IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

A machine learning framework for interdisciplinary research recommendation via researcher clustering

Chaisiri Sanitphonklang[1], Bunthida Chunngam[2*]

[1]Department of Computer Science and Artificial Intelligence, Faculty of Science, Chandrakasem Rajabhat University, Bangkok, Thailand.
[2]Department of Computer Engineering, Faculty of Industrial Education, Rajamangala University of Technology Suvarnabhumi, Suphan Buri, Thailand.

Corresponding author: Bunthida Chunngam (*Email: bunthida.c@rmutsb.ac.th*)

Abstract

Interdisciplinary research (IDR) plays a pivotal role in addressing complex global challenges, yet forming productive cross-disciplinary teams remains a persistent obstacle. This study proposes a reproducible machine-learning framework to identify potential interdisciplinary collaborators by clustering researchers using semantic features from abstracts, author keywords, and reference lists. We assembled a Scopus corpus (2019–2023) of 9,160 publications (16,110 authors), preprocessed text with standard tokenization and stop-wording, and derived topic/semantic vectors via LDA and LSA (optimal k = 11). We compared three feature configurations—(A) abstracts, (B) abstracts+keywords, (C) abstracts+keywords+references—and two clustering algorithms: K-means (k via elbow) and DBSCAN (ε via k-distance). Evaluation combined internal validity (Silhouette = 0.559 reported for best configuration) with domain-expert assessment (expert agreement = 29%) and robustness checks. Results show systematic improvement in cluster cohesion as feature sets were enriched: K-means on the fused three-feature representation produced 11 coherent clusters, while DBSCAN revealed numerous fine-grained niche communities aligned with emerging interdisciplinary themes. The fused feature approach yields practical gains for institutional researcher profiling, collaboration recommendation, HR planning, and policy design. Code and processed data will be made available to support reproducibility; limitations and future testing with contextual embeddings (e.g., SPECTER family) are discussed.

Keywords: DBSCAN, Interdisciplinary research, K-means, Recommendation, Researcher Clustering, Topic Modeling.

DOI: 10.53894/ijirss.v8i6.10119
Funding: This work was supported by the Office of Research Development and the Faculty of Science, Chandrakasem Rajabhat University and Rajamangala University of Technology Suvarnabhumi (RUS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
History: Received: 18 July 2025 / **Revised:** 22 August 2025 / **Accepted:** 26 August 2025 / **Published:** 19 September 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Acknowledgements: The authors gratefully acknowledge the financial support from the Office of Research and Development, Chandrakasem Rajabhat University, and Rajamangala University of Technology Suvarnabhumi (RUS). Their help and support made this research a success.
Publisher: Innovative Research Publishing

1. Introduction

Interdisciplinary research (IDR) is increasingly central to tackling climate resilience, public health, and sustainability, yet institutional silos, geographic dispersion, and information overload still impede the discovery of suitable collaborators [1, 2]. The effectiveness of IDR also hinges on teams' capacity to articulate shared goals that translate expertise into workable coordination mechanisms [3]. Evidence shows that platforms and policies which nurture institutional support and teamwork are associated with gains in researcher skill development [4] innovation capacity [5] and scholarly influence [6].

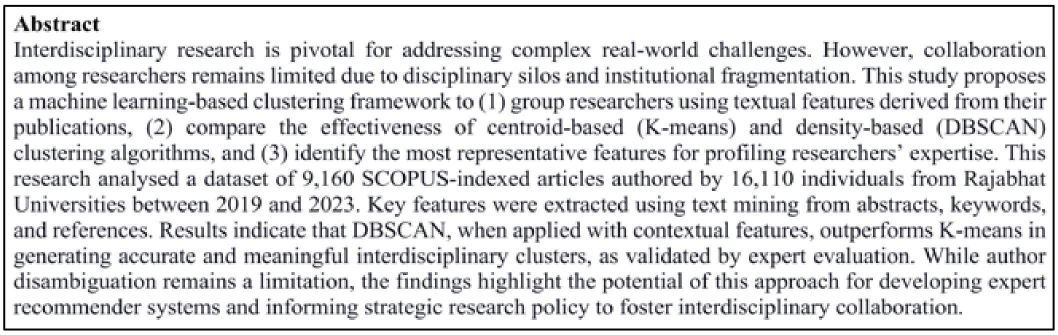
To address these barriers, recent studies have advanced analytical tools for understanding and supporting IDR. Bibliometric and network analyses that integrate co-citation and co-authorship data help assess interdisciplinarity and reveal structural linkages that can improve collaboration efficiency and knowledge exchange [7, 8] while knowledge-mapping approaches surface emerging cross-disciplinary themes [9]. Institutions have also piloted researcher clusters organized around shared interests (e.g., San José State University's initiative) to mobilize expertise on complex problems [10]. Methodologically, network-based recommendation systems with dimensionality reduction have been proposed [11]; clustering and similarity methods—such as K-means applied to abstract-derived keywords [12] and cosine-based matching across topics and outlets [13]—have shown promise. Frameworks combining NLP with network analysis support institutional portfolio development [14] with recommender variants like "ComBSAGE" increasing novelty and diversity in suggestions [15]. Despite this progress, comparative evaluations of clustering families on real-world publication corpora remain scarce, and many studies rely on a single textual element (commonly abstracts), potentially omitting disciplinary signals embedded in reference lists.

Building on this gap, we develop a researcher-level profiling and clustering framework that (i) integrates three textual sources—abstracts, author keywords, and references; (ii) compares centroid- and density-based clustering (K-means [16] vs. DBSCAN [17]); and (iii) evaluates quality using internal indices alongside domain-expert assessment. Using a SCOPUS-derived corpus (2019–2023) of 9,160 articles and 16,110 authors from Rajabhat Universities (Thailand), we find that feature enrichment improves cluster coherence under K-means. In contrast, DBSCAN uncovers fine-grained niches indicative of emerging interdisciplinary themes. We discuss methodological implications and outline a pathway toward embedding-based representations and standardized evaluation to strengthen generalizability. These insights inform research policy instruments, human-resource planning, and collaboration programs at national and international levels.

2. Materials and Methods

2.1. Dataset

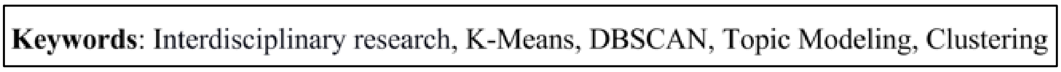
Over the 2019–2023 period, we assembled a corpus of 9,160 SCOPUS-indexed publications authored by 16,110 scholars affiliated with Rajabhat Universities. For each article, we retrieved the abstract—a single paragraph of roughly 150–250 words—summarizing the study's objectives, methodology, key findings, and conclusions (Figure 1). Author-provided keywords, typically three to eight terms, were collected to capture the main themes and support thematic classification (Figure 2). We also gathered complete reference lists, usually 10–50 entries per article, formatted in standard citation styles such as APA or IEEE, offering insight into each work's intellectual foundations (Figure 3). These elements were aggregated into enriched researcher-level profiles using Scopus Author IDs to minimize homonym and alias issues, with affiliation-constrained fuzzy matching applied when IDs were unavailable. The study preprocesses all textual data by applying lowercasing, punctuation removal, stopword filtering, and lemmatization. It then analyzes the processed corpus in Python, employing spaCy/scispaCy for linguistic processing and scikit-learn for vectorization and clustering.



Abstract

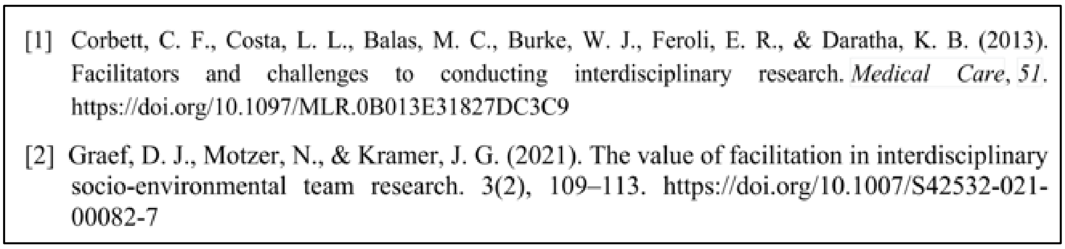
Interdisciplinary research is pivotal for addressing complex real-world challenges. However, collaboration among researchers remains limited due to disciplinary silos and institutional fragmentation. This study proposes a machine learning-based clustering framework to (1) group researchers using textual features derived from their publications, (2) compare the effectiveness of centroid-based (K-means) and density-based (DBSCAN) clustering algorithms, and (3) identify the most representative features for profiling researchers' expertise. This research analysed a dataset of 9,160 SCOPUS-indexed articles authored by 16,110 individuals from Rajabhat Universities between 2019 and 2023. Key features were extracted using text mining from abstracts, keywords, and references. Results indicate that DBSCAN, when applied with contextual features, outperforms K-means in generating accurate and meaningful interdisciplinary clusters, as validated by expert evaluation. While author disambiguation remains a limitation, the findings highlight the potential of this approach for developing expert recommender systems and informing strategic research policy to foster interdisciplinary collaboration.

Figure 1.
Structural and Characteristics of Abstract.



Keywords: Interdisciplinary research, K-Means, DBSCAN, Topic Modeling, Clustering

Figure 2.
Structural and Characteristics of Keywords.



[1] Corbett, C. F., Costa, L. L., Balas, M. C., Burke, W. J., Feroli, E. R., & Daratha, K. B. (2013). Facilitators and challenges to conducting interdisciplinary research. *Medical Care*, *51*. https://doi.org/10.1097/MLR.0B013E31827DC3C9

[2] Graef, D. J., Motzer, N., & Kramer, J. G. (2021). The value of facilitation in interdisciplinary socio-environmental team research. 3(2), 109–113. https://doi.org/10.1007/S42532-021-00082-7

Figure 3.
Structural and Characteristics of Reference.
Source: Corbett, et al. [18] and Graef, et al. [19].

2.2. Research Framework

The framework comprises four stages: (i) data acquisition and normalization; (ii) researcher-level aggregation; (iii) feature engineering via topic vectors; and (iv) clustering and evaluation. We investigated three feature configurations: A) abstracts only, B) abstracts + keywords, and C) abstracts + keywords + references. Figure 4 conceptually summarizes the pipeline from text ingestion to cluster interpretation and recommendation use-cases.

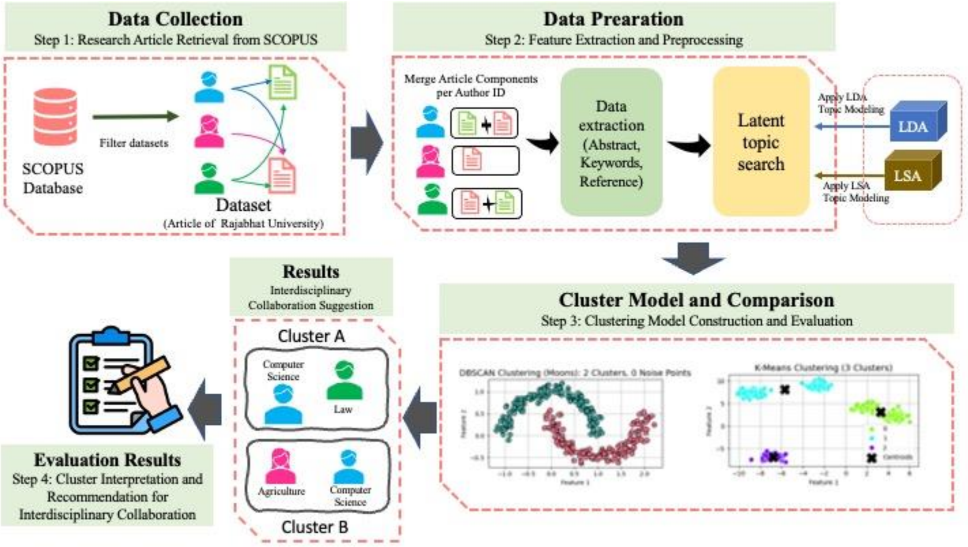

Data Collection
Step 1: Research Article Retrieval from SCOPUS
SCOPUS Database
Filter datasets
Dataset
(Article of Rajabhat University)
Data Prearation
Step 2: Feature Extraction and Preprocessing
Merge Article Components per Author ID
Data extraction (Abstract, Keywords, Reference)
Latent topic search
Apply LDA Topic Modeling
LDA
LSA
Apply LSA Topic Modeling
Cluster Model and Comparison
Step 3: Clustering Model Construction and Evaluation
DBSCAN Clustering (Moons): 2 Clusters, 0 Noise Points
K-Means Clustering (3 Clusters)
Centroids
Results
Interdisciplinary Collaboration Suggestion
Cluster A
Computer Science
Law
Agriculture
Computer Science
Cluster B
Evaluation Results
Step 4: Cluster Interpretation and Recommendation for Interdisciplinary Collaboration

Figure 4.
Research framework for clustering research articles and suggesting interdisciplinary collaboration.

2.3. Methodology

In order to enable multidisciplinary cooperation, this work clusters scholars depending on textual characteristics from their papers using an organized method. Four primary phases define the approach: data collecting, feature extraction and preprocessing, clustering model creation and comparison, and result interpretation. Every stage is meant to methodically convert unorganized study material into significant clusters reflecting underlying knowledge and joint paths based on underlying expertise and possible collaborative routes.

2.3.1. Method of Data Collection

The first step in the data-collecting process was institutional subscription access to the SCOPUS database. Filters using SCOPUS's sophisticated search interface helped to find research papers written by people connected to Thai Rajabhat Universities. The search was limited even more to papers published between 2019 and 2023. Selected metadata fields—author name(s), institutional affiliation, journal title, abstract, keywords, references, and publication year—ensure complete depiction of researcher profiles. These disciplines were selected for their capacity to record intellectual output with both semantic and bibliographic aspects. The final dataset was produced in CSV form following the filter application. Text normalization, feature extraction, and clustering are among the preprocessing techniques used in later phases that this organized approach helps to be compatible with. Thus, the gathered dataset provides the basis for building semantic researcher profiles and performing machine learning–based clustering for interdisciplinary research suggestions. Table 1 shows the ordered dataset derived from SCOPUS following predefined filters. The next rounds of preparation and analysis depend on this dataset.

Table 1.
Raw data loaded from the SCOPUS database.

Authors	**Title**	**Year**	**Affiliation**	**Abstract**	**Keyword**	**References**
Author A, B, C	Title 1	2023	Chandrakasem Rajabhat ……….,	Abstract A1	Key X, Y, Z	Ref A, M, C
Author A, M	Title 2	2022	Phranakhon Rajabhat……….,	Abstract A2	Key A, B, C	Ref B, O, M
…	…	…	…	…	…	…
Author S, O, M, C	Title N	2019	Suan Sunandha Rajabhat……….	Abstract An	Key N, M, O, X	Ref A, X, Y

2.3.2. Data Preparation Procedure

The data preparation process begins with data collection and feature extraction, followed by topic modeling and clustering to facilitate multidisciplinary researcher profiling. The following methods describe the methodological stages used in this work.

Step 1: Researcher-Level Article merging: Since a single researcher may be connected to several publications, authors must cluster research papers to create reasonable researcher profiles. Author identities from every article are first broken out using this method, then records matching the same person are located. Articles credited to the same researcher are then combined, keeping each textual element—such as title, abstract, and keywords—in separate columns to maintain the structural integrity of the dataset. The following feature selection and theme modeling will be built on this combined dataset. Figure 5 illustrates the running flow of this process.

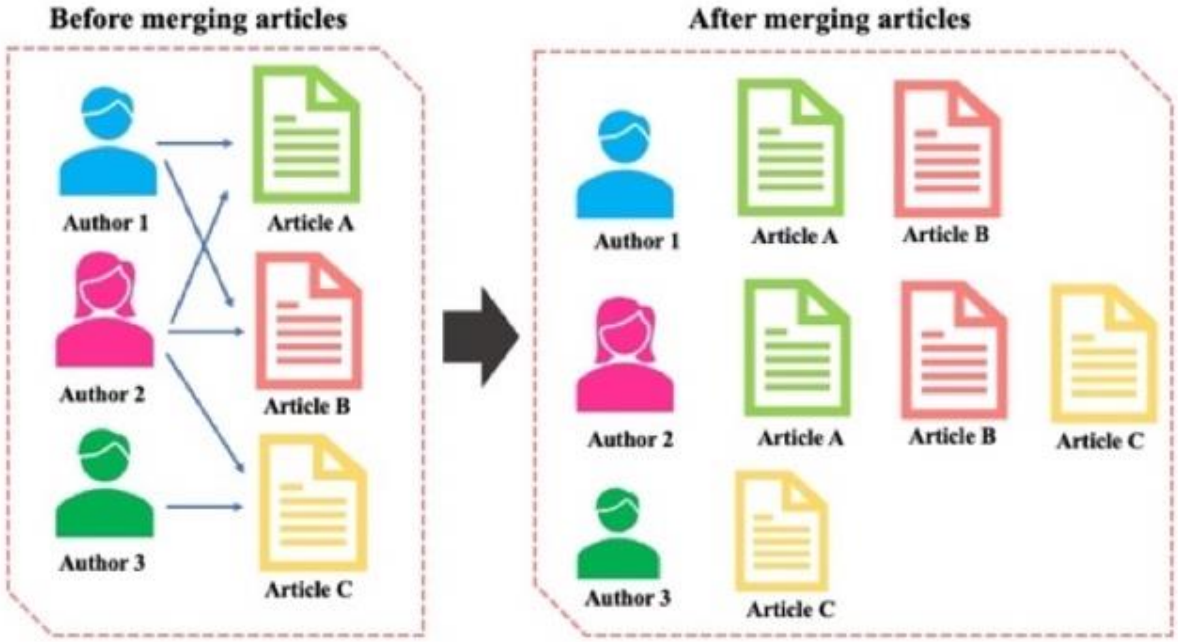


Figure 5.
The process of integrating the data streams of research articles.

The merged dataset, referred to as the "researcher profile," is presented in Table 2. Each row represents a unique researcher, and the columns capture key attributes extracted from their associated articles, including Author, Title, Authors with Affiliations, Abstract, Author Keywords, and References, all structured in CSV format. Following the completion of this integration step, a total of 16,110 researcher profiles were generated.

Table 2.
Researcher profile information.

Author	**Title**	**…**	**Abstract**	**Keyword**	**References**
Author A	Title 1, 2	…	Abstract A1, A2	Key X, Y, Z, A, B, C	Ref A, M, C, B, O, M
Author B	Title 1	…	Abstract A1	Key X, Y, Z	Ref A, M, C
Author C	Title 1, N		Abstract A1, An	Key X, Y, Z, N, M, O, X	Ref A, M, C A, X, Y
…	…	…	…	…	…
Author S	Title N		Abstract An	Key N, M, O, X	Ref A, X, Y

An algorithmic procedure (Table 3, Algorithm 1) to enable researcher-centric analysis that aggregates article-level metadata into structured author profiles. The process begins by splitting the 'Authors' field and merging the index and author keywords. Key metadata fields—title, source title, citation count, abstract, and keywords—are extracted and assigned to each listed author for each publication record. An author-level dictionary is incrementally constructed to collect all associated metadata. This aggregation facilitates accurate profiling of individual researchers and supports subsequent analyses such as feature extraction, clustering, or trend identification at the researcher level.

Table 3.
Author-Centric Metadata Aggregation.

```
Algorithm 1: Author-Centric Metadata Aggregation
Input: A CSV file containing publication metadata
Output: A structured table summarizing publication information per author

1. For each record, split the 'Authors' field by "; "
2.Concatenate 'Index Keywords' and 'Author Keywords' into a single 'Keywords' field
3. Initialize an empty dictionary to store per-author aggregated data
4. For each record in the dataset:
     a. Extract: Authors, Title, Source Title, Cited By, Abstract, and Keywords
     b. For each author in the list:
          i. If author not in dictionary, initialize a new entry
          ii. Append current record's metadata to the author's entry Store all fields as sets to
            ensure uniqueness
5. Transform the dictionary into a DataFrame:Join set elements using appropriate delimiters (e.g.,
   "; " for titles, "," for keywords)
6. Export the final DataFrame to a CSV file
```

Step 2: Choosing features for researcher profiling. To lower data dimensionality and increase calculating performance, we developed researcher profiles using a feature selection procedure [20]. There were 9,160 SCOPUS-indexed research papers in the dataset, with author names, article titles, abstracts, keywords, and references, among other metadata. Every element was tested statistically for its semantic contribution to profiling. The most instructive were abstracts, which captured goals, techniques, and results. Keywords exhibited improved clustering effectiveness when paired with abstracts, even if they were brief and domain-specific. Keyword-based clustering accomplished 27% expert validation using one-hot encoding and K-means clustering with elbow-based optimization (k = 5). Using the same method, references—especially titles, journal names, and conference proceedings—were also examined and produced 32% expert agreement [21].

Based on their combined semantic depth and clustering capacity in identifying researcher profiles [22] three main variables were thus chosen: Abstract, Keyword, and Reference. Table 4 lists the chosen characteristics applied to create researcher profiles, providing input for the next analytical stage.

Table 4.
Researcher Profile after feature selection.

Author	**Abstract**	**Keyword**	**References**
Author A	Abs A1, A2	Key X, Y, Z, A, B, C	Ref A, M, C, B, O, M
Author B	Abs A1	Key X, Y, Z	Ref A, M, C
Author C	Abs A1, An	Key X, Y, Z, N, M, O, X	Ref A, M, C A, X, Y
…	…	…	…
Author S	Abs An	Key N, M, O, X	Ref A, X, Y

Step 3: Tokenization Procedure Researchers the Textual Features: SpaCy In this work, we tokenized researcher profile data using the SpaCy (v3.5) toolkit. The dataset was first loaded, including references, abstracts, and keywords. By lowering all entries to lowercase and eliminating leading and trailing whitespace with the strip () and lower () functions, text preprocessing was used to provide consistency. We examined the dataset's structure and chose three primary textual

columns—Abstract, Keyword, and Reference—as input elements for additional linguistic investigation. For each researcher, every text field was handled consecutively. SpaCy's English language model segmented text into linguistically meaningful units (tokens) for tokenization. This method helps later tasks, including vectorization and semantic similarity analysis, and allows consistent parsing across several academic writing styles. Reliable downstream natural language processing applications depend on the pretreatment pipeline, improving computing efficiency and data quality. As Table 5 shows, the researcher follows Algorithm 2 during this operation.

Table 5.
Scholarly Text Normalization.

```
Algorithm 2: Scholarly Text Normalization
Initialize spaCy English model
Load English stopwords from NLTK
   Step 1: Load and Prepare Data
      Read CSV file "file_data.csv" into dataframe
      Normalize column names (strip and convert to lowercase)
      Verify required columns: "author", "abstracts", "keywords", "references"
      If any required column is missing:
         Raise an error
   Step 2: Define Text Preprocessing Function
      Define preprocess_text(text):
         If text is null:
            Return empty list
         Remove HTML tags and special characters
         Convert text to lowercase
         Tokenize using spaCy
         For each token:
            Lemmatize token
            If token is a stopword, non-alphabetic, or shorter than 3 characters:
               Skip token
         Return list of valid tokens
   Step 3: Apply Preprocessing to Selected Features
      For each row in dataframe:
         Apply preprocess_text() to "abstracts","keywords", and "references"
         Store results in new columns:"abstract_tokens","keyword_tokens",and "reference_tokens"
```

Step 4: Topic Modeling and Vector Representation Based on the coherence score analysis (Figure 6), the optimal number of latent topics was identified as 11. The coherence technique evaluates the semantic similarity among high-probability words within each topic, ensuring a balance between specificity and interpretability. This number was subsequently used to configure the Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) models to compare their effectiveness in extracting latent topics from the abstract data.

LDA is a probabilistic generative model that assumes each document is a mixture of multiple topics, where a distribution of semantically related words represents each topic [23]. In contrast, LSA employs Singular Value Decomposition (SVD) on the document-term matrix to uncover latent semantic structures by reducing dimensionality [24]. LDA is effective for interpretable topic modeling across large and multilingual datasets, while LSA is simpler but less accurate in topic inference. Both techniques have been applied to tasks such as summarizing medical abstracts using PyLDAvis and ROUGE metrics [25] and analyzing Thai news corpora [26]. A major limitation of both methods is the manual selection of topic numbers, which may introduce bias. To address this, LDA often uses perplexity and coherence scores [27] while LSA relies on the cumulative variance of singular values [28] to improve accuracy and semantic interpretation.

This step constitutes a critical component of the data preparation process, particularly in data transformation, which aligns with the study's objective of extracting latent topics from academic texts. The application of topic modelling techniques to abstracts or scholarly content has been validated in prior studies, such as those by Vu [29] and Jelodar, et al. [30] which demonstrated the utility of LDA and LSA in identifying meaningful research themes from large-scale textual datasets.

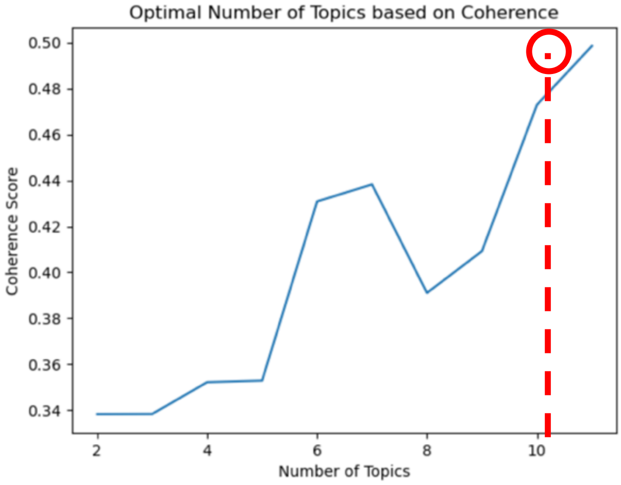


Figure 6.
Coherence Scores by Number of Topics for Optimal Topic Selection.

The output of Step 4 is structured as a matrix, where columns represent the 11 identified topics and rows correspond to individual researchers. Each cell indicates the degree of similarity between a researcher's text and a given topic (between 0 and 1), as determined through LDA and LSA. This matrix serves as a foundational structure for subsequent analysis, enabling topic-based profiling of researchers.

Table 6 visualizes the transformed data matrix following topic modelling, facilitating comparative evaluation and interpretation of thematic alignment across the dataset.

Table 6.
Latent topic matrix in documents.

Author	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	…	Topic11
Author A	0.995	0.000	0.00	0.000	0.000	0.005	…	0.000
Author B	0.960	0.028	0.00	0.000	0.010	0.000	…	0.000
……	…	…	…	…	…	…	…	…
Author S	0.000	0.981	0.00	0.017	0.000	0.017	…	0.017

2.3.3. Clustering Procedure and Performance Comparison

This research adopted a three-step procedure to explore interdisciplinary researcher grouping and compare K-means and DBSCAN clustering performance based on textual features from academic publications.

Step1: Dimensionality Reduction and Visualization: The topic matrix was reduced to two dimensions using Principal Component Analysis (PCA) to enable visual inspection of data distribution. The resulting scatter plot revealed irregular and non-spherical patterns, guiding the choice of clustering algorithms accordingly as shown in Figure 7.

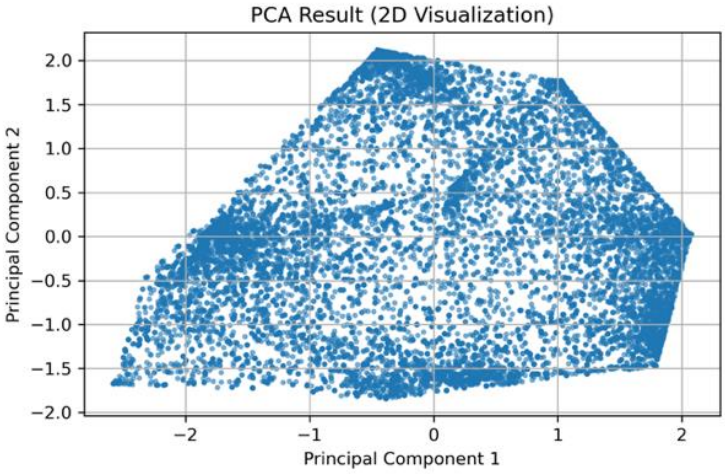


Figure 7.
PCA Scatter 11 Features 16,110 Items.

Step 2: Clustering Implementation: This study applies K-means and DBSCAN to three different text feature configurations: (1) abstracts only, (2) abstracts with keywords, and (3) abstracts, keywords, and references. K-means, known for its efficiency with large datasets [31] requires cluster number identification via the Elbow and Silhouette scores methods. In contrast, DBSCAN detects clusters based on density, effectively handling noise and irregular shapes [32].

Previous research has emphasized the adaptability of K-means to content data and the strength of DBSCAN in noisy or sparse text, which provides references for its application to unstructured research document analysis.

Step 3: Evaluation and Expert Validation Clustering quality was assessed using internal validation metrics, including the Silhouette Coefficient introduced by Rousseeuw [33] which assesses clustering quality by measuring how similar an object is to its cluster compared to others. Additionally, domain experts qualitatively evaluated the coherence and interdisciplinary relevance of the resulting group compositions. Details of clustering results, performance metrics, and expert feedback are presented in the following section.

3. Results and Discussion

3.1. Results

This section presents the research findings, organized according to the study's three primary objectives: (1) to develop a machine learning-based approach for clustering researchers using textual features extracted from their publications, (2) to compare the effectiveness of centroid-based and density-based clustering methods, and (3) to identify the most representative features for profiling researchers' expertise. The results are structured into four main subsections as follows:

3.2. Data Preparation for Topic Modeling

Data preparation is central to the text mining pipeline, particularly when applying latent topic modeling techniques such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). These techniques enable the transformation of large-scale textual data into topic-based representations suitable for downstream machine-learning tasks.

In this study, both LDA and LSA were implemented to assess their effectiveness in extracting latent topics from a heterogeneous set of academic documents. Based on probabilistic modeling, LDA assumes that each document comprises multiple topics, each defined by a word distribution. In contrast, LSA employs Singular Value Decomposition (SVD) to reduce the dimensionality of the term-document matrix, capturing underlying semantic structures.

The preprocessing workflow involved text cleaning, tokenization, vectorization, and dimensionality reduction. The input corpus consisted of documents from multiple disciplines. Topic vectors generated via LDA and LSA were subsequently used as inputs for the clustering stage.

Experimental results revealed that LDA produced more distinct and interpretable topic groupings due to its probabilistic nature, while LSA offered greater flexibility in identifying semantically similar terms across contexts. These findings are consistent with cluster interdisciplinary researcher profiles.

This process supports Objective 1—to develop a machine learning-based framework for grouping researchers based on features derived from their publications. The topic modelling pipeline involved the following stages: (1) researcher-level document aggregation, (2) feature selection, (3) tokenization, (4) latent topic modelling, (5) clustering model, and (6) evaluating the model. This process is illustrated in Figure 8.

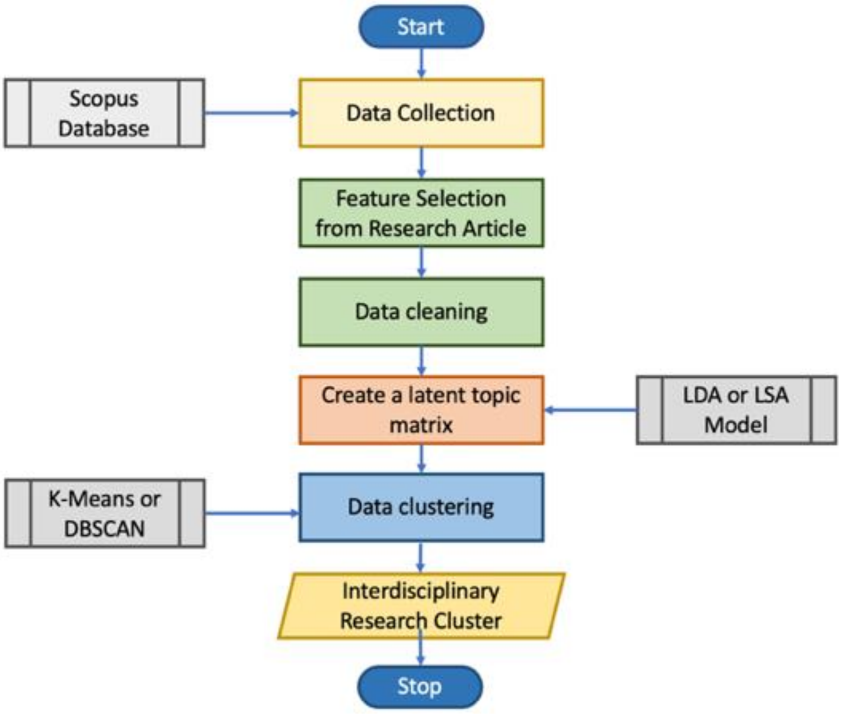

Start
Scopus Database
Data Collection
Feature Selection from Research Article
Data cleaning
Create a latent topic matrix
LDA or LSA Model
K-Means or DBSCAN
Data clustering
Interdisciplinary Research Cluster
Stop

Figure 8.
Machine learning algorithms for clustering researchers.

3.3. Feature Selection for Researcher Profiling

Feature selection is critical for identifying relevant textual elements that represent researchers' expertise and similarity. Three key features were evaluated: abstracts, keywords, and references, as summarized in Table 7. This addresses Objective 3—to determine the most representative features for effective clustering.

These selected features were used to generate topic representations that informed the clustering of researchers across disciplines, particularly those without prior collaboration.

Table 7.
Principles for selecting research article characteristics for creating researcher profiles.

Feature	Description	Rationale
Abstract	A 200–300 word summary highlighting research objectives, methods, results, and conclusions.	Captures academic focus and research scope; supports topic-based similarity and domain classification.
Keywords	Author-defined terms summarizing content scope.	Proven useful through K-means clustering and expert evaluation with 27% validation by domain experts.
References	Includes article titles and journal names from citation lists.	Reflects scholarly influence and domain expertise; aids in identifying disciplinary alignment through cited sources.

3.4. Clustering Models and Evaluation Setup

To address Objective 2—comparing the effectiveness of centroid-based and density-based clustering approaches—two clustering algorithms were employed to analyze researcher similarity based on topic distributions: K-means (centroid-based) and DBSCAN (density-based). Both models were evaluated using three configurations of input features: 1) abstracts only, 2) abstracts combined with keywords, and 3) abstracts, keywords, and references
The clustering aimed to identify interdisciplinary groupings, especially those without prior collaboration.

In K-means clustering, the number of clusters must be predefined, which can introduce bias or subjectivity. To ensure methodological rigor, this study employed the Elbow Method to determine the optimal number of clusters. The within-cluster sum of squared errors (WCSS) was computed for various k values and plotted to identify the inflection point. As shown in Figure 9, the "Elbow" appears at k = 11, which was selected as the optimal cluster count for subsequent analysis. This approach is consistent with prior research in wireless sensor networks, where a hybrid LEACH–K-means model, guided by the Elbow Method, improved clustering efficiency by dynamically optimizing the number of clusters and reducing redundant routing overhead [34].

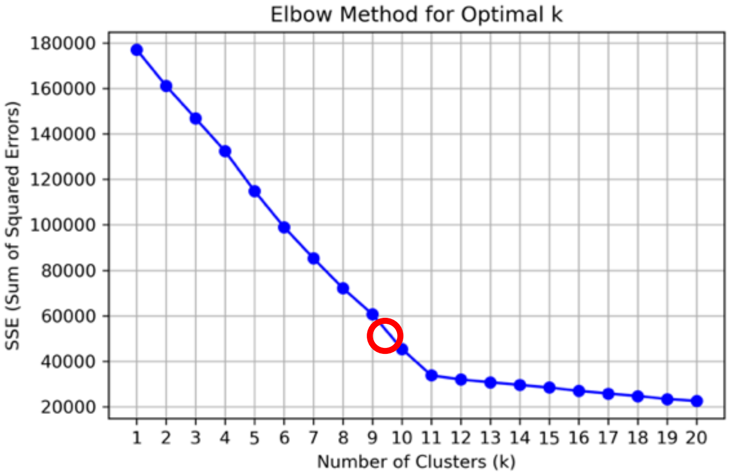
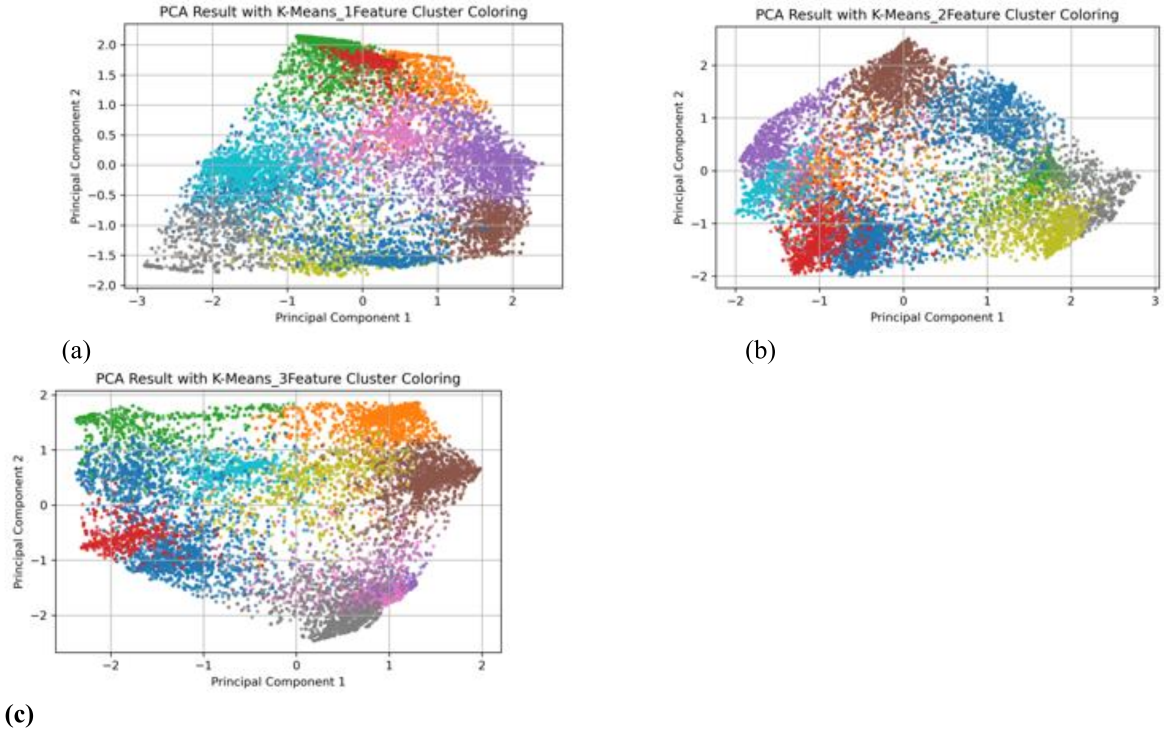


Figure 9.
the Elbow Method for Optimal *k* Selection.

Figure 9 illustrates the K-means clustering results based on three different feature configurations. The clustering was performed with the number of clusters set to k = 11, determined using the Elbow Method. The centroids were initialized using random selection, and the algorithm was executed 20 times (n_init = 20) to ensure convergence stability. Other parameters followed the default settings in Scikit-learn.

In Figure 10(a), clustering based on abstracts only yields relatively indistinct cluster boundaries, suggesting limited semantic separability. Figure 10(b), which includes abstracts and keywords, shows enhanced structure and more precise separation between clusters, indicating the added value of keywords in representing topical information. Figure 10(c) incorporates abstracts, keywords, and references, producing the most coherent and well-separated clusters. This configuration captures a richer semantic profile of documents, enabling a more accurate grouping of researchers. The observed progression from (a) to (c) highlights the impact of feature enrichment on clustering quality and the identification of interdisciplinary research groupings.



(a)

(b)

(c)

Figure 10.
PCA Projection of Topic Vectors Colored by K-means Cluster Assignments.

Figure 12 presents the DBSCAN clustering results for three feature configurations. The algorithm used Euclidean distance, with ε = 0.23 determined via the k-distance graph method, as shown in Figure 11. The parameter min_samples was set to 5, and all other settings followed Scikit-learn's default configuration.

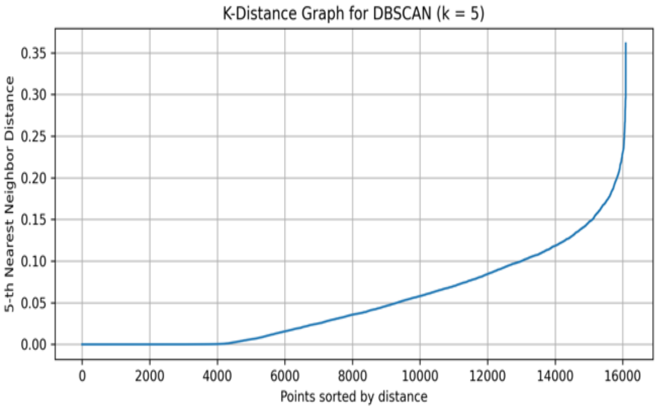
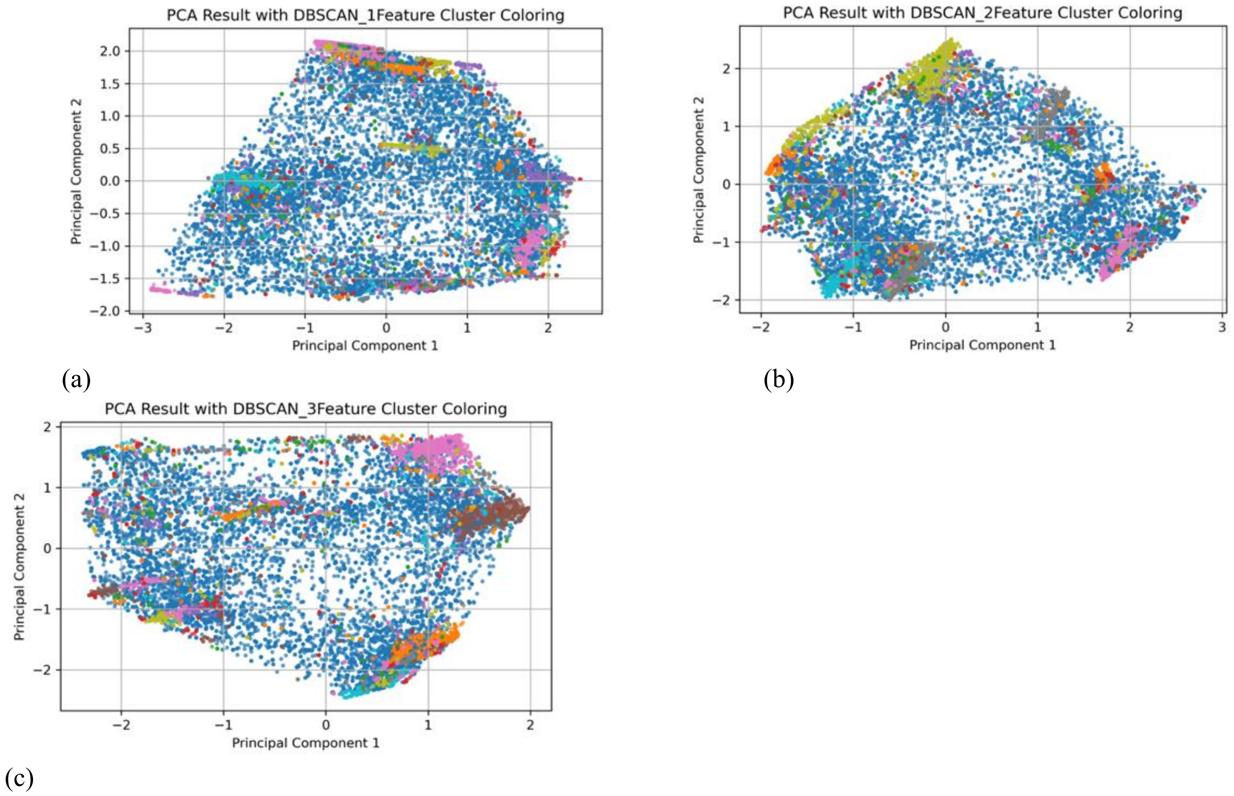


Figure 11.
K- Distance Graph for DBSCAN (k=5).

In Figure 12(a), using abstracts only, DBSCAN identified 480 clusters, many of which appeared sparse or fragmented due to limited semantic content. Figure 12(b), incorporating abstracts and keywords, yielded 466 clusters with improved cohesion and fewer noise points. Figure 12(c), combining abstracts, keywords, and references, produced 466 well-formed clusters—the highest among all. This progression underscores the benefit of semantic feature enrichment in enhancing DBSCAN's capability to capture fine-grained interdisciplinary groupings, particularly in high-density regions of the data space



(a)

(b)

(c)

Figure 12.
PCA Projection of Topic Vectors Colored by DBSCAN Cluster Assignments.

The comparative results and performance metrics of both algorithms are presented and discussed in the following section.

3.5. Clustering Performance Evaluation

This section summarizes the clustering outcomes across three feature configurations: (1) abstracts only, (2) abstracts with keywords, and (3) abstracts, keywords, and references, as shown in Table 5.

For K-means, the Elbow Method consistently identified 11 clusters for all configurations. Cluster quality improved with feature enrichment, yielding the highest Silhouette score (0.559) and expert correctness rating (29%) in the three-feature scenario. These improvements indicate that incorporating additional semantic features enhances clustering precision and interpretability.

In contrast, DBSCAN generated more clusters (480, 466, and 466, respectively) with consistently negative Silhouette scores (−0.414 to −0.347), suggesting weak cohesion and high boundary ambiguity. Expert validation scores also remained lower (15–16%), as presented in Table 5, supporting the interpretation of reduced clustering quality. Despite this, DBSCAN successfully uncovered small, specialized groups—particularly among new or interdisciplinary researchers—highlighting its capacity to detect localized patterns overlooked by centroid-based models.

Table 8 provides a comparative summary of both algorithms across all feature sets, illustrating that K-means offers more stable and coherent clustering. In contrast, DBSCAN contributes exploratory value by revealing granular structures potentially relevant to interdisciplinary collaboration.

Table 8.
Internal and Expert Evaluation Metrics for K-means and DBSCAN Across Feature Configurations.

Method	**K-means**			**DBSCAN**		
	1 Feature	**2 Features**	**3 Features**	**1 Feature**	**2 Features**	**3 Features**
Number of Clusters	11	11	11	480	466	466
Silhouette	0.498	0.509	0.559	-0.414	-0.398	-0.347
Expert (Correctness)	27%	27%	29%	15%	16%	16%

Usually, by building researcher profiles based on abstract keywords and grouping them using the K-means algorithm, few earlier studies have employed clustering approaches to facilitate multidisciplinary research. In about 27% of situations, these methods showed potential by effectively grouping scholars from several disciplines. Still, depending on abstract-based traits could restrict our capacity to detect more complex semantic connections [35].

Some academics have suggested improvements to K-means, including the Grey Wolf Optimization (GWO) algorithm, addressing the difficulties of high-dimensional text data and displaying better clustering performance. Still, it is difficult to distinguish groups of marginal or sophisticated researchers properly.

In this work, we suggest a hybrid strategy using DBSCAN (a density-based method) and K-means (a centroid-based method), applied over three different textual features: abstracts, keywords, and reference lists. Internal validation tests like the Silhouette score and professional assessments help to evaluate the clustering results. Our results show that means beans may effectively identify up to 29% of researchers with potential for multidisciplinary cooperation by using enhanced feature sets, providing a larger framework for research networking.

4. Discussion

This study explored the application of machine learning techniques—specifically K-means and DBSCAN—for clustering interdisciplinary researchers based on textual features derived from academic publications. The results clearly show that clustering quality is influenced by both the algorithm and the richness of features used.

K-means, a centroid-based algorithm, consistently identified 11 clusters across all three feature configurations. As feature enrichment progressed—from abstracts only to abstracts combined with keywords and references—both Silhouette scores and expert evaluations improved, indicating enhanced internal consistency and semantic cohesion. These findings suggest that incorporating multiple textual features leads to more meaningful groupings of researchers with similar expertise.

In contrast, DBSCAN, a density-based method, detected much larger clusters and yielded negative Silhouette values across all configurations, pointing to boundary ambiguity and overlapping groups. Visual inspection via PCA and expert scoring confirmed these observations. Despite its lower quantitative performance, DBSCAN successfully revealed small and nuanced clusters that included new researchers and those working across disciplines, supporting its value in identifying collaboration opportunities.

While K-means proved effective for generating coherent clusters with enriched data, DBSCAN offered complementary insights, especially in capturing localized, interdisciplinary patterns. Future work should consider hybrid approaches and explore alternative evaluation metrics—such as DBCV—and refined parameter tuning, including adaptive ε selection, to better align clustering outcomes with specific research collaboration objectives.

5. Conclusions

This study explored machine learning techniques—specifically K-means and DBSCAN—for clustering researchers based on textual features extracted from academic publications. The research demonstrated that feature enrichment significantly enhances clustering quality by employing three different feature configurations—abstracts only, abstracts with keywords, and references. K-means produced higher internal validation scores and expert agreement, especially with enriched feature sets, indicating stronger cohesion and more precise segmentation. DBSCAN, despite its lower silhouette values, revealed smaller and more specialized groups, including previously unlinked or interdisciplinary researchers, thus offering complementary insights.

Nevertheless, our work should acknowledge several limitations. The dataset was composed primarily of documents from specific domains, which may constrain the generalizability of findings. The moderate data scale may not fully represent broader research ecosystems. Moreover, the domain-specific nature of the textual content may affect the performance of clustering models in more heterogeneous or cross-disciplinary settings. Future studies should examine the scalability of the approach using larger, more diverse datasets and investigate alternative validation metrics such as DBCV. Further improvements in parameter tuning, particularly for DBSCAN (e.g., adaptive ε selection), and integrating citation networks or co-authorship data may enhance the reliability and utility of interdisciplinary research clustering.

References

[1] G. Vladova, J. Haase, and S. Friesike, "Why, with whom, and how to conduct interdisciplinary research? A review from a researcher's perspective," *Science and Public Policy,* vol. 52, no. 2, pp. 165-180, 2025. https://doi.org/10.1093/scipol/scae070

[2] R. Shahid, K. Farrukh, and A. Ayesha, "Investigating the benefits and challenges of interdisciplinary education in higher education settings," *Journal of Social Research Development,* vol. 5, no. 1, pp. 87-100, 2024. https://doi.org/10.53664/JSRD/05-01-2024-08-87-100

[3] G. Beaird, M. Baernholdt, and K. R. White, "Perceptions of interdisciplinary rounding practices," *Journal of Clinical Nursing,* vol. 29, no. 7-8, pp. 1141-1150, 2020. https://doi.org/10.1111/jocn.15161

[4] A. L. Lanier *et al.*, "Facilitating integration in interdisciplinary research: Lessons from a South Florida water, sustainability, and climate project," *Environmental Management,* vol. 62, pp. 1025-1037, 2018. https://doi.org/10.1007/s00267-018-1099-1

[5] C. A. Mahringer, F. Baessler, M. F. Gerchen, C. Haack, K. Jacob, and S. Mayer, "Benefits and obstacles of interdisciplinary research: Insights from members of the Young Academy at the Heidelberg Academy of Sciences and Humanities," *Iscience,* vol. 26, no. 12, p. 108508, 2023. https://doi.org/10.1016/j.isci.2023.108508

[6] E. Leahey, "The perks and perils of interdisciplinary research," *European Review,* vol. 26, no. S2, pp. S55-S67, 2018. https://doi.org/10.1017/S1062798718000261

[7] C. S. Wagner *et al.*, "Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature," *Journal of Informetrics,* vol. 5, no. 1, pp. 14-26, 2011. https://doi.org/10.1016/j.joi.2010.06.004

[8] M. Vantard, C. Galland, and M. Knoop, "Interdisciplinary research: Motivations and challenges for researcher careers," *Quantitative Science Studies,* vol. 4, no. 3, pp. 711-727, 2023. https://doi.org/10.1162/qss_a_00265

[9] I. Rafols and M. Meyer, "Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience," *Scientometrics,* vol. 82, no. 2, pp. 263-287, 2010. https://doi.org/10.1007/s11192-009-0041-y

[10] San José State University, "Research clusters initiative. Office of Research," 2024. https://www.sjsu.edu/research/about/rsca-initiatives/research-clusters/index.php

[11] M. W. Rodrigues, M. A. J. Song, and L. E. Zárate, "Effectively clustering researchers in scientific collaboration networks: Case study on ResearchGate," *Social Network Analysis and Mining,* vol. 11, p. 71, 2021. https://doi.org/10.1007/s13278-021-00781-9

[12] C. Sanitphonklang and N. Soonthornphisaj, "The discovery of experts for multidisciplinary research using data mining approach," in *2018 22nd International Computer Science and Engineering Conference (ICSEC), pp. 1-4. IEEE*, 2018.

[13] B. Lund, "Researchers matching for collaboration: A novel algorithm-based approach using cosine similarity," *Available at SSRN 4346965,* 2023. http://dx.doi.org/10.2139/ssrn.4346965

[14] H. H. Lathabai, A. Nandy, and V. K. Singh, "Institutional collaboration recommendation: An expertise-based framework using NLP and network analysis," *Expert Systems with Applications,* vol. 209, p. 118317, 2022. https://doi.org/10.1016/j.eswa.2022.118317

[15] E. Cunningham, D. Greene, and B. Smyth, "Facilitating interdisciplinary knowledge transfer with research paper recommender systems," *Computer Science,* 2025. https://doi.org/10.48550/arXiv.2309.14984

[16] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Statistics.University of California Press*, 1967.

[17] M. Ester, P. Kriegel, J. Sander, and X. X., "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceeding KDD-96, Institute for Computer Science, University of Munich, Germany*, 1966.

[18] C. F. Corbett, L. L. Costa, M. C. Balas, W. J. Burke, E. R. Feroli, and K. B. Daratha, "Facilitators and challenges to conducting interdisciplinary research," *Medical Care,* vol. 51, pp. S23-S31, 2013. https://doi.org/10.1097/MLR.0b013e31827dc3c9

[19] D. J. Graef, N. Motzer, and J. G. Kramer, "The value of facilitation in interdisciplinary socio-environmental team research," *Socio-Ecological Practice Research,* vol. 3, no. 2, pp. 109-113, 2021. https://doi.org/10.1007/s42532-021-00082-7

[20] M. G. Parsa, H. Zare, and M. Ghatee, "Unsupervised feature selection based on adaptive similarity learning and subspace clustering," *Engineering Applications of Artificial Intelligence,* vol. 95, p. 103855, 2020. https://doi.org/10.1016/j.engappai.2020.103855

[21] C. Zhang, L. Zhao, M. Zhao, and Y. Zhang, "Enhancing keyphrase extraction from academic articles with their reference information," *Scientometrics,* vol. 127, no. 2, pp. 703-731, 2022.

[22] M. Ostendorff, T. Ruas, T. Blume, B. Gipp, and G. Rehm, "Aspect-based document similarity for research papers," *arXiv preprint arXiv:2010.06395,* 2020. https://doi.org/10.48550/arXiv.2309.14984

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research,* vol. 3, no. Jan, pp. 993-1022, 2003.

[24] S. T. Dumais, "Latent semantic analysis," *Annual Review of Information Science and Technology,* vol. 38, pp. 189-230, 2004.

[25] D. F. Onah, E. L. Pang, and M. El-Haj, "A data-driven latent semantic analysis for automatic text summarization using lda topic modelling," in *2022 IEEE International Conference on Big Data (Big Data) (pp. 2771-2780). IEEE*, 2022.

[26] J. Cheevaprawatdomrong, A. Schofield, and A. T. Rutherford, "More than words: Collocation tokenization for latent Dirichlet allocation models," *arXiv preprint arXiv:2108.10755,* 2021. https://doi.org/10.48550/arXiv.2108.10755

[27] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, 2013.

[28] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 2015.

[29] M. Vu, "Building topic modelling on theses abstracts data: Thesis supervisors finder for students," Master's Thesis, Finnish Universities of Applied Sciences, 2021.

[30] H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools and Applications,* vol. 78, pp. 15169-15211, 2019. https://doi.org/10.1007/s11042-018-6894-4

[31] J. Wang, C. Pan, and J. F. Shi, "K-means text clustering method based on Decision Grey Wolf Optimization," *ACM Transactions on Asian and Low-Resource Language Information Processing,* vol. 23, no. 4, p. Article 3689210, 2024. https://doi.org/10.1145/3689210

[32] F. Andriyani and Y. Puspitarani, "Performance comparison of K-means and DBScan algorithms for text clustering product reviews," *Sinkron: jurnal dan penelitian teknik informatika,* vol. 6, no. 3, pp. 944-949, 2022.

[33] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, 1987. https://doi.org/10.1016/0377-0427(87)90125-7

[34] P. Bholowalia and A. Kumar, "EBK-means: A clustering technique based on elbow method and k-means in WSN," *International Journal of Computer Applications,* vol. 105, no. 9, pp. 17-24, 2014.

[35] C. Sanitphonklang and N. Soonthornphisaj, "Mexpert: An algorithm for finding cross-disciplinary experts using data mining techniques," *ECTI Transactions on Computer and Information Technology,* vol. 17, no. 4, pp. 544–553, 2023.