International Journal of Computing Sciences Research (ISSN print: 2546-0552; ISSN online: 2546-115X)
Vol. 9, pp. 4000-4023
doi: 10.25147/ijcsr.2017.001.1.257
https://stepacademic.net

Short Paper

Context-Aware PDF Clustering Using Social Spider Algorithm and Agglomerative Hierarchical Techniques

Wellanie M. Molino
Computer Studies Department, Technological University of the Philippines
wellanie_molino@tup.edu.ph
(corresponding author)

Francis A. Alfaro
Computer Studies Department, Technological University of the Philippines
francis_alfaro@tup.edu.ph

Jonathan M. Caballero
Computer Studies Department, Technological University of the Philippines
jonathan_caballero@tup.edu.ph

Francis L. Dela Cruz
Computer Studies Department, Technological University of the Philippines
Francis_delacruz@tup.edu.ph

Jan Eilbert L. Lee
Computer Studies Department, Technological University of the Philippines
janeilbert_lee@tup.edu.ph

Jasmin D. Niguidula
Computer Studies Department, Technological University of the Philippines
jasmin_niguidula@tup.edu.ph

Fernando L. Renegado
Computer Studies Department, Technological University of the Philippines
fernando_renegado@tup.edu.ph

Ariel L. Tomagan
Computer Studies Department, Technological University of the Philippines
ariel_tomagan@tup.edu.ph

Darwin C. Vargas
Computer Studies Department, Technological University of the Philippines
darwin_vargas@tup.edu.ph

Date received: June 25, 2025
Date received in revised form: November 30, 2025
Date accepted: December 3, 2025

Recommended citation:

Molino, W. M., Alfaro, F. A., Caballero, J.M., Dela Cruz, F.L., Lee, JE.L., Niguidula, J.D., Renegado, F.L., Tomagan, A.C., & Vargas, D. C. (2025). Optimized Context-Aware PDF Clustering Using Social Spider Algorithm and Agglomerative Techniques. *International Journal of Computing Sciences Research*, 9, 4000-4023. https://doi.org/10.25147/ijcsr.2017.001.1.257

Abstract

Purpose – This study primarily aims to develop an automated PDF document clustering system using a hybrid approach that combines Social Spider Optimization (SSO) with Agglomerative Clustering. It seeks to improve the efficiency of unsupervised document classification while addressing limitations commonly associated with traditional techniques when processing large-scale data sets.

Method – In this study, the researchers adopted a metaheuristic optimization based on SSO to refine the initialization and parameter optimization process for Agglomerative Clustering. Fifteen unclassified PDF documents were preprocessed using text extraction and TF-IDF vectorization. The hybrid model's performance was evaluated using Silhouette Coefficient, Dunn Index, and Cluster Purity through cluster distribution diagrams and Principal Component Analysis (PCA) aided by visualization analysis.

Results – Using the hybrid SSO–Agglomerative Clustering model, fifteen unclassified PDF documents were efficiently partitioned into thematically coherent clusters. The system's performance was evaluated using internal cluster validity metrics to evaluate the performance of the model compared to the Agglomerative method. A notable improvement is observed in Silhouette Coefficient with values from 0.63 to 0.82. The Dunn Index likewise demonstrated a substantial gain from 0.49 to 0.68. The increase in Cluster Purity from 84.6% to 94.3% indicates a more accurate clustering of documents. Although, computation time increased from 9.4 to 11.9 seconds, the performance were meaningfully consistent and visually distinct, representing each document category.

Conclusion – Applying SSO with Agglomerative Clustering demonstrate a strong potential for automating PDF categorization by combining metaheuristic optimization with hierarchical clustering to handle unstructured textual information.

Recommendations - Future research should investigate the model's scalability and be validated using bigger datasets.

Research Implications – This study provides a valuable contribution to data mining and machine learning through a novel framework for document clustering. This technique supports practical benefits in information retrieval and aiding decision process, particularly for organizations managing large document repositories.

Keywords – social spider optimization, text clustering, automated PDF clustering, agglomerative clustering, document clustering

INTRODUCTION

In the fields of machine learning and data modelling, document clustering has become a key research area because of its crucial role in handling large volumes of unstructured textual information. This approach involves grouping of related contents for efficient indexing, systematic storage, and information retrieval. According to Jain et al.(1999), clustering technique is essential for identifying underlying datasets enabling data-driven insights without the need for labeling inputs. In academic repositories and enterprise information systems, clustering serves as a foundation for recommendation systems and topic analysis. Its success depends on identifying deeper semantic patterns rather than relying mainly on surface linguistic features. The goal is that the objects within a group share strong similarities while separating them from unrelated objects in other clusters. The higher the similarity within a cluster and the higher the difference between groups, the better or more distinct the clustering. In many contexts, the notion of a cluster is not well defined. Traditional algorithms frequently struggle to reach this level of semantic precision. Clustering algorithm such as K-Means require prior knowledge of the number of clusters and is considered sensitive to initialization and the choice of distance metrics. This often leads to inconsistent results. In addition, language ambiguity such as words with various or overlapping meanings can lead to misclassification of documents. Moreover, the challenges of noise, outliners and constantly changing text make clustering more challenging as mathematically derived groups do not always align with meaningful semantic relationships.

To address these limitations, this study presents a novel framework that integrates Social Spider Optimization with Agglomerative Clustering for PDF document clustering optimization (Yang, 2012; Dorigo & Stützle, 2004). The main objective is to combine the strengths of both metaheuristic optimization and hierarchical clustering approach to produce more contextually accurate grouping of textual information. This study strengthens existing work in the fields of data mining and unsupervised learning. By analyzing empirical evaluations on real-world datasets, the study also aims to evaluate the model's effectiveness in terms of clustering scalability, accuracy and relevance for document processing.

LITERATURE REVIEW

Document clustering represents a fundamental function in information retrieval, designed to group documents with similar thematic characteristics. Standard clustering techniques such as k-means and Agglomerative Clustering are frequently rely on fixed distance metrics, which may limit their adaptability. Such metrics can fail to capture deeper semantic connections across documents, specifically when dealing with unstructured textual data analytics (Jain, Murty, & Flynn, 1999).

To address these challenges, nature-inspired metaheuristic algorithms have been utilized to improved clustering outcomes by optimizing initialization and refining cluster centroids. Methods including Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) can enhance clustering outcomes by adjusting initial parameter settings and improving centroid locations (Kennedy & Eberhart, 1995; Chen & Ye, 2004). In recent years, SSO, a swarm intelligence algorithm has been introduced as a novel swarm metaheuristic, based on the cooperative behavior of social spiders (Yang, 2012).

Traditional Clustering Techniques

Among traditional clustering techniques are K-means and Agglomerative Clustering, which remains the most popular method because of its simplicity and computational efficiency. The k-means algorithm organizes dataset into a specified number of clusters by repeatedly assigning each document to the centroid based on Euclidean or cosine similarity. Similarly, Agglomerative Clustering uses hierarchical technique to create a tree-like structure of clusters by combining the most similar pairs of documents (Jain et al., 1999). While it produces interpretable cluster results and does not require predefining the number of clusters, it is still computationally demanding for large datasets and is highly influence by the linkage and distance metrics. When dealing with high dimensional text information, both algorithms depend on similarity distance metric which becomes a core limitation when dealing with unstructured data. These methods frequently fail to identify the latent semantic relationships among concepts that determine the true contextual meaning of a document. Two documents may discuss “treatments” and “disease,” but unless the algorithm identifies their conceptual closeness, they may consequently be allocated to different clusters. This semantic representation has encourage the adoption of data-driven clustering approach for capturing deeper contextual patterns. To address these limitations in traditional methods, researchers have studied nature-based metaheuristic algorithms that emulate biological evolution such as swarm behavior, and collective intelligence. These are developed to achieve efficient optimization and more balanced clustering results in multifaceted search spaces.

Genetic Algorithm (GA) were among the earliest first metaheuristics approach that emulates the process of natural collection by refining groups of solution over multiple

generations. GA optimizes clustering criteria such as the number of centroid position through mutation (Chen & Ye, 2004). Likewise, PSO (Kennedy & Eberhart, 1995), imitates the cooperative behavior of bird flocks with particles representing potential solution candidates moving through search space under the influence of their neighbors. Both approaches have proven to improve performance by providing optimized initial conditions calibrating distance measure.

However, these algorithms also have their limitations. GA often involves careful tuning of parameters to find optimal solution to prevent early convergence, while PSO's performance may depend heavily on the number of key parameters that control the behavior of the particles in the search space. This has led to the development of new framework by combining SSO and Agglomerative clustering algorithms with improved balance between searching and refining locally.

Social Spider Optimization (SSO) and Agglomerative Techniques

The SSO is a novel approach and one of the more recent advancements in swarm intelligence based on the cooperative characteristics of social spiders in a colony (Yang, 2012). Figure 1 illustrates the complete evolutionary process of SSO algorithm. In SSO, these agents collectively move according to the biological behavior through vibrations on their web to find potential mates. In contrary, SSO defines two distinct search agents namely: male and female, each represents an individual solution with different communication strategies by using different evolutionary operation that replicates biological role in the colony to achieve more stable convergence. Once a new spider member is generated, it is compared with the remaining population. If one of the spider agent has a better fitness than the worst spider, it is replaced, otherwise, it is discarded. (Erik Cuevas, 2015). See figure below.

Agglomerative Clustering is the most frequently algorithm used to find structures or groups to classify similar data points for discovering nested document structures were each data points initially forms an independent cluster, and the algorithm iteratively merges the most similar pairs based on a linkage parameter. It follows bottom-up hierarchical method and continues until all points' form a set of number of clusters. It employs distance metrics like Manhattan or Euclidean distance to measure similarity visualized using dendogram, which shows the hierarchy of clusters. (Jain et al., 1999).

Although it is easy to execute and effective in many cases, its performance can be sensitive to the initial parameter settings to identify structures, known as clusters. The clusters includes data vectors defined by multiple characteristic features. Data vectors in the same cluster exhibit strong similarities, although they differ significantly from data outside that cluster. The number of clusters starts with as many clusters of data vectors in agglomerative hierarchical cluster. Different agglomerative clustering methodologies are important for analyzing data that may produce varying outcomes (Aljumily, 2016).

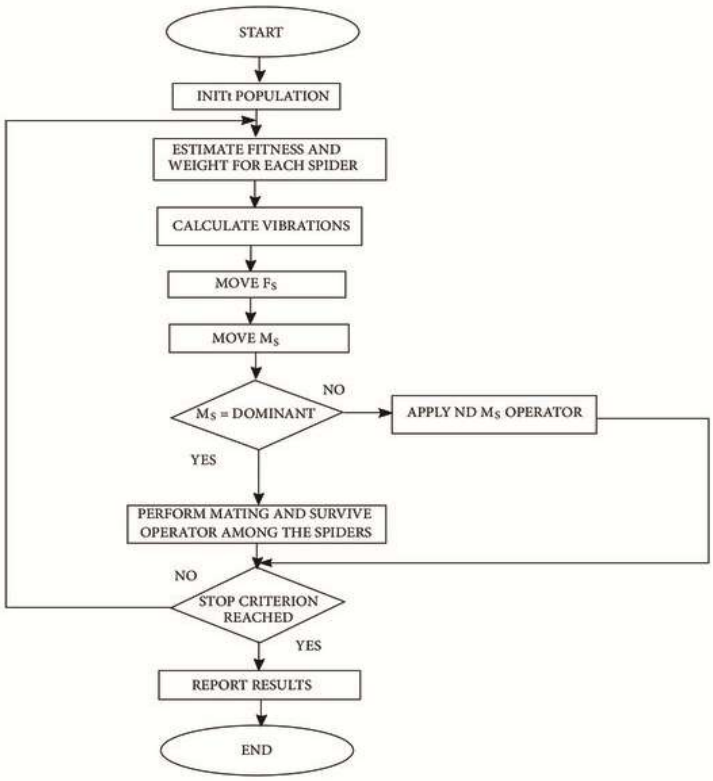

Figure 1. SSO Flowchart

The combination of SSO with Agglomerative Clustering framework, serves as a metaheuristic-driven optimizer that dynamically determines the best clustering parameters. While Agglomerative Clustering performs merging of documents based on optimized parameters, this approach allows the search capability to perform local refinement that produces the formation of coherent and semantically meaningful clusters. Li et al. (2019) note that refining optimization procedure is essential to attain clustering accuracy compared to traditional algorithms. This principle relies on intra-cluster similarity ensuring that text data within each cluster share the same underlying context.

Term Frequency (TF) and Inverse Document Frequency (IDF) Method

Information retrieval systems are fundamental in identifying significant words for extracting term patterns within documents. It is intended to enable effective searching and clustering extensive data corpora. On the other hand, robust text-mining methods are necessary for generating informative patterns in the presence of vast amount of data resources. TF-IDF resolves this by evaluating the significance of the given terms within an individual document and within a broader collection of datasets, supporting the system's ability to identify relevant abstracts. TF-IDF computes the term's relevance by multiplying two metrics: Term Frequency (TF) and Inverse Document Frequency (IDF). As stated by

Sebastiani (2002), TF-IDF remains to be a common term-weighting method in automated text categorization and machine learning applications.

TFIDF is a technique for converting text data into vectorized form suitable for computational analysis. Term frequency (TF) measures the number of occurrences of term *t* within document *d*. As frequencies increase, the value of *tf* also increases. Inverse document frequency (IDF) quantifies the document frequency within a corpus that holds the phrase *t*. TFIDF values are computed as follows:

$$tfidf\,(t,d,D) = tf\,(t,d).\ idf\,(t,D) \qquad \textit{Equation 1}$$

Cosine Similarity

Techniques for determining semantic relationship, including cosine similarity, go beyond simple token matching. The cosine of the angle between document vectors in a multi-dimensional feature space is evaluated using this technique. It facilitates the detection of semantic relationships, including reformulated text. The computational complexity when applied to lengthy documents or extensive datasets may be a problem, despite its capability to identify more textual alterations of plagiarism, such as paraphrasing.

Cosine similarity has long played a central role in embedding-based machine learning systems. Its geometric simplicity, scale invariance, and alignment with the training goals of most modern models have made it a common selection across various fields. Cosine similarity holds inherent limits (Manning et al., 2008).

$$\cos(\theta) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \hat{\mathbf{x}}^\top \hat{\mathbf{y}} \qquad \textit{Equation 2}$$

Optimization and Robustness in Modern Clustering Models

Optimization-based clustering techniques as highlighted by Alam et al, (2008) and Das, et al.,(2009), have shown strong performance in multi-objective optimization by sustaining a varied populations of solution that allows for both broad searching and focused refinement. Such variation helps avoid early convergence and strengthens the algorithms ability to handle different types of datasets. In text analysis, such adaptability is critical because documents exhibit significant variations in length., word usage and contextual representation.

METHODOLOGY

The system architecture diagram presented in Figure 2 illustrates the complete workflow of the hybrid-clustering model for automated PDF document clustering, which combines SSO with Agglomerative Clustering. It shows the process of converting raw PDF documents into meaningful clusters based on semantic content. The figure also emphasizes the sequential workflow of data acquisition, text preprocessing, feature extraction, SSO-based clustering optimization, evaluation of cluster quality, and visual representation of results.

The process begins by uploading of fifteen PDF documents, which represents the dataset of unstructured files obtained from multiple digital repositories. These unstructured set of data are necessary for processing prior to data analysis for effective evaluation of the method. Metaheuristic algorithms including SSO, PSO, and Genetic Algorithms are usually evaluated on small datasets in the early stages of experimentation. Yang (2012) states that using small, controlled datasets allow researchers to evaluate the performance of the parameter. Talbi (2009) and Blum and Roli (2003) give emphasis to hybrid metaheuristic systems which undergo initial validation on small data before scalability assessment. In the context of text mining research, small document corpora are acceptable for exploratory clustering to adjust similarity metrics and analyze semantic groupings. Aggarwal and Zhai (2012).

In the second stage, text preprocessing converts the unstructured content into a clean, machine-readable representation suitable for machine processing. This process includes text extraction using libraries with PyMuPDF, followed by tokenization, lowercasing, stop words removal including punctuations and lemmatization for further processing. Eliminating irrelevant and redundant content while ensuring meaningful textual features, plays a crucial part in improving clustering performance. The study of Raghavarao, Sravankumar & Madhu (2012) provides a foundational study on hierarchical clustering methods for document datasets, highlighting similarity measures and linkage methods. The study also points out that effective preprocessing plays a crucial role in ensuring meaningful clustering of textual documents based on their semantic relationships.

Once preprocessing is complete, the feature extraction stage converts the documents into numerical representations using the TF-IDF (Term Frequency–Inverse Document Frequency) vectorizer from the scikit-learn library (Pedregosa et al., 2011).

TF-IDF measures the importance of terms in each document relative to the entire corpus. It generates high-dimensional vectors that produce semantic information. These vectors serve as the input for the subsequent clustering and optimization stages. To ensure accurate vectorization, common stop words such as articles, pronouns, prepositions, and conjunctions are removed during data processing to avoid distortion of weights and minimize noise in the dataset. This process produce a spare of TF-IDF

representation that captures the relevance of terms throughout the corpus. Research shows the effectiveness of TF-IDF in enhancing clustering performance particularly when dealing with imbalanced datasets (Lemaître, Nogueira, & Aridas, 2017).

SSO is a swarm intelligence technique modeled after the cooperative foraging patterns of social spiders (Yang, 2012). In this system, SSO determines optimal initial cluster centroids based on the TF-IDF matrix prepared in the preprocessing stage. The algorithm begins with the initialization of parameters such as the number of clusters, population size, number of iterations, and communication coefficients (α and β). SSO algorithm was executed with a swarm size of 30 spiders and was iterated 100 times, using vibration attenuation values of $\alpha = 0.6$ and $\beta = 0.3$. These values were identified through initial testing to balance rapid convergence speed and cluster quality. The fitness of each spider is evaluated through cosine similarity between its position and the document vectors. The spider with the highest fitness value is recognized as the global best solution. Through stochastic movement rules, spiders are attracted to or repelled from the global best spiders allowing the swarm to search broadly to maintain population diversity. This process continues until stabilization is achieved producing centroid vectors that direct the initial grouping of documents. To refine and stabilize the preliminary clusters produced by SSO, Agglomerative Clustering technique merges document vectors based on cosine similarity. By initializing the hierarchical clustering process with the optimized centroids obtained from SSO, this refinement process improves coherence and cluster quality.

Using scikit-learn (Pedregosa et al., 2011), Agglomerative Clustering refines the local clusters while maintaining the global structure produced by SSO. This hybrid approach achieves balance between fine-tuned local groupings and global optimization, resulting in more reliable and meaningful semantics. Prior research highlights the benefits and limitations of hierarchical clustering methods in large-scale document analysis, particularly with respect to scalability and computational efficiency (Croft et al., 2010). Once document clustering is complete, the quality of the resulting clusters is then evaluated. Silhouette Coefficient, Dunn Index, and Cluster Purity are used to measure the overall accuracy of each data point with respect to each own cluster. These metrics provides basis for comparing hybrid approach to traditional clustering methods. All testing procedures were run on Windows 11 operating system, with Intel Core i7-14700 CPU and 32 GB RAM. The system's hardware provided adequate computational resources to enable efficient execution of the hybrid-clustering algorithm. Python was used to implement all algorithms with NumPy, scikit-learn, and matplotlib libraries.

In the final stage, results are shown in a structured format that includes the creation of subfolders for each cluster. Dendrogram was used to display the visualization of clustered documents. Dimensionality reduction approaches, including PCA (Principal Component Analysis) and UMAP (Uniform Manifold Approximation and Projection) were applied to convert high-dimensional vectors into two-dimensional components for meaningful visualization of clusters. This stage supports the interpretation of clustering results, enabling researchers to understand and gain insights into cluster composition.

Overall, the system architecture diagram presents the structured overview of the entire workflow consisting of the coordinated stages of preprocessing, feature extraction, metaheuristic optimization, hierarchical clustering, evaluation, and visualization. It highlights how SSO enhances the clustering process by optimizing parameters, which are then leveraged by Agglomerative Clustering to produce high-quality, semantically coherent clusters.

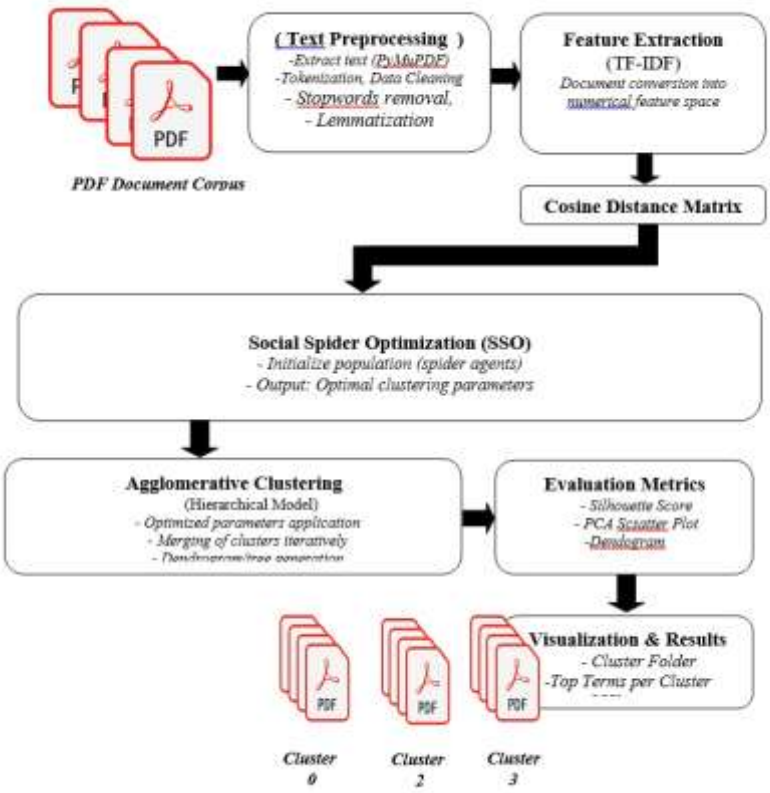

Figure 2. System architecture of the hybrid Social Spider Optimization (SSO) and Agglomerative Hierarchical Clustering algorithm

RESULTS

The main objective of the study is to classify PDF documents into relevant topics using SSO algorithm and Agglomerative clustering. Shown in Figure 3 are the uncategorized files containing fifteen PDF documents with different topics including Science and Technology, Health and Medicine, and Business and Economics.

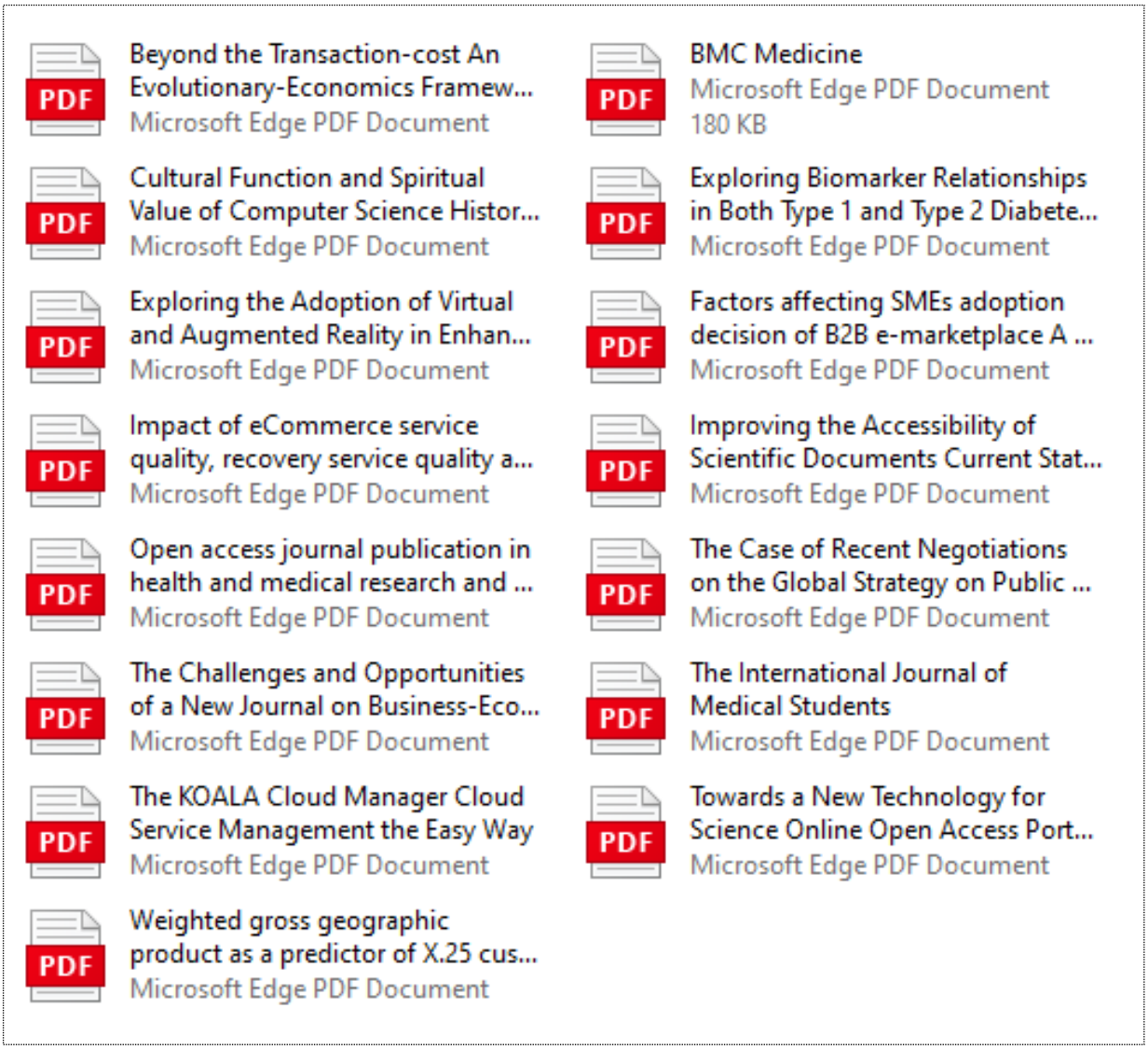

Figure 3. Uncategorized Files

The developed system was created using Python programming language where it aggregates documents by converting PDFs into TF-IDF. The application of TF-IDF converts documents into feature-dense vectors that represent content significance rather than frequency using Cosine Similarity prior to clustering. Moreover, it utilizes SSO method to identify the three optimal centroids. These centroids are implemented to evaluate the cosine similarity of each PDF document and assign it to the nearest semantic cluster. SSO-based clustering is superior to conventional K-means for clustering text content because it seeks optimal centroid location throughout the entire space rather than presuming a homogenous feature distribution. Once executed, fifteen uncategorized PDF documents were successfully clustered into three different subfolders namely cluster 0, cluster 1, and cluster 2. Each folder contained files with the same topic. This outcome illustrates the effectiveness of combining SSO and Agglomerative Clustering. Figure 4 presents the

program code for obtaining semantic similarity to refine document clustering using a precomputed distance matrix.

```
# ================================
# SOCIAL SPIDER OPTIMIZATION
# ================================
def sso_optimize(dist_matrix, n_clusters, population=20, iterations=30):
    n = dist_matrix.shape[0]
    spiders = [np.random.randint(0, n_clusters, n) for _ in range(population)]

    def fitness(labels):
        total = 0
        for k in range(n_clusters):
            idx = np.where(labels == k)[0]
            if len(idx) > 1:
                sub = dist_matrix[np.ix_(idx, idx)]
                total += sub.mean()
        return total

    best = spiders[0]
    best_fit = fitness(best)

    for _ in range(iterations):
        new_spiders = []
        for s in spiders:
            mutant = s.copy()
            for i in range(n):
                if random.random() < 0.1:
                    mutant[i] = random.randint(0, n_clusters - 1)
            if fitness(mutant) < fitness(s):
                new_spiders.append(mutant)
            else:
                new_spiders.append(s)
        spiders = new_spiders
        for s in spiders:
            fit_s = fitness(s)
            if fit_s < best_fit:
                best_fit = fit_s
                best = s.copy()

    return best
```

Figure 4. Social Spider Optimization Program Code for Semantic Similarity

SSO and Agglomerative Clustering Results

As illustrated in Figure 5, SSO and Agglomerative Clustering results were obtained using Python code for the execution of the agglomerative clustering algorithm. To evaluate the closeness of documents based on their text content, cosine similarity was used. The cluster distance is determined by the most distant document between clusters.

This produces compact clusters and reduces noise. The documents are categorized according to semantic similarity.

```
model = AgglomerativeClustering(n_clusters=n_clusters, metric="cosine", linkage="average")
model.fit(tfidf_matrix.toarray())
final_labels = model.labels_
```

Figure 5. Python code that runs the Agglomerative Clustering Algorithm

As shown in Figure 6, is the scatterplot of the fifteen uncategorized PDF files used to observe relationships between variables. A scatter plot uses dots to represent values for two distinct numeric parameters. The position of each dot on the vertical and horizontal axis specifies values for each data point. Scatter plots are often used to visualize these data points onto principal components helping to interpret the relationships among observations. In this study, Principal Component (PC1) represents the x-axis while the y-axis indicates the Second Principal Component. To achieve further insights, two plots were generated to determine the movement from initial clustering using SSO to the final clustering process. The PCA scatterplot reveals that the fifteen PDF files were categorized into two distinct clusters. The first main component (PCI1) differentiates the PDFs according to significant content variations, creating left and right clusters. The second main component (PC2) reveals additional adjustment within each cluster, highlighting which PDFs is contextually within their group. PCA simplifies complex datasets while preserving their most important structures.

Point	PC1	PC2
1	-0.485	-0.325
2	-0.48	-0.32
3	-0.21	-0.32
4	-0.19	-0.31
5	-0.22	-0.29
6	0.39	-0.01
7	0.45	0
8	0.55	0
9	0.54	0.04
10	0.6	0.04
11	-0.21	0.29
12	-0.22	0.32
13	-0.29	0.34
14	-0.31	0.49
15	-0.31	0.52

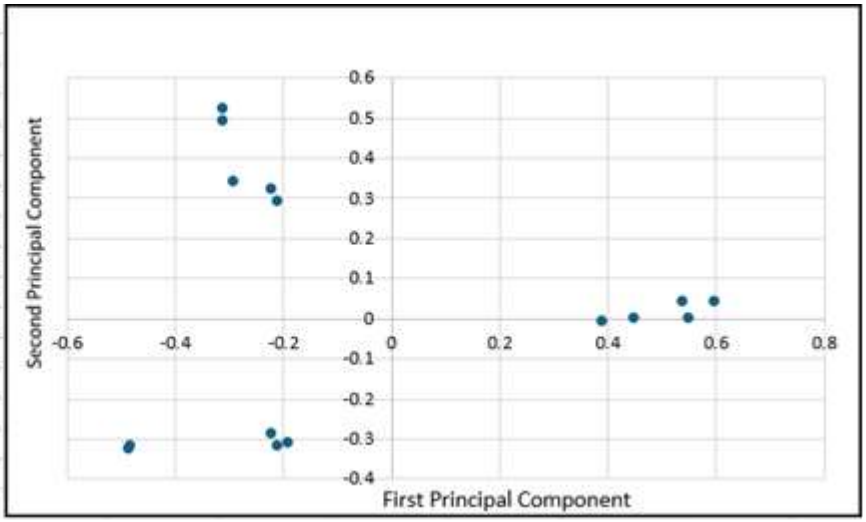

Figure 6. Scatterplot Diagram

The final clustering results were generated using Agglomerative Clustering algorithm as shown in Figure 7. The comparison between the first and final clustering plot clearly demonstrates the effectiveness of using the Agglomerative Clustering obtained from the

SSO algorithm's cluster centroids as the initial clusters. This integration of SSO and Agglomerative Clustering algorithm enhances the initial cluster formation yielding greater clustering accuracy and more consistent group structure. The visualization shows distinct cluster boundaries, with minimal overlap between document groups. Cluster 0 (Science and Technology) is visually distant from Cluster 1 (Health and Medicine), signifying distinct semantic features identified during feature extraction.

Figure 7. Scatter Plot Diagram of Final Clustering Results

Table 1 summarizes clustering results with three clusters were formed using cosine similarity. Points 1 to 5 belong to cluster 0, showing moderate spread while maintaining similar direction. Points 6–10 are assigned to Cluster 1, showing a largely uniform orientation, suggesting a significant degree of consistency among PDF documents. Lastly, Cluster 2 displays strong cohesion among its points.

Table 1. Cluster Summary based on Cosine Similarity

Cluster	Points	Characteristics
Cluster 0	1, 2, 3, 4, 5	Similar direction, moderate spread
Cluster 1	6, 7, 8, 9, 10	Extremely high similarity, almost uniform
Cluster 2	11, 12, 13, 14, 15	Extremely tight similarity

Final clustered PDF documents as shown in Figure 8 illustrates how the system successfully clustered PDF documents into semantically coherent clusters based on their extracted textual content and comparison patterns. The clustered document with closely related topics be likely to appear within the same group, indicating the model's capability to recognize underlying thematic structures across unrelated PDF files. Those with weaker semantic relationships were grouped in different clusters while documents that share similar keywords or domain-specific terminologies are positioned together. This

visual distribution highlights the efficiency of SSO in controlling the search function for optimal linkage configuration, distance thresholds, and cluster boundaries.

Furthermore, the clear separation between clusters specifies that the integration of SSO improved the hierarchical clustering mechanism by reducing overlap and minimizing classification ambiguity. This resulted in more distinguishable cluster formations, higher cluster purity, and enhanced interpretability of the grouped documents.

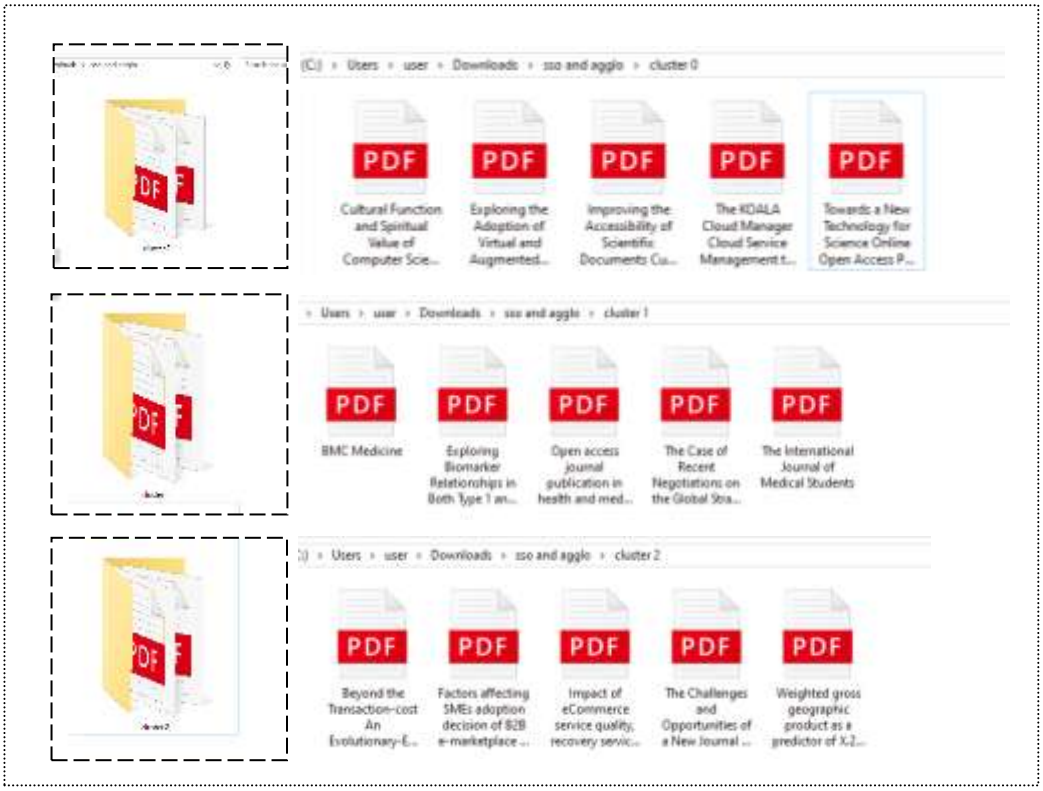

Figure 8. Final Clustered PDF Files Generated by the Model

Quantitative Evaluation

The system's performance was evaluated using internal cluster validity metrics, including Silhouette Coefficient, Dunn Index, and Cluster Purity. These metrics were used to evaluate the performance of the SSO-Agglomerative Clustering algorithm instead of using the Agglomerative Clustering model without optimization. Moreover, Table 2 shows the integration of SSO, which significantly improved clustering performance. The Silhouette Coefficient increased from 0.63 to 0.82, indicating more closely grouped clusters. When clustering documents, a higher Silhouette Coefficient shows that documents within a cluster share strong semantic similarity, whereas, documents in different clusters are comparatively distinct. This metric provides an effective performance of different clustering approaches on text data. There is also a substantial increase in the Dunn Index metric. It gained a score from 0.49 to 0.68, demonstrating more clearly defined cluster boundaries. Its main purpose is to quantify the compactness

and separation between clusters in a clustering solution. Equally, Cluster Purity increased from 84.6% to 94.3%, indicating that the SSO-Agglomerative method produced more accurate topic-based groupings. Cluster Purity is measured according to their ground-truth category. Given the number of clusters, computation time increased from 9.4 to 11.9 seconds, due to the use of more complex algorithm and the need for iterative optimization to achieve better results.

Table 2. Comparison of Clustering Performance Metrics

Metric	Agglomerative Clustering	SSO + Agglomerative
Silhouette Coefficient	0.63	0.82
Dunn Index	0.49	0.68
Cluster Purity	84.6%	94.3%
Computation Time (s)	9.4	11.9

Cluster Distribution and Semantic Coherence

Table 3 illustrates the cluster distribution using Principal Component Analysis (PCA) applied to TF-IDF space for comparing and grouping of related text.

Table 3. Cluster Distribution

Cluster	Distribution	Summary
Cluster 0	Science & Technology	5 documents
Cluster 1	Health & Medicine	5 documents
Cluster 2	Business & Economics	5 documents

Based on the analysis, hybrid approach effectively showed semantic relationships within unstructured text content. Each cluster contained thematically consistent content. Specifically, documents in Cluster 1 contained frequent terms such as "patients," "treatments," and "diseases," while Cluster 0 featured terms like "innovation," "model," and "computer." Cluster 2 included words such as "operations," "investments," and "product," and ,"strategy" which strengthened its economic orientation.

Comparative Performance Analysis

Table 4 presents the results for comparative evaluation with traditional methods such as Hierarchical Clustering without optimization. It further highlighted the effectiveness of the hybrid model. This shows that while Agglomerative Clustering gained moderate performance with a silhouette score of 0.62, the integration of SSO produced the best outcomes with a score 0.76. Compared to Particle Swarm Optimization (PSO), the SSO achieved the highest score of 0.82 highlighting its cluster stability and semantic consistency. According to Yang (2012), SSO's unique cooperative search mechanism solves optimization problems, where male and female agents represent varied search agents that balance exploration and exploitation in the solution space.

Table 4. Comparative Analysis with Other Existing Clustering Techniques

Algorithm	Silhouette Score
Agglomerative Clustering	0.62
PSO + Agglomerative Clustering	0.76
SSO + Agglomerative Clustering	0.82

Visualization and System Output

Dendrogram was used to depict the hierarchical relationship between objects. This diagrammatic representation is like a family tree of clusters showing how individual data points or group frequently used in different context. In a cluster dendrogram as in Figure 9, clusters exhibit higher semantic coherence. The height of the linkage represents the distance between clusters. The longer the line represents greater difference. These results show how SSO dynamically influenced the formation of clusters resulting in more closely grouped clusters while ensuring well-separated hierarchical structures.

Moreover, SSO user interface serves as in input mechanism for the uploading of new PDF files. Once uploaded, the system can generate dendrogram after extracting text documents in preparation for the classifications of appropriate clusters. This process checks the model's potential scalability for real-world applications, such as academic repository management, enterprise knowledge organization, and automated literature review systems.

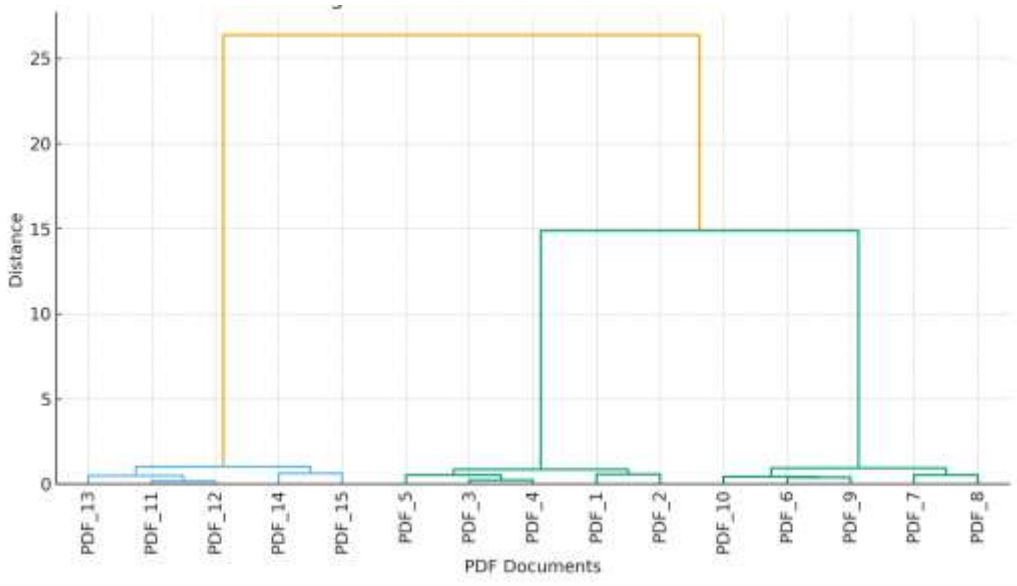

Figure 9. Cluster Dendrogram

DISCUSSION

The results of this study highlight the effectiveness of integrating SSO with the Agglomerative hierarchical Clustering technique for automated Portable Document Format (PDF) document. Fifteen unclassified PDF documents were classified into three different clusters each containing related text content. Compared to conventional clustering techniques, this hybrid optimization approach showed significant performance in both semantic coherence and accuracy. High semantic coherence indicates that each cluster contains documents that are meaningfully related.

The collective behavior of social spiders produced meaningful initial clusters where each spider in the algorithm represents candidate solution. These clusters are considered inputs for Agglomerative approach to further refine the document grouping based on similarity measures. The hybrid approach enabled the system to automatically identify the most suitable and accurate merging techniques in improving document clustering process.

Li et al. (2019) and Das et al. (2009), highlighted the potential of hybrid approach in improving hierarchical clustering among textual entities resulting in more accurate and semantically coherent classifications.

The clustering patterns align with the findings of Chen and Ye(2004). The researchers highlighted that optimization-guided clustering can be meaningfully improve topic insight in large textual datasets. Furthermore, Jain et al. (1999) emphasizes that hierarchical clustering can be enhanced further through metaheuristic integration.

However, a notable limitation of this study is the relatively small dataset consisting of only fifteen PDF documents as it restricts the extent to which it may not fully measure the performance of the hybrid SSO-Agglomerative clustering model in a more diverse document collection. While the method produced notable results within this scope, this approach remains technically valid to which the results can be generalized. However, Agglomerative Clustering despite its simplicity in smaller datasets, faces computational scalability issues when applied to high-dimensional data as it increasingly demanding as the dataset grows (Li, Zhang, Xu, Zhu, & Sharafaldin, 2019).

Additionally, Khan et al. (2022) highlighted that hybrid clustering techniques, improves the interpretability of unstructured text data such as PDFs using TF-IDF for feature extraction, combined with optimization-oriented parameter selection to identify hidden topic relationships. Likewise, Ghosh et al. (2023) stated that hierarchical clustering technique yields optimal performance when supported by adaptive optimization methods, as it enables adjustments of similarity metrics for documents of varying complexities.

The clustering performance observed in this study, support these conclusions with documents on Science & Technology, Health & Medicine, and Business & Economics that were properly classified into their respective clusters. Overall, the findings underscore the system's robustness and potential use in organizational document management. Future studies may consider extending this framework to support different document formats such as HTML web pages, Word document file to assess its accuracy and scalability.

In addition, inclusion of two metaheuristic approaches Swarm Intelligence techniques namely: Particle Swarm Optimization and Ant Colony Optimization could provide further probabilistic technique for solving computational clustering problems.

Finally, ethical and societal considerations must also be considered. Document clustering system that processes sensitive information should follow data privacy regulations and safeguard anonymity.

CONCLUSIONS AND RECOMMENDATIONS

In conclusion, the study demonstrates the successful integration of SSO algorithm with Agglomerative Clustering to facilitate clustering of PDF documents.

Findings from this analysis underscore the superiority of this integrated approach compared to conventional methods. Leveraging the cooperative behavior of social spiders, the SSO algorithm efficiently explored and exploited the search space, resulting in significantly improved accuracy and efficiency in clustering PDF documents.

The significance of this research lies in its potential implications it helps practitioners, researchers and institutions dealing with high-volume of PDF documents. The combination of SSO and Agglomerative Clustering techniques, serves as a powerful resource for organization of documents and knowledge extraction. This contribution strengthens data mining automation by enabling more accurate text-based and efficient decision-making. It can also serve as a significant improvement in automating PDF document clustering using the combined method of SSO and Agglomerative Clustering. New methods have been established by the researchers to enable more in-depth analysis of PDF documents.

Lastly, the use of SSO and Agglomerative Clustering to automate PDF document clustering yields promising results, boosting accuracy and efficiency over existing techniques. A new study is needed to investigate scalability, including the integration of new optimization techniques while addressing ethical issues. The researchers can further improve information retrieval, knowledge extraction, and document organization across disciplines by exploring new frameworks while assuring responsible and ethical deployment.

IMPLICATIONS

Applying the SSO with Agglomerative Clustering for automated PDF document clustering brings meaningful implications in terms of systems performance and societal domains. From a computational perspective, the approach strengthens clustering effectiveness by applying swarm-based algorithms, providing a more accurate method for semantic grouping of PDF documents. The study also contributes to the growing body of work on integrated optimization techniques, encouraging future development fostering deeper exploration in machine learning applications. However, the adoption of such automated systems entails ethical concerns related to data privacy, potential impacts on job displacement in handling manual document classification roles, and promoting transparency in algorithmic decision-making. Overall, this study establishes a basis for scalable and ethically aware document clustering systems that can provide benefit across sectors.

ACKNOWLEDGEMENT

The researchers sincerely extend their heartfelt gratitude to the Technological University of the Philippines (TUP) for its continuous support to academic excellence and the institution's provision of resources and professional guidance played an essential role in the successful achievement of this research. The authors also wish to acknowledge academic advisers and experts whose expertise significantly contributed to the development of the study.

DECLARATIONS

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this research.

Informed Consent

Not applicable. This study did not involve human participants, interviews, or personal data requiring informed consent.

Ethics Approval

Not applicable. As the research focused solely on the development and evaluation of an automated document clustering system using publicly available datasets, ethics approval was not required.

REFERENCES

Alam, M., Dobbie, G., & Riddle, P. (2008). A hybrid genetic algorithm for clustering mixed datasets. *Proceedings of the 2008 IEEE Congress on Evolutionary Computatio*n (pp. 1889–1895). https://doi.org/10.1109/CEC.2008.4631087.

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer.

Aljumily, R. (2016). Agglomerative Hierarchical Clustering: An Introduction to Essentials. *Global Journal of Human-Social Science: G Linguistic & Education*.

Blum, C., & Roli, A. (2003). Metaheuristics in combinatorial optimization. *ACM Computing Surveys, 35*(3), 268–308.

Chen, Y., & Ye, X. (2004). Particle swarm optimization algorithm and its application to clustering analysis. *Proceedings of the 2004 International Conference on Machine Learning and Cybernetics*, *2,* 1336–1341. https://doi.org/10.1109/ICMLC.2004.1380188

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice*. Pearson Education.

Das, S., Chowdhury, A., & Abraham, A. (2009). A survey on swarm inspired clustering techniques. *Swarm, Evolutionary, and Memetic Computing*, 318–324. https://doi.org/10.1007/978-3-642-10877-8_39

Erik Cuevas, M. C. (2015). A Computational Intelligence Optimization Algorithm Based on the Behavior of the Social-Spider. In M. C. Erik Cuevas, *Computational Intelligence Applications in Modeling and Control, Studies in Computational Intelligence* (pp. 123-145). Springer International Publishing Switzerland.

Dorigo, M., & Stützle, T. (2004). *Ant colony optimization*. MIT Press.

Ghosh, D., Singh, A., & Mishra, P. (2023). Optimization-driven hierarchical clustering for large-scale unstructured text data. *IEEE Access, 11*, 59822–59835.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323. https://doi.org/10.1145/331499.331504

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Proceedings of ICNN'95 - *International Conference on Neural Networks,* 4, 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Khan, R., Ahmad, I., & Nasir, M. (2022). A hybrid document clustering framework integrating swarm optimization and semantic similarity. *Applied Soft Computing, 118*, 108484.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research, 18*(1), 559–563.

Li, W., Zhang, Y., Xu, Y., Zhu, Y., & Sharafaldin, I. (2019). Clustering analysis using hybrid optimization algorithms: A review. *Applied Soft Computing, 81*, 105492. https://doi.org/10.1016/j.asoc.2019.105492

Liu, J., & Wang, S. (2020). Enhanced hierarchical text clustering with metaheuristic parameter tuning. *Information Sciences, 512*, 354–370.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Raghavarao, N., Sravankumar, K., & Madhu, P. (2012). A survey on document clustering with hierarchical methods and similarity measures. *International Journal of Engineering Research & Technology (IJERT), 1*(7), 1-9.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47. https://doi.org/10.1145/505282.505283

Talbi, E. G. (2009). *Metaheuristics*: *From design to implementation*. Wiley.

Yang, X. S. (2012). *Nature-inspired metaheuristic algorithms (2nd ed.).* Luniver Press.

Zhao, X., Li, Q., & Chen, Y. (2021). Hybrid metaheuristic-based text clustering using social spider optimization and k-means. *Expert Systems with Applications, 175*, 114824.

Author's Biography

Dr. Wellanie M. Molino is an Associate Professor III at the Computer Studies Department of the Technological University of the Philippines – Manila. She holds a Doctor of Information Technology degree and has established a research portfolio in the areas of local governance, data analytics, e-governance, and emerging technologies in education. Dr. Molino has co-authored several scholarly publications focusing on virtual-reality classrooms, NFC-based mobile solutions, citizen e-participation, and predictive analytics. Experienced in teaching mobile and web application to learners of varying abilities.

Dr. Francis A. Alfaro received his Doctor of Education degree at the Technological University of the Philippines. He is a Faculty of the University Research and Development Services of TUP. A passionate faculty with excellent presentation, research and communication skills. His research focused on education, mathematics and information technology especially in prototype development and database management systems. He is also in-charge in the electronic data processing (SPSS, Turnitin plagiarism checker) of all the research manuscripts of the University.

Dr. Jonathan M. Caballero is a faculty member at the Technological University of the Philippines – Manila, specializing in computing and management. He holds a Bachelor of Science in Computer Science and a Master of Science in Computer Science, and has earned a Doctor of Technology degree, demonstrating a strong foundation in both theoretical and applied aspects of the discipline. With his academic background and leadership role, he actively contributes to curriculum development, research initiatives, and instructional delivery. His expertise spans computing technologies and academic management, positioning him as a key figure in advancing the university's IT education programs.

Prof. Francis L. Dela Cruz is an Assistant Professor III in the Computer Studies Department at the Technological University of the Philippines – Manila. He earned his Bachelor of Science in Computer Science and a Master in Information Technology, and is currently pursuing a PhD in Technology Management (ongoing) in TUP Manila. His permanent faculty status reflects his stable commitment to the university's academic mission. As a mid-career academic, he contributes to both teaching and curriculum delivery in IT-related courses. His educational background positions him well to support both undergraduate and graduate-level students in the Computer Studies program.

Prof. Jan Eilbert L. Lee is a faculty member in the Computer Studies Department at the Technological University of the Philippines – Manila. He holds a Bachelor of Science in Mathematics and a Master of Science in Mathematics from the Eulogio "Amang" Rodriguez Institute of Science and Technology (EARIST), reflecting a strong foundation in analytical and computational disciplines. At TUP, he teaches advanced subjects such as Artificial Intelligence and Data Analytics, where he integrates mathematical theory with practical computing applications. His expertise supports the department's goal of producing graduates skilled in intelligent systems and data-driven decision-making. Through his academic background and teaching focus, Mr. Lee contributes significantly to fostering innovation and critical thinking among IT students.

Dr. Jasmin D. Niguidula is a faculty member in the Computer Studies Department at the Technological University of the Philippines – Manila, with expertise in computing and academic management. She holds a Bachelor of Science in Computer Science, a Master in Information Technology, and a Doctor of Technology degree, reflecting her strong academic and professional foundation. Dr. Niguidula actively contributes to the enhancement of curriculum and instruction in the field of information technology, while also engaging in research and institutional initiatives. Her work integrates both technical knowledge and leadership, making her a valuable resource in shaping future-ready IT professionals. With her extensive experience and dedication to educational advancement, she plays a pivotal role in fostering academic excellence and innovation at TUP Manila.

Associate Professor III Fernando L. Renegado is a distinguished faculty member in the Computer Studies Department at the Technological University of the Philippines– Manila, specializing in computing, administration, and quality management. He earned his Bachelor of Science in Computer Science and a Master in Engineering Management, and has undertaken doctoral coursework in Technology, reflecting his dedication to higher learning and academic excellence. As a Professional Board Examination Test (PBET) lecturer, he contributes significantly to professional certification preparation in computing and engineering fields. Known for his strong command of data structures, he is regarded as an expert in teaching complex programming concepts with clarity and depth. Through his leadership in curriculum, quality assurance, and scholarly contributions, he plays a vital role in advancing both academic rigor and educational innovation within TUP-Manila's Computer Studies program.

Prof Ariel C. Tomagan is a faculty member in the Computer Studies Department at the Technological University of the Philippines – Manila. He holds a Bachelor of Science in Computer Science and a Master of Science in Information Science, and is currently pursuing a Doctor of Technology degree, with 18 academic units completed including 12 through online learning. As a permanent faculty member, he contributes to the department through both classroom instruction and academic advising. His teaching focuses on core subjects such as Networking 1 and 2 and Programming Languages, reflecting his expertise in foundational and applied computing. With a commitment to continuous learning and technical excellence, Mr. Tomagan plays a vital role in shaping the competencies of future IT professionals at TUP.

Prof. Darwin C. Vargas is a faculty member in the Computer Studies Department at the Technological University of the Philippines – Manila. He holds a Master's degree in Information Technology and has a strong interest in programming, hardware embedding, and software–hardware integration. Vargas co-authored a landmark April 2024 paper on geographic trend analysis for predicting crop-yield success rates in the Philippines. He also contributed to research on developing virtual-reality classroom environments as alternative learning platforms during the pandemic. Within campus life, he plays an active role as a faculty adviser to TUP's robotics group, Graybots, working to mentor and inspire student innovators.