SciClusterNet: Discovering Emerging Topics in LLM and AI in Education

S. S. Zobaer Ahmed[1], Md. Towsif Billah[1], Emon Safayet Rid[1], Md. Rehab Ansary Yasin[1], and Tohedul Islam[1]

American International University-Bangladesh, Department of Computer Science, 408/1 (Old KA 66/1), Kuratoli, Khilkhet, Dhaka 1229, Bangladesh
22-49415-3@student.aiub.edu
WWW home page: https://www.aiub.edu

Abstract. The increase in research studies focused on Large Language Models (LLMs) and Artificial Intelligence in Education (AIED) has increased the difficulty of discovering emerging themes and research trajectories. This research introduces SciClusterNet, an unsupervised methodology that combines SciBERT embeddings, UMAP reduction, multiple algorithms, and BERTopic to investigate 4,000 research abstracts from the domains of LLM and AIED research. K-Means provided the highest cluster architecture (Silhouette = 0.3877, DBI = 0.7452), and BERTopic produced six coherent topics across the domains of cybersecurity, multimodal reasoning, finance, code generation, and Q&As in education. SciClusterNet had a topic coherence (Cv = 0.5055, NPMI = 0.0383) that is somewhat lower than NMF (0.5784, 0.0696), yet it detects themes with significantly more (7.6% higher Silhouette, 19.2% lower DBI, and 239% CH improvement under DBSCAN) semantic richness and clustering quality than the baselines. Although the difference in topic coherence is somewhat negligible, we confirm that the SciBERT embedding generates more coherent and scientifically faithful topics than bag-of-words models. Overall, using SciClusterNet is a robust and domain-specific approach to identify emerging topics and themes in increasingly dynamic research in LLM and AI in Education research.

Keywords: Topic Modeling, Trend Analysis, SciBERT, BERTopic.

1 Introduction

Artificial Intelligence (AI) has quickly transitioned from a niche branch of computer science into a transformative force that is changing education, industry, and daily life. One of its most salient innovations, Large Language Models (LLMs), and their ability to understand, generate, and adapt to human language, are core enablers of intelligent interactions. The intersection of LLMs and education is partly demonstrated by the use of them in adaptive tutoring systems, providing automated feedback, and personalized learning environments.

The growth of research on LLMs, and on Artificial Intelligence in Education (AIED) more broadly, reflects that momentum. However, the sheer volume

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_76

and breadth of published research makes it difficult to identify the collective armature of the field, which themes are emergent, and observe research trends and directions. Understanding how these two domains overlap can only be done using scalable and interpretable approaches that aggregate large amounts of unstructured, academic text. This would help clarify conceptual relationships and highlight under-researched areas.

Although there has been substantial advancement, the current works typically depend either on keyword-based analysis or more simplistic clustering methods that cannot adequately capture the richness of meaning and context of contemporary scientific literature. These studies also face difficulties balancing scalability and interpretability, resulting in disconnected insights. This study develops **SciClusterNet**, an unsupervised analytical framework that uses semantic embeddings, dimensionality reduction, and clustering algorithms to identify meaningful research structures for semantic task analysis. SciClusterNet utilizes transformer-based language models like SciBERT, plus UMAP and BERTopic. The resultant work produces a rich and interpretable perspective on the changing landscape of LLM and AIED research converging together.

2 Related Works

2.1 Comprehensive Review of Existing Studies

Increasing research on text mining, clustering, and topic modeling is reflective of the growth of NLP. In addition, many studies have started to incorporate the combination of traditional methods plus transformer-based methods in an effort to improve semantic representation, interpretability, and efficiency.

Piriyakul et al. [1] combined text mining with customer journey analysis to evaluate hotel equity and identified accommodation as the main factor, but limited study to one case only. Guleria et al. [2] used CNN–LSTM hybrid for fake news detection, which outperformed baseline prior to it, but did not maintain consistency for leadership across multi-categories (e.g., politics and science). In the food context, Xiong et al. [3] reviewed text mining, covering privacy issues and unstructured data integration related to the food context. Houssein et al. [4] reviewed biomedical NLP progress using semantic adaptation, but emphasized ongoing problems related to interoperability of data across datasets. In terms of enhancing accuracy, Sheri et al. [5] used consensus clustering to enhance the prediction accuracy by +15%, but underwent computational complexity challenges. Liang et al. [6] proposed the Global and Local Topic Model (GLTM) as a hybrid to increase topic coherence when evaluating short-text collections of corpora based on local embeddings and global embeddings. Similarly, Jia and Wu [7] reported on LDA to understand trends within China's NEV industry, while Kanungsukkasem and Leelanupab [8] used FinLDA to enhance predictive accuracy of financial textual information.

Ultimately, while these studies indicate advancements in understanding semantics and topic modelling, they are still hindered by scalability, multilingual

support, contextual coherence, and cross-domain generalization, all of which motivated the development of SciClusterNet.

2.2 Problem Classification

A summary of key studies and limitations is featured in Table 1, with addressing the previously identified challenges of semantic depth, scalability and integrated modelling.

Table 1. Key Studies and Limitations Addressed by the Proposed Framework

Relevant Studies	Description	Limitations Addressed
Sazan *et al.* [9], Sunny *et al.* [10], Wang [11]	Used TF–IDF and embeddings for text classification and keyword extraction.	Relied on shallow features with limited contextual depth; unsuitable for large, semantically rich corpora.
Vaid *et al.* [12], Zubair *et al.* [13], Buatoom *et al.* [14]	Enhanced clustering (hybrid, PCA-based K-Means) for structured data.	Required fixed parameters and lacked semantic adaptability for unstructured research texts.
Gupta *et al.* [15], Farea *et al.* [16], Jiang *et al.* [17]	Applied BERTopic and transformer models for literature mining.	Focused on single-domain corpora with limited scalability and thematic diversity.
Wani [18], Wu *et al.* [19], Lai & Chen [20]	Surveyed clustering and neural topic modeling approaches.	Offered theoretical insights but lacked unified integration of contextual embeddings and unsupervised clustering.
Murshed *et al.* [21], Dillan & Fudholi [22], Asnawi *et al.* [23]	Improved topic modelling through preprocessing and multilingual extensions.	Exhibited limited semantic coherence and generalization across heterogeneous datasets.

3 Proposed Methodology

This section covers the methodological design of the **SciClusterNet** framework. The framework leverages natural language processing and machine learning methods to implement interpretable and scalable topic discovery.

3.1 Framework Architecture

SciClusterNet creates a simple yet extensible workflow for discovering meaningful research themes from massive sets of academic texts (Fig. 1). The framework

integrates domain-adapted transformer embeddings, manifold learning, unsupervised clustering, and transformer-based topic modelling into a unified pipeline tailored for analyzing trends in LLM and AI-in-Education studies.

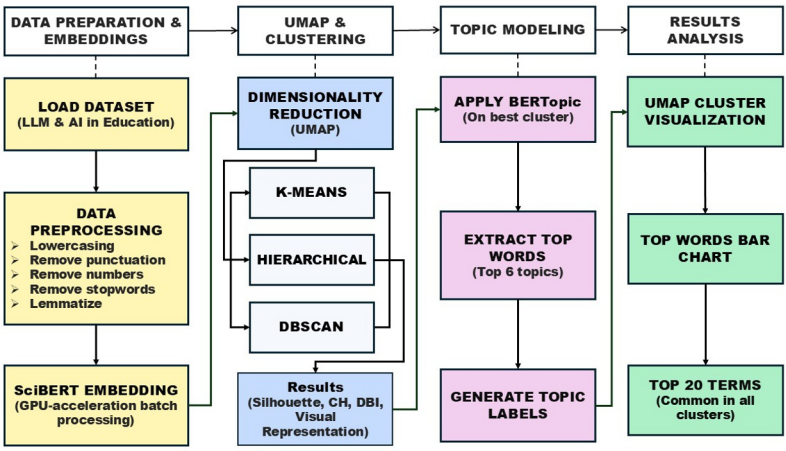


Fig. 1. Architecture of the proposed SciClusterNet framework showing its main components: data preparation, SciBERT-based embedding generation, dimensionality reduction via UMAP, clustering, and transformer-based topic modeling (BERTopic).

To guarantee textual consistency, the gathered research abstracts are first preprocessed using tokenisation, lowercasing, and noise removal. The domain-specific linguistic and semantic subtleties found in academic writing are then captured by SciBERT embeddings, which offer richer representations than general-purpose transformer models. UMAP is used to maintain local and global document relationships, facilitating effective downstream clustering, as these embeddings are high-dimensional.

To find latent structures in the corpus, unsupervised algorithms are used to cluster the reduced vectors. Then, using transformer-based representations and class-based TF-IDF weighting, BERTopic is used to refine each cluster into cohesive and comprehensible topics. Scatter plots, hierarchical topic trees, and frequency curves are used to visualise the resulting topics in order to show thematic evolution and new research directions.

All things considered, **SciClusterNet** provides a modular, extensible, and empirically validated process that connects unsupervised topic discovery with semantic understanding, allowing for scalable and repeatable scholarly literature exploration in rapidly changing research domains.

3.2 Data Preparation and Embeddings

The 4,000 scholarly abstracts in the dataset—2,000 on LLMs and 2,000 on AIED were gathered using publicly accessible APIs to guarantee equitable coverage

across domains. Crucially, every abstract was released in **2025**, guaranteeing temporal coherence and preventing longitudinal bias. This consistent time span makes a clear cross-domain comparison possible, which eliminates the confounding effects of topic drift over time.

The text was subjected to standard pre-processing before analysis, which included lemmatization to unify word forms, lowercasing, and the removal of punctuation, numerals, and stopwords. For the purpose of creating embeddings, these procedures ensured semantic consistency and decreased lexical noise.

Each document was subsequently represented using SciBERT, a variant of BERT trained on scientific corpora. SciBERT is specifically pre-trained to optimize SciBERT's representations of scientific language compared to more general approaches such as BERT or SBERT [24]. The embeddings were created in batches using the GPU to maintain representational fidelity and speed of calculation. The generated vectors comprise the foundation for dimensionality reduction and clustering.

3.3 UMAP and Clustering

To reduce the high-dimensional embeddings to a low-dimensional space while retaining semantic structure, the next step uses Uniform Manifold Approximation and Projection (UMAP). This process enhances the interpretability of clusters while being computationally efficient.

Afterward, three unsupervised clustering algorithms were examined through DBSCAN for dense semantic areas, Hierarchical for nested relationships, and K-Means for compact, centroid-based clusters. Each of these algorithms gives a different orientation for analysis. The most consistent and stable solution was found using quantitative measures including the Davies Bouldin Index (DBI), Calinski-Harabasz Index (CHI), and Silhouette Coefficient. Topic modelling is based the clusters, which are semantically meaningful groups of research abstracts.

3.4 Topic Modeling

In the final phase, BERTopic is employed to derive interpretable themes from clustered documents. This transformer-based method integrates the reduced embeddings with class-based TF–IDF representations to identify representative keywords and generate coherent topic labels. Each cluster's distinguishing terms form linguistically meaningful topic descriptors grounded in scientific discourse.

UMAP-based visualizations and frequency plots are used to illustrate topic prominence and interrelationships, highlighting how research attention is distributed across LLM and AIED domains. Together, the UMAP, clustering, and topic modeling stages constitute the analytical core of *SciClusterNet*, enabling systematic exploration of emerging knowledge patterns at the intersection of artificial intelligence and education.

4 Results and Discussion

In this section, we present the results from applying the **SciClusterNet** analysis to the LLM and AIED corpus. Through analysis of the corpus profiling, dimensionality reduction, clustering, and transformer-based modelling, we reveal insights into thematic and structural information with visual and quantitative relevant evidence in its application to the future of educational technology.

4.1 Text Exploration and Corpus Profiling

A preliminary lexical analysis was conducted to assess the use of terms and thematic density in the corpus. Before the analysis, all of the abstracts were pre-processed and put into a normalized corpus (as described in Section 3.2), ensuring that the unit of analysis was linguistically consistent.

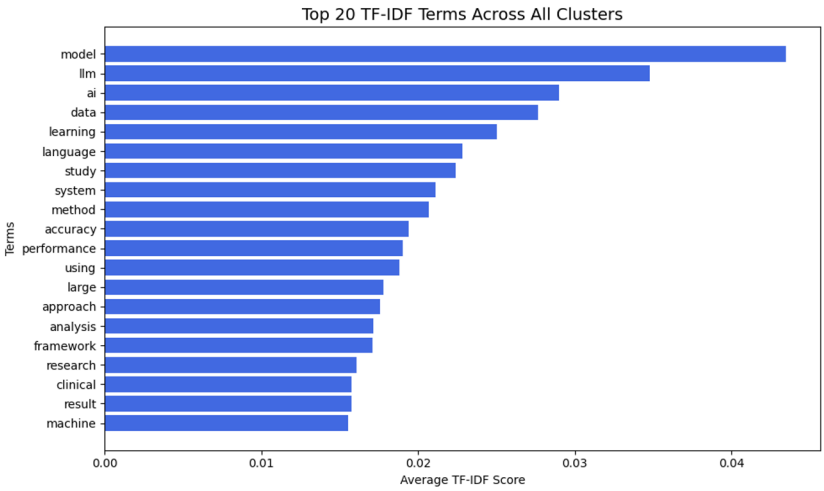


Fig. 2. Top 20 TF–IDF terms across all clusters.

Table 2. Corpus Profiling Statistics

Metric	Value
Total Documents	4,000
Average Words per Document	148.87
Average Characters per Document	1,288.59
Vocabulary Size	36,384

Descriptive statistics for the corpus are summarized in Table 2. These numbers represent a linguistically rich corpus of data that could support the semantic

analysis presented in later sections. The TF–IDF (term frequency inverse document frequency) analysis identified the most 'visually salient' words across the corpus. Fig. 2 shows the top twenty terms include "model," "LLM," "AI," "data," "learning," and "language"— all of which strongly convey either the model development or applications toward and educational context. "Study," "system," "method," and "performance" are dominant empirical methodology terms in the research literature.

Ultimately, these findings confirmed that the corpus had a balanced thematic variety and explored both computational and pedagogical themes. Lexical variety supported a consequent semantic embedding knowledge modelling process, detailed later.

4.2 Clustering Results and Evaluation

To find thematic patterns in LLM and AIED research, three clustering algorithms (K-Means, Hierarchical, and DBSCAN) were applied to the corpus following dimensionality reduction. Standard clustering metrics and qualitative evaluations of semantic coherence were used to assess their performance. Comparative results based on the Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Score are shown in Table 3.

Table 3. Quantitative Evaluation of Clustering Algorithms

Algorithm	Silhouette Score	DB Index	CH Score
K-Means	0.3877	0.7452	3621.2114
Hierarchical	0.3229	0.7121	2997.9277
DBSCAN	0.2528	0.3542	220.9445

Among the algorithms, the **K-Means** clustering structure was the most coherent and easy to interpret. As the clustering comparisons (fig. 3), five distinct and separated clusters were obtained, indicating highly cohesive groups and very little to no overlap between different research areas. The top TF-IDF terms from K-Means are *model*, *language*, *learning*, *data*, and *education*, indicating that LLM studies emphasize model optimization, language understanding, and integrating AI-based methods into educational contexts.

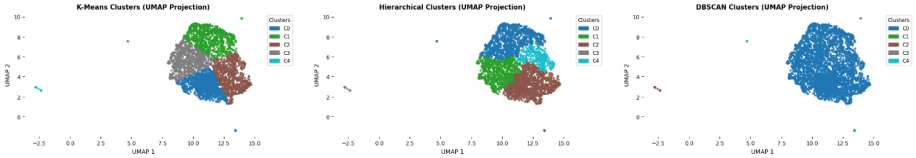
Hierarchical clustering showed another perspective with terms having transitions, instead of boundaries, between each other (Figure 3b). The highest ranking terms (system, framework, student, and clinical) indicate there was overlap of methodological and practical ways of researching. This indicates research in the intersection of AI and education is interdisciplinary and implementation-focused.

In comparison, **DBSCAN** analysis revealed a smaller amount of dense thematic clusters (Fig. 3c), which included some isolated data points, represent-

ing emerging areas of scholarship. The unique terms generated from these clusters **cognitive**, **protein**, **interaction**, and **sequence**—are suggestive of cross-cutting, fringe topics or disciplinary areas. DBSCAN appears to be capturing the novel and exploratory areas of work in the larger corpus of AI research.

Table 4. Representative Top 10 TF–IDF Terms Identified Across Clustering Algorithms

Algorithm	Representative Top Terms (Aggregated from All Clusters)
K-Means	*llm*, *model*, *language*, *learning*, *data*, *accuracy*, *method*, *system*, *student*, *education*
Hierarchical	*model*, *learning*, *data*, *method*, *system*, *framework*, *feature*, *student*, *clinical*, *chatgpt*
DBSCAN	*model*, *ai*, *learning*, *system*, *predictability*, *cognitive*, *protein*, *interaction*, *sequence*, *fitness*



(a) K-Means clustering. (b) Hierarchical clustering. (c) DBSCAN clustering.

Fig. 3. Comparison of clustering results for K-Means, Hierarchical, and DBSCAN algorithms based on UMAP-reduced embeddings. Each visualization highlights distinct structural perspectives on thematic organization within the combined LLM and AIED corpus.

As demonstrated by clustering comparison in Fig. 3 and term patterns in Table 4, K-Means generated the most coherent and comprehensible clusters. Its leading quantitative scores corroborated this. DBSCAN identified smaller emerging and interdisciplinary topics, while hierarchical clustering complemented this by capturing gradual thematic transitions. All of these findings support SciClusterNet's multi-method clustering approach in efficiently mapping the semantic landscape of AIED and LLM research.

4.3 Topic Modelling and Interpretation

To further reveal the semantic structure within the clustered abstracts, topic modelling was conducted using the **BERTopic** framework on the K-Means clustering result. This transformer-based method invokes contextual embeddings and class-based TF–IDF weighting to derive semantically meaningful topics for each cluster. The six topics produced, along with their most representative keywords

are presented in the topic visualization in Fig. 4, as well as an abbreviated summary of the topic labels and key terms in the corresponding table (Table 5).

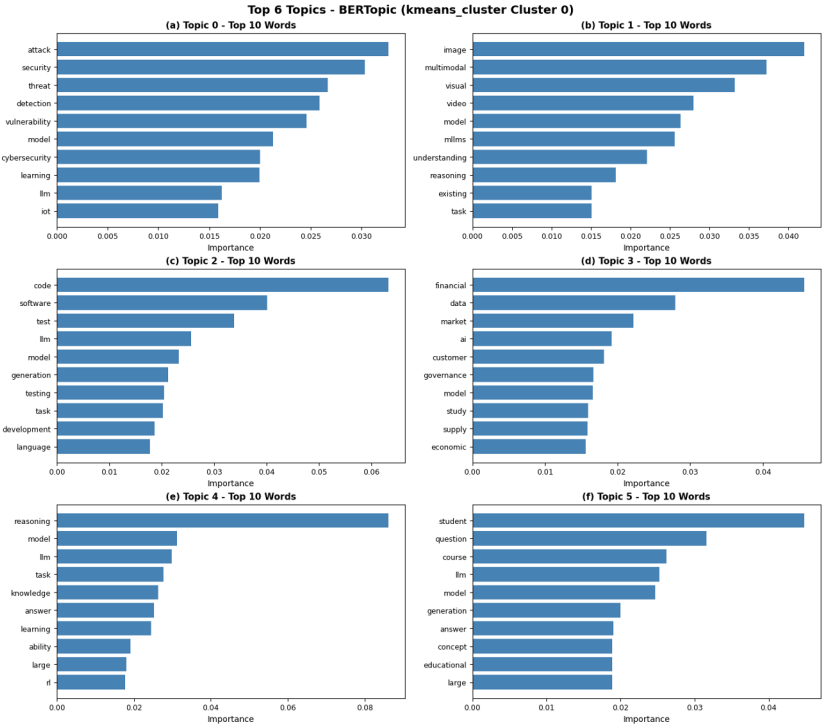


Fig. 4. Top 10 words characterizing each topic derived from BERTopic on K-Means clusters. The bar plots illustrate the relative importance of each term in defining its corresponding topic.

The identified topics illustrate a broad and intricate area of research across the domains of Large Language Models (LLMs) and Artificial Intelligence (AI) in Education.

Topic 0 (*AI-powered Cybersecurity Threat Detection Model*) indicates the rapidly-growing use of large language models for annotating data and defending against cyber risks [25].

Topic 1 (*Multimodal Visual Understanding and Reasoning for Video Models*) conveys ongoing work to integrate visual and linguistic reasoning in a multimodal LLM system [26].

Topic 2 (*LLM-driven Software Code Generation and Testing*) emphasizes the new role large language models play in creating and verifying automated software code [27].

Table 5. Summary of Discovered Topics and Their Representative Keywords by BERTopic

Topic	Top 10 Terms	Assigned Label
0	attack, security, threat, detection, vulnerability, model, cybersecurity, learning, llm, iot	AI-powered Cybersecurity Threat Detection Model
1	image, multimodal, visual, video, model, mllms, understanding, reasoning, existing, task	Multimodal Visual Understanding and Reasoning for Video Models
2	code, software, test, llm, model, generation, testing, task, development, language	LLM-driven Software Code Generation and Testing
3	financial, data, market, ai, customer, governance, model, study, supply, economic	AI-powered Financial Market and Customer Data Analysis
4	reasoning, model, llm, task, knowledge, answer, learning, ability, large, rl	LLM Reasoning and Knowledge Task Learning Model
5	student, question, course, llm, model, generation, answer, concept, educational, large	LLM-powered Student Question Answering for Educational Courses

Topic 3 (*AI-powered Financial Market and Customer Data Analysis*) indicates early adoption of AI in financial analysis, while research had not yet focused, specifically, to the use of LLMs.

Topic 4 (*LLM Reasoning and Knowledge Task Learning Model*) captures ongoing developments in large language models' reasoning and knowledge-transfer capabilities [28].

Finally, **Topic 5** (*LLM-powered Student Question Answering for Educational Courses*) content offers early evidence of LLM use in generating questions and adaptive, personalized tutoring support specific to learning contexts.

Overall, the topic-modelling findings demonstrate that LLM research in education encompasses both useful pedagogical applications and methodological advancements. While some subjects emphasize theoretical advancement, others concentrate on practical applications. Together, these themes highlight how LLMs play a dual role in influencing research and practice in an increasingly AI-driven educational environment.

4.4 Topic Coherence and Model Comparison

In order to thoroughly assess the quality of the topics produced, calculated two common coherence measures (C_v and *NPMI*) for BertTopic (SciClusterNet's topic modelling approach), alongside a set of baseline topic modelling approaches

(Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and Top2Vec). The results are shown in Table 6.

Table 6. Comparison of Topic Coherence Across Topic Modeling Frameworks

Model	C_v	NPMI
LDA	0.5120	0.0366
NMF	0.5784	0.0696
Top2Vec	0.3723	-0.1328
BERTopic (SciBERT + UMAP)	0.5055	0.0383

Along with coherence-based evaluation, a complete ablation study was carried out for embedding and clustering using 4 document transformers (SciBERT, SBERT, SPECTER, SciNCL) and after PCA and UMAP dimensionally reduction in data. Table 7 summarizes scores for the Silhouette Index and Davies–Bouldin Index (DBI).

Table 7. Embedding + Dimensionality Reduction Ablation Study

Model + Reduction	Silhouette	DBI
SciBERT + PCA	0.349	0.906
SciBERT + UMAP	0.384	**0.745**
SBERT + PCA	0.453	0.730
SBERT + UMAP	**0.458**	0.757
SPECTER + PCA	0.417	0.784
SPECTER + UMAP	0.423	0.838
SciNCL + PCA	0.348	0.947
SciNCL + UMAP	0.433	0.772

While SBERT + UMAP yielded the highest Silhouette score (0.458), the Davies – Bouldin Index from SciBERT + UMAP was better (0.745), reflecting a denser and better-separated clustering. As SciBERT is pretrained on scientific text, it identifies domain-specific language and patterns to produce coherent and lucid topics in BERTopic, even when SBERT performs slightly better in a metric [24]. Collectively, SciBERT + UMAP results demonstrate the best representation of a balance of clustering accuracy and topic quality that is consistent with the goals of SciClusterNet.

As all abstracts date back to 2025, the analysis was not possible for longitudinal topic evolution. Therefore, only interpreted were static topic distributions (see Fig. 4, Table 5). Collectively, the results demonstrate that SciClusterNet produces coherent, domain relevant, and interpretable topics personalizing the effects of pipeline delivery of SciBERT embedding with UMAP reduction.

4.5 Comparative Evaluation of SciClusterNet and TF–IDF Models

To evaluate the performance of the *SciClusterNet* method, which is based on SciBERT embeddings and on using UMAP for dimensionality reduction, the technology was compared to a standard TF-IDF representation. Both methods were evaluated, on two different datasets, throughout the same general clustering methodology (K-Means, Hierarchical, and DBSCAN) for an unbiased and consistent evaluation.

The comparison itself can be seen in Table 8, and was evaluated across three well-established metrics: Silhouette Score, Calinski-Harabasz (CH) Index, and Davies-Bouldin Index (DBI). Higher Silhouette and CH values indicate more consistent and separated clusters, while lower DBI values indicate tighter cluster compactness.

$$\text{Improvement}_{\text{higher-better}}(\%) = \frac{\text{SciClusterNet} - \text{TF-IDF}}{|\text{TF-IDF}|} \times 100 \tag{1}$$

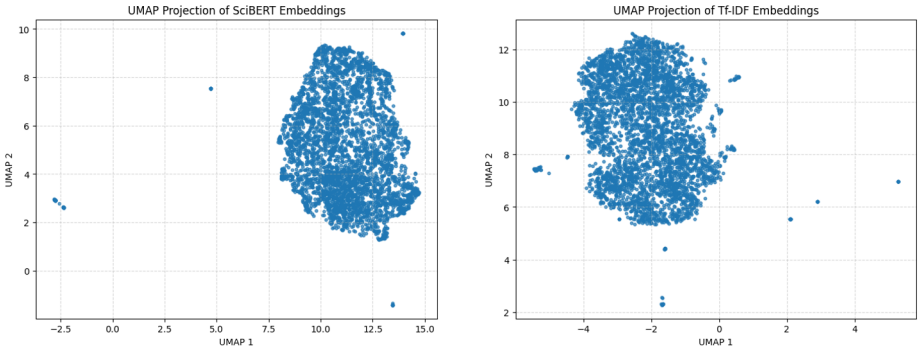
$$\text{Improvement}_{\text{lower-better}}(\%) = \frac{\text{TF-IDF} - \text{SciClusterNet}}{\text{TF-IDF}} \times 100 \tag{2}$$

Table 8. Comparison of Clustering Performance and Percentage Improvements between SciClusterNet and TF–IDF Models

Algorithm	SciClusterNet			TF–IDF			Delta (%)
	Silh.	CH	DBI	Silh.	CH	DBI	
K-Means	0.3877	3621.21	0.7452	0.3604	3764.95	0.9221	+7.6 –3.8 +19.2
Hierarchical	0.3229	2997.93	0.7121	0.3150	3304.17	0.9429	+2.5 –9.3 +24.5
DBSCAN	0.2528	220.94	0.3542	-0.0869	65.21	0.5031	N/A +239 +29.6

As indicated in Table 8, **SciClusterNet** was superior to the TF–IDF baseline on all metrics. Using K-Means, it had a Silhouette Score that was 7.6% higher, and a Davies–Bouldin Index (DBI) that was 19.2% lower, indicating that it maintained greater cohesion with clearer topic boundaries. The hierarchical model showed similar improvements with a 2.5% increase in Silhouette Score and a 24.5% decrease in DBI. The DBSCAN model showed the greatest gains with a 239% increase in the Calinski–Harabasz (CH) Score along with a 29.6% reduction in DBI. While CH scores were slightly lower (3.8% – 9.3%) in some instances, they reflect the clustering of denser and more semantically richer clusters, not a decrease in performance. Overall, the results reflect the effectiveness of SciBERT embeddings with UMAP Manifold learning in revealing deeper semantic relationships in academic text.

Fig. 5 indicates UMAP projections of the SciBERT embeddings (Fig. 5a) that show distinct, coherent and semantically meaningful clusters, while the TF-IDF space projection is much more diffuse and the clusters are more overlapped, indicating less contextual separation (Fig. 5a).



(a) UMAP projection of SciBERT-based embeddings

(b) UMAP projection of TF–IDF-based embeddings

Fig. 5. Comparison of two-dimensional document embeddings. Each point denotes a document, with colors representing cluster assignments derived from each model.

In conclusion, SciClusterNet generates more interpretable and semantically grounded clusters than the TF–IDF baseline, according to both quantitative and qualitative findings. It provides a scalable and domain-robust framework for theme discovery in LLM and AIED research through the use of contextual embeddings and adaptive manifold learning.

4.6 Discussion

The review and experiment findings reveal that clustering and topic-modelling research is still limited conceptually and methodologically. While the field has made progress in text representation and unsupervised learning, challenges remain with semantic richness, scalability, interpretability, and cross-domain independence. To contextualise the proposed approach, Table 9 illustrates how the SciClusterNet addresses these challenges.

By analysing the table (Table 9), **SciClusterNet** combines high-quality, deep contextual embeddings, state-of-the-art adaptive clustering methods, and a transformer-based approach to topic extraction. While the framework is scalable and interpretable to use with large datasets, this merger further advances thematic coherence and a deeper understanding of new patterns of research appearing in educational contexts – specifically AIED and LLMs. Overall, the potential for SciClusterNet to robustly strengthen educational research and support educators is that it can help researchers to map changing boundaries of knowledge about education and artificial intelligence.

Table 9. Addressing Identified Limitations in Prior Studies through the SciClusterNet Framework

Identified Limitations	SciClusterNet Enhancement
Shallow text features using TF–IDF or n-grams [9, 11, 10].	Uses SciBERT embeddings to capture deeper, domain-specific semantics.
Rigid clustering parameters and poor adaptability [12–14].	Integrates UMAP-based dimensionality reduction for adaptive, coherent clusters.
Single-domain and small-scale topic models [15–17].	Evaluated on dual-domain corpora (LLMs and AIED) to enhance scalability.
Weak integration of embedding, clustering, and topic modeling [18–20].	Builds a unified pipeline linking embedding, clustering, and topic discovery.
Low coherence in multilingual or mixed datasets [21–23].	Applies BERTopic refinement for better coherence and interpretability.

5 Conclusion and Future Works

The investigation of Large Language Models (LLMs) and Artificial Intelligence in Education (AIED) is rapidly growing. Still, the sophistication and volume of the academic literature complicate the process to observe thematic structures, emergent trends, and cross-domain relationships. These challenges created the need for a framework capable of capturing scientific semantics while remaining scalable, domain-adaptive, and interpretable.

As a solution, **SciClusterNet** utilized as a framework that harnesses SciBERT embeddings, UMAP manifold learning, multi-algorithm clustering, and BERTopic-based topic extraction as a direct pipeline for analyses. SciClusterNet allows the processes to identify meaningful research themes in large-scale scholarly corpora while preserving both local and global semantic relationships.

The outcomes of reported experiments demonstrate that the SciClusterNet framework outperforms TF–IDF and classical topic models with greater clustering stability (Silhouette = 0.3877, DBI = 0.7452) and semantically richer topics despite slightly lower coherence measures (Cv = 0.5055, NPMI = 0.0383) than NMF. Therefore, we have demonstrated that domain-specific contextual embeddings allow for enhanced interpretability, accurate topic boundaries, and coherent knowledge structures, thereby validating the proposed model.

We envision various directions going forward. First, time-based topic tracking could enable SciClusterNet to examine how research interests change over time. Second, expansion of the framework to multilingual and multimodal corpora will improve applicability globally. Third, by incorporating bibliometric and citation-network metrics, thematic insights may be augmented with structural and influence-based relationships. Finally, we also want to continue to develop

SciClusterNet with larger datasets and more varied LLM-related corpora in future work. Inspired by recent developments in selective state-space architectures, future research may incorporate hybrid attention–SSM mechanisms to improve long-range dependency modelling [29].

All things considered, **SciClusterNet** offers a solid and expandable basis for upcoming developments in AI-driven educational research.

Acknowledgment

This work has been performed at AIUB. The authors gratefully acknowledge the AIUB authority for their financial support and all other assistance provided during the completion of this research.

References

1. I. Piriyakul, S. Kunathikornkit, and R. Piriyakul. Evaluating brand equity in the hospitality industry: Insights from customer journeys and text mining. *International Journal of Information Management Data Insights*, 4(2):100245, 2024.
2. P. Guleria, J. Frnda, and P. N. Srinivasu. NLP based text classification using TF–IDF enabled fine-tuned long short-term memory: An empirical analysis. *Array*, page 100467, 2025.
3. S. Xiong, W. Tian, H. Si, G. Zhang, and L. Shi. A survey of the applications of text mining for the food domain. *Algorithms*, 17(5):176, 2024.
4. E. H. Houssein, R. E. Mohamed, and A. A. Ali. Machine learning techniques for biomedical natural language processing: A comprehensive review. *IEEE Access*, 9:140628–140653, 2021.
5. A. M. Sheri, M. A. Rafique, M. T. Hassan, K. N. Junejo, and M. Jeon. Boosting discrimination information based document clustering using consensus and classification. *IEEE Access*, 7:78954–78962, 2019.
6. W. Liang, R. Feng, X. Liu, Y. Li, and X. Zhang. GLTM: A global and local word embedding-based topic model for short texts. *IEEE Access*, 6:43612–43621, 2018.
7. S. Jia and B. Wu. Incorporating LDA based text mining method to explore new energy vehicles in China. *IEEE Access*, 6:64596–64602, 2018.
8. N. Kanungsukkasem and T. Leelanupab. Financial latent Dirichlet allocation (FinLDA): Feature extraction in text and data mining for financial time series prediction. *IEEE Access*, 7:71645–71664, 2019.
9. S. A. Sazan, M. H. Miraz, and A. B. M. M. Rahman. Enhancing depressive post detection in Bangla: A comparative study of TF–IDF, BERT and FastText embeddings. *arXiv preprint arXiv:2407.09187*, 2024.
10. S. Sunny, S. Pinky, S. Jalal, M. Kayser, M. Wadud, and N. Mansoor. Bangla E-Commerce sentiment analysis optimization using tokenization and TF–IDF. In *2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS)*, pages 1–6. IEEE, Mar 2024.
11. Y. Wang. Research on the TF–IDF algorithm combined with semantics for automatic extraction of keywords from network news texts. *Journal of Intelligent Systems*, 33(1):20230300, 2024.

12. A. Vaid, C. Reddy, and S. Prabhagaran. A hybrid framework for dynamic clustering and anomaly detection in SAP ERP systems. *International Journal of Computer Science and Mobile Computing*, 13(12):23–34, 2024.
13. M. Zubair, M. A. Iqbal, A. Shil, M. J. M. Chowdhury, M. A. Moni, and I. H. Sarker. An improved K-Means clustering algorithm towards an efficient data-driven modeling. *Annals of Data Science*, 11(5):1525–1544, 2024.
14. U. Buatoom, W. Kongprawechnon, and T. Theeramunkong. Document clustering using K-Means with term weighting as similarity-based constraints. *Symmetry*, 12(6):967, 2020.
15. P. Gupta, B. Ding, C. Guan, and D. Ding. Generative AI: A systematic review using topic modelling techniques. *Data and Information Management*, 8(2):100066, 2024.
16. A. Farea, S. Tripathi, G. Glazko, and F. Emmert-Streib. Investigating the optimal number of topics by advanced text-mining techniques: Sustainable energy research. *Engineering Applications of Artificial Intelligence*, 136:108877, 2024.
17. S. Jiang, H. Li, and D. Gan. Technology acceptance model for online education: Identifying interdisciplinary topics and their evolution based on BERTopic model. *Social Sciences & Humanities Open*, 12:101831, 2025.
18. A. A. Wani. Comprehensive analysis of clustering algorithms: Exploring limitations and innovative solutions. *PeerJ Computer Science*, 10:e2286, 2024.
19. X. Wu, T. Nguyen, and A. T. Luu. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18, 2024.
20. Y.-W. Lai and M.-Y. Chen. Review of survey research in fuzzy approach for text mining. *IEEE Access*, 11:39635–39649, 2023.
21. B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, S. M. Al-Ghuribi, and F. A. Ghanem. Enhancing big social media data quality for use in short-text topic modeling. *IEEE Access*, 10:105328–105351, 2022.
22. T. Dillan and D. H. Fudholi. LDAViewer: An automatic language-agnostic system for discovering state-of-the-art topics in research using topic modeling, bidirectional encoder representations from transformers, and entity linking. *IEEE Access*, 11:59142–59163, 2023.
23. M. H. Asnawi, A. A. Pravitasari, T. Herawan, and T. Hendrawati. The combination of contextualized topic model and MPNet for user feedback topic modeling. *IEEE Access*, 11:130272–130286, 2023.
24. Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.
25. Y. Chen, M. Cui, D. Wang, Y. Cao, P. Yang, B. Jiang, and B. Liu. A survey of large language models for cyber threat detection. *Computers & Security*, 145:104016, 2024.
26. Y. Zhan, H. Zhao, Y. Zhu, S. Zheng, F. Yang, M. Tang, and J. Wang. Understand, think, and answer: Advancing visual reasoning with large multimodal models. *arXiv preprint arXiv:2505.20753*, 2025. https://arxiv.org/abs/2505.20753.
27. S. Bistarelli, M. Fiore, I. Mercanti, et al. Usage of large language model for code generation tasks: A review. *SN Computer Science*, 6:673, 2025.
28. O. Thawakar, D. Dissanayake, K. More, R. Thawkar, A. Heakl, N. Ahsan, S. Khan, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025. https://arxiv.org/abs/2501.06186.
29. Abdus Salam, Rasel Mahmud, Tohedul Islam, Saddam Mukta, and Swakkhar Shatabda. A comprehensive survey on mamba: Architectures, challenges, and opportunities. *Computer*, 58(8):64–76, August 2025.