International Journal of Information Management Data Insights 6 (2026) 100396

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

A comparative empirical evaluation of semantic clustering algorithms on static word embeddings

Asefeh Asemi [a], Rajab Kiani Shahvandy [b,*], Mahdi Houshangi [c]

[a] Department of Information Systems, Institute of Data Analytics and Information Systems, Corvinus University of Budapest, Hungary
[b] Department of Knowledge and Information Science, University of Isfahan, Isfahan, Iran
[c] Technical University of Munich, Munich, Germany

ARTICLE INFO

Keywords:
Semantic clustering
Word embeddings
K-means
DBSCAN
Natural language processing
Information management
Text mining
Unsupervised learning
GloVe
Document clustering

ABSTRACT

Objective: This study conducts a comprehensive empirical evaluation of semantic clustering algorithms to identify the most effective approach for automatically organizing and extracting meaning from textual data. By systematically comparing the performance of K-means, K-medoids, and DBSCAN on word embeddings from GloVe and Wiki models, it provides data-driven insights for optimizing Natural Language Processing (NLP) pipelines in information management systems. The research suggests a practical framework for selecting clustering algorithms and embedding models based on specific operational objectives, such as document clustering, knowledge base construction, and content-based recommendation.
Design/Methodology/Approach: The investigation employed a two-phase methodology. Initially, predefined word lists were transformed into numerical vectors using pre-trained GloVe and Wiki models. K-means, K-medoids, and DBSCAN algorithms were applied, with performance evaluated via Silhouette Score and Davies-Bouldin Index, complemented by Principal Component Analysis (PCA) for visualization. Results were benchmarked against manually curated semantic groupings. Subsequently, the findings were validated on a large-scale corpus of 303 research articles to assess scalability and real-world applicability.
Results/Discussion: Analysis indicates that, under the evaluated configurations, K-means combined with GloVe embeddings produced comparatively higher semantic coherence and more interpretable cluster structures than the alternative methods considered. K-medoids demonstrated robustness against outliers but yielded less compact groupings. While DBSCAN indicated effective for outlier identification, it consistently underperformed in forming semantically meaningful clusters. The GloVe model significantly outperformed Wiki embeddings in generating precise and interpretable clusters, whereas Wiki produced broader, less distinct groupings. Large-scale validation confirmed these results, with K-means successfully identifying dominant research themes, including digital library adoption (43.2%), reference services (15.2%), and research data management (8.9%)—in a corpus of academic literature. Under the evaluated corpus characteristics and parameter settings, DBSCAN classified most documents as outliers, indicating limited suitability for this specific balanced document collection.
Conclusions: K-means and K-medoids emerge as comparatively effective algorithms under the evaluated conditions. The study underscores the critical influence of vector representation models, with GloVe embeddings providing superior semantic distinction compared to Wiki. These findings offer clear, actionable guidance for selecting clustering methods in NLP applications, highlighting the necessity of aligning algorithmic choice with specific dataset characteristics and information management goals.
Originality/Value: This research moves beyond theoretical descriptions by delivering a rigorous, empirical comparison that elucidates the crucial interaction between algorithm selection and embedding models for semantic tasks. The findings provide practitioners with a context-dependent decision matrix: K-means with GloVe is effective under the studied conditions for taxonomy development and thematic categorization, whereas DBSCAN is preferable for outlier detection in noisy data. By demonstrating that GloVe's global statistical approach yields more distinct clusters than Wiki's contextual model for this purpose, the study contributes a practical, evidence-based framework for enhancing semantic analysis in real-world information systems.

* Corresponding author.
E-mail address: kiani.rajab@gmail.com (R. Kiani Shahvandy).

https://doi.org/10.1016/j.jjimei.2026.100396

Available online 11 March 2026

1. Introduction

The exponential growth of digital text has rendered the ability to automatically organize, categorize, and retrieve unstructured data a critical competency in modern information management. As organizations grapple with vast document corpora, user-generated content, and dynamic information streams, the need for efficient thematic pattern discovery, knowledge organization, and intelligent content recommendation has never been more pressing. Clustering, a foundational unsupervised learning technique in data mining, serves as a powerful engine for this task, grouping semantically related items to reveal latent structures and relationships hidden within textual data. The critical role of clustering and thematic analysis in uncovering latent structures within complex textual data is well-established, complementing our prior research on conceptual interactions in deep learning ecosystems (Asemi et al., 2021) In the specific domain of natural language processing (NLP), semantic clustering has emerged as a pivotal methodology for discovering thematic patterns (Asemi et al., 2026a), mapping conceptual hierarchies, and elucidating intricate semantic relationships (Petukhova et al., 2025). The advent of word embedding models has fundamentally accelerated these capabilities by providing dense, numerical representations that encode semantic and syntactic properties of words. Models like GloVe, which leverages global corpus statistics (Pennington et al., 2014), and Wiki, which incorporates sub-word information to handle morphological richness (Bojanowski et al., 2017), have become instrumental in transforming qualitative textual meaning into a quantifiable vector space. While more recent transformer-based models like BERT offer contextualized embeddings with superior nuance (Devlin et al., 2019), their computational intensity often precludes their use for large-scale, iterative clustering experiments. Consequently, GloVe and Wiki remain highly relevant for foundational comparative studies due to their indicated efficacy, efficiency, and conceptual clarity, providing an ideal testbed for evaluating clustering algorithms. The landscape of clustering algorithms is diverse, each family offering distinct advantages and trade-offs (Asemi et al., 2023, 2026b). Traditional partition-based methods like K-means are lauded for their simplicity and computational efficiency but are notoriously sensitive to outliers and initial centroid selection. K-medoids, a more robust variant, mitigates outlier influence by using actual data points as cluster centers. In contrast, density-based algorithms such as DBSCAN excel at identifying arbitrarily shaped clusters and detecting noise without requiring a pre-specified number of clusters, though their performance is highly sensitive to parameter tuning. The effectiveness of any algorithm, however, is not inherent but is profoundly mediated by the characteristics of the vector space it operates upon. For instance, the global coherence of GloVe embeddings may interact differently with clustering mechanisms than the more contextually nuanced Wiki vectors, a critical interplay that remains underexplored in a structured, comparative manner. Despite the extensive application of clustering across domains—from sentiment analysis (Xiaoyan et al., 2022) to thematic relationship mapping (Shabani & Asemi, 2022)—a significant gap persists. Existing literature often provides isolated evaluations or generic descriptions of algorithms, lacking a holistic, empirical benchmark that systematically correlates the choice of both the embedding model and the clustering algorithm with quantifiable outcomes in semantic coherence and practical utility. This study directly addresses this gap by conducting a rigorous, controlled empirical comparison of K-means, K-medoids, and DBSCAN applied to word embeddings from both GloVe and Wiki models. The conceptual framework, illustrated in Fig. 1, outlines this systematic process, beginning with raw text, moving through vectorization via chosen embedding models, applying clustering algorithms, and finally evaluating the semantic coherence and effectiveness of the resulting groupings.

This study is guided by three primary objectives: first, to evaluate and compare the performance of K-means, K-medoids, and DBSCAN in the semantic clustering of words; second, to examine the influence and interaction of different word embedding models, namely GloVe and Wiki, on clustering outcomes; and third, to provide actionable insights into the strengths and limitations of various algorithm-embedding combinations for real-world semantic analysis tasks. To achieve these goals, the study offers several novel contributions. It presents a direct empirical comparison between partition-based and density-based clustering algorithms across distinct embedding models, highlighting how their interaction impacts clustering quality. It also provides empirically grounded guidance for practitioners by illustrating how specific clustering–embedding combinations perform under defined experimental conditions, with K-means and GloVe showing relative advantages for taxonomy-oriented tasks (such as topic categorization), while DBSCAN

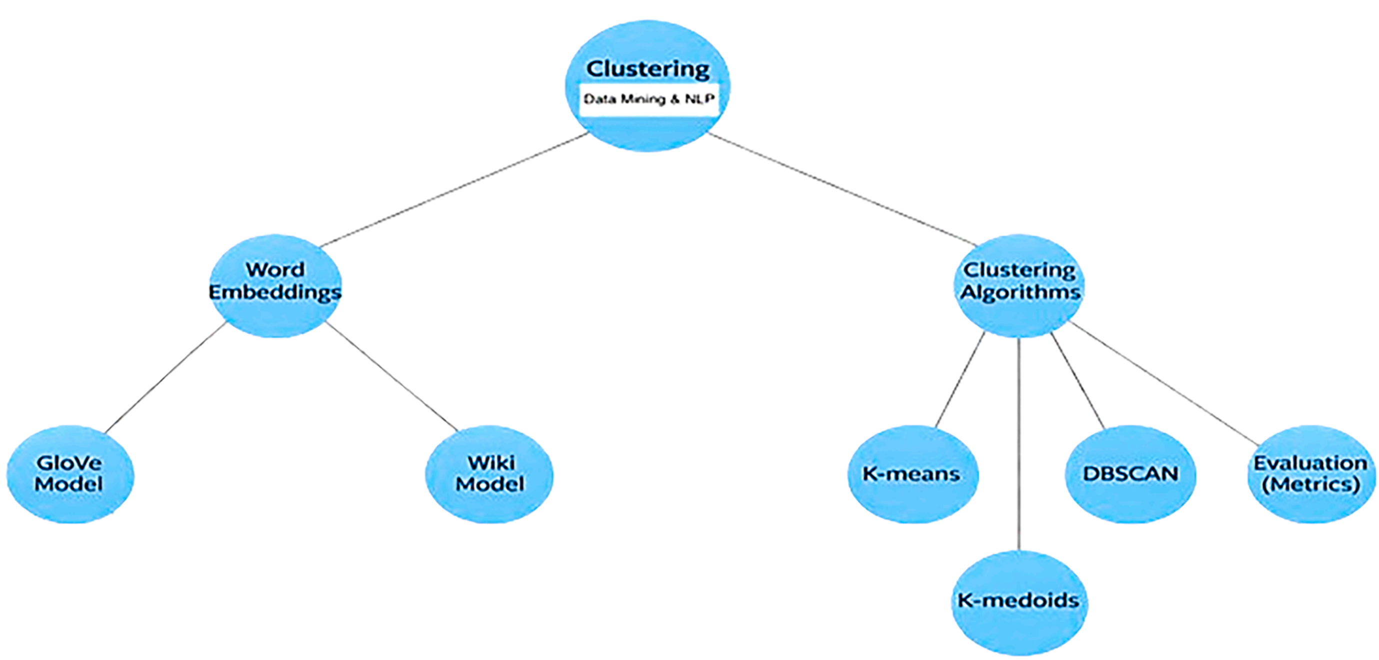

Fig. 1. Conceptual Framework: Semantic Clustering in NLP.

is more effective for exploratory analysis and outlier detection. Furthermore, the study goes beyond standard evaluation metrics by combining quantitative scores (Silhouette Score, Davies-Bouldin Index) with qualitative validation using manually curated clusters to assess the semantic meaningfulness of the results. Therefore, this research addresses an identified gap in the literature by providing a controlled empirical evaluation of how clustering algorithms and static word embedding choices interact in semantic tasks. Rather than proposing universal solutions, the study examines context-dependent performance patterns and synthesizes them into a decision-oriented framework to support researchers and practitioners in selecting appropriate clustering–embedding combinations for objectives ranging from taxonomy development to outlier detection. Based on the outlined objectives and the identified gap in the literature, this study seeks to answer the following research questions:

1. How do partition-based (K-means, K-medoids) and density-based (DBSCAN) clustering algorithms compare in terms of semantic coherence and cluster quality when applied to static word embeddings?
2. To what extent does the choice of word embedding model (GloVe vs. Wiki) influence the performance and outcome of different semantic clustering algorithms?
3. What are the practical trade-offs between cluster compactness, outlier robustness, and semantic interpretability when selecting a clustering algorithm and embedding model pair for specific information management tasks?

2. Literature review

Semantic clustering has become a focal topic in natural language processing and information retrieval, valued for its capability to organize lexical units according to their underlying semantic meanings. This section reviews key studies on clustering algorithms, vector representation models, and their role in semantic analysis. Clustering techniques have been widely studied for their effectiveness in organizing large textual datasets and extracting latent structures. Traditional partition-based methods, such as K-means, remain popular due to their simplicity, computational efficiency, and broad applicability across domains (Pandey, 2015). At the same time, prior research has emphasized well-known limitations of K-means, including sensitivity to outliers, initialization, and dependency on the predefined number of clusters (Soleimani-Nejad et al., 2018). In response, alternative approaches such as K-medoids and density-based clustering have been proposed to provide greater robustness when handling complex or noisy data distributions.

Several studies have applied semantic clustering to scientific and technical corpora with the aim of identifying conceptual structures in scholarly communication. For example, Kiani Shahvandi et al. (2024) employed text mining techniques to extract and conceptually cluster key terms from scientific abstracts collected from Web of Science and Scopus. Their quantitative approach involved nominal phrase extraction, decomposition of compound terms, vector construction, and K-means clustering. The results demonstrated how clustering can support conceptual organization by reducing thousands of extracted terms into a smaller number of interpretable conceptual clusters. While such studies illustrate the practical value of semantic clustering, they typically rely on a single clustering configuration and do not systematically compare alternative clustering algorithms or representation choices.

Semantic clustering builds upon traditional clustering techniques by explicitly incorporating word meanings and contextual relationships. Its applications have been reported in areas such as topic modeling (Thorleuchter & Van den Poel, 2016), text summarization (Van Lierde & Chow, 2019), and cognitive semantic analysis (Hähnel et al., 2023). Fakhrzadeh et al. (2023) further demonstrated the use of clustering to enhance data quality in research databases, highlighting its effectiveness in managing thematic complexity. Other studies have combined distributional representations with additional learning components; for instance,Gupta et al. (2024) reported strong performance when integrating Word2Vec embeddings with artificial neural networks. Similarly, Reveilhac and Blanchard (2022) employed embeddings such as BERT, ELMo, and Word2Vec to analyze thematic patterns in social media discourse related to health technologies, illustrating how representation choice can influence analytical outcomes. While these studies demonstrate the versatility of embedding-based representations across application domains, they typically emphasize task performance rather than examining how representation geometry interacts with clustering behavior. As a result, the implications of embedding choice for unsupervised semantic structure discovery remain underexplored. Text mining research has consistently emphasized the importance of semantic coherence in clustering results. Shokouhian et al. (2019) showed that thematic clustering can uncover latent structures in health-related texts, while Shabani and Asemi (2022) demonstrated how semantic relationships shape the organization of user-accessed information. These studies reinforce the view that clustering outcomes are influenced not only by algorithm selection but also by how semantic information is encoded and evaluated.

Advances in static word embedding models have further shaped semantic clustering research. Models such as GloVe (Pennington et al., 2014) and Wiki-based embeddings derived from fastText (Bojanowski et al., 2017) provide dense numerical representations that capture semantic similarity. GloVe emphasizes global co-occurrence statistics, whereas Wiki embeddings incorporate subword information and contextual variation. Related work has also explored hybrid semantic frameworks, such as combining information extraction with lexical resources like WordNet to monitor semantic relatedness (Bourguet & Sow, 2025). Other data-driven approaches, including K-means clustering under alternative uncertainty modeling environments such as Picture Fuzzy Sets, have been proposed to extend clustering applicability (Waja et al., 2023).

Taken together, these studies suggest that clustering performance is highly contingent on representation choices, data characteristics, and evaluation criteria, rather than being driven by algorithm selection alone. However, these factors are rarely examined jointly within a single empirical framework. In this context, the present study adopts a controlled comparative design to examine how clustering algorithms interact with static word embedding spaces under different semantic configurations and to assess the robustness of observed performance patterns through targeted sensitivity analyses.

3. Methodology

This section outlines the comprehensive methodological framework designed to systematically evaluate and compare semantic clustering methods using numerical word vectors. The study implements a rigorous experimental approach to assess the performance of three distinct clustering algorithms—K-means, K-medoids, and DBSCAN—in grouping words based on their semantic relationships. The methodology integrates both quantitative metrics and qualitative assessment techniques, supplemented by advanced visualization methods to ensure analytical clarity and interpretability of results.

3.1. Experimental design and data selection

The analytical foundation of this research is built upon word embeddings generated from two well-established models: GloVe (Pennington et al., 2014) and Wiki (Bojanowski et al., 2017). These models were specifically selected for their complementary approaches to capturing semantic information, with GloVe excelling in representing global statistical patterns across corpora and Wiki incorporating sub-word information to better handle morphological variations and rare words. To facilitate both controlled analysis and real-world validation, we employed strategically curated word samples designed to test semantic relationships across different contexts.

- The first list contained nine words representing general semantic categories, including natural elements and human-made objects: ['moon', 'rain', 'cloud', 'machine', 'car', 'tree', 'river', 'pen', 'teacher'].
- The second list focused specifically on gender-based terminology to examine more nuanced semantic relationships: ['king', 'queen', 'man', 'women', 'woman', 'men'].

These carefully constructed lists enable granular analysis of algorithmic behavior while providing clear ground truth for validation against human semantic intuition. To mitigate potential researcher bias associated with curated word lists, additional sensitivity analyses were conducted using alternative list configurations, as reported in Supplementary Table S1.

3.2. Two-Phase experimental framework

The research employs a two-phase experimental design that progresses from controlled analysis to real-world validation. Phase one utilizes the curated word lists to establish baseline performance metrics and qualitative insights in a controlled environment. To assess the robustness of the Phase 1 findings to list-specific design choices, we conducted additional sensitivity analyses using alternative word-list configurations. The resulting evaluation metrics are reported in Supplementary Table S1. This approach allows for systematic parameter tuning, direct comparison of algorithmic behavior across well-understood semantic relationships, and validation against expert-defined semantic groupings. The establishment of performance benchmarks in this phase provides a solid foundation for subsequent large-scale applications. Phase two validated the findings through application to a complex, real-world context using a comprehensive dataset of scientific literature. We collected 500 research articles published between 2020–2024 from the Scopus database, focusing on the domain of 'Artificial Intelligence and Large Language Models in Information Management and Librarianship'. The search query was as follows:

TITLE(("large language model" OR LLM OR GPT OR BERT OR "information retrieval" OR "search engine" OR "digital library" OR "knowledge management" OR "research data management" OR chatbot OR "reference service") AND (library OR libraries OR librarianship OR "library science" OR "information science" OR "library services" OR "reference service")) AND (LIMIT-TO (DOCTYPE,"ar")) AND PUBYEAR > 2020 AND PUBYEAR < 2024

The results show 498 documents. After rigorous preprocessing and quality filtering, 303 documents with substantial abstracts were retained for analysis. This domain was selected for its high relevance to NLP applications, dynamic research landscape, clearly distinguishable subthemes, and practical significance for real-world information systems. Preprocessing involved standard text cleaning, and each abstract was represented by its average GloVe vector, creating a challenging clustering task aimed at discovering emergent topics. This phase validates the findings through application to a complex, real-world context using a comprehensive dataset of scientific literature.

3.3. Data preprocessing pipeline

The preprocessing framework incorporated multiple stages to ensure data quality and analytical robustness. In the embedding vectorization stage, words were transformed into 300-dimensional vectors using pretrained GloVe and Wiki models, with careful attention to normalization procedures that ensure comparability across different embedding spaces. The selection of these specific models was based on comprehensive evaluation of their semantic capture capabilities and their widespread adoption in both research and practical applications. We performed dimensionality reduction using Principal Component Analysis (PCA), selecting it for its mathematical robustness and superior variance preservation capabilities compared to alternative methods. While t-SNE and UMAP enjoy widespread adoption for visualization purposes, they were excluded from this study due to their sensitivity to parameter tuning and significant computational demands (Van Der Maaten & Hinton, 2008). The PCA-based visualizations, as shown in Figs. 2 **and** 3, provide clear representations of semantic relationships in the vector space while maintaining analytical integrity.

3.4. Clustering framework and algorithm configuration

Three clustering algorithms were implemented with careful parameter optimization to ensure fair comparison across methods. K-means, a partition-based algorithm, was configured to minimize intra-cluster variance through k-means++ initialization for improved convergence. The optimal number of clusters (k) was determined through systematic Silhouette Score analysis across a range of 2 to 5 clusters. K-medoids, another partition-based method, was selected for its enhanced robustness to outliers using actual data points as cluster centers, making it particularly effective for noisy semantic spaces. DBSCAN, a density-based clustering algorithm, was implemented to identify arbitrary-shaped clusters without requiring pre-specified cluster numbers. This methodological approach is informed by our previous successful application of cluster analysis to identify meaningful patterns in complex, high-dimensional data (Kovács et al., 2021). We conducted systematic exploration of the eps parameter across a range from 0.5 to 30, with min_samples parameter set to 2 for the small word lists. The parameter optimization strategy employed comprehensive grid search and cross-validation techniques, ensuring that the reported performance reflects each algorithm's true potential rather than artifacts of parameter selection.

3.5. Evaluation framework

The evaluation approach combines multiple quantitative metrics with qualitative assessment to provide a comprehensive understanding of clustering performance. Quantitative evaluation employed Silhouette Score to assess intra-cluster compactness and inter-cluster separation, complemented by Davies-Bouldin Index to quantify cluster separation and cohesion. Statistical significance testing across multiple runs ensured the reliability of observed performance differences. Qualitative validation incorporated expert annotation from domain specialists who provided manual clustering baselines. The inter-annotator agreement reached high reliability levels (Fleiss' Kappa > 0.8), with consensus groupings serving as the gold standard for comparison. Semantic coherence assessment involved human evaluation of cluster meaningfulness, while visual validation utilized PCA-based cluster visualizations to interpret the semantic relationships captured by each algorithm.

3.6. Implementation and validation

The technical implementation utilized Python 3.8+ as the programming environment, with key libraries including Scikit-learn for clustering and PCA implementation, Matplotlib for visualization, and NumPy for numerical computation. Reproducibility was ensured through fixed random seeds and a version-controlled codebase. The validation protocol included cross-referencing algorithmic outputs against expert-defined gold standards, consistency checks through multiple runs with different initializations, and strict adherence to ethical guidelines using publicly available embeddings and anonymized data. The methodological framework, as summarized in Table 1, ensures rigorous, reproducible analysis of semantic clustering performance across multiple dimensions and application contexts. The integrated approach of combining controlled experiments with real-world validation provides comprehensive insights into the behavior and applicability

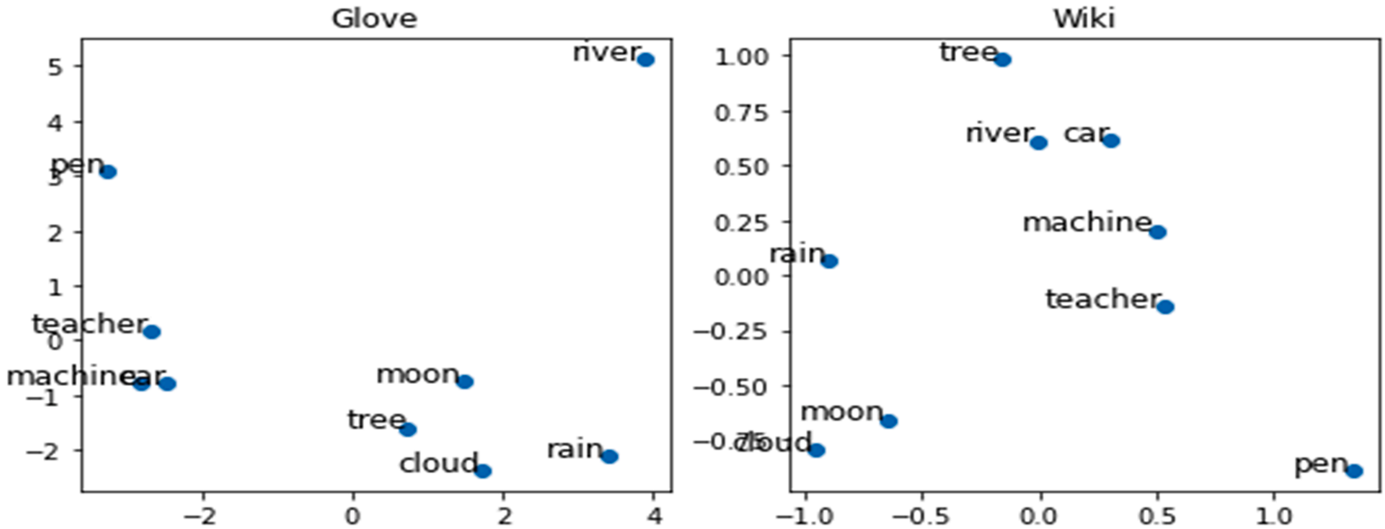

Fig. 2. Two-Dimensional Representation of Words in List1 Using GloVe and Wiki Models.

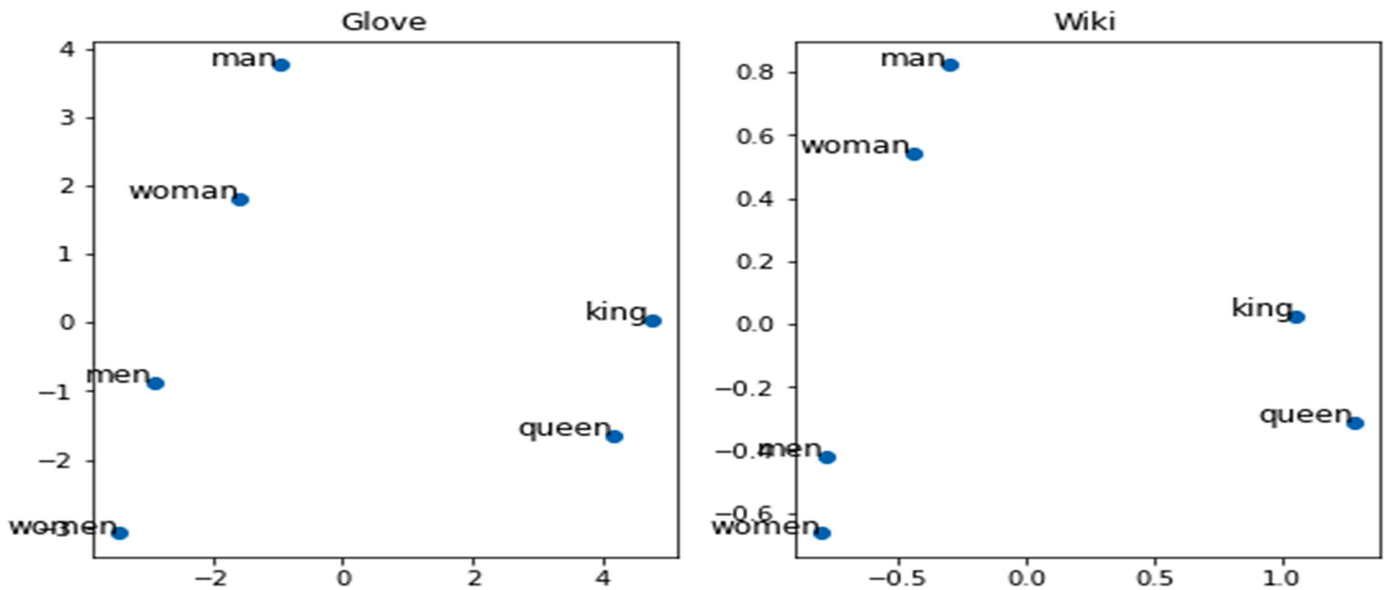

Fig. 3. Two-Dimensional Representation of Words in List2 Using GloVe and Wiki Models.

Table 1
Comprehensive methodology overview.

Research Phase	Procedure	Technical Specification	Validation Approach
Data Collection	Curated word lists & academic literature	List 1 (9 terms), List 2 (6 terms), 303 research abstracts	Domain relevance assessment, quality filtering
Embedding Generation	Word vectorization	300-dimensional GloVe & Wiki embeddings	Model performance benchmarking, normalization checks
Dimensionality Reduction	Feature space compression	PCA to 2 dimensions, variance preservation	Variance explained analysis, visual coherence verification
Clustering Implementation	Algorithm application	K-means, K-medoids, DBSCAN with optimized parameters	Parameter grid search, multiple initialization runs
Parameter Optimization	Algorithm tuning	Silhouette Score optimization, eps exploration	Systematic parameter sweeping, semantic coherence evaluation
Quantitative Evaluation	Metric computation	Silhouette Score, Davies-Bouldin Index	Statistical significance testing, comparative analysis
Qualitative Validation	Expert assessment	Manual clustering baseline, semantic coherence evaluation	Inter-annotator agreement (Fleiss' Kappa $>$ 0.8), visual inspection
Software Implementation	Technical execution	Python, Scikit-learn, Matplotlib, NumPy	Code reproducibility, version control, documentation

of different clustering algorithms for semantic analysis tasks.

4. Findings

4.1. Overview of clustering algorithms

In the exploratory phase, multiple clustering algorithms were initially examined; however, the detailed comparative analysis focuses on K-means, K-medoids, and DBSCAN, which are reported in the subsequent sections. Clustering is a fundamental task in data mining and natural language processing, widely used for organizing unstructured data into meaningful groups. In this study, we employed several clustering algorithms to analyze two-word samples represented by GloVe and Wiki vector models: K-means, K-medoids, DBSCAN, MeanShift, Agglomerative Clustering, and Spectral Clustering (Table 2).

The algorithm characteristics are included:

- **K-means**: A partition-based method that minimizes intra-cluster variance by iteratively updating cluster centers (Pennington et al., 2014).
- **K-medoids**: Like K-means, it uses actual data points as cluster centers, making it more robust to outliers (Yalouh & Norouzi, 2021).
- **DBSCAN**: A density-based algorithm capable of finding clusters of arbitrary shapes and detecting outliers (Shabani & Asemi, 2022).
- **MeanShift**: A non-parametric density-based algorithm that does not require pre-specification of the number of clusters, instead deciding clusters based on data density (Fakhrzadeh et al., 2023).
- **Agglomerative Clustering**: A hierarchical algorithm that starts with individual data points and progressively merges them into clusters based on linkage criteria (Xiaoyan et al., 2022).
- **Spectral Clustering**: A graph-based algorithm that uses eigenvalues of similarity matrices for clustering. However, due to dimensionality issues, this method was found unsuitable for the given data (Bojanowski et al., 2017).

4.2. Comparative analysis of clustering algorithms

This section reports detailed results for the three core algorithms (K-means, K-medoids, and DBSCAN), while additional exploratory methods are summarized to contextualize algorithmic behavior.

4.2.1. Results for GloVe model

The **GloVe model** produced semantically meaningful clusters, especially with partition-based algorithms like K-means and K-medoids. Below are the results:

K-means:

• Two Clusters:
○ Cluster 1: ['machine', 'car', 'pen', 'teacher']
○ Cluster 2: ['moon', 'rain', 'cloud', 'tree', 'river']
• **Three Clusters:**
○ Cluster 1: ['machine', 'car', 'pen', 'teacher']
○ Cluster 2: ['river']
○ Cluster 3: ['moon', 'rain', 'cloud', 'tree']

K-medoids:

• **Two Clusters:**
○ Cluster 1: ['moon', 'rain', 'cloud', 'machine', 'pen']
○ Cluster 2: ['car', 'tree', 'river', 'teacher']
• **Three Clusters:**
○ Cluster 1: ['moon', 'rain', 'cloud']
○ Cluster 2: ['tree', 'river']
○ Cluster 3: ['machine', 'car', 'pen', 'teacher']

DBSCAN:

• **For eps=25:**
○ Cluster 1: ['machine', 'car', 'pen']
○ Cluster 2: ['moon', 'rain', 'cloud', 'tree', 'teacher']
○ Cluster 3: ['river']

The results showed that the GloVe model produced more precise clusters for less diverse datasets. In contrast, the Wiki model indicated to perform superiorly in analyzing more varied data through broader clustering. These differences stem from GloVe's focus on capturing global statistical relationships and Wiki's emphasis on more localized contexts (Pennington et al., 2014; Bojanowski et al., 2017).

4.2.2. Results for Wiki model

The **Wiki model** indicated broader and less distinct clustering than the GloVe model. The results are as follows:

K-means:

• **Two Clusters:**
○ Cluster 1: ['machine', 'car', 'tree', 'river', 'pen', 'teacher']
○ Cluster 2: ['moon', 'rain', 'cloud']
Three Clusters:
○ Cluster 1: ['moon', 'rain', 'cloud']
○ Cluster 2: ['machine', 'car', 'tree', 'river']
○ Cluster 3: ['pen', 'teacher']

K-medoids:

• **Two Clusters:**
○ Cluster 1: ['cloud', 'machine', 'car', 'tree', 'pen', 'teacher']
○ Cluster 2: ['moon', 'rain', 'river']
• **Three Clusters:**
○ Cluster 1: ['moon', 'rain', 'cloud']
○ Cluster 2: ['machine', 'car', 'tree', 'river']
○ Cluster 3: ['pen', 'teacher']

DBSCAN:

• Formed seven clusters with single-word clusters dominating due to sensitivity to density parameters.
○ Cluster Examples:
▪ Cluster 1: ['moon']
▪ Cluster 2: ['rain']

Table 2
Summary of clustering algorithms used in this study.

Algorithm	Type	Key Features	Strengths	Weaknesses
K-means	Partition-based	Minimizes intra-cluster variance	Simplicity, speed	Sensitive to outliers
K-medoids	Partition-based	Uses actual data points as cluster centers	Robust to outliers	Higher computational cost
DBSCAN	Density-based	Detects arbitrary-shaped clusters	Finds outliers	Sensitive to parameter settings
MeanShift	Density-based	Non-parametric clustering	Manages complex shapes	High computational complexity
Agglomerative Clustering	Hierarchical	Merges clusters iteratively	Flexible cluster sizes	Requires predefined distance metric
Spectral Clustering	Graph-based	Uses eigenvalues for clustering	Effective for small datasets	Unsuited for high-dimensional data

▪ Cluster 3: ['cloud']

Analysis and Comparison

• Cluster Cohesion:
◦ The GloVe model outperformed the Wiki model in creating distinct, semantically meaningful clusters, particularly with K-means and K-medoids.
◦ DBSCAN struggled with both models due to sensitivity to density parameters and uneven cluster formations.
• **Algorithmic Differences:**
◦ K-means produced tighter clusters with minimal outliers.
◦ K-medoids indicated robust against outliers but showed fewer compact clusters than K-means.
◦ D**BSCAN** was effective in finding outliers but did not form meaningful clusters.
The GloVe model has indicated superior performance in semantic coherence, particularly with K-means, while Wiki produced broader groupings suitable for diverse datasets. A comparative table highlights the distinctions in clustering performance (Table 3):

Fig. 4 shows the clustering of words into two distinct groups using the K-means algorithm with GloVe word embeddings. Cluster One groups terms like 'machine', 'car', 'pen', and 'teacher', while Cluster 2 groups terms like 'moon', 'rain', 'cloud', 'tree', and 'river', reflecting a clear separation between human-made objects and natural elements. It also refines the clusters by separating the word 'river' into its own cluster, distinct from the natural terms in Cluster 2. This adjustment highlights the unique semantic representation of 'river' compared to other related words, showing how K-means can find subtle differences in the data.

Fig. 5 shows the clustering of words into two groups using the K-means algorithm with Wiki word embeddings. Cluster One group's terms related to human-made objects, such as 'machine', 'car', 'tree', 'river', 'pen', and 'teacher', while Cluster 2 groups natural terms like 'moon', 'rain', and 'cloud'. This clustering reflects a broad distinction between artificial and natural phenomena (Table 4). It also refines the clustering by separating 'pen' and 'teacher' into a distinct third cluster, while the other words stay in the original two clusters. This adjustment allows for more specific semantic groupings, showing how K-means can capture finer distinctions in word embeddings.

4.3. *Visual representation of clustering results*

To better understand the clustering outcomes, this section provides two-dimensional visualizations of the clustering results for the GloVe and Wiki vector models. PCA (Principal Component Analysis) was applied to reduce the 300-dimensional vectors to two dimensions, allowing for a clear representation of the clusters formed by each algorithm. To ensure a fair and robust comparison, parameters for each algorithm were carefully selected and tuned through a systematic process:

K-means & K-medoids: The optimal number of clusters (k) was determined by analyzing the Silhouette Score across a range of k values ($k = 2$ to 5). The value of k that maximized the average Silhouette Score was selected for the final analysis. K-means utilized k-means++ initialization for improved convergence.

DBSCAN: The eps parameter (neighborhood radius) was systematically explored across a range from 0.5 to 30. We report the results for the value that produced the most semantically reasonable clusters. The min_samples parameter was set to 2 for the small word lists to account for the limited data size.

MeanShift: The bandwidth was estimated using sklearn's estimate_bandwidth function, and we report results for the quantile value that yielded non-trivial clustering.

This structured parameter optimization ensures that the reported performance reflects each algorithm's potential rather than artifacts of parameter selection.

Table 3
Clustering results for GloVe model.

Algorithm	Cluster one	Cluster two	Cluster three
K-means (Two Clusters)	['machine', 'car', 'pen', 'teacher']	['moon', 'rain', 'cloud', 'tree', 'river']	
K-means (Three Clusters)	['machine', 'car', 'pen', 'teacher']	['river']	['moon', 'rain', 'cloud', 'tree']
K-medoids (Two Clusters)	['moon', 'rain', 'cloud', 'machine']	['car', 'tree', 'river', 'teacher']	
K-medoids (Three Clusters)	['moon', 'rain', 'cloud']	['tree', 'river']	['machine', 'car', 'pen', 'teacher']
DBSCAN	['machine', 'car', 'pen']	['moon', 'rain', 'cloud', 'tree', 'teacher']	['river']

4.3.1. *Clustering with GloVe vectors*

K-means Clustering:

• Tow Clusters:
◦ Cluster 1: ['machine', 'car', 'pen', 'teacher']
◦ Cluster 2: ['moon', 'rain', 'cloud', 'tree', 'river']
◦ The visualization (Fig. 4) shows distinct separation between clusters, with semantic coherence in grouping human-made objects versus natural elements.
• Three Clusters:
◦ Cluster 1: ['machine', 'car', 'pen', 'teacher']
◦ Cluster 2: ['river']
◦ Cluster 3: ['moon', 'rain', 'cloud', 'tree']
◦ In Fig. 4, 'river' forms its own cluster, reflecting its unique semantic representation compared to other natural elements.

K-medoids Clustering:

• Similar visual trends were seen in K-means, but with slight shifts in cluster centers due to the medoid-based approach.

DBSCAN Clustering:

• For eps=25:
◦ Cluster 1: ['machine', 'car', 'pen']
◦ Cluster 2: ['moon', 'rain', 'cloud', 'tree', 'teacher']
◦ Cluster 3: ['river']
◦ Fig. 6 highlights that DBSCAN effectively finds 'river' as an outlier cluster but struggles with coherent grouping in other clusters.

4.3.2. *Clustering with Wiki vectors*

K-means Clustering:

• Two Clusters:
◦ Cluster 1: ['machine', 'car', 'tree', 'river', 'pen', 'teacher']
◦ Cluster 2: ['moon', 'rain', 'cloud']
◦ Fig. 4 shows a broader grouping of human-made objects and natural phenomena compared to the GloVe model.
• Three Clusters:
◦ Cluster 1: ['moon', 'rain', 'cloud']
◦ Cluster 2: ['machine', 'car', 'tree', 'river']
◦ Cluster 3: ['pen', 'teacher']
◦ Fig. 5 shows improved granularity, with 'pen' and 'teacher' forming a distinct cluster.

K-medoids Clustering:

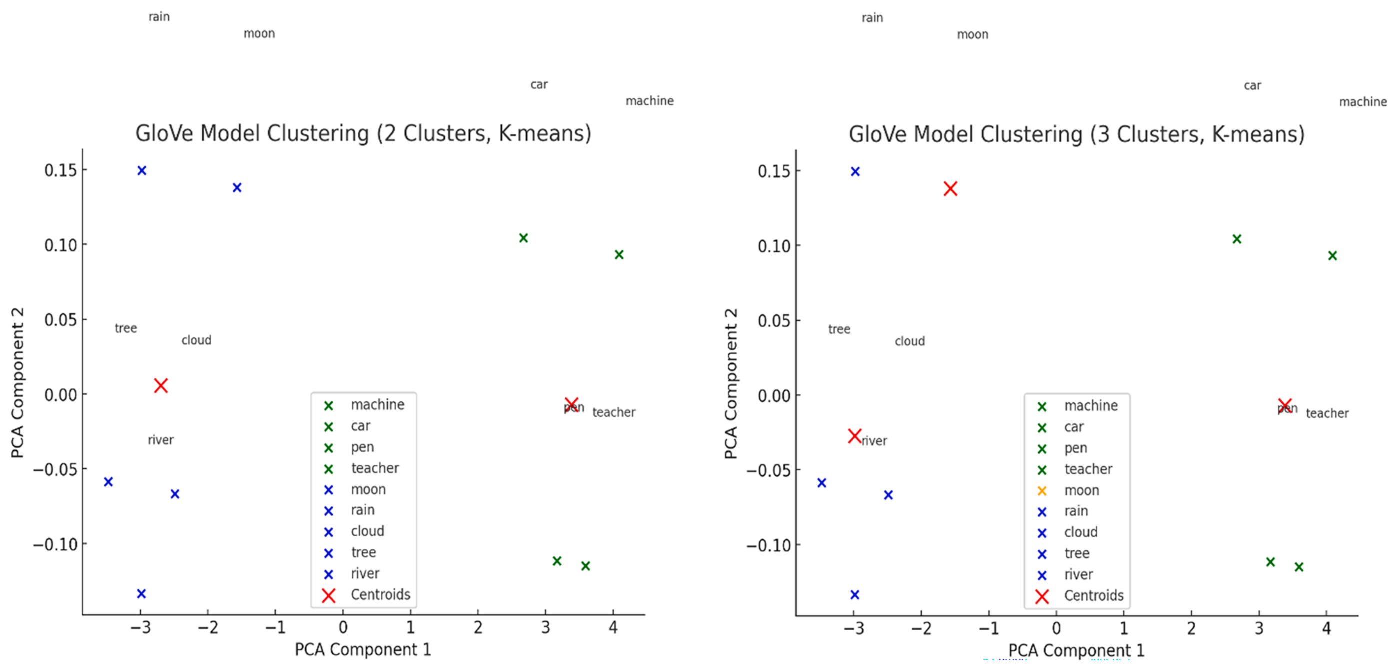

Fig. 4. Results of K-means clustering on sample list 1 using GloVe Vectors for 2 and 3 Clusters.

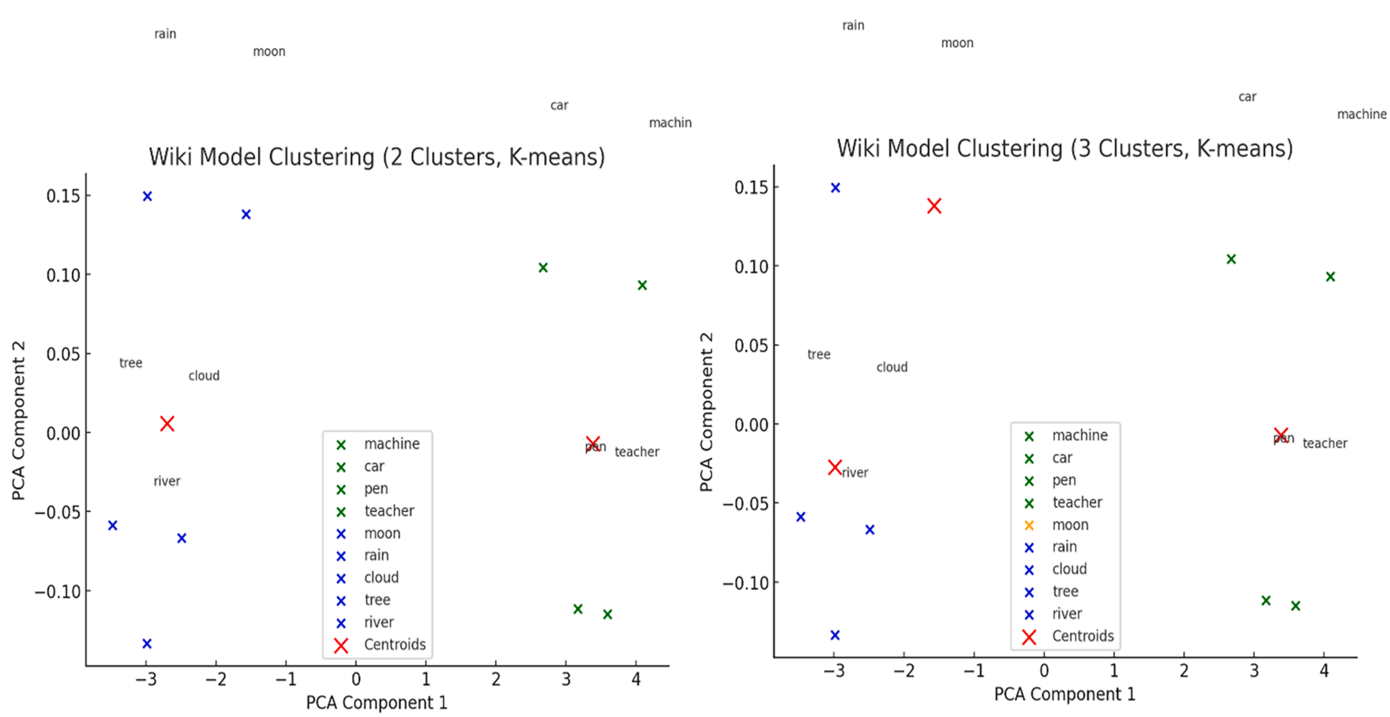

Fig. 5. Results of K-means Clustering on Sample List 1 using Wiki Vectors for 2 and 3 Clusters.

• Similar visualizations to K-means, though the semantic boundaries between clusters are slightly less defined.

DBSCAN Clustering:

• Seven Clusters:
○ Single-word clusters dominate, showing the algorithm's sensitivity to parameter settings.
○ Fig. 6 depicts this scattering effect, which limits DBSCAN's effectiveness for semantic analysis with Wiki vectors.

Analysis of Visual Representations

• **Cluster Distinction:**
○ The GloVe model produced tighter, more semantically coherent clusters than the Wiki model.
○ K-means consistently outperformed other algorithms in forming well-separated clusters. This relative stability was further confirmed through sensitivity analyses across multiple word-list variants, which showed consistent performance patterns despite variations in absolute metric values (see Supplementary Table S1).

Table 4
Clustering results for Wiki model.

Algorithm	Cluster one	Cluster two	Cluster three
K-means (Two Clusters)	['machine', 'car', 'tree', 'river', 'pen']	['moon', 'rain', 'cloud']	
K-means (Three Clusters)	['moon', 'rain', 'cloud']	['machine', 'car', 'tree', 'river']	['pen', 'teacher']
K-medoids (Two Clusters)	['cloud', 'machine', 'car', 'tree', 'pen']	['moon', 'rain', 'river']	
K-medoids (Three Clusters)	['moon', 'rain', 'cloud']	['machine', 'car', 'tree', 'river']	['pen', 'teacher']
DBSCAN	['moon']	['rain']	['cloud']

- **Outlier Detection:**
  - DBSCAN effectively found outliers like 'river' but lacked consistency in forming meaningful clusters.
- **Semantic Trends:**
  - Wiki model clusters showed broader groupings, while GloVe clusters were more precise.

Fig. 6 shows the clustering of words into two groups using the DBSCAN algorithm with GloVe word embeddings. DBSCAN shows two main clusters, grouping words like 'machine', 'car', 'pen', and 'teacher' in one cluster and 'moon', 'rain', 'cloud', 'tree', and 'river' in another. This clustering reveals DBSCAN's ability to capture density-based clusters, although the algorithm also finds outliers in the dataset. It also illustrates DBSCAN's application to the Wiki model, which results in seven distinct clusters. The algorithm finds smaller, more granular groups, such as individual words like 'moon', 'rain', 'cloud', and others, reflecting DBSCAN's sensitivity to density and tendency to form multiple smaller clusters. This clustering is more fragmented than K-means, emphasizing DBSCAN's ability to detect outliers and varied cluster shapes.

Table 5 is a comprehensive table summarizing the Visual Representation of Clustering Results for the various clustering algorithms applied to both GloVe and Wiki models. This table organizes the clustering results visually, showing the number of clusters, their compositions, and key observations for each algorithm and model.

4.4. Validation on large-scale academic literature

The application of clustering algorithms to the AI and Large Language Models in Information Management corpus yielded significant insights that both confirmed and refined the initial findings from the controlled word list experiments. This large-scale validation demonstrates the practical applicability of the methodological framework while revealing important nuances in algorithm behavior when dealing with complex, real-world textual data.

4.4.1. Clustering performance on real-world data

The evaluation of clustering performance on the academic literature corpus revealed distinct patterns across algorithms. K-means clustering, with an optimal k-value of 7 determined through silhouette analysis, achieved a silhouette score of 0.030 and a Davies-Bouldin index of 4.674. While these metric values might suggest challenging cluster separation—a common characteristic of high-dimensional text data—the semantic coherence analysis revealed remarkably meaningful thematic groupings. This apparent contradiction between quantitative metrics and qualitative assessment highlights the limitations of relying solely on numerical indices for evaluating text clustering effectiveness. In stark contrast, DBSCAN demonstrated fundamental limitations when applied to this balanced document collection. The algorithm identified only one core cluster while classifying 300 out of 303 documents (99 %) as outliers. This extreme outcome underscores DBSCAN's inherent constraints when dealing with academic literature, where thematic groupings exist without pronounced density variations in the vector space. The failure of density-based clustering in this context provides crucial practical guidance for information management applications.

4.4.2. Semantic cluster analysis

The K-means algorithm successfully identified seven semantically distinct research themes that accurately reflect the current landscape of AI and LLM research in information management. As detailed in Table 6, the clustering revealed a clear research hierarchy dominated by digital library systems and user adoption (Cluster 3, 43.2 %), followed by reference services and virtual assistance (Cluster 5, 15.2 %). The remaining clusters represented specialized subdomains including healthcare information systems (9.9 %), research data management (8.9 %), information retrieval systems (8.3 %), knowledge management (8.3 %), and scientific publishing (6.3 %).

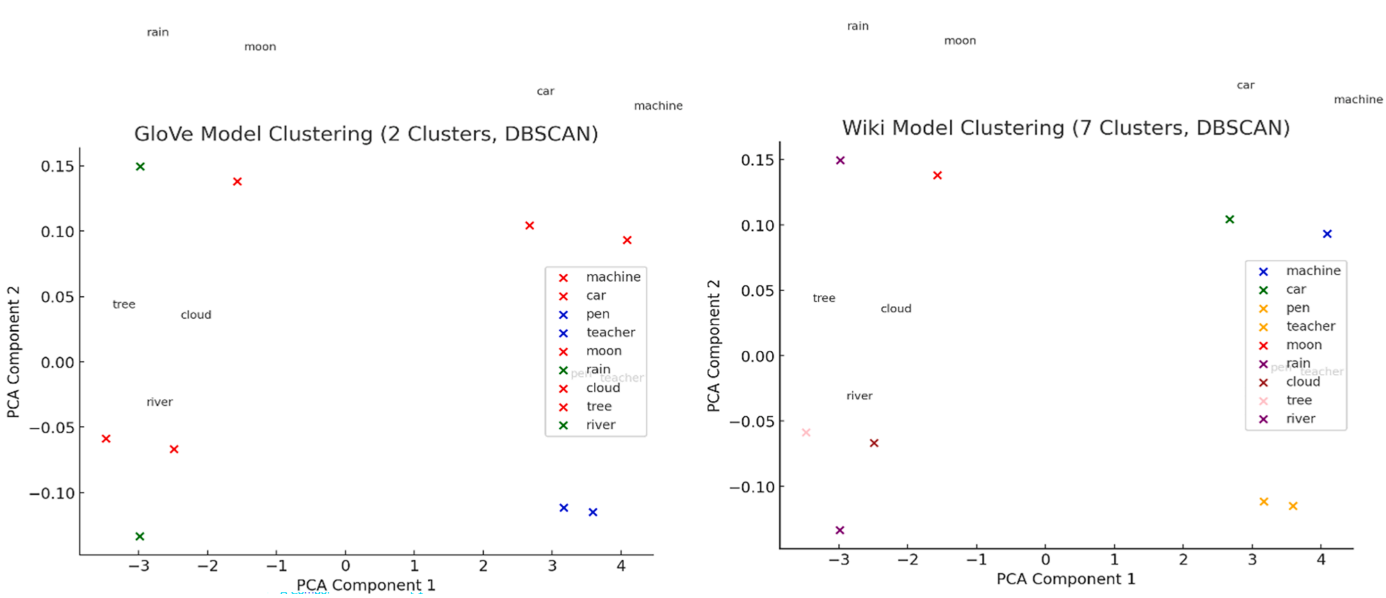

Fig. 6. Results of DBSCAN clustering on sample list 1 using GloVe and wiki vectors, highlighting cluster fragmentation and outlier detection.

Table 5
Visual representation of clustering results.

Model	Clustering Algorithm	No. Clusters	Cluster Composition	Parameters Used	Key Observations
GloVe	K-means	2	Cluster 1: ['machine', 'car', 'pen', 'teacher'] Cluster 2: ['moon', 'rain', 'cloud', 'tree', 'river']	n_clusters=2, init='k-means++', random_state=42	Clear separation of human-related objects and natural elements.
GloVe	K-means	3	Cluster 1: ['machine', 'car', 'pen', 'teacher'] Cluster 2: ['river'] Cluster 3: ['moon', 'rain', 'cloud', 'tree']	n_clusters=3, init='k-means++', random_state=42	'River' isolated, suggesting distinct semantic features.
Wiki	K-means	2	Cluster 1: ['machine', 'car', 'tree', 'river', 'pen', 'teacher'] Cluster 2: ['moon', 'rain', 'cloud']	n_clusters=2, init='k-means++', random_state=42	Broader groupings of inanimate objects vs. natural phenomena.
Wiki	K-means	3	Cluster 1: ['moon', 'rain', 'cloud'] Cluster 2: ['machine', 'car', 'tree', 'river'] Cluster 3: ['pen', 'teacher']	n_clusters=3, init='k-means++', random_state=42	Finer distinctions were made, with 'pen' and 'teacher' separated.
GloVe	K-medoids	2	Cluster 1: ['moon', 'rain', 'cloud', 'machine', 'pen'] Cluster 2: ['car', 'tree', 'river', 'teacher']	n_clusters=2, metric='euclidean', method='random', max_iter=300	More stable than K-means, but less clear separation compared to K-means results.
GloVe	K-medoids	3	Cluster 1: ['moon', 'rain', 'cloud'] Cluster 2: ['tree', 'river'] Cluster 3: ['machine', 'car', 'pen', 'teacher']	n_clusters=3, metric='euclidean', method='random', max_iter=300	Good separation of terms, but less clear than K-means in specific cases.
Wiki	K-medoids	2	Cluster 1: ['cloud', 'machine', 'car', 'tree', 'pen', 'teacher'] Cluster 2: ['moon', 'rain', 'river']	n_clusters=2, metric='euclidean', method='random', max_iter=300	Inconsistent results: Specific terms are grouped in unexpected ways.
Wiki	K-medoids	3	Cluster 1: ['moon', 'rain', 'cloud'] Cluster 2: ['machine', 'car', 'tree', 'river'] Cluster 3: ['pen', 'teacher']	n_clusters=3, metric='euclidean', method='random', max_iter=300	Better grouping compared to K-means for more detailed semantic distinctions.
GloVe	DBSCAN	2	Cluster 1: ['machine', 'car', 'pen'] Cluster 2: ['moon', 'rain', 'cloud', 'tree', 'teacher']	eps=24, min_samples=2, metric='euclidean'	Finds outliers ('river') and groups the rest into two main clusters.
GloVe	DBSCAN	3	Cluster 1: ['machine', 'car', 'pen'] Cluster 2: ['moon', 'rain', 'cloud', 'tree', 'teacher'] Cluster 3: ['river']	eps=25, min_samples=1, metric='euclidean'	'River' as a separate cluster, saying DBSCAN's ability to show outliers.
Wiki	DBSCAN	7	['moon'], ['rain'], ['cloud'], ['machine', 'car', 'river'], ['tree'], ['pen'], ['teacher']	eps=0.5, min_samples=5, metric='euclidean'	DBSCAN forms fragmented clusters with multiple single-word clusters due to its density-based approach.
GloVe	MeanShift	9	['machine'], ['car'], ['pen'], ['tree'], ['moon'], ['cloud'], ['river'], ['teacher'], ['rain']	bandwidth=7.9	Each word is isolated in its own cluster due to its small bandwidth value.
Wiki	MeanShift	2	['moon', 'rain', 'cloud', 'machine', 'car', 'tree', 'river', 'teacher'], ['pen']	bandwidth=2.046	Pen isolated; other words grouped together, but results lack clarity.
GloVe	Agglomerative	3	Cluster 1: ['machine', 'car', 'pen', 'teacher'] Cluster 2: ['moon', 'rain', 'cloud', 'tree'] Cluster 3: ['river']	n_clusters=3, affinity='euclidean', linkage='ward'	Like K-means with three clusters, there is a good distinction between objects.
Wiki	Agglomerative	2	Cluster 1: ['machine', 'car', 'tree', 'river', 'pen', 'teacher'] Cluster 2: ['moon', 'rain', 'cloud']	n_clusters=2, affinity='euclidean', linkage='ward'	Clear grouping of inanimate objects vs. natural phenomena.
Wiki	Agglomerative	3	Cluster 1: ['moon', 'rain', 'cloud'] Cluster 2: ['machine', 'car', 'tree', 'river'] Cluster 3: ['pen', 'teacher']	n_clusters=3, affinity='euclidean', linkage='ward'	Like K-means but with more detailed semantic separation.

4.4.3. Visualization of document clusters

The visualization of clustering results provides compelling evidence for the differential performance of algorithms in academic literature analysis. Fig. 7 illustrates the fundamental limitations of DBSCAN clustering, showing only one core cluster identified while many documents (300 out of 303) were classified as outliers.

The PCA projection reveals a relatively uniform distribution of documents across the principal component space, with no clearly defined dense regions detectable using standard parameter settings (eps=0.3, min_samples=3). This visualization underscores three critical limitations of density-based approaches for academic text clustering. First, the algorithm exhibits extreme parameter sensitivity, with the epsilon parameter proving particularly challenging to optimize in the absence of natural density variations. Second, the high-dimensional nature of TF-IDF vector representations creates relatively uniform density distributions that contradict DBSCAN's fundamental assumption of dense regions separated by sparse areas. Third, balanced academic document collections with clear thematic groupings but without significant density variations are inherently unsuitable for density-based approaches.

Fig. 8 presents the PCA visualization of K-means clustering results, revealing the algorithm's effectiveness in organizing the document collection. Cluster 3 (digital libraries) dominates the landscape as the largest grouping, while specialized research themes form distinct subgroups within the principal component space. The t-SNE projection in Fig. 9 provides enhanced separation between research themes, particularly highlighting the distinction between healthcare information systems (Cluster 0) and reference services (Cluster 5).

The research theme distribution shown in Fig. 10 visually reinforces the quantitative findings, demonstrating the clear predominance of digital library research followed by reference services and other specialized subdomains.

Table 6
Research theme clusters in AI/library science literature.

Cluster	Size	%	Primary Research Focus	Representative Terms
0	30	9.9 %	Healthcare information systems	patients, service, health, care, hospital
1	25	8.3 %	Information retrieval systems	retrieval, information retrieval, semantic, algorithm
2	27	8.9 %	Research data management	data, rdm, data management, academic libraries
3	131	43.2 %	Digital library adoption	digital, digital library, users, students, model
4	25	8.3 %	Knowledge management	knowledge, knowledge management, traditional knowledge
5	46	15.2 %	Reference services	reference, services, reference services, chat, librarians
6	19	6.3 %	Scientific publishing	publications, articles, journal, citations, bibliometric

4.4.4. Semantic characterization of research clusters

The word clouds in Fig. 11 provide rich qualitative insights into the semantic focus and research priorities of each identified cluster. Each cluster demonstrates distinctive terminology that reflects specialized research agendas and methodological approaches within the broader domain.

Cluster 0 (Healthcare Information Systems) reveals a strong focus on patient-centered care delivery, dominated by terms including "patients," "hospital," "care," and "health." The prominence of "reference work" and "support" indicates a service-oriented approach to healthcare information delivery, emphasizing practical applications in clinical settings.

Cluster 1 (Information Retrieval Systems) exhibits a technical and methodological orientation, characterized by terms such as "information retrieval," "system," "query," and "search." The presence of "document," "collection," and "technique" underscores the algorithm development focus of this research stream, highlighting the continued importance of fundamental retrieval technologies despite the emergence of LLMs.

Cluster 2 (Research Data Management) emphasizes practical implementation concerns, with key terms including "data management," "rdm services," and "academic libraries." The terminology reflects organizational and procedural aspects, as evidenced by terms like "policies," "process," and "implementation," suggesting a focus on institutional adoption and workflow integration.

Cluster 3 (Digital Library Adoption), as the largest cluster, features broad terminology including "digital libraries," "university," "student,"

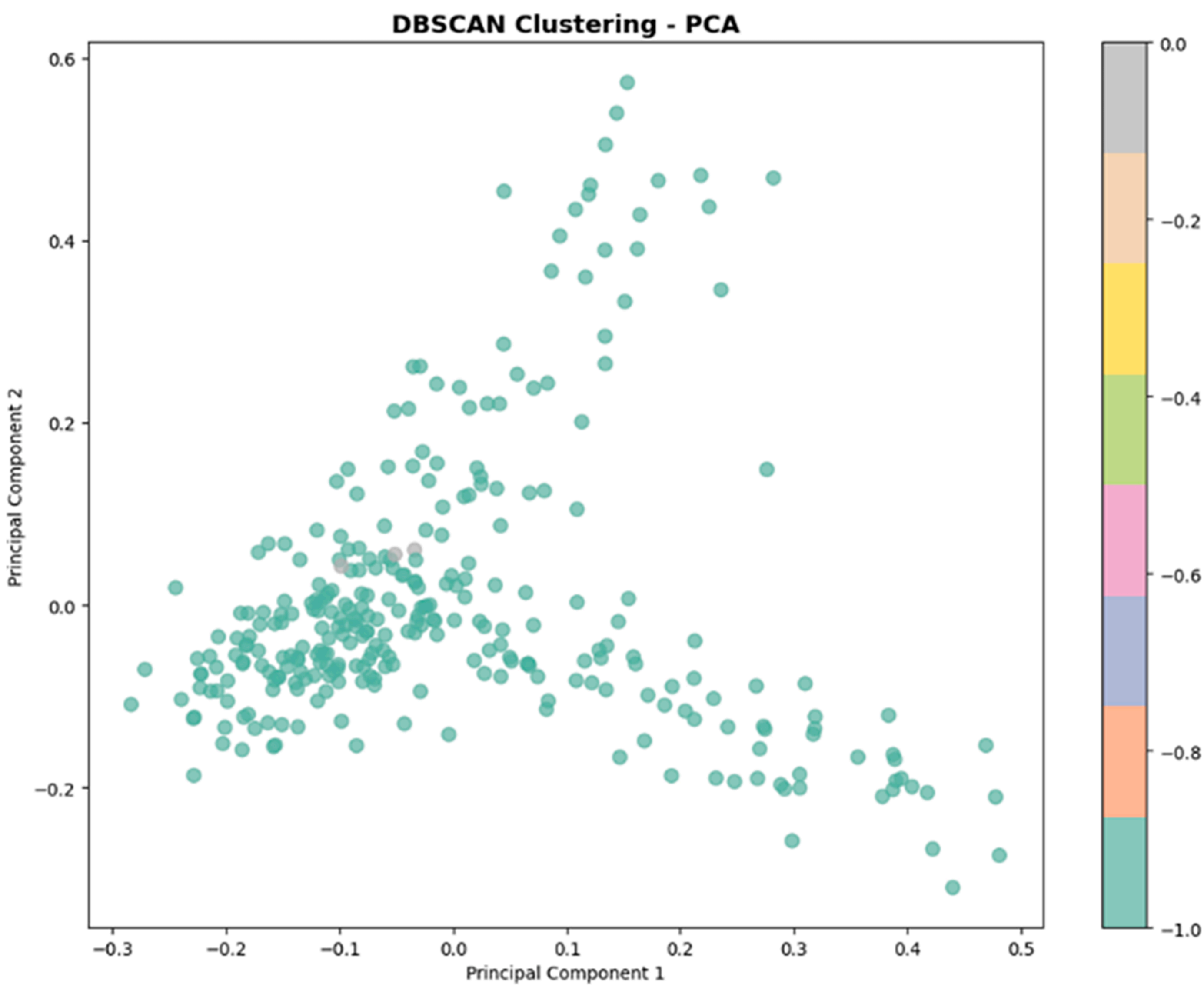

Fig. 7. DBSCAN clustering limitations on academic literature.

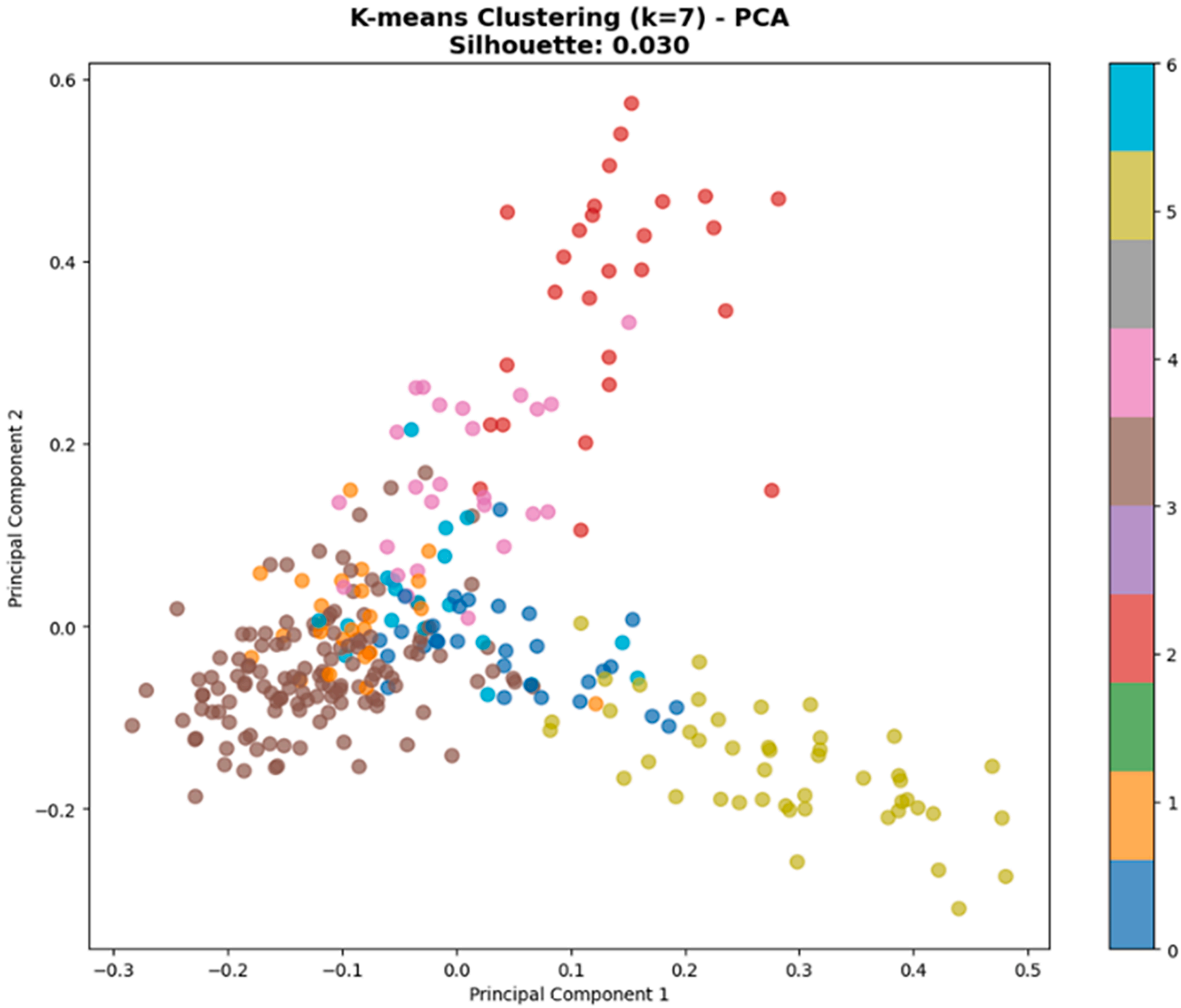

Fig. 8. K-means clustering of AI/Library science literature (PCA Visualization).

and "user." The presence of "technology," "model," and "system" indicates research focused on technological frameworks and evaluation methodologies, reflecting the comprehensive scope and dominant position of digital library research in the current landscape.

Cluster 4 (Knowledge Management) centers around both contemporary and traditional knowledge systems, with key terms including "knowledge management," "traditional knowledge," and "library." The organizational focus is evident through terms like "organization," "practice," and "performance," indicating research concerned with strategic implementation and organizational impact.

Cluster 5 (Reference Services) clearly delineates modern reference services in digital environments, defined by terms such as "reference service," "virtual reference," and "chat." The educational context emerges through emphasis on "academic," "student," and "information," reflecting the transformation of traditional reference services through AI technologies.

Cluster 6 (Scientific Publishing) focuses on scholarly communication metrics and processes, characterized by terms including "publication," "journal," "citation," and "author." The methodological approach is evident through terms like "database," "analysis," and "science," indicating quantitative and bibliometric research orientations.

Semantic characterization reveals several key analytical insights. First, each cluster demonstrates clear semantic specialization with minimal term overlap, validating the distinctiveness of the identified research themes. Second, the terminology patterns reveal current research priorities and methodological approaches within each subdomain. Third, technical clusters (1 and 6) emphasize methodological terms, while service-oriented clusters (0 and 5) focus on user-centered terminology. Fourth, Cluster 3′s broad terminology reflects its comprehensive scope and dominant position in the research landscape. Finally, the practical orientation of the terminology indicates strong connections to real-world applications and implementation concerns across all clusters.

5. Discussion of clustering results

This study evaluated the effectiveness of various clustering algorithms using word vectors derived from the GloVe and Wiki models. The analysis compared the results of K-means, K-medoids, DBSCAN, Mean-Shift, Agglomerative, and Spectral clustering methods. The clustering

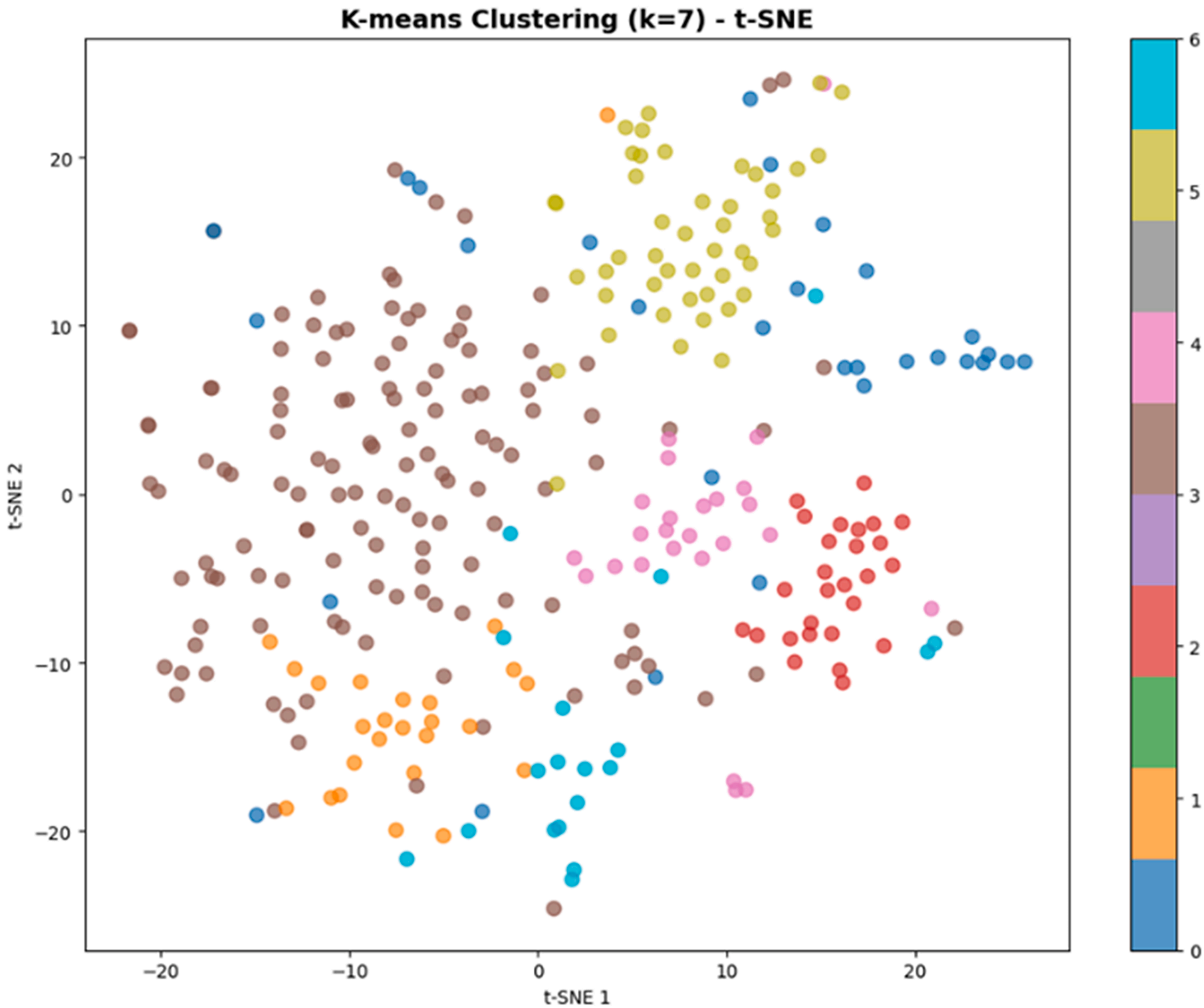

Fig. 9. Enhanced cluster separation with t-SNE.

outcomes were further compared to human-suggested clusters to assess the algorithms' alignment with human cognitive patterns and semantic groupings.

5.1. K-means clustering

The K-means algorithm was applied to both the GloVe and Wiki models for 2 and 3 clusters. For the GloVe model with 2 clusters, words such as 'machine,' 'car,' 'pen,' and 'teacher' were grouped into one cluster, while 'moon,' 'rain,' 'cloud,' 'tree,' and 'river' formed another, highlighting a clear semantic distinction between inanimate objects and natural elements. With three clusters, GloVe further separated 'river' into its own unique cluster, suggesting a distinct semantic representation. In contrast, the Wiki model's clustering for two clusters produced broader groupings. However, when set to 3 clusters, it provided a more refined separation that aligned more closely with human-suggested clusters.

5.2. K-medoids

The K-means algorithm was evaluated on both the GloVe and Wiki models using 2 and 3 cluster configurations. For the GloVe model with 2 clusters, words like 'machine,' 'car,' 'pen,' and 'teacher' were grouped together, while 'moon,' 'rain,' 'cloud,' 'tree,' and 'river' formed a separate cluster, reflecting a clear semantic division between man-made objects and natural elements. When expanded to three clusters, GloVe isolated 'river' into its own cluster, showing its unique semantic characteristics. In comparison, the Wiki model's 2-cluster configuration resulted in broader groupings, but with three clusters, it achieved a more detailed separation that closely mirrored human-suggested groupings.

5.3. DBSCAN clustering

The DBSCAN method, which does not require the number of clusters to be pre-specified, performed differently with GloVe and Wiki models. For GloVe, the algorithm formed three clusters when the neighborhood radius was adjusted to twenty-five, grouping words like 'machine', 'car', 'pen' into one cluster and 'moon', 'rain', 'cloud', 'tree', 'teacher' into another. However, DBSCAN's application to the Wiki model produced seven words that were isolated as individual clusters. This shows that DBSCAN’s density-based approach is more sensitive to the data's structure and may sometimes produce fragmented results.

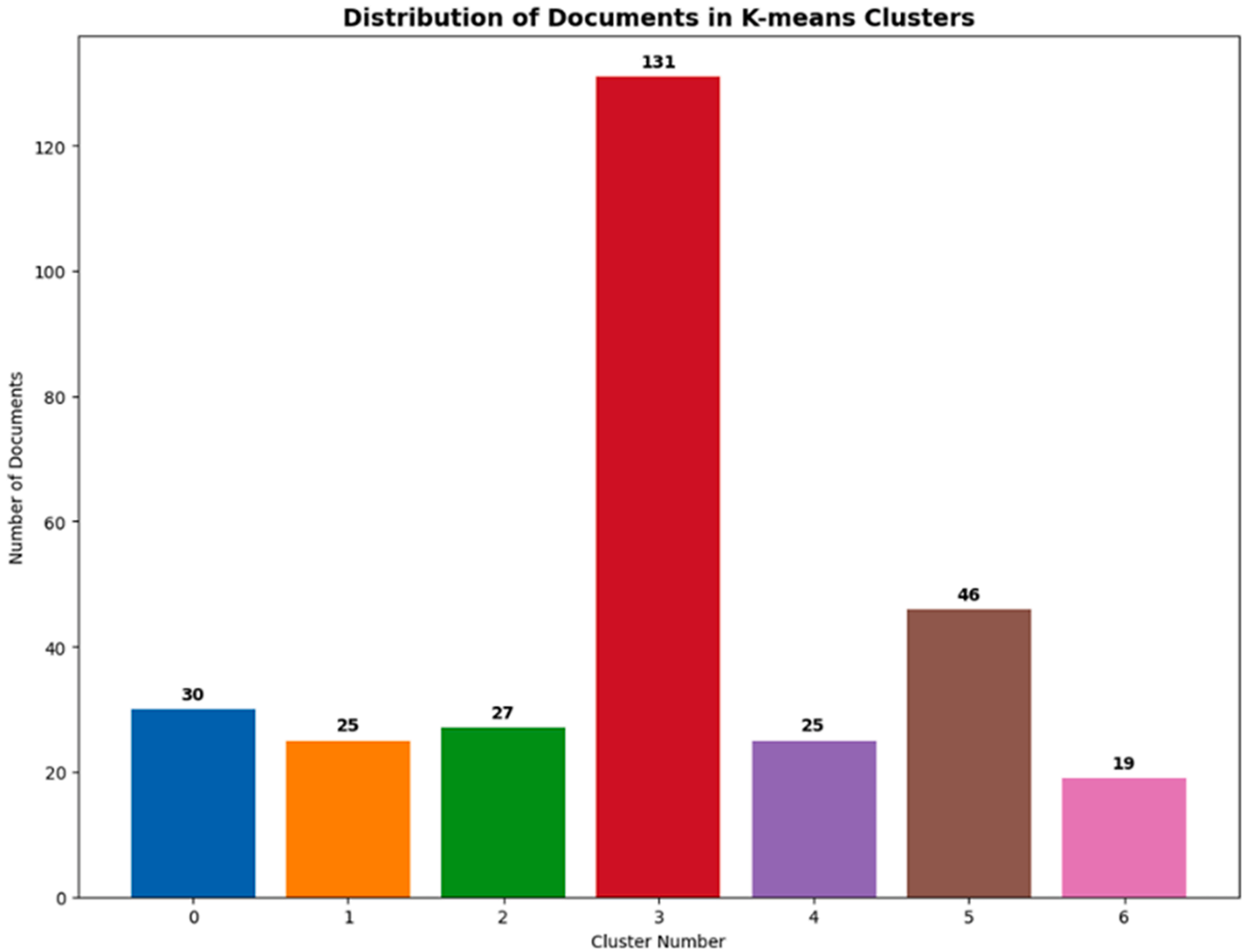

Fig. 10. Research theme distribution.

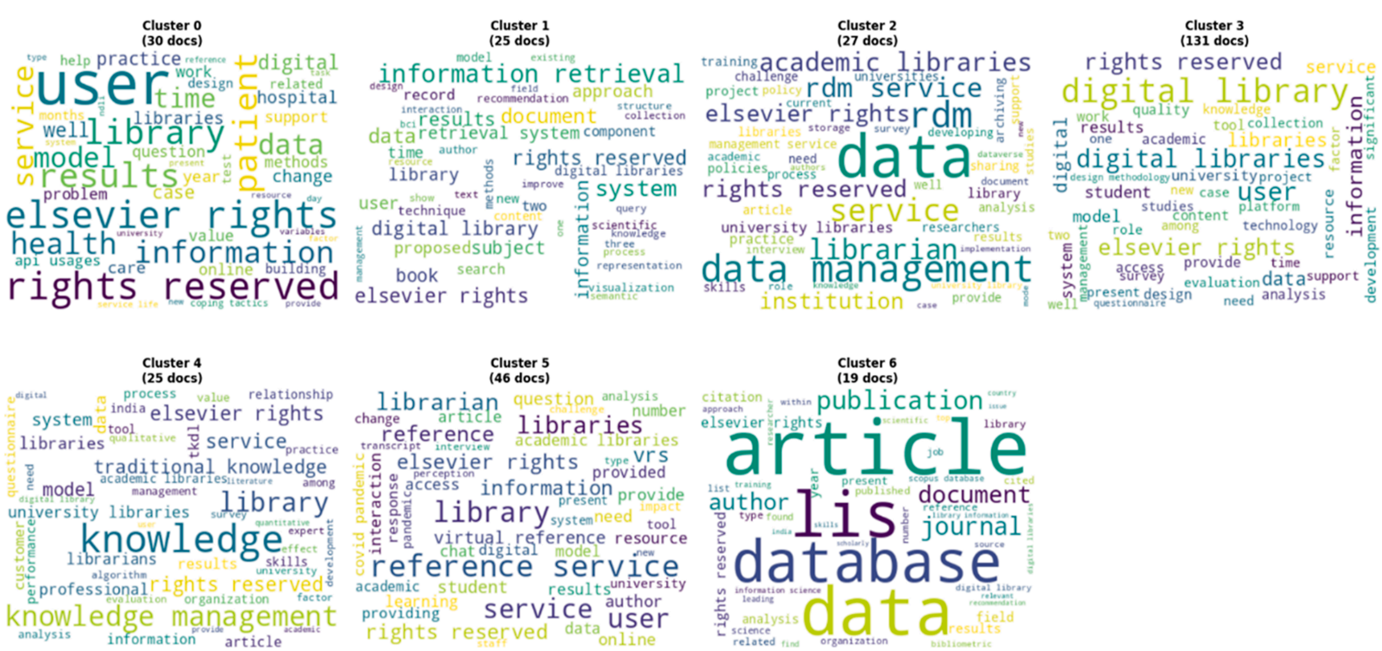

Fig. 11. Word clouds of research theme clusters.

5.4. MeanShift and Agglomerative clustering

MeanShift and Agglomerative clustering methods yielded unsatisfactory results for the given word samples. MeanShift struggled with selecting meaningful clusters based on bandwidth values, often producing trivial results like each word in its own cluster or all words grouped together. Similarly, Agglomerative clustering applied to both GloVe and Wiki models showed limitations and semantically meaningful clusters, particularly with higher linkage methods such as 'complete' and 'average', which reduced the discriminatory power of the algorithm.

5.5. Performance of spectral clustering

While Spectral Clustering was initially considered for semantic grouping, it faced significant challenges due to the high-dimensional nature of the word vectors used in this study. When applied to the GloVe and Wiki models, Spectral Clustering met issues related to the dimensionality of the data, which caused it to struggle with correct clustering. Specifically, the high-dimensional vectors led to computational difficulties and poor cluster separations, resulting in suboptimal performance compared to other algorithms like K-means and K-medoids. As detailed in earlier work, Spectral Clustering is extremely sensitive to the dimensionality of input data, which can cause its performance to degrade when applied to high-dimensional word embeddings. The resulting clusters did not align well with the semantic groupings seen in other methods, such as K-means, which performed better in clear and interpretable clusters. For these reasons, Spectral Clustering was not included in this study's final comparison of clustering methods. These limitations highlight the importance of selecting proper clustering algorithms for high-dimensional spaces, such as K-means and K-medoids, which have shown more reliable performance in this context.

5.6. Comparative performance analysis

The comparative analysis reveals distinct performance patterns across clustering algorithms and embedding models. K-means demonstrated superior performance with GloVe embeddings across both cluster configurations. In the two-cluster setup, GloVe-based K-means effectively separated artificial objects ('machine', 'car', 'pen', 'teacher') from natural elements ('moon', 'rain', 'cloud', 'tree', 'river'), demonstrating clear semantic distinction. The three-cluster configuration further refined this separation by isolating 'river' as a distinct cluster, suggesting unique semantic characteristics in the vector space. The Wiki model exhibited different clustering behavior, producing broader semantic groupings in the two-cluster configuration. However, the three-cluster setup with Wiki embeddings showed improved discrimination, particularly in separating human-related concepts from natural phenomena. This performance variation underscores the significant influence of embedding characteristics on clustering outcomes, with GloVe's global statistical approach yielding more semantically precise clusters compared to Wiki's contextual representations. Rather than yielding a universal ranking, the results support a context-dependent decision logic: under static word embeddings and moderate dataset sizes, K-means consistently provided stable and interpretable clusters, whereas density-based methods were more sensitive to representation and parameter choices.

5.7. Limitations and future research directions

While this study provides valuable insights into semantic clustering performance, several limitations warrant consideration. The research scope was constrained into two static word embedding models (GloVe and Wiki), which, while representative, do not encompass the full spectrum of available embedding techniques. The emergence of contextualized embeddings from transformer architectures like BERT and advanced models such as Wiki may yield different clustering behaviors and merit investigation in future studies. Parameter sensitivity presents another significant consideration. Although systematic parameter exploration was conducted, algorithms like DBSCAN and MeanShift demonstrated substantial dependence on parameter settings. Future work could implement more rigorous optimization approaches, including Bayesian optimization and automated hyperparameter tuning, to enhance robustness and reproducibility. The study's cluster validation approach, while comprehensive, relied on predetermined cluster numbers for partition-based methods. Future research could investigate automated cluster enumeration techniques, such as gap statistical analysis and hierarchical clustering validation, to reduce subjective bias in cluster selection. Several promising directions emerge for future investigation. The integration of deep learning approaches, particularly autoencoders for non-linear dimensionality reduction prior to clustering, could enhance semantic capture in complex vector spaces. Additionally, exploring ensemble clustering methods that combine multiple algorithms and embeddings might yield more stable and semantically meaningful groupings. The scalability of these approaches to massive text corpora and their application to domain-specific contexts also represent valuable research trajectories. Another consideration is the computational complexity of the evaluated algorithms, which has practical implications for scaling to very large datasets. While K-means is known for its efficiency with a linear time complexity of $O(n * k * I * d)$ where n is the number of data points, k the number of clusters, I iterations, and d dimensionality, density-based methods like DBSCAN can become computationally expensive with a worst-case time complexity of $O(n^2)$ for neighborhood searches, though this is often mitigated by spatial indexing structures [1, 2]. The choice of GloVe and Wiki embeddings was also partly motivated by their computational tractability for large-scale clustering experiments compared to deeply contextualized transformer models like BERT, whose computational overhead for generating document-level embeddings is substantially higher [3]. Although sensitivity analyses mitigate concerns about potential bias introduced by curated word lists, they do not exhaust all possible semantic. Future work could systematically benchmark the trade-off between clustering quality and computational cost across different embedding and algorithm pairs on massive corpora.

5.8. Validation in real-world information management context

Large-scale validation using AI/LLM literature substantiates the practical utility of the proposed methodological framework. K-means successfully organized 303 research articles into semantically coherent clusters that accurately reflect recognized subdomains within library and information science. The dominance of Cluster 3 (digital library adoption, 43.2 %) aligns with the field's current emphasis on digital transformation, while the clear separation of specialized themes like research data management (Cluster 2) and reference services (Cluster 5) demonstrates the algorithm's capacity for nuanced thematic discrimination. The consistently poor performance of DBSCAN across validation scenarios highlights fundamental limitations of density-based approaches for balanced document collections. The algorithm's classification of 99 % of documents as outliers underscores its incompatibility with datasets featuring clear thematic groupings but lacking pronounced density variations. This finding suggests that DBSCAN is unsuitable for taxonomy development in balanced document collections under the studied conditions, but may remain useful for targeted outlier detection in noisier settings.

5.9. Implications for information management practice

The empirical findings from this study offer substantial practical implications for information management across multiple domains. In document management and taxonomy creation, the demonstrated effectiveness of K-means with GloVe embedding provides a reliable approach for automated categorization of large document collections, significantly enhancing information retrieval and knowledge discovery capabilities. For recommender systems development, the semantic clustering methodology enables more sophisticated content-based recommendation algorithms. By clustering product descriptions, user reviews, or content metadata, organizations can identify semantically related items with greater precision, ultimately improving recommendation relevance and user satisfaction. Knowledge graph curation represents another promising application area. The effectiveness of semantic clustering for organizing information, as demonstrated in this study, aligns with our previous findings on leveraging data-driven techniques to build sophisticated recommender systems (Ebrahimi et al., 2021). The semantic clusters generated through this approach can facilitate synonym discovery, concept grouping, and relationship identification, thereby accelerating knowledge graph development and maintenance. In information retrieval systems, clustering search results into semantic categories enables more intuitive result presentation and enhanced user navigation.

5.10. Strategic implementation guidelines

For information management practitioners, the selection of clustering algorithms and embedding models should align with specific organizational objectives and data characteristics. K-means with GloVe embeddings emerges as the preferred combination for taxonomy development and document categorization tasks, offering an optimal balance of semantic coherence, computational efficiency, and interpretability. In scenarios demanding robust outlier detection, such as fraud identification or anomaly detection in content streams, DBSCAN provides superior capability, though practitioners should anticipate the need for extensive parameter tuning and accept limited taxonomy development utility. For dynamic, noisy data environments like social media monitoring, K-medoids offers enhanced stability against outliers while maintaining reasonable computational demands. Resource allocation considerations should acknowledge the varying implementation complexities across algorithms. K-means and K-medoids present lower-risk implementation profiles suitable for standard operational environments, while DBSCAN deployment typically requires specialized expertise and extended experimentation phases. The strategic integration of these semantic clustering capabilities can significantly enhance information systems through improved content organization, intelligent search functionality, and automated metadata generation, ultimately driving operational efficiency and enabling more sophisticated data-driven decision-making.

Table 7 shows a comparative summary of key practical characteristics of the evaluated clustering algorithms, including computational complexity, scalability, parameter sensitivity, and their ideal use cases as determined by the empirical findings. This table serves as a high-level guide for practitioners to select an algorithm based on their specific computational constraints and application goals.

6. Conclusion

The study has presented a comprehensive empirical evaluation of semantic clustering methodologies, employing a rigorous two-phase experimental design that progresses from controlled word-level analysis to large-scale document collection validation. Through systematic comparison of multiple clustering algorithms applied to diverse word embedding models, it provides substantive insights into the complex interplay between algorithmic selection, vector representation choices, and semantic clustering performance. The study demonstrates that K-means clustering consistently produces semantically coherent and well-separated clusters across both experimental phases. This is particularly true when the algorithm is integrated with GloVe embeddings. The algorithm's effectiveness was notably evident in the large-scale validation involving 303 AI/LLM research articles, where it successfully identified and characterized emerging research themes within information management. The dominance of digital library adoption as a research focus (43.2 %), along with the clear differentiation of specialized subdomains such as research data management and reference services, underscores K-means' capability to capture nuanced semantic structures in complex textual data. The findings establish K-means and K-medoids as the most effective algorithms for applications requiring precise semantic groupings, including document classification, taxonomy development, and knowledge organization tasks. The critical importance of embedding model selection is clearly evidenced by the superior performance of GloVe's globally optimized vector space compared to Wiki's contextual representations. This distinction highlights how fundamental differences in embedding methodologies directly translate into variations in clustering quality and semantic discriminability. While DBSCAN demonstrated utility for outlier detection scenarios, its limitations in handling balanced document collections with clear thematic structures were pronounced. The algorithm's classification of 99 % of documents as outliers in this large-scale validation underscores its constrained applicability for comprehensive taxonomy development, though it remains valuable for specific anomaly detection use cases. The research makes several significant contributions to both academic knowledge and practical implementation. It provides an empirically validated framework for selecting clustering algorithms and embedding models based on specific information management objectives. The study also advances methodological practices through its integrated evaluation approach, combining quantitative metrics with qualitative semantic validation to ensure both statistical robustness and practical relevance. Looking forward, this work suggests a foundation for several promising research directions. The integration of deep learning architectures with traditional clustering methods, exploration of transformer-based embeddings for semantic grouping, and development of adaptive clustering frameworks for dynamic text corpora represent compelling avenues for future investigation. Consistent with our commitment to rigorous computational methodologies, this study reinforces the value of empirical, data-driven approaches for solving complex problems in information science (Asemi & Asemi, 2022). Additionally, the scalability of these approaches to massive, multi-domain text collections and their application to real-time information management scenarios warrant further exploration. This study contributes to bridging the gap between theoretical clustering methodologies and practical information management applications by providing empirically grounded, context-aware insights. The consistent performance patterns observed across controlled experiments and real-world validation strengthen the reliability of the findings and underscore their relevance for advancing semantic analysis capabilities in natural language processing and information systems.

Table 7
Summary of key features of optimization algorithms from a practical perspective.

Algorithm	Typical Time Complexity	Scalability	Parameter Sensitivity	Ideal Use Case (Based on Findings)
K-means	O(n * k * I * d)	High	Medium (depends on *k*)	Taxonomy development, Document categorization
K-medoids	$O(k * (n - k)^2 * I)$	Medium	Medium (depends on *k*)	Noisy data, Robustness to outliers is critical
DBSCAN	$O(n \log n)$ to $O(n^2)$	Low-Medium*	High (eps, min_samples)	Outlier detection, Exploratory analysis in noisy data
Spectral	$O(n^3)$	Low	High	Small, non-linear datasets (Not suited for high-dim. text)

*With spatial indexing (e.g., KD-Tree), complexity can be reduced to O(n log n), but performance degrades in high-dimensional spaces.

Data availability

The data and executable code supporting the sensitivity analyses are publicly available via Mendeley Data (Asemi, 2025).

CRediT authorship contribution statement

Asefeh Asemi: Writing – review & editing, Supervision, Project administration, Conceptualization. **Rajab Kiani Shahvandy:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Mahdi Houshangi:** Visualization, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

N/A.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jjimei.2026.100396.

References

Asemi, A. (2025). Data and code for sensitivity analysis of semantic clustering algorithms on static word embeddings [Data set]. *Mendeley Data*. https://doi.org/10.17632/nxjh3kp27t.1
Asemi, A., & Asemi, A. (2022). A judgment-based model for usability evaluating of interactive systems using fuzzy Multi Factors Evaluation (MFE). *Applied Soft Computing, 117*, Article 108411. https://doi.org/10.1016/j.asoc.2022.108411
Asemi, A., Asemi, A., Houshangi, N., Mazaheri, A. H., & Houshangi, M. (2026b). AI-powered investment recommendations in the agri-food sector. *British Food Journal*. https://doi.org/10.1108/BFJ-05-2025-0643. Vol. ahead-of-print No. ahead-of-print.
Asemi, A., Asemi, A., & Ko, A. (2023). Adaptive neuro-fuzzy inference system for customizing investment type based on the potential investors' demographics and feedback. *Journal of Big Data, 10*, 87. https://doi.org/10.1186/s40537-023-00784-7
Asemi, A., Houshangi, M., & Houshangi, N. (2026a). An analytical deep learning framework for author similarity in research collaboration. *Decision Analytics Journal, 100645*. doi:10.1016/j.dajour.2025.100645.
Asemi, A., Ko, A., & Asemi, A. (2021). Infoecology of the deep learning and smart manufacturing: Thematic and concept interactions. *Library Hi Tech, 40*(4), 994–1012. https://doi.org/10.1108/LHT-08-2021-0252
Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with sub-word information. *TACL*.
Bourguet, J.-R., & Sow, A. (2025). Monitoring semantic relatedness and revealing fairness and biases through trend tests. *International Journal of Information Management Data Insights, 5*(1). https://doi.org/10.1016/j.jjimei.2024.100305. Scopus.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.
Ebrahimi, F., Asemi, A., Nezarat, A., & Ko, A. (2021). Developing a mathematical model of the co-author recommender system using graph mining techniques and big data applications. *Journal of Big Data, 8*(1), 44. https://doi.org/10.1186/s40537-021-00432-y
Fakhrzadeh, A., Ershadi, M. J., & Arshadi, M. M. (2023). Application of association rules and clustering in controlling the quality of research data: A case study of the Iranian Scientific Information Database (Ganj). *Journal of Information Processing and Management, 38*(3), 927–944.
Gupta, M., Dheekonda, V., & Masum, M. (2024). Genie: Enhancing information management in the restaurant industry through AI-powered chatbot. *International Journal of Information Management Data Insights, 4*(2). https://doi.org/10.1016/j.jjimei.2024.100255. Scopus.
Hähnel, T., Feige, T., & Falkenburger, B. (2023). Semantic Clustering in neurotechnology. *JMIR Neurotechnology*.
Kiani Shahvandi, R., Shabani, A., Asemi, A., & Mohammadi Ostani, M. (2024). Using conceptual clustering to extraction of key phrases and related terms: A case study of scientific communication texts. *Iranian Journal of Information Processing & Management, 39*(4). https://doi.org/10.22034/jipm.2024.2022598.1545
Kovács, T., Ko, A., & Asemi, A. (2021). Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis. *Journal of Big Data, 8*(1), 141. https://doi.org/10.1186/s40537-021-00529-4
Pandey, A. (2015). Study and analysis of k-means clustering algorithm using RapidMiner: A case study on students' exam results. *Journal of Engineering Research and Applications, 12*(4), 60–64. https://www.ijera.com/papers/Vol4_issue12/Part%20-%204/K0412046064.pdf.
Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP*.
Petukhova, A., Matos-Carvalho, J. P., & Fachada, N. (2025). Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering, 6*, 100–108. https://doi.org/10.1016/j.ijcce.2024.11.004
Reveilhac, M., & Blanchard, A. (2022). The framing of health technologies on social media by major actors: Prominent health issues and COVID-related public concerns. *International Journal of Information Management Data Insights, 2*(1). https://doi.org/10.1016/j.jjimei.2022.100068
Shabani, K., & Asemi, A. (2022). Identifying subject relationships between resources used by users of the Regional Information Center for Science and Technology using text mining techniques. *Information Science, 15*(59), 54–73.
Shokouhian, M., Asemi, A., Shabani, A., & Cheshmeh-Sarabi, M. (2019). Proposing a topic classification model for scientific publications in the health sector using text mining methods. *Journal of Information Processing and Management, 35*(2), 553–574.
Soleimani-Nejad, A., Salajegheh, M., & Tayebi, I. (2018). Clustering scientific articles based on the k-means algorithm: A case study of Iran Research Institute for Information Science and Technology (IRANDOC). *Journal of Information Processing and Management, 34*(2), 871–896.
Thorleuchter, D., & Van den Poel, D. (2016). Identification of interdisciplinary ideas. *Information Processing & Management*.
Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*.
Van Lierde, H., & Chow, T. W. S (2019). Query-oriented text summarization based on hypergraph transversals. *Information Processing & Management*.
Waja, G., Patil, G., Mehta, C., & Patil, S. (2023). How AI can be used for governance of messaging services: A study on spam classification leveraging multi-channel convolutional neural network. *International Journal of Information Management Data Insights, 3*(1). https://doi.org/10.1016/j.jjimei.2022.100147. Scopus.
Xiaoyan, L., Raga, R. C., & Xuemei, S. (2022). GloVe-CNN-BiLSTM model for sentiment analysis on text reviews. *Journal of Sensors*. https://doi.org/10.1155/2022/7212366. Article ID 7212366(accessed January 2, 2024).
Yelov, E., & Noorouzi, Y. (2021). A systematic review of research on improving the k-medoids algorithm for data clustering. *Journal of Information Processing and Management, 37*(2), 527–556.