Suan Sunandha Science and Technology Journal
©2026 Faculty of Science and Technology, Suan Sunandha Rajabhat University

Semantic Clustering of Research Papers Using SBERT and K-Means to Support the Discovery of Potential Academic Collaborations

Werayuth Charoenruengkit

Department of Computer Science, Faculty of Science, Srinakharinwirot University, 114 Sukhumvit 23, Wattana, Bangkok 10110, Thailand
*Corresponding author e-mail: werayuth@g.swu.ac.th

Received: *5 May 2026* / **Revised:** *25 August 2026* / **Accepted:** *20 October 2026*

Abstract

Identifying potential research collaborations is critical for fostering interdisciplinary innovation and academic productivity. Traditional approaches, such as co-authorship network analysis, are limited in scope because they rely solely on past collaborations and fail to capture latent thematic connections among researchers. This paper proposes an integrated analytical framework that applies semantic content analysis to research publications to support the discovery and exploration of potential academic collaborations. We compare SBERT-based K-Means clustering with a traditional TF-IDF K-Means baseline and observe superior performance in capturing semantic relationships, as evidenced by higher Silhouette Scores and lower Davies-Bouldin Index values. The framework was evaluated using a text corpus comprising the titles, abstracts, and keywords of 1,335 articles authored or co-authored by researchers affiliated with a university over the past five years. Using the SBERT-based configuration, the approach identified twelve thematic clusters that correspond to interpretable research domains within the institution. Representative terms extracted for each cluster provide clear thematic boundaries and indicate the primary research focus of each domain. Furthermore, multi-dimensional analysis reveals that departmental affiliation patterns exhibit both expected disciplinary concentrations and cross-departmental research overlaps, while network visualization illustrates the interconnected research landscape and highlights authors who act as bridges between clusters.

Keywords: SBERT, K-Means clustering, Authorship network, Topic modeling

1. Introduction

The increasing volume of academic publications, coupled with the growing need for interdisciplinary collaboration, presents significant opportunities to strengthen partnerships among university departments and research laboratories (Ferligoj et al., 2015). However, identifying suitable collaborators remains a major challenge, as researchers often rely on existing personal networks or conduct literature reviews to discover potential partners. These traditional approaches are typically inefficient, time-consuming, and prone to overlooking valuable collaboration opportunities beyond one's immediate network. This challenge highlights a critical need for automated and intelligent systems that can facilitate the discovery of research collaboration opportunities (Ullah et al., 2022).

Prior work has attempted to create author networks by building co-author network graphs (Glänzel & Schubert, 2005; Newman, 2001). While this approach effectively captures existing collaborations and provides insight into established research communities, it does not account for latent or potential collaborations between authors who have not yet published together. Co-authorship graphs are inherently retrospective and limited to observable relationships, failing to reveal underlying thematic connections or shared research interests that may exist across different publication clusters. This paper proposes a novel approach that leverages unsupervised machine learning techniques, specifically research paper clustering, to support the systematic exploration of potential academic collaborations. By analyzing the content and authorship patterns within past academic publications, we aim to develop an approach that assists in identifying groups of research articles that exhibit strong conceptual similarities, thereby indicating shared research interests among their authors. The core

hypothesis is that authors whose papers cluster together, even if they have not previously collaborated, share thematic interests that make them candidates worth exploring for future joint research. This method moves beyond simple co-authorship analysis to infer potential collaborations based on the underlying intellectual alignment of researchers' published works.

The contribution of this study presents an integrated analytical framework that combines semantic publication clustering, topic interpretation, author-cluster mapping, departmental affiliation analysis, and network visualization to support the discovery of potential interdisciplinary research connections within an institution. Unlike studies that focus primarily on evaluating document clustering performance, the proposed framework extends the analysis from document-level semantic grouping to researcher- and organizational-level interpretation. However, it does not attempt to propose a new document clustering algorithm, nor to develop a recommendation system for research collaboration.

The remainder of this paper is structured as follows: Section 2 reviews existing literature on collaboration recommendation systems and text clustering techniques. Section 3 details our proposed methodology for research paper clustering and the subsequent identification of collaboration opportunities. Section 4 presents the experimental setup, experimental result validation, and discusses the results obtained from applying our method to a real-world dataset of academic publications. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Background

The problem of identifying potential research collaborators has been addressed from various angles in recent years, drawing upon advancements in network analysis, text mining, and recommendation systems (de Paula Fonseca e Fonseca et al., 2016; Fagan et al., 2018; Färber & Jatowt, 2020; Kumar, 2015; Schäfermeier et al., 2023). This section provides an overview of existing literature relevant to our proposed approach, focusing on collaboration recommendation systems and the application of text clustering techniques in academic contexts.

2.1 Collaboration recommendation systems

Traditional approaches to identifying research collaborators often rely on co-authorship networks (de Paula Fonseca e Fonseca et al., 2016; Newman, 2001). These methods analyze existing collaboration patterns to suggest new links. For instance, early work focused on constructing co-authorship graphs, where nodes represent authors and edges indicate co-authored publications (Newman, 2001). Metrics like shortest path, common neighbors, and preferential attachment have been used to predict future collaborations within these networks. However, a significant limitation of co-authorship-based systems is their inability to recommend collaborations between researchers who have not yet published together, even if their research interests are highly aligned. They primarily reinforce existing connections rather than discovering novel ones.

More sophisticated systems have begun to incorporate content-based analysis to overcome this limitation. These systems analyze the textual content of publications (e.g., titles, abstracts, keywords, full text, citation) to infer thematic similarities between researchers (Cao & Gao, 2005). Techniques such as Latent Dirichlet Allocation (LDA) (Gropp et al., 2019; Yau et al., 2014), and more recently, word embeddings (e.g., Word2Vec, Doc2Vec) (Mu et al., 2022) have been employed to represent papers and authors in a semantic space. By measuring the similarity between these representations, systems can recommend collaborators whose research topics are conceptually close (Griffiths & Steyvers, 2004). While these content-based methods are a significant improvement, many still focus on individual paper-to-paper (Griffiths & Steyvers, 2004) or author-to-author similarity (de Paula Fonseca e Fonseca et al., 2016), which can be computationally intensive for large datasets and may miss broader thematic convergences.

2.2 Text clustering in academic contexts

Text clustering, a fundamental task in unsupervised machine learning, groups documents based on their textual content. In the academic domain, it has been widely applied for various purposes, including topic modeling, document organization, and trend analysis. Common clustering algorithms include K-means (Kim & Gil, 2019), Hierarchical Clustering (Grootendorst, 2022), and DBSCAN (Ikonomakis et al., 2019). The

effectiveness of these algorithms heavily depends on the quality of document representation (e.g., TF-IDF, semantic embeddings) and the chosen similarity metric.

In the context of research papers, clustering has been used to identify research topics, categorize papers into thematic areas, and visualize intellectual structures within scientific fields. For instance, some studies cluster papers to identify emerging topics or to map the evolution of a discipline over time (Kim et al., 2016). However, the direct application of paper clustering to the exploration of potential collaboration has received less attention. While clusters inherently group thematically similar papers, the challenge lies in translating these paper-level groupings into actionable insights for author-level collaboration discovery, especially when considering authors who have not yet interacted. Our work aims to bridge this gap by specifically focusing on how paper clusters can reveal underlying research interests that transcend existing co-authorship links, thereby facilitating the identification of novel potential collaboration opportunities. By combining document clustering with author-cluster mapping, this study illustrates thematic proximity between researchers to highlight potential paths for collaboration.

For document clustering, widely adopted algorithms such as K-means have been used extensively in the past research (Probierz et al., 2022; Radu et al., 2020). The algorithm aims to partition N data points, each with D dimensions $X = \{x_1,x_2,...,x_n\}$ into K clusters by minimizing an objective function. These algorithms necessitate specifying the number of clusters, denoted as K, as a hyperparameter. In the case of conventional K-means, during each iteration, x_i is assigned to the nearest cluster centroid c_j based on a chosen distance metric (e.g., Euclidean distance), and the cluster centroid c_j is updated as the mean of all data points assigned to it. The objective function J and the cluster update process can be represented as follows:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{K} w_{ij} \|x_i - c_j\|^2$$
$$c_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i \quad (1)$$

, where S_j is the set of data points assigned to cluster j and w_{ij} is an indicator variable that equals 1 if data point x_i is assigned to cluster j, and 0 otherwise. In our case, x_i is the sentence embedding of a research article.

2.3 Sentence embedding

Sentence embedding is essential for tasks involving sentence semantic similarity, as they transform sentences into fixed-length vectors capturing semantic information (Moghadasi & Zhuang, 2020). This approach has revolutionized the process of vectorizing text into numerical representations, surpassing the traditional TF-IDF method that was widely adopted prior to the word embedding and deep learning revolution (Subakti et al., 2022). Building on the capabilities of sentence embedding models, pre-trained models like BERT and its derivatives, including SBERT, have set new standards in semantic similarity tasks. SBERT fine-tunes BERT for sentence embeddings, optimizing it specifically for tasks like semantic textual similarity by using a Siamese network structure to produce sentence embeddings that can be directly compared using cosine similarity (Reimers & Gurevych, 2019). For instance, SBERT variants such as all-mpnet-base-v2 offer robust performance with high accuracy in capturing sentence semantics (Song et al., 2020). In comparison, the all-MiniLM-L6-v2 model, a smaller and more efficient variant, strikes a balance between performance and computational efficiency (Wang et al., 2020). The model has been utilized by advanced topic modeling algorithm such as BERTopic, in which the documents are clustered into groups and meaningful topics are extracted from the textual data (Grootendorst, 2022).

3. Methods

This section outlines the proposed methodology for exploring potential research collaborations through author-publication clustering. The dataset used in this study comprises bibliographic records, including authors, titles, keywords, and abstracts of conference papers and journal articles authored or co-authored by researchers

affiliated with Srinakharinwirot University (SWU), published between 2019 and 2024. The dataset downloaded from the Scopus database includes a total of 1,335 articles. For the clustering process, the textual corpus was constructed by concatenating the title, abstract, author-provided keywords, and system-generated keywords of each publication. To uncover meaningful research groupings, we employed an SBERT-based K-Means clustering algorithm, chosen for its ability to capture deep semantic relationships and produce robust and interpretable clusters (Subakti et al., 2022).

In the absence of ground truth labels and number of actual clusters, a heuristic approach based on intrinsic clustering evaluation metrics is commonly employed to determine the optimal number of clusters (Wang, 2010). These metrics assess the quality of clustering by analyzing the internal characteristics of the data, such as compactness within clusters and separation between clusters. For this experiment, we utilize two of the most widely adopted intrinsic evaluation metrics: the Davies-Bouldin Index (DBI) and the Silhouette Score (SIL). The DBI measures the average similarity between each cluster and its most similar counterpart, with lower values indicating better clustering solutions (Davies & Bouldin, 1979). The DBI is calculated using the following formula:

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{S_i + S_j}{d(c_i, c_j)}\right) \tag{2}$$

, where k is the number of clusters, S_i is the average distance between each point in cluster i and the centroid of cluster i, and $d(c_i,c_j)$ is the distance between the centroids of clusters i and j.

In contrast, the Silhouette Score quantifies how well each data point fits within its assigned cluster compared to other clusters, with higher scores reflecting more coherent and well-separated clusters (Rousseeuw, 1987). The silhouette score for a single data point i is calculated using the following equation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

, where s(i) is the SIL for data point a(i) is the average distance from data point i to other data points within the same cluster. b(i) is the smallest average distance from data point i to data points in a different cluster, minimized over clusters.

For each cluster, representative keywords can be extracted using the class-based Term Frequency Inverse Document Frequency (c-TFIDF), which aggregates all documents in each cluster into a single composite document and computes term weights based on their significance across all clusters (Grootendorst, 2022). This method highlights terms that are both frequent within a specific cluster and relatively rare in others, thereby providing a more interpretable summary of the dominant themes in each cluster. The c-TFIDF score for a term t in cluster c is computed as follows:

$$c - TFIDF_{t,c} = \frac{f_{t,c}}{\sum_{t'\in c} f_{t',c}} \cdot \log\left(\frac{N}{n_t}\right) \tag{4}$$

where: $f_{t,c}$ is the frequency of term t in cluster c, $f_{t',c}$ is the total frequency of all terms in cluster c, N is the total number of clusters, and n_t is the number of clusters in which term t appears.

Following the clustering process, the next step involves associating individual authors with the identified clusters. This association is essential for mapping researchers to specific thematic areas and for uncovering potential collaboration opportunities based on shared research interests. To achieve this, we extract co-authorship information and author affiliations from the bibliographic metadata using a combination of rule-based pattern matching and Named Entity Recognition (NER) techniques. Specifically, we employ SpaCy, a natural language processing library, to identify and classify named entities such as institutional affiliations from the publication records and performing keyword matching on country names (Fantechi et al., 2021). This step results in a structured mapping between authors and their corresponding publication clusters, forming the foundation for subsequent analysis of collaboration patterns and academic network structures.

In this study, sentence embeddings were generated using the sentence-transformers library with the pre-trained all-MiniLM-L6-v2 model. To facilitate efficient clustering, the embedding dimensionality was reduced to five using UMAP with n_neighbors=15 and min_dist=0.1. K-Means clustering was then executed using the k-means++ initialization method, with a fixed random seed set to ensure experimental reproducibility. Following the clustering process, representative keywords for each cluster were extracted using class-based TF-IDF (c-TFIDF), implemented via Scikit-Learn's CountVectorizer. The vectorizer was configured to extract unigrams and bigrams, removing standard English stop words while filtering terms with a minimum document frequency (min_df) of 0.01 and a maximum document frequency (max_df) of 0.95. It should be noted that fine-tuning or optimizing specific hyperparameters (such as dimensionality reduction thresholds) was beyond the scope of this study; rather, our primary objective is to demonstrate an adaptable framework that institutions can tune based on their specific datasets.

Because no ground-truth topic labels exist for this corpus, clustering quality cannot be assessed against a reference partition. We therefore adopt a multi-level evaluation strategy consisting of three complementary levels of evidence. First, internal clustering evaluation uses the Silhouette Score and the Davies-Bouldin Index to assess the geometric structure of the candidate solutions and to guide the choice of K. Second, semantic interpretation examines the representative terms extracted for each cluster using c-TFIDF, in order to establish whether each cluster corresponds to a coherent and nameable research theme. Third, organizational contextual analysis compares the resulting clusters with the departmental affiliations of the contributing authors, which provides external contextual evidence that is independent of the clustering procedure. In addition to these sources of evidence, we present an author-level mapping that groups researchers within specific clusters, enabling network exploration and the identification of authors who span multiple research themes.

4. Results and Discussion

4.1 Cluster results

This subsection analyzes the clustering results by examining the keywords and author affiliations within each topic. We compare clustering metrics of the SBERT-based K-Means clustering with those obtained using the traditional TF-IDF based K-Means approach. To evaluate performance, we varied the number of clusters from 3 to 19 to explore the clustering behavior under different granularities. The DBI and SIL for each K are shown in Figure 1.

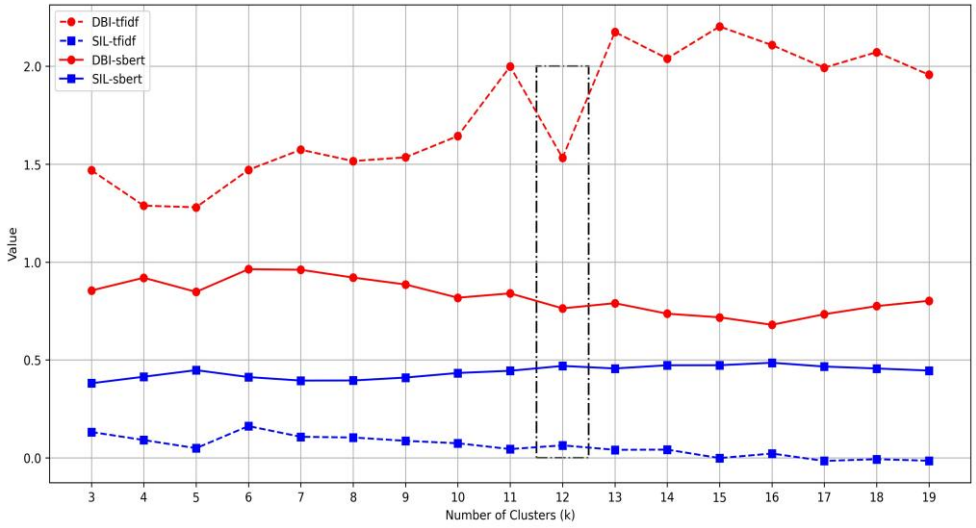


Figure 1. Davies-Bouldin Index (DBI) and Silhouette Score (SIL) for cluster sizes ranging from K=3 to K=19.

Table 1. Top-10 representative keywords extracted using c-TFIDF for each cluster at K=12, along with the number of authors affiliated with SWU in each clustered topic. Note that AuthorsCount are counted once per cluster in which they appear, so an author contributing to more than one cluster is counted in each; the column therefore does not sum to the number of distinct authors.

Topic	Keywords	Authors Count
0	patients, female, treatment, cancer, outcome, risk, gene, clinical study, skin, middle aged	179
1	protein, drug, antioxidant, cell line, apoptosis, cytotoxicity, cancer, oil, stem, extracts	174
2	electrochemical, cellulose, carbon, adsorption, spectroscopy, graphene, biomass, electrode, paper based, voltammetry	120
3	cooling, thermoelectric, classification, electrical, images, deep learning, neural network, segmentation, energy, impedance tomography	159
4	students, thai, teaching, development, nursing, skills, learners, learning management, qualitative, cultural	160
5	health, exercise, thai, training, elderly, questionnaire, health literacy, cognitive, smoking, stroke	172
6	concrete, reinforced, glass, composites, mechanical, fiber reinforced, compressive strength, finite element, beams, recycled	81
7	gene, sequence, rna, streptomyces, listeria, virus, bacterium, salmonella, loop mediated, mediated isothermal	128
8	tourism, social, brand, innovation, thai, factors, consumer, perceived, marketing, consumption	77
9	waste, forest, land use, environmental, growth, agricultural, asean, tropical, climate, risk	42
10	perovskite, dielectric, temperature, superconductors, ceramics, perovskite solar, scn, solar cells, optical, tetragonal	15
11	compressive sensing, compressed, compressed sensing, compressive, norm, hard thresholding, iterative hard, sparse, minimization, signal reconstruction	28

Figure 1 highlights the peak and valley of clustering metrics for both SBERT-based and TF-IDF-based K-Means clustering. As anticipated, the SBERT-based approach demonstrates better performance, evidenced by higher SIL scores and lower DBI values. For SBERT-based clustering, the candidate values for K are observed at K = 5, K = 12, and K = 16, which correspond to local maxima in the SIL score and local minima in the DBI score. Given that SWU supports over 270 curricula across diverse disciplines, a higher number of clusters could potentially reflect the multidisciplinary nature of the research. However, fewer clusters may offer better manageability for administrative tasks such as funding allocation and group coordination. We selected K = 12, as indicated by the dashed rectangle, based on a balance between clustering quality, thematic interpretability, and practical manageability. A smaller value of K, such as K = 5, may combine multiple distinct research areas, making the resulting thematic topics too coarse for meaningful interpretation. Conversely, a larger value of K, such as K = 16, may fragment related research areas into too many narrowly defined clusters, making the results more difficult to interpret and manage at the institutional level. Further analysis of the clustering results at K = 12 was conducted by extracting the top 10 representative keywords for each cluster using the class-based TF-IDF (c-TFIDF) method, as shown in Table 1. The table also shows the number of authors affiliated with SWU, in which

we extracted individual author names and their corresponding affiliations from each article to form an author-affiliation-topic relationship.

The keywords associated with each topic suggest that most articles within a topic share common knowledge field. Some topics are clearly concentrated within specific disciplines, while others reflect an interdisciplinary mix. For example, Topic 0 is related to medical field and Topic 8 is relevant to tourism and economics.

4.2 Result validation

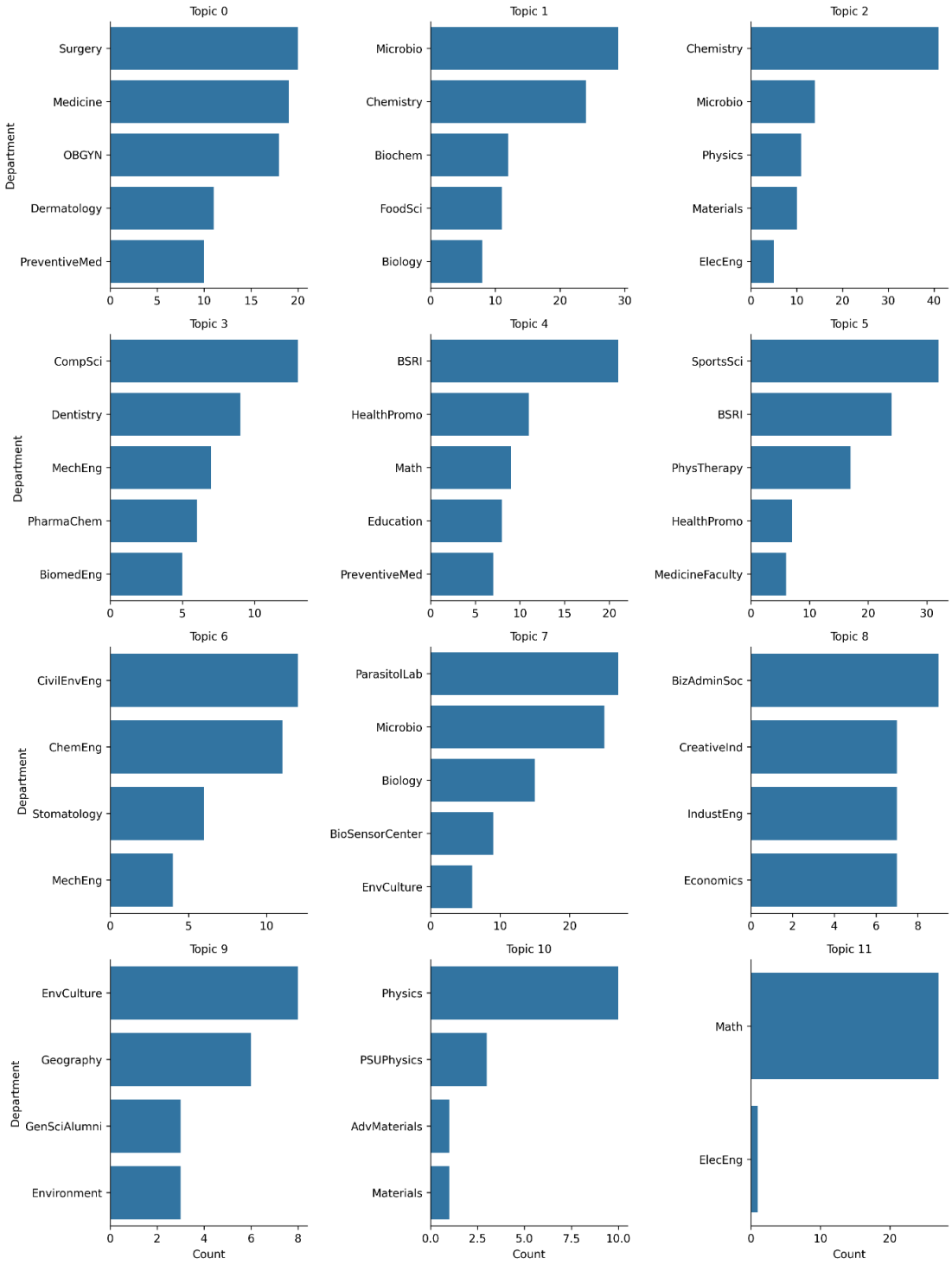


Figure 2. Top five departments affiliated with SWU authors for each topic.

To assess whether the clustering results are interpretable and consistent with the institutional structure, we examined the departmental affiliations of authors within each topic group. The affiliation patterns are used in this context as external contextual evidence that complements internal validation metrics, helping to assess whether the semantic clusters reflect recognized disciplinary patterns within the institution. Ideally, departments grouped within a cluster (Figure 2) should belong to related fields, and their corresponding keywords should align with the primary thematic topic (Table 1). We emphasize that the departmental affiliations are not treated as ground-truth topic labels and agreement between a cluster and a department is neither necessary nor sufficient for the cluster to be meaningful. Affiliation patterns are used here only as external contextual evidence. Specifically, we identified the top five departments contributing the highest number of publications to each cluster. Figure 2 illustrates the distribution of authors' departmental affiliations across the topics based on clustering their publication data. The y-axis lists the departments specified in the Scopus bibliographic data, while the x-axis represents the number of authors affiliated with each unit per topic. It should be noted that some authors only provide faculty-level data. In such cases, the faculty name is used instead.

There are some topics where authors are predominantly affiliated with departments that are thematically related, supporting the coherence of the clustering approach. For instance, Topic 0 is mainly composed of authors from the Department of Surgery (Surgery) and Department of Medicine (Medicine), suggesting a clinical or medical focus. Topic 9 combines affiliations from the Faculty of Environmental Culture and Ecotourism (EnvCulture) with various science-based departments, which may reflect research in sustainability or environmental education. In contrast, Topic 11 displays a sharp dominance by the Department of Mathematics (Math), with minimal representation from Department of Electrical Engineering (ElecEng). This might suggest a narrowly defined research area or limited interdisciplinary overlap, mainly focusing on signal compression and sensing, as shown in Table 1.

Nevertheless, there are topics where cross-disciplinary affiliations are observed. For example: Topic 4 includes contributors from both behavioral sciences and medical-related departments, suggesting interdisciplinary research between psychology, rehabilitation, and healthcare. Topic 6 includes departments such as Civil and Environmental Engineering (CivilEnvEng) and Chemical Engineering (ChemEng), which are commonly aligned with infrastructure and environmental research. Interestingly, this topic also includes articles from dentistry-related affiliations, notably the Department of Stomatology (Stomatology). Upon further examination of the article contents, these publications are found to be related to themes such as restorative materials, ultraviolet radiation, and mechanical strength, which likely intersect with materials science and engineering domains. Table 2 presents examples of articles assigned to the same thematic cluster whose authors belong to different departments. In these cases, the publications were produced independently by different research units, and within this dataset the corresponding author groups do not appear as co-authors of any shared publication. Such overlaps suggest areas in which researchers from different departments are working on thematically related topics, and therefore represent candidate starting points for cross-departmental collaboration, subject to further exploration. For example, Topic 4 (T4) includes publications related to psychology and healthcare, with authors affiliated with the Behavioral Science Research Institute, the Faculty of Education, and the Department of Preventive and Social Medicine. Similarly, Topic 6 (T6) includes publications related to substances and chemicals authored by researchers from the Department of Stomatology, the Department of Chemical Engineering, and the Department of Civil Engineering. None of the authors across these respective groups have co-authored publications within the observed dataset.

The findings of this study indicate that the clustering approach effectively captures disciplinary boundaries while also revealing latent connections between departments, particularly in interdisciplinary research areas. These insights can serve as a valuable starting point for exploring and identifying cross-departmental collaboration opportunity. It is also important to note that the results presented correspond to the selection of $K = 12$, which represents a balance between the granularity of topic themes and interpretability. However, in practice, the choice of K can be adjusted to align with institutional requirements on thematic granularity and strategic research objectives. Furthermore, the level of detail can be enhanced in cases of incomplete bibliographic data by integrating author profiles from the institution's human resources database.

4.3 Author-cluster network

This subsection presents a network visualization using Gephi to represent the author-cluster relationships (Bastian et al., 2009). In this visualization, author nodes are connected to topic cluster nodes through weighted edges, where edge thickness corresponds to the strength of the connection and distinct colors differentiate each thematic cluster. Figure 3 demonstrates an author-cluster visualization across research themes Topic 0 to Topic 11 (T0- T11). Each colored node represents a distinct thematic cluster, while the connecting nodes represent individual authors who may have contributed to multiple clusters, thereby serving as intellectual bridges across different research domains. Because the network contains a large number of author nodes, Figure 3 is intended to convey the overall structure of the author-cluster relationships rather than to allow individual researchers to be identified from the figure. For institutional use, the network is intended to be explored interactively rather than read as a static image: the graph can be filtered by affiliation, research keyword or title phrase, which are stored as node attributes.

Table 2. Examples of interdisciplinary research publications assigned to the same thematic cluster, produced independently by different research units, and representing potential starting points for cross-departmental collaboration.

Cluster number	Title	Department/Research unit
T4	Development of an Intervention to Foster Post-traumatic Growth and Perceived Social Support Among Economically Disadvantaged Students…	Behavioral Science Research Institute
T4	The Impact of Resilience and Coping Strategies on Techno-Stress in Chinese University Students	Faculty of Education
T4	The effectiveness of community-based interprofessional education for undergraduate medical and health promotion students	Department of Preventive and Social Medicine
T6	Increased Fluorohydroxyapatite across Dentin after Fluoride Iontophoresis	Department of Stomatology
T6	Development of a magnified sunlight responsive shape memory bio-composite: effects of titanium nitride (TiN) nanoparticles on a bio-based benzoxazine/epoxy copolymer	Department of Chemical Engineering
T6	Development of stress-strain models for glass fiber reinforced polymer composites confined sustainable concrete …	Department of Civil Engineering
T9	Spatial Patterns of Cumulative Hotspots and Their Relationships with Topographical Factors and Land Use…	Department of Geography
T9	Greenhouse Gas Emission Estimation Using Extended Input–Output Tables for Thailand's Biomass Pellet Industry	Faculty of Economics
T9	Study on Low Carbon Tourism Model at Ancient Town of Si Thep, a World Heritage City, Thailand	Department of Tourism

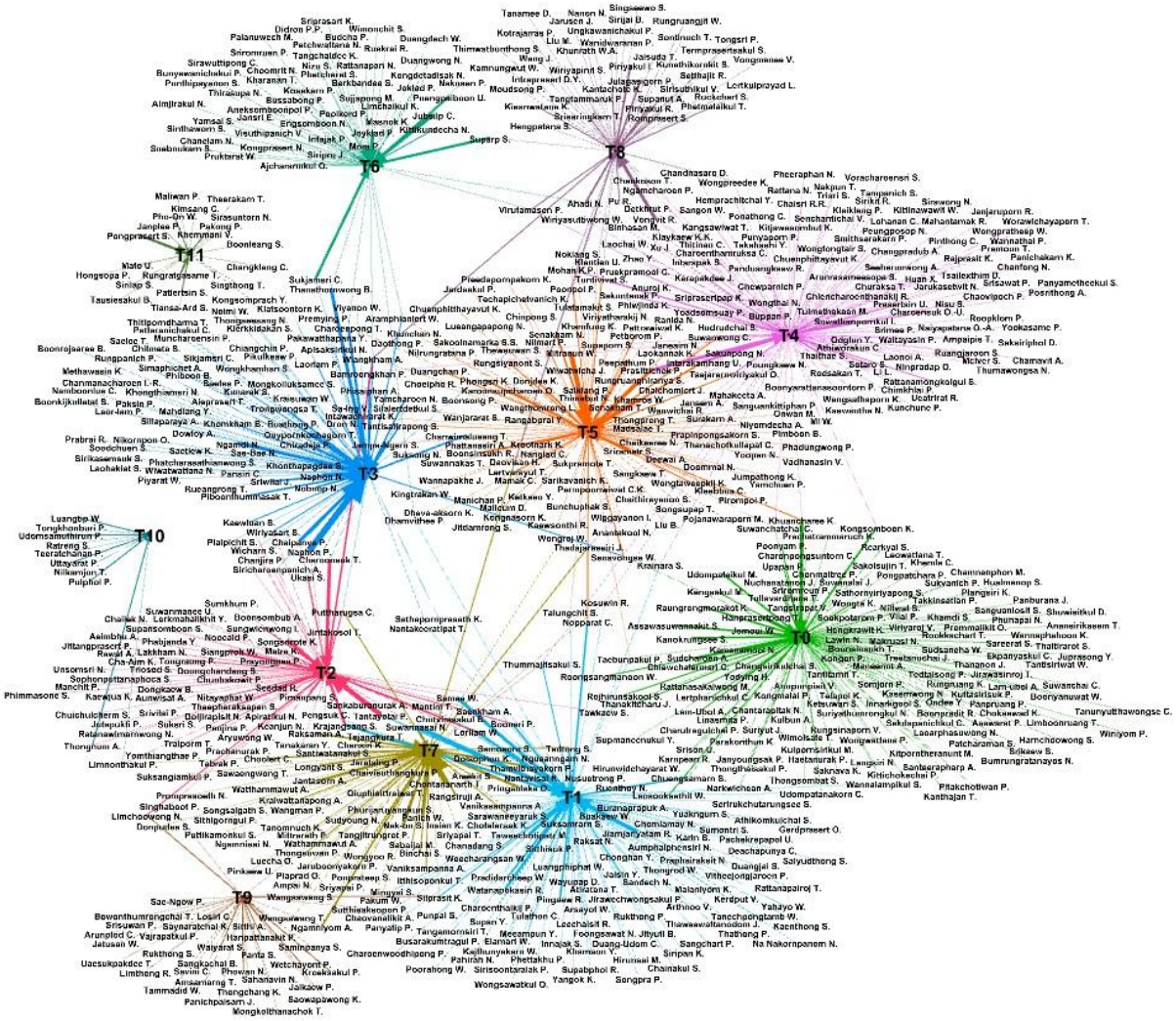


Figure 3. Author-cluster network visualization created in Gephi showing the network structure between individual authors and thematic clusters. Edge connections indicate authorship within specific research domains, with multi-connected authors representing cross-thematic research engagement patterns.

The network structure reveals several key insights. Authors within the same cluster may share related research interests without having a prior publication record together. The framework allows authors to discover potential collaborators by applying graph filters based on affiliation, research keywords, or title phrases. These filterable attributes are embedded as node properties within the network. Certain clusters exhibit high interconnectivity through shared authorship, suggesting thematic or topical relationships between these research areas. For instance, connections are observed between clusters T2 and T3, and between T1 and T2. In contrast, cluster T11 appears isolated, indicating that authors within this field may engage in relatively limited cross-disciplinary collaboration. Certain author nodes are connected to multiple, otherwise unrelated clusters (e.g., T1 and T4), indicating that these authors may be engaged in multidisciplinary research. This finding is particularly significant, as it reveals potential linkages between distant research fields that may, in fact, converge around similar topics. Particularly, the density and positioning of author nodes between clusters provide insights into the interdisciplinary nature of the field, with authors appearing as central connectors who facilitate knowledge transfer across different research themes. Furthermore, through graph-based network analysis, it becomes possible to systematically identify authors with the highest degree of inter-cluster connectivity, which can be interpreted as indicators of multi-disciplinary research engagement and cross-thematic scholarly influence. The visualization effectively captures the dual nature of academic research communities, highlighting both the specialization within clusters and the collaborative bridges that enable scientific progress across disciplinary boundaries.

5. Conclusions

This study demonstrates the effectiveness of semantic-based clustering in assisting the discovery of potential research collaborations beyond the limitations of traditional co-authorship analysis. We show that the SBERT-based K-Means clustering algorithm exhibits superior performance in capturing semantic relationships compared to traditional TF-IDF K-Means approaches, as evidenced by higher Silhouette Scores and lower Davies–Bouldin Index values. By applying an SBERT-based K-Means clustering algorithm to a corpus of SWU-affiliated publications, we illustrated twelve distinct thematic clusters that reflect meaningful research domains within the institution. Our analysis revealed significant insights across multiple dimensions: the keyword analysis for each cluster provided clear thematic boundaries and research focus areas, while the departmental affiliation patterns exposed both expected disciplinary concentrations and surprising cross-departmental research overlaps. The network visualization further illuminated the interconnected nature of the research landscape, identifying key authors who serve as bridges between clusters and facilitate interdisciplinary knowledge transfer. The results reveal latent connections between departments and research domains, providing opportunities to strategically discover potential collaborations among authors in related fields by highlighting thematic proximity across articles and researchers. Although prospective validation of collaboration was not performed, it represents a crucial direction for future work. This approach offers a scalable and systematic framework for institutional research assessment, enabling data-driven decisions regarding research strategy, resource allocation, and the promotion of interdisciplinary and cross-institutional collaboration. The methodology can be readily adapted to other institutional contexts based on specific requirements.

6. Appendix

In Figure 2, we use abbreviated names to represent academic departments and research groups for brevity and clarity. Table 3 provides the name mapping between the short names and their corresponding official department names. These abbreviations were designed to remain recognizable and unambiguous, ensuring readability while avoiding overlapping labels in the Figure.

Table 3. Mapping of Short name to Department/Research unit.

Short name	Department/Research unit
Surgery	Department of Surgery
Medicine	Department of Medicine
OBGYN	Department of Obstetrics and Gynecology
Dermatology	Department of Dermatology
PreventiveMed	Department of Preventive and Social Medicine
Microbio	Department of Microbiology
Chemistry	Department of Chemistry
Biochem	Department of Biochemistry
FoodSci	Division of Food Science and Nutrition
Biology	Department of Biology
Physics	Department of Physics
Materials	Department of Materials Science
ElecEng	Department of Electrical Engineering
CompSci	Department of Computer Science
Dentistry	Faculty of Dentistry
MechEng	Department of Mechanical Engineering
PharmaChem	Department of Pharmaceutical Chemistry
BiomedEng	Department of Biomedical Engineering
BSRI	Behavioral Science Research Institute
HealthPromo	Department of Health Promotion

Short name	Department/Research unit
Math	Department of Mathematics
Education	Faculty of Education
SportsSci	Department of Sports Science
PhysTherapy	Department of Physical Therapy
MedicineFaculty	Faculty of Medicine
CivilEnvEng	Department of Civil and Environmental Engineering
ChemEng	Department of Chemical Engineering
Stomatology	Department of Stomatology
ParasitolLab	Applied Parasitology Research Laboratory
BioSensorCenter	Center of Excellence in Biosensors
EnvCulture	Faculty of Environmental Culture and Ecotourism
BizAdminSoc	Faculty of Business Administration for Society
CreativeInd	College of Creative Industry
IndustEng	Department of Industrial Engineering
Economics	Faculty of Economics
EconIntl	The Faculty of Economics and International Col...
Geography	Department of Geography
GenSciAlumni	Alumni of Department of General Science
Environment	Department of Environment
PSUPhysics	Prasanmitr Physics Research
AdvMaterials	Advanced Materials Research Unit

Acknowledgements

This work was supported by the Faculty of Science, Srinakharinwirot University grant number 372/2567.

Conflict of Interest

The authors declare no conflict of interest.

Ethical Approval

The study was approved by SWUEC-672515 (Date 16 Aug. 2567).

Publication Ethic

The manuscript has not been submitted nor under consideration in any print or online journal or source.

References

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, *3*, 361–362. https://doi.org/10.1609/icwsm.v3i1.13937

Cao, M. D., & Gao, X. (2005). Combining contents and citations for scientific document classification. *Lecture Notes in Computer Science*, *3809*, 143–152.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

de Paula Fonseca e Fonseca, B., Sampaio, R. B., de Araújo Fonseca, M. V., & Zicker, F. (2016). Co-authorship network analysis in health research: Method and potential use. *Health Research Policy and Systems*, *14*, Article 34. https://doi.org/10.1186/s12961-016-0104-5

Fagan, J., Eddens, K. S., Dolly, J., Vanderford, N. L., Weiss, H., & Levens, J. S. (2018). Assessing research collaboration through co-authorship network analysis. *Journal of Research Administration*, *49*(1), 76–99.

Fantechi, A., Gnesi, S., Livi, S**.,** & Semini, L. (2021). A spaCy-based tool for extracting variability from NL requirements. In *Proceedings of the 25th ACM International Systems and Software Product Line Conference — Volume B* (p. 32–35).

Färber, M., & Jatowt, A. (2020). Citation recommendation: Approaches and datasets. *International Journal on Digital Libraries*, *21*(4), 375–405. https://doi.org/10.1007/s00799-020-00288-2

Ferligoj, A., Kronegger, L., Mali, F., Snijders, T. A., & Doreian, P. (2015). Scientific collaboration dynamics in a national scientific system. *Scientometrics*, *104*(3), 985–1012. https://doi.org/10.1007/s11192-015-1585-7

Glänzel, W., & Schubert, A. (2005). Analysing scientific networks through co-authorship. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 257–276). Springer. https://doi.org/10.1007/1-4020-2755-9_12

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv. https://doi.org/10.48550/arXiv.2203.05794

Gropp, C., Herzog, A., Safro, I., Wilson, P. W., & Apon, A. W. (2019). Clustered latent Dirichlet allocation for scientific discovery. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 4503–4511). IEEE. https://doi.org/10.1109/BigData47090.2019.9005964

Ikonomakis, E. K., Tasoulis, D. K., & Vrahatis, M. N. (2019). Density based text clustering. In *Recent Progress in Computational Sciences and Engineering* (2 vols, pp. 218–221). CRC Press.

Kim, J., Tao, L., Lee, S.-H., & Diesner, J. (2016). Evolution and structure of the scientific co-publishing network in Korea between 1948–2011. *Scientometrics*, *107*(1), 27–41. https://doi.org/10.1007/s11192-016-1878-5

Kim, S.-W., & Gil, J.-M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*, *9*, Article 30. https://doi.org/10.1186/s13673-019-0192-7

Kumar, S. (2015). Co-authorship networks: A review of the literature. *Aslib Journal of Information Management*, *67*(1), 55–73. https://doi.org/10.1108/AJIM-09-2014-0116

Moghadasi, M. N., & Zhuang, Y. (2020). Sent2Vec: A new sentence embedding representation with sentimental semantic. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 4672–4680). IEEE. https://doi.org/10.1109/BigData50022.2020.9378337

Mu, W., Lim, K. H., Liu, J., Karunasekera, S., Falzon, L., & Harwood, A. (2022). A clustering-based topic model using word networks and word embeddings. *Journal of Big Data*, *9*, Article 38. https://doi.org/10.1186/s40537-022-00585-4

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409. https://doi.org/10.1073/pnas.98.2.404

Probierz, B., Kozak, J., & Hrabia, A. (2022). Clustering of scientific articles using natural language processing. *Procedia Computer Science*, *207*, 3443–3452. https://doi.org/10.1016/j.procs.2022.09.403

Radu, R.-G., Rădulescu, I.-M., Truică, C.-O., Apostol, E.-S., & Mocanu, M. (2020). Clustering documents using the document to vector model for dimensionality reduction. In *2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)* (pp. 1–6). IEEE. https://doi.org/10.1109/AQTR49680.2020.9129967

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

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

Schäfermeier, B., Hirth, J., & Hanika, T. (2023). Research topic flows in co-authorship networks. *Scientometrics*, *128*, 5051–5078. https://doi.org/10.1007/s11192-022-04529-w

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 16857–16867).

Subakti, A., Murfi, H., & Hariadi, N. (2022). The performance of BERT as data representation of text clustering. *Journal of Big Data*, *9*, Article 15. https://doi.org/10.1186/s40537-022-00564-9

Ullah, M., Shahid, A., ud Din, I., Roman, M., Assam, M., Fayaz, M., Ghadi, Y., & Aljuaid, H. (2022). Analyzing interdisciplinary research using co-authorship networks. *Complexity*, *2022*, Article 2524491. https://doi.org/10.1155/2022/2524491

Wang, J. (2010). Consistent selection of the number of clusters via cross-validation. *Biometrika*, *97*(4), 893–904. https://doi.org/10.1093/biomet/asq061

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 5776–5788).

Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767–786. https://doi.org/10.1007/s11192-014-1321-8