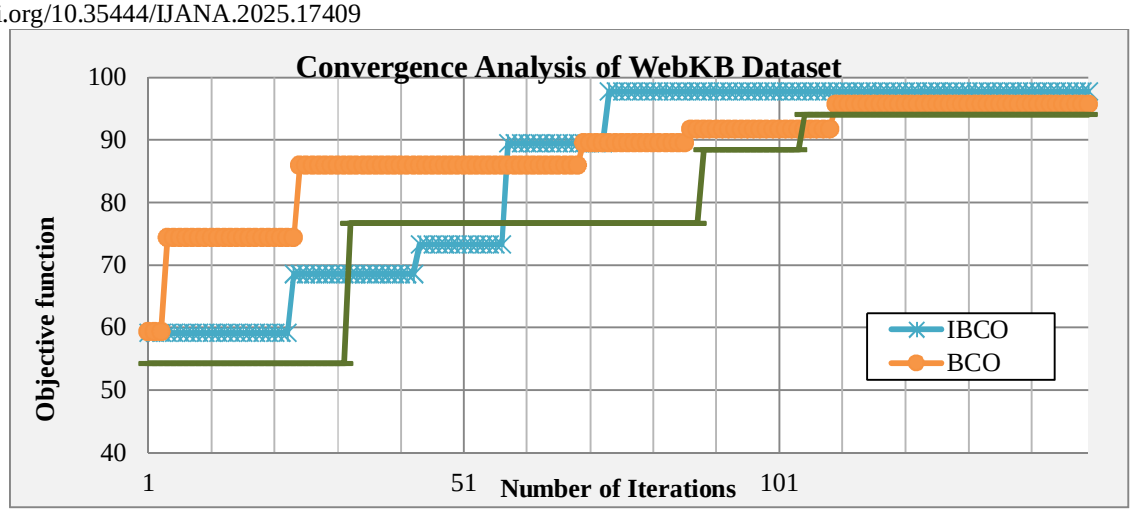

Figure 4 : Convergence Analysis of WebKB Dataset

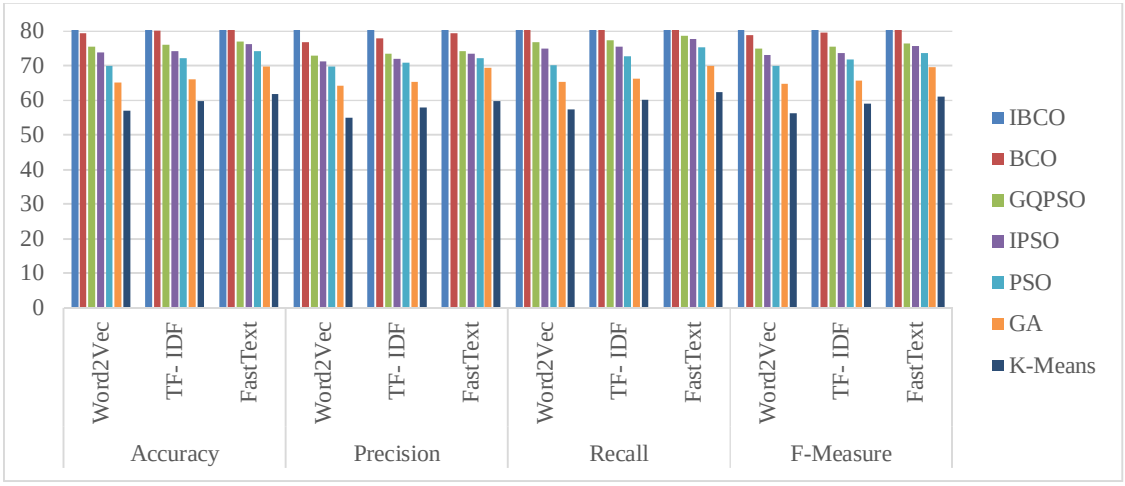

Figure 5: Comparative results analysis for the Classic4 dataset

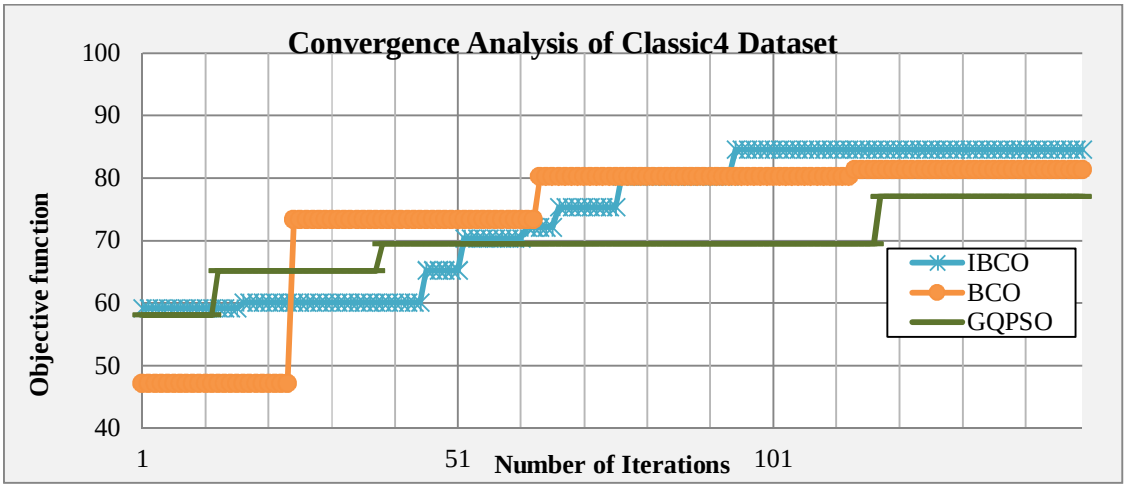

Figure 6 : Convergence Analysis of Classic4 Dataset

Int. J. Advanced Networking and Applications
Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

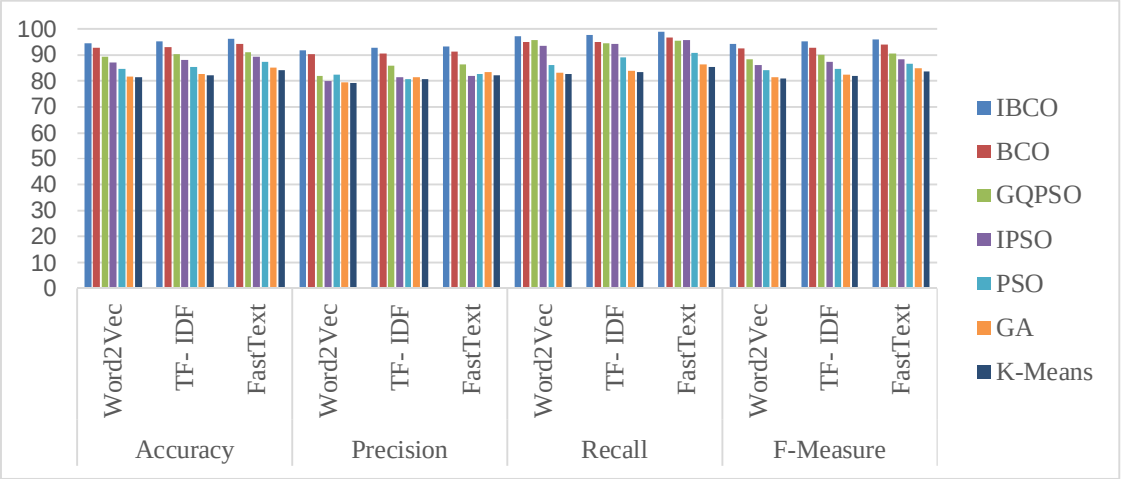

Figure 7 : Comparative results analysis for the BBC dataset

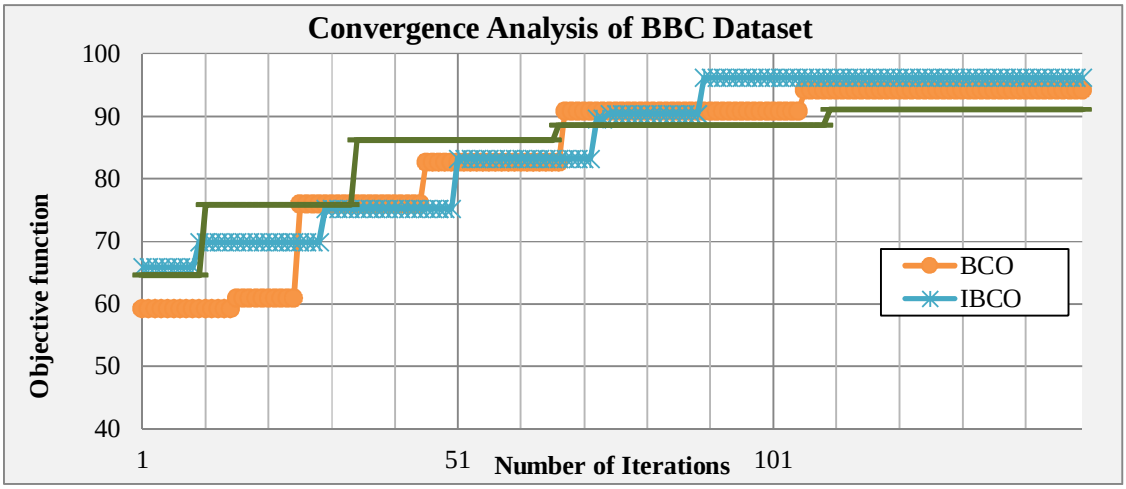

Figure 8 : Convergence Analysis of BBC Dataset

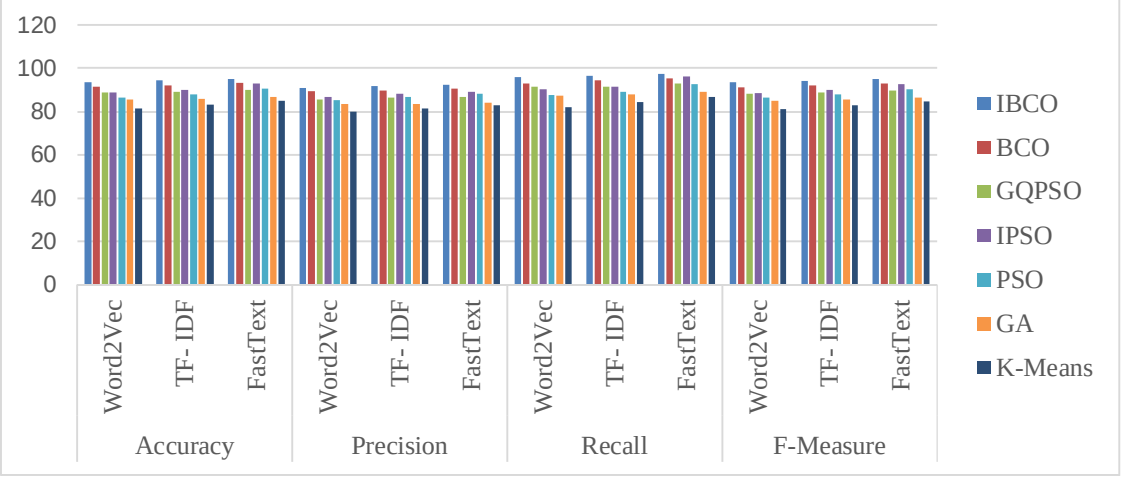

Figure 9 : Comparative results analysis for the SMS dataset

Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

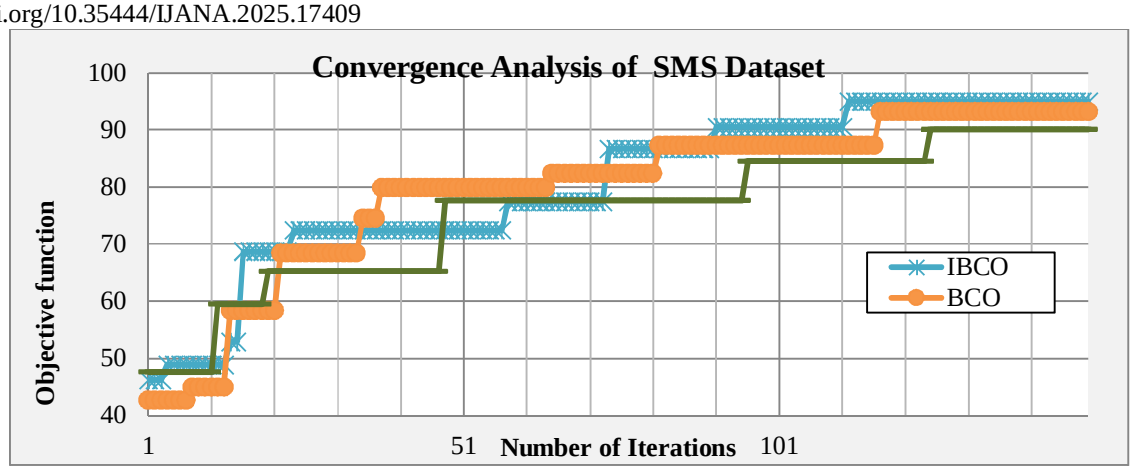

Figure 10 : Convergence Analysis of SMS Dataset

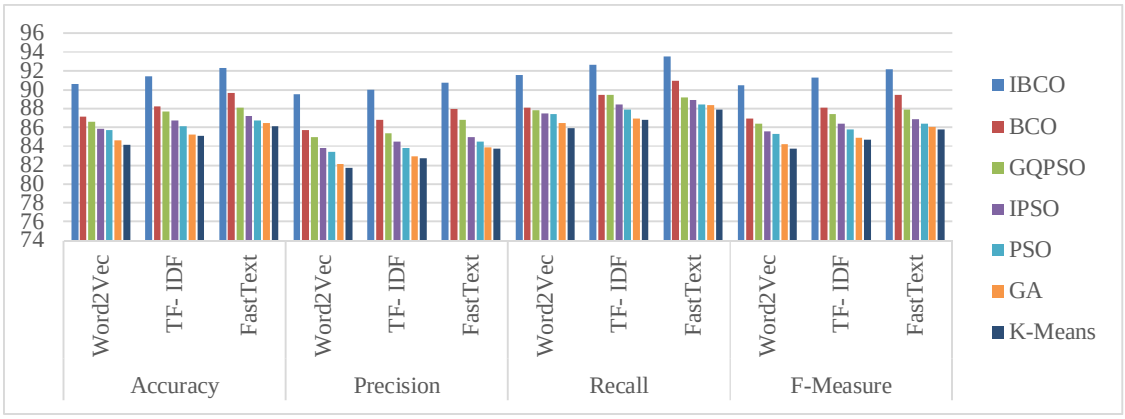

Figure 11 : Comparative results analysis for the DMOZ dataset

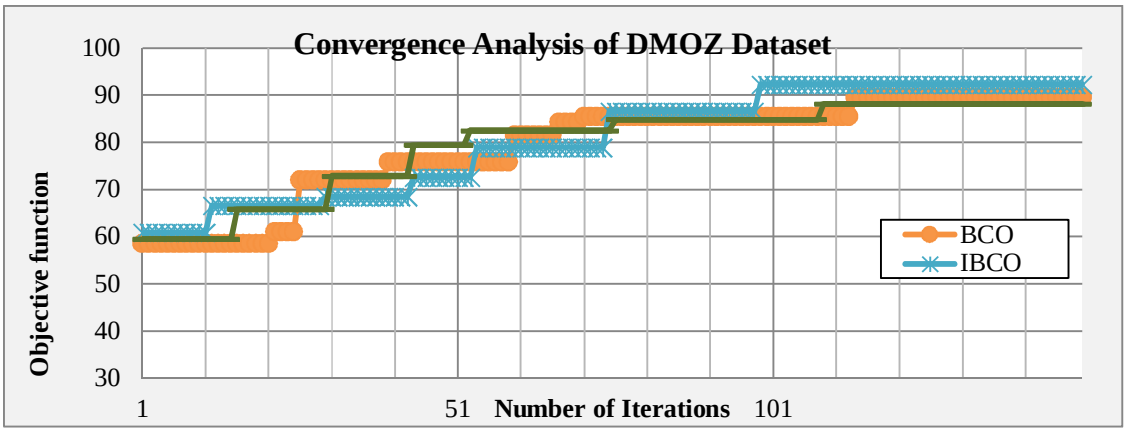

Figure 12 : Convergence Analysis of DMOZ Dataset

Generally, the article highlights the theoretical significance of hybrid metaheuristic optimization and its feasibility in the digital library, spam filtering, academic text classification as well as web document analysis. Though the suggested IBCO approach showed better results in all the datasets, there are limitations to it. This was only tested on benchmark datasets, which might not be a complete representation of real-time or domain-specific applications like social media or text streams in multiple languages. The algorithm also relies on the quality of pre-trained embeddings, and the optimization procedure, which, although computationally inexpensive compared to other models, can remain computationally expensive with limited sizes of data. In the future, more sophisticated embeddings like BERT or GPT-related models can be

Volume: 17 Issue: 04 Pages: 7062-7075 (2026) ISSN: 0975-0290
https://doi.org/10.35444/IJANA.2025.17409

incorporated to achieve more sophisticated semantic associations. The approach may also be generalised to multi-objective optimization to trade off accuracy and efficiency, and scaled to distributed or parallel computing to process large-scale text data. Besides, cross-lingual and multilingual clustering and combining IBCO with deep learning frameworks might be explored to enhance robustness and applicability in a variety of real-life situations.

Reference

[1] J. M. Sanchez-Gomez, S. Santander-Jiménez, and M. A. Vega-Rodríguez, "Multi-objective swarm-intelligence algorithm for document clustering," *Expert Systems with Applications*, vol. 289, p. 128348, 2025.

[2] S. A. Mousumi, A. Asaduzzaman, M. M. Islam, and M. R. Hoque, "Context-Aware Tourist Guide: A Rule based Approach," *International Journal of Advanced Networking and Applications*, vol. 16, no. 1, pp. 6270-6274, 2024.

[3] J. Rautaray, S. Panigrahi, A. K. Nayak, P. Sahu, and K. Mishra, "Deep learning based feature extraction technique for single document summarization using hybrid optimization technique," *IEEE Access*, 2025.

[4] A. Haque, "Hate speech detection in social media using the ensemble learning technique," *Eswar Publications*, 2023.

[5] R. Priyadarshi and R. R. Kumar, "Evolution of swarm intelligence: a systematic review of particle swarm and ant colony optimization approaches in modern research," *Archives of Computational Methods in Engineering*, pp. 1-42, 2025.

[6] B. Zeng, X. Shang, R. Lu, and Y. Zhang, "Particle swarm optimization-based NLP methods for optimizing automatic document classification and retrieval," *PloS one*, vol. 20, no. 7, p. e0325851, 2025.

[7] R. Dodda and A. S. 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.

[8] W. A. ABUAIN, "IMPROVED SALP SWARM ALGORITHM FOR TEXT DOCUMENT CLUSTERING," *Journal of Theoretical and Applied Information Technology*, vol. 102, no. 14, 2024.

[9] E. Hassan *et al.*, "A Hybrid K-Means++ and Particle Swarm Optimization Approach for Enhanced Document Clustering," *IEEE Access*, 2025.

[10] M. A. Al-Betar, A. K. Abasi, G. Al-Naymat, K. Arshad, and S. N. Makhadmeh, "Bare-bones based salp swarm algorithm for text document clustering," *IEEE Access*, vol. 11, pp. 100010-100028, 2023.

[11] R. Dodda and A. S. Babu, "Text document clustering using chaotic northern goshawk optimization with K-means algorithm," *Int. J. Intell. Eng. Syst*, vol. 17, no. 3, pp. 630-642, 2024.

[12] R. Kumar and S. Thirumaran, "Developed Hybrid Self-Organizing Map, Weighed Quantum Particle Swarm Optimization with Improved K-Means Clustering for Document Clustering," in *2024 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2024: IEEE, pp. 1-6.

[13] P. Pitchandi and M. Balakrishnan, "Document clustering analysis with aid of adaptive Jaro Winkler with Jellyfish search clustering algorithm," *Advances in Engineering Software*, vol. 175, p. 103322, 2023.

[14] S. Larabi-Marie-Sainte, M. Bin Alamir, and A. Alameer, "Arabic Text Clustering Using Self-Organizing Maps and Grey Wolf Optimization," *Applied Sciences*, vol. 13, no. 18, p. 10168, 2023.

[15] M. Nachaoui, I. Lakouam, and I. Hafidi, "Hybrid particle swarm optimization algorithm for text feature selection problems," *Neural Computing and Applications*, vol. 36, no. 13, pp. 7471-7489, 2024.

[16] A. K. Abdulsahib, M. Balafar, and A. Baradarani, "DGBPSO-DBSCAN: An Optimized Clustering Technique based on Supervised/Unsupervised Text Representation," *IEEE Access*, 2024.

[17] M. Azarshab, M. Fathian, and B. Amiri, "An enhanced Teaching-Learning-Based Optimization (TLBO) with Grey Wolf Optimizer (GWO) for text feature selection and clustering," *arXiv preprint arXiv:2402.11839*, 2024.

[18] D. Debnath, R. Das, and P. Pakray, "Single document text summarization addressed with a cat swarm optimization approach," *Applied Intelligence*, vol. 53, no. 10, pp. 12268-12287, 2023.

[19] S. S. Aote, A. Pimpalshende, A. Potnurwar, and S. Lohi, "Binary Particle Swarm Optimization with an improved genetic algorithm to solve multi-document text summarization problem of Hindi documents," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105575, 2023.

[20] R. Das, D. Debnath, P. Pakray, and N. C. Kumar, "A binary grey wolf optimizer to solve the scientific document summarization problem," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23737-23759, 2024.

[21] M. A. Sheikh, M. Bashir, and M. K. Sudddle, "Advancing automatic text summarization: Unleashing enhanced binary multi-objective grey wolf optimization with mutation," *Plos one*, vol. 19, no. 5, p. e0304057, 2024.

[22] B. Niu and H. Wang, "Bacterial colony optimization: principles and foundations," in *Emerging Intelligent Computing Technology and Applications: 8th International Conference, ICIC 2012, Huangshan, China, July 25-29, 2012. Proceedings 8*, 2012: Springer, pp. 501-506.

[23] M. Bey, P. Kuila, B. B. Naik, and S. Ghosh, "Quantum-inspired particle swarm optimization for efficient IoT service placement in edge computing systems," *Expert Systems with Applications*, vol. 236, p. 121270, 2024.

[24] K. K. Bharti and P. K. Singh, "Chaotic gradient artificial bee colony for text clustering," *Soft Computing*, vol. 20, pp. 1113-1126, 2016.

[25] A. C. Cachopo. "Datasets for single-label text categorization." https://ana.cachopo.org/datasets-for-single-label-text-categorization (accessed.

[26] "Classic3 and Classic4 Datasets." ftp://ftp.cs.cornell.edu/pub/smart/ (accessed.

[27] B. Bose. "BBC News Classification." https://storage.googleapis.com/dataset-uploader/bbc/bbc-text.csv. (accessed.

[28] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 377-384.

[29] J. H. Tiago Almeida. "SMS Spam Collection." https://archive.ics.uci.edu/dataset/228/sms+spam+collection (accessed.

[30] A. Shawon. "URL Classification Dataset [DMOZ]." https://www.kaggle.com/datasets/shawon10/url-classification-dataset-dmoz