Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Original Article

Fine-tuning SBERT for Semantic Research Title Classification in Trilingual University Repository

Havan Wahid Rashid [a] *, **Sarkar Hasan Ahmed** [b]

[a] Information Technology Department, Technical College of Informatics, Sulaimani Polytechnic University, Sulaymaniyah, Iraq.
[b] Computer Network Department, Technical College of Informatics, Sulaimani Polytechnic University, Sulaymaniyah, Iraq.

Submitted: 23 May 2025
Revised: 18 June 2025
Accepted: 1 August 2025

* **Corresponding Author**: havan.wahid.tci@spu.edu.iq

Keywords: Recommendation System, Sentence-BERT, Content-based filtering, Kurdish higher education, Paper classification.

How to cite this paper: H.W. Rashid, S. H. Ahmed, "Fine-tuning SBERT for Semantic Research Title Classification in Trilingual University Repository", KJAR, vol. 10, no. 2, pp: 119-135, Dec 2025, doi: 10.24017/science.2025.2.9

Abstract: Recommendation systems are essential for automatically surfacing relevant content from large datasets, reducing search time, and facilitating discovery. In academia, content-based recommendation systems are beneficial when only brief titles are available and multilingual text is standard. Universities in the Kurdistan Regional Government currently lack a centralized research repository, with records scattered across different institutions and often manually maintained. This makes it difficult for students and faculty to find related topics, potential supervisors, or cross-disciplinary connections. This paper presents a trilingual (English, Arabic, and Kurdish) recommendation system for academic research titles. Three key contributions are made: (1) the creation of the first integrated dataset of 4,257 research titles from Sulaimani Polytechnic University publicly available; (2) the development of a web-based platform for semantic search and title-level recommendations to support research discovery and student–supervisor matching; and (3) an evaluation between Sentence-BERT models—all-MiniLM-L6-v2 and paraphrase-multilingual-MiniLM-L12-v2—before and after fine-tuning with a domain-specific taxonomy and cosine embedding loss. Performance is assessed using Precision@5, Mean Reciprocal Rank, and NDCG@5 with expert-annotated relevance judgments for 20 query titles. Fine-tuning resulted in performance improvements, with paraphrase-multilingual-MiniLM-L12-v2 achieving Precision@5 of 0.94 and NDCG@5 of 0.991. The English-only model also showed improvements, Precision@5: 0.79→0.82; NDCG@5: 0.885→0.922.

1. Introduction

Universities in the Kurdistan Regional Government of Iraq currently face significant challenges due to the absence of a unified, systematically maintained repository of scholarly research titles. Each university independently manages its academic records, often employing manual or fragmented digital methods, severely limiting information accessibility and sharing. Consequently, students and faculty members experience difficulty discovering relevant research topics, identifying potential supervisors, and engaging in interdisciplinary collaborations [1]. This siloed approach to manage research data contributes to delays in student graduation and hinders the advancement of cross-departmental scholarly work [2].

Recommendation systems are a technology that can be integrated into this proposed system to make it smarter, enhance user engagement, and personalize the user experience. Using learned relevance signals, recommendation systems predict user interests and rank items from large catalogs. Core approaches include content-based filtering, which represents items and queries with text features and

computes similarity, and collaborative filtering, which infers preferences from historical interactions; hybrid designs combine both to improve accuracy and coverage. In academic settings, recommendation systems map research titles and queries into vector representations, retrieve semantically similar works, and re-rank results using domain cues, thereby accelerating topic discovery, supervisor matching, and cross-department collaboration. When only titles are available and texts are multilingual, content-based methods with sentence embeddings effectively reduce search time and increase the visibility of prior research across departments and universities [3].

Most of the university websites in KRG come with Kurdish language in addition to English and Arabic. Kurdish language is an Indo-Iranian language with two major standardized varieties—Kurmanji and Sorani—used across Iraq, Iran, Turkey, and Syria. Kurmanji is typically written in a Latin-based alphabet of about 31 letters, while Sorani employs an Arabic-derived script with roughly 33 letters [4, 5]. In the Kurdistan Region of Iraq, Kurdish has official status; however, orthographic variation, diacritics, and Unicode-specific issues (e.g., zero-width non-joiner, hamza variants) complicate normalization and tokenization, especially for short titles [5, 6]. Code-switching with Arabic and English and limited high-quality Natural Language Processing (NLP) resources further increase sparsity and out-of-vocabulary rates, making robust multilingual sentence embeddings and domain-specific fine-tuning important for effective retrieval and recommendation [6, 7]. However, despite the proven effectiveness of recommendation systems in enhancing retrieval and personalization, there is a notable absence of such systems tailored for the Kurdish language, mainly due to its linguistic challenges and limited NLP resources.

In this research, a model is proposed to fill this gap by (1) creating a dataset comprising 4,257 publication titles collected from various colleges and departments within Sulaimani Polytechnic University (SPU). This dataset represents the first integrated academic dataset for research titles for SPU in the Kurdistan Region, with provisions for daily automatic updates, (2) developing a web-based content-based recommendation system to support student–supervisor matching; and (3) conducting an empirical evaluation of the two content-based recommendation models, paraphrase-multilingual-MiniLM-L12-v2, and all-MiniLM-L6-v2.

The remainder of this paper is organized as follows. Section 2 introduces recommendation systems and reviews related work. Section 3 details the methodology and the proposed system. Section 4 reports results for the two recommendation models, and sections 5 and 6 present the discussion and conclusions, respectively.

2. Related Works

Recommender systems have been widely utilized in higher education to match students effectively with appropriate resources, such as project topics or thesis supervisors. Ensuring an optimal student–supervisor fit significantly impacts students' academic success and timely graduation [2]. Traditionally, universities manually assign thesis advisors, a method that can be inefficient, biased, and prone to overlooking student interests, especially without a centralized mechanism to compare research topics across departments.

Ismail *et al*. [8] Used Euclidean distance similarity to match students with supervisors for their final-year project and introduced content-based recommended systems to automate and improve this process by aligning student interests with faculty expertise and availability. Meanwhile, Amaad *et al.* [9] Evolved an academic recommendation system from simple keyword-matching techniques to sophisticated machine learning approaches. Traditional content-based filtering (CBF) and collaborative filtering (CF) methods often fail to incorporate contextual information and sequential access behavior, making them prone to generating irrelevant recommendations in scholarly environments. These limitations have led researchers to explore hybrid approaches that combine multiple recommendation strategies to improve accuracy and relevance.

Stergiopoulos *et al*. [10] Have focused on addressing the scalability challenges of academic recommendation systems and presented a multi-staged recommendation system based on clustering, graph modeling, and deep learning that manages to operate on datasets containing millions of users and

papers. Their approach demonstrates how to bridge the gap between academic state-of-the-art systems and real-world applications, achieving improvements compared to traditional methods.

Church *et al*. [11] Argue that content-based filtering and graph-based methods complement academic search recommendations, with CBF using abstracts to infer authors' positions and graph-based methods using citations to capture audience responses. Their work with Specter (BERT-like encodings) and ProNE (spectral clustering) demonstrates the synergistic potential of hybrid approaches.

Mohamed *et al*. [12] Have developed specialized recommendation systems for various academic contexts, proposing concept-based methods for representing researchers' interests, where concept generation depends on the semantics of words in articles related to researchers, while Albusac *et al*. [13] Focus on expert finding systems that recommend researchers based on their published articles and query matching.

Content-based filtering has been extensively studied in academic contexts, with researchers exploring various approaches to represent and match document content. Content-based models combining deep neural networks and factorization machines for scientific article recommendations address the challenge of high-order feature interactions that existing models often miss [14]. A Bayesian nonparametric hybrid filtering approach combines user preferences modeling through the infinite relational model with topic modeling for item features. Rodriguez and Vuppala [15] demonstrate how content-based and collaborative filtering can be effectively integrated through shared item partitions.

Advancements in content-based filtering have focused on improving text representation methods, developing a scalable end-to-end content-based scientific paper recommendation system capable of recommending papers based on abstracts or contextual information [16]. The integration of semantic understanding has become increasingly crucial for semantic relationship embeddings for text classification, considering synonymy, hyponymy, and hypernymy relationships extracted from Wikipedia to improve content representation beyond traditional word-based models [17]. Sentence-BERT (SBERT) has emerged as a powerful tool for generating semantically meaningful sentence embeddings. The effectiveness of SBERT in educational contexts, using similarity learning to optimize teacher report embeddings for academic performance prediction, is seen in achieving 73% accuracy in detecting strong performance [18]. Another work is a comprehensive approach using SBERT in research literature recommendation systems, combining it with latent dirichlet allocation models in a semi-supervised methodology. Their work shows how SBERT can capture contextual information while maintaining global topic information, leading to improved recommendation performance over traditional baselines [19].

Another work used supervised transfer learning with supervised fine-tuning using meta-/few-shot strategies; Models: BERT, BERT+BiLSTM, all-MiniLM-L6-v2; Precision: all-MiniLM-L6-v2 improved from 0.74 to 0.91 after fine-tuning [20].

Mohamad *et al.* [21] Carried out a comparative evaluation of five pre-trained sentence encoders (USE, BERT, InferSent, ELMo, and SciBERT) for research paper recommendation. Their findings indicate that while semantic information from these encoders alone does not improve performance over traditional BM25 techniques, their integration enables the retrieval of relevant papers that conventional ranking functions may not capture. The effectiveness of SBERT in news recommendation systems is seen in achieving 99.14% precision, 92.48% recall, and 95.69% F1-score when combined with neural collaborative filtering approaches [22]. Recent studies have explored sophisticated applications of SBERT in various domains. Cascaded transformer mechanisms using SBERT derived from large language models (T5 and BERT) for apparel recommendation systems achieved 94.76% similarity scores [23]. BERT approaches that combine SBERT with RoBERTa, treating sentences as tokens to capture both intra-sentence and inter-sentence relations [24].

Several studies have investigated fine-tuning in domain-based settings, such as Li *et al*. [25], who applied fine-tuning transformer models for domain-specific tasks, which became a critical area of research. Fine-tuning strategies for BERT found that multitask fine-tuning with smoothness-induced adversarial regularization achieves the best overall results across sentiment classification, paraphrase detection, and semantic text similarity tasks. Meanwhile, Zhang *et al.* [26]. Applied an XR-Transformer, a recursive approach, to accelerate fine-tuning procedures through multi-resolution objectives. Their method demonstrates 20x faster training than X-Transformer while improving Precision@1 from 51%

to 54% on the Amazon-3M dataset. Zhang *et al.* [27] applied domain-specific fine-tuning, which has shown particular promise in academic and scientific contexts, and developed label attention-based architectures for biomedical text classification, injecting semantic label descriptions into the fine-tuning process of pretrained models. Their approach outperforms conventionally fine-tuned models on medical datasets. Mücke *et al.* [28] focused on fine-tuning language models for scientific writing support, training models on sentences from peer-reviewed papers to determine scientific section classification and paraphrasing suggestions. Their work demonstrates how domain-specific fine-tuning can achieve specialized performance in academic contexts. However, Dhamecha *et al.* [29] explored multilingual fine-tuning in Indo-Aryan languages, showing that careful selection of related languages can significantly improve performance over individual language models. Their work reveals that low-resource languages like Oriya and Punjabi benefit most from multilingual fine-tuning, with relative performance improvements.

Ma *et al.* [30] Traditional classification approaches evolved from manual methods to sophisticated automated systems. The ERNIE-BiGRU-Attention model fuses titles, abstracts, and keywords into pretrained models, using bidirectional gated recurrent units and attention mechanisms to improve classification accuracy. Traditional approaches often relied on static word embedding methods like TF-IDF or Word2Vec. Liu *et al*. analyzed document classification based on Word2Vec and TF-IDF, proposing improved weighting methods to address the limitation that Word2Vec embeddings ignore word importance within documents [31]. Modern classification systems leverage sophisticated neural architectures. Liu *et al*. [29] developed text classification models based on label embedding and attention mechanisms, using convolutional neural network to learn compatibility between words and labels as attention values for Bi-LSTM outputs.

Constructing datasets for academic literature relevance ranking using BERT fine-tuning demonstrates how learnable search ranking models can improve literature search efficiency across different disciplinary fields [32]. The complexity of research paper classification has led to hierarchical approaches. The HFT-ONLSTM model addresses hierarchical multi-label text classification using fine-tuning techniques where upper-level classification results contribute to lower-level decisions, reducing computational costs while achieving superior performance [33].

Educational technology systems in low-resource contexts face unique challenges related to computational limitations, data scarcity, and infrastructure constraints. Educational recommender system techniques indicate that traditional approaches combined with deep learning networks can improve recommendations and provide higher accuracy results [34].

Hybrid recommendation methodologies for educational video systems, incorporating both content-based filtering and collaborative filtering algorithms, are needed to address the diverse needs of students in informal education settings [35]. Research has focused on developing adaptive systems that can function effectively with limited resources. A systematic review examined recommender systems in higher education, finding that various approaches, including content-based filtering, collaborative filtering, and knowledge-based methods, can be effectively combined for better results [36]. Muzdybayeva *et al*. [37] used matrix factorization-based collaborative filtering frameworks for course recommendations in higher education, analyzing data from 603 students to provide personalized course recommendations based on individual preferences and academic performance.

Recent developments have explored AI-based personalized systems for educational contexts, developing AI-based personalized research paper recommendation systems using content-based filtering algorithms and academic term dictionaries, achieving 50.2% precision for recommendation results and 52.2% for extracted research keywords [38]. Zhao and Ma [39] adopted recommendation systems for higher education based on deep learning, using autoencoder advantages for dimension reduction and achieving 0.90 efficiency on their dataset, while enabling the scoring of different recommended articles.

In summary, prior research on academic recommendation systems highlights a steady progression from traditional keyword-matching and content-based filtering approaches to advanced hybrid and deep learning methods incorporating semantic embeddings, clustering, graph modelling, and fine-tuned transformer models. These studies demonstrate the advantages of integrating contextual signals and domain-specific adaptations to improve scalability, accuracy, and personalization in educational

and research environments. However, despite these advancements, challenges remain in adapting such methods to low-resource settings and multilingual contexts, where data scarcity, linguistic complexity, and infrastructural constraints continue to limit the robustness and applicability of current approaches. Despite the related work above, most of the studies focus on well-resourced languages such as English and Arabic; however, to date, no work or dataset is suitable for a recommendation system in the Kurdish language. There are some examples of existing datasets in the Kurdish language, which are related to creating a corpus for fake news [40], and there is another dataset for sentiment analysis [41].

A synthesized overview of the literature is provided in table 1, which contrasts techniques, models, datasets, results, and stated limitations.

Table 1: Summary of some prior work on academic recommendation systems: techniques, models, datasets, reported metrics, and limitations.

Reference	Model / Technique	Dataset Name	Results
[8]	Content-based recommender system, Euclidean distance similarity, Content-based filtering	Not mentioned	Similarity 44.95%
[9]	Hybrid recommendation system combining CBF and CF, Context-aware sequential pattern mining	Not mentioned	Recall 99%, precision 13%
[10]	Multi-staged recommendation system, Clustering, Graph modeling, Deep learning	DBLP Citation-network-v13	Recall 54.94%, NDCG 24.77%
[11]	BERT-like encodings (Spectre), Spectral clustering (ProNE), Content-based filtering, Graph-based methods	Semantic Scholar (S 2)	cosine between a query and the top candidate recommendation, got about 10% better
[12]	Semantic concept generation model, Concept-based representation	CiteULike	Recall 30%
[13]	Content-based expert recommendation system, Expert finding, Query matching	PMSC-UGR	Precision 81.6%
[15]	Infinite Relational Model with topic modeling, Bayesian nonparametric filtering, Topic modeling	CiteULike	Accuracy 81.4%
[16]	End-to-end content-based system, Content-based filtering, Text representation	Not mentioned	average response time of 1.4 seconds
[17]	Text classification model with semantic relationships, Semantic relationship embeddings	MNIST, CIFAR-10, WOS, IMDB, Reuters, and 20-Newsgroup	accuracy 79%, recall 87%
[18]	Similarity learning, SBERT	Not mentioned	accuracy 73%
[20]	SBERT-based embedding optimization, Supervised fine-tuning	Not mentioned	0.91 precision score
[21]	USE, BERT, InferSent, ELMo, SciBERT	CiteULike	Precision 56%
[22]	SBERT + Neural collaborative filtering	Microsoft news dataset	99.14% precision, 92.48% recall, 95.69% F1-score
[23]	T5 and BERT-derived SBERT	Amazon review dataset	94.76% similarity scores
[24]	SBERT + RoBERTa combination	Goodreads	Precision 78.89%
[25]	Fine-tuned BERT with smoothness-induced regularization	Not mentioned	Not mentioned
[26]	XR-Transformer (recursive approach)	Amazon-3M dataset	20x faster training, Precision 54%
[27]	Fine-tuned pretrained models with semantic labels	Disease-5	Fine-tuned BERT F1 83%, precision 80%
[28]	Language models for scientific writing	IDM-DS, Pegasus-DS, GYAFC	F1 score 90%
[29]	Multilingual language models	Indo-Aryan languages dataset	F-score 0.8848

3. Materials and Methods

3.1. Dataset Construction

3.2. Data Collection

The dataset used a custom web crawler targeting Google Scholar entries associated with @spu.edu.iq email domains to represent SPU research output. The crawler systematically queries Google Scholar for SPU-authored publications, extracting titles based on the institutional email identifier. The crawler periodically updates the dataset with new titles as they become available, ensuring continuous and automatic data enrichment. This automated method significantly reduces manual effort and maintains a real-time snapshot of the university's research activities.

3.3. Preprocessing

After data harvesting, the raw list of titles was checked manually to ensure that the titles were the same as those added by researchers in Google Scholar. The full link of the title from the scholar was not fetched; it was added manually by this code (https://scholar.google.com. $ record ->publication_Path).

An Excel equation identified duplicate titles, but duplicates were retained. Specifically, duplicates representing legitimate repetitions—such as publications co-authored by multiple SPU members appearing under different profiles were preserved such as Knowledge, attitude and practice toward breast cancer among Kurdish women in Sulaimani Governorate/Iraq with ID: 1555 and Knowledge, attitude and practice toward breast cancer among Kurdish women in Sulaimani Governorate/Iraq with ID: 2668; these two tittles were fetched from two different profiles and two different researchers. These were retained because one of the objectives of this work is to discover researchers, and removing any of these titles means less chance of collaboration.

3.3.1. Dataset Statistics

The dataset consists of six columns (Id, Title, User, Publication path, Authors, Publisher details) and it has 4,257 rows of data. It is publicly available on GitHub at this link (https://github.com/havan2018/academic-tittles-dataset). Figure 1 shows the Statistics of Academic Domain Distributions in the dataset. In contrast, figure 2 shows the number of records in the different languages, which are English, Arabic, Kurdish, and others like Persian, Russian, and Urdu.

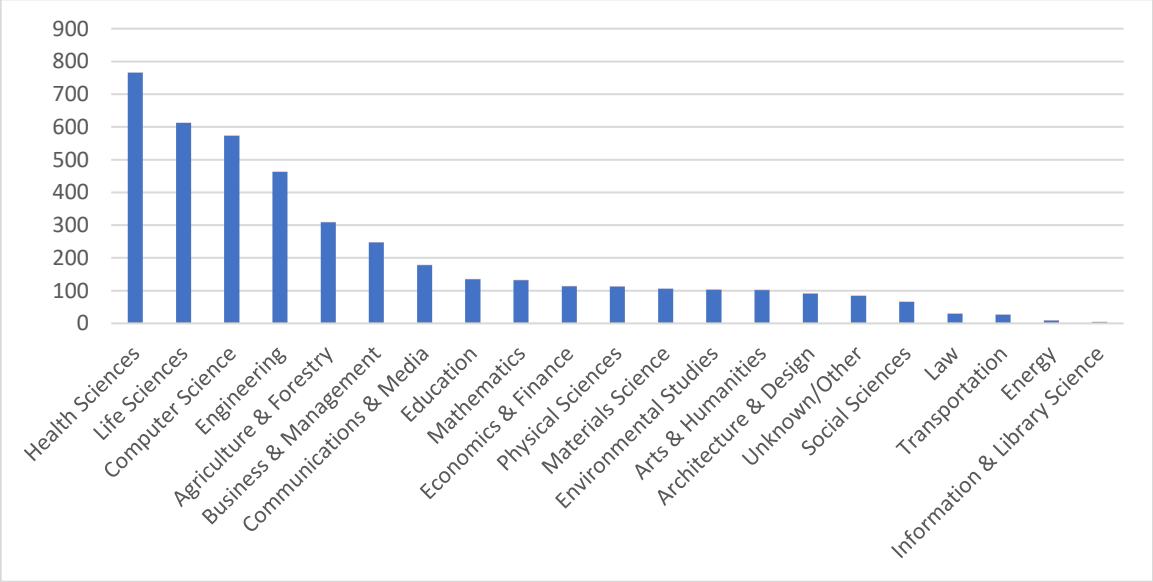

Figure 1: Statistics of academic domain distributions in the dataset.

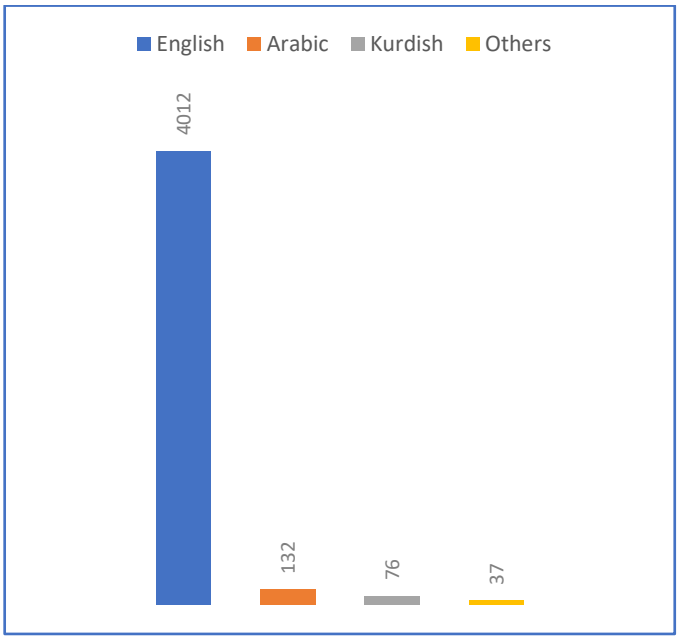

Figure 2: Title distribution per language.

3.4. Labelling Protocol

To facilitate evaluation of the recommendation models, titles were manually labelled by expert annotators from various academic fields, including computer science, medical sciences, and business studies, ensuring accurate domain-specific categorization and enhancing reliability. Due to the dataset's diverse nature and limited size, we adopted a three-level taxonomy—domain, sub-domain, and specific topic. For example, a title could be categorized as engineering / civil engineering / concrete materials, as shown in table 2. These labels were used exclusively for evaluation purposes and training recommendation models.

A quantitative scoring system has been implemented to assess model recommendations based on taxonomy alignment:

- Domain match: +0.3
- Sub-domain match: +0.3 (additional)
- Specific topic match: +0.3 (additional)

A perfect score (0.9) indicates complete alignment, whereas lower scores reflect partial relevance. This hierarchical scoring provided consistent relevance metrics, facilitating accurate comparison among recommendation approaches.

Table 2: Title labeling with three-level taxonomy.

Recommended Title	Label	Domain Match	Sub-domain Match	Specific Topic Match	Total Score
A novel encryption method for secure cloud communication	Computer Science / Cybersecurity / Cloud Security	+0.3	+0.3	+3.0	0.9
A review of image steganography	Computer science / Cybersecurity / Image steganography techniques	+0.3	0	0	0.9
A novel approach for stock price prediction using gradient Boosting machine with feature engineering (GBM-wFE)	Computer science / Artificial intelligence / GBM for stock prediction	0	0	0	0
The English Language Teachers in Iraq Face Challenges and Have Opportunities Beyond Just Giving Knowledge	Education / Language education / Social media and intercultural skills	0.3	0.3	0	0.6

Table 2: Continue

Knowledge, Attitudes, Practices, and the Factors that Influence Breastfeeding among Mothers attending a Primary Health Center in Sulaimani City	Health sciences / Public health / Breastfeeding practices	0.3	0	0.3	0.6
استخدام طلبة الجامعة للفيسبوك وحدود الادمان عليه- دراسة مسحية على عدد من طلاب جامعتي بغداد والسليمانية	Communications & Media / Social media studies / Facebook addiction	0	0	0.3	0.3
ڕۆڵی تیکتۆك لەسەر ڕەفتاری گەنجان لە شاری سلێمانی	Communications & Media / Social media studies / TikTok's impact on youth behavior	0	0	0	0

3.5. SBERT Models

Baseline SBERT is an extension of the original BERT architecture that repackages the encoder in a Siamese (or triplet) configuration and fine-tunes it on semantic-similarity objectives—typically natural language inference (NLI) and paraphrase detection—so that whole sentences (rather than individual tokens) are mapped directly into a fixed-length vector space where cosine distance reflects meaning. At inference time, SBERT passes each sentence once through a transformer encoder, applies a pooling operation (mean or CLS-token), and outputs a dense embedding that can be compared with simple dot-product operations, enabling efficient semantic search and clustering. Two compact SBERT checkpoints are especially popular. all-MiniLM-L6-v2 uses a 6-layer MiniLM backbone (33 M parameters, 384-dimensional embeddings) trained on a large English paraphrase dataset. It offers near-BERT semantic quality while remaining light enough for CPU-level deployment and real-time inference. paraphrase-multilingual-MiniLM-L12-v2 doubles the depth to 12 MiniLM layers (55 M parameters, 768-dimensional embeddings) and is trained on parallel and paraphrase data covering 50+ languages. [42, 43] The larger capacity and multilingual pre-training yield higher retrieval scores—especially for non-English text—but at roughly four times the memory footprint and latency of the 6-layer model. Consequently, all-MiniLM-L6-v2 is often chosen when computational efficiency and web scalability are paramount, whereas paraphrase-multilingual-MiniLM-L12-v2 is preferable when maximum cross-lingual accuracy outweighs resource constraints.

The primary objective was to enable the model to learn fine-grained semantic distinctions among research titles such that conceptually related titles, for example, "feature engineering by machine learning" in computer science, are closer to "machine learning for forecast data" than to "machine engineering in Kurdistan: case study." To do that, every research title in our dataset is tagged with a domain / sub-domain / specific topic. An 80-20 split (3,404 items) was grouped by the concatenated domain + sub-domain key. Inside every group, adjacent titles formed positive pairs. In contrast, each title was additionally paired with a random title from a different group to create a negative pair, yielding approximately 30,000 training pairs. In every mini-batch (size = 16), these pairs were processed in parallel through two weight-shared copies of both models' encoders, producing 384-dimensional embeddings as in the first equation, and optimization employed the cosine-embedding loss.

$$L = \begin{cases} 1 - cos(h1, h2), & y < 1 \\ max(0, cos(h1, h2), & x \geq 0 \end{cases} \quad (1)$$

Where **y** denotes positive (identical *domain/sub-domain*) or negative pairs, and 0.5 is the margin. Five training epochs with Adam (200 warm-up steps) minimized this loss, forcing embeddings of discipline-matched titles toward a cosine similarity of 1 while pushing mismatched pairs below the margin [44].

3.6. System Architecture

After the dataset was created and labeled, a web application was developed using Laravel and Filament, providing an intuitive interface for searching and browsing academic titles. Users input

keywords to retrieve relevant titles, displaying results with metadata linking to detailed pages. Each detail page includes metadata such as authors, publication link, year, and citations, along with a "Find Similar Titles" button. Clicking this button invokes the back-end recommendation engine, which returns related titles. Figure 3 shows the system interface where the user can input the keyword, figure 4 shows the search results, and figure 5 shows the details of each result when the user clicks on the item title. Figure 6 shows the workflow of the system.

The back-end recommendation system comprises independent Python microservices built with Flask, each serving a distinct similarity model of SBERTs when users request recommendations. The Laravel front-end simultaneously queries these services, receiving JSON responses containing recommended titles and similarity scores. Results from each model are displayed separately for easy comparison. This microservice architecture supports modular, scalable development and simplifies maintenance by isolating recommendation algorithms, allowing independent updates and scalability.

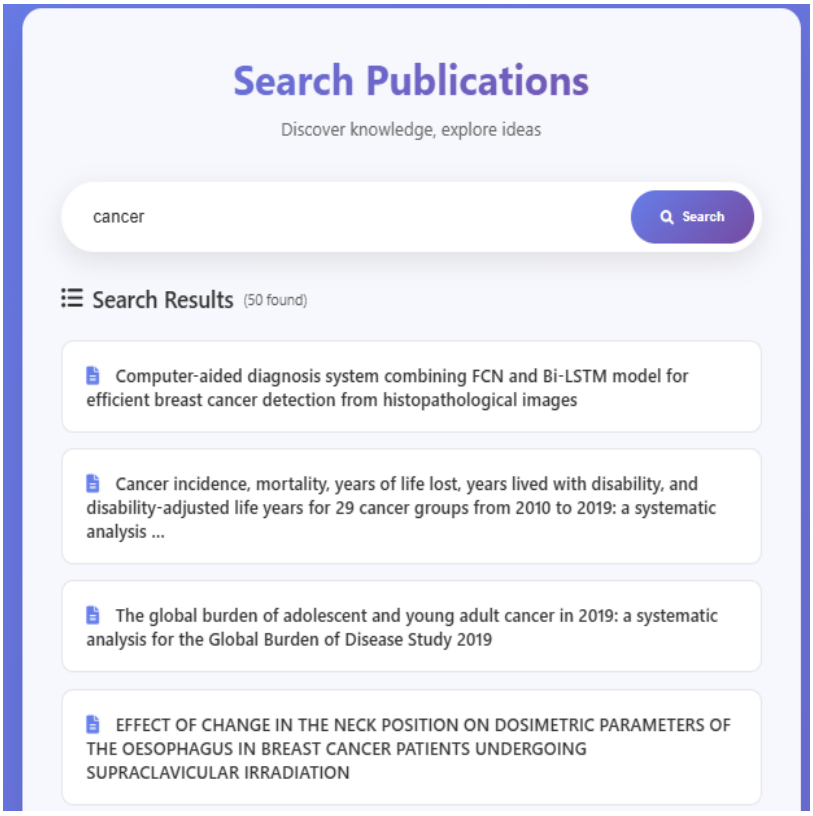

Figure 3: Search engine with the results of the search.

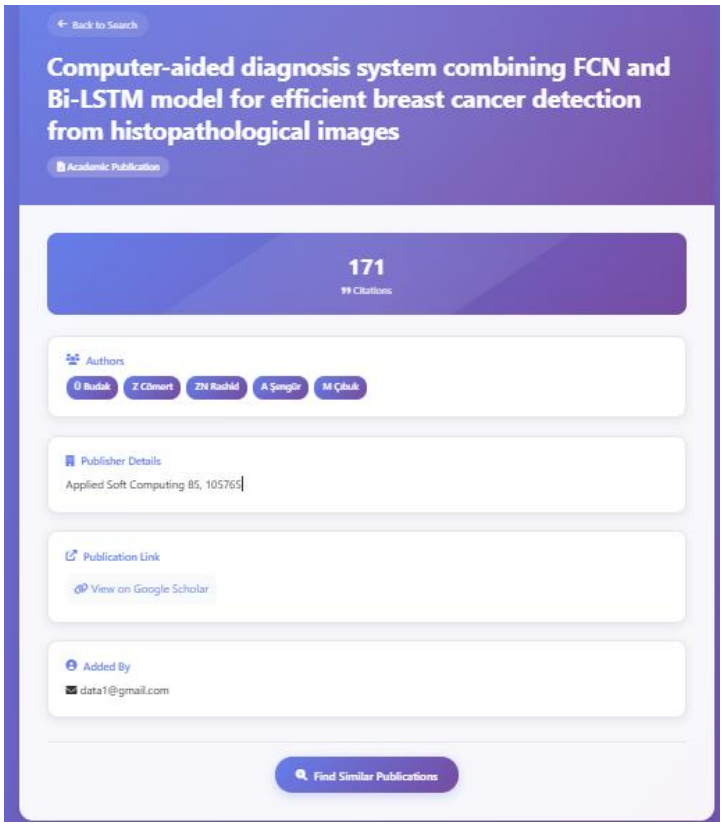

Figure 4: Details of the title.

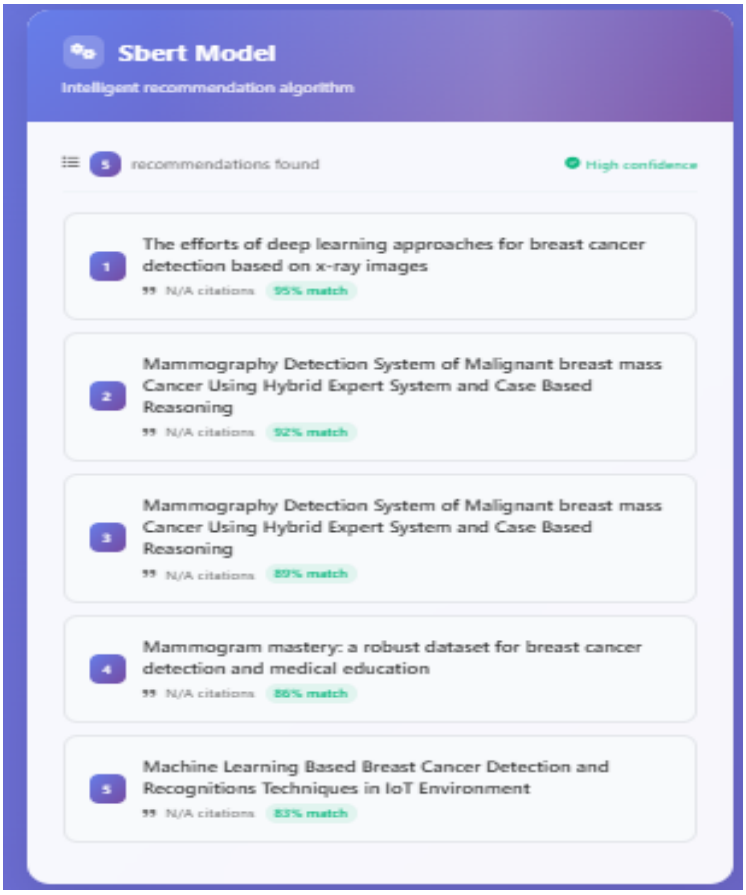

Figure 5: Recommended titles by SBERT model.

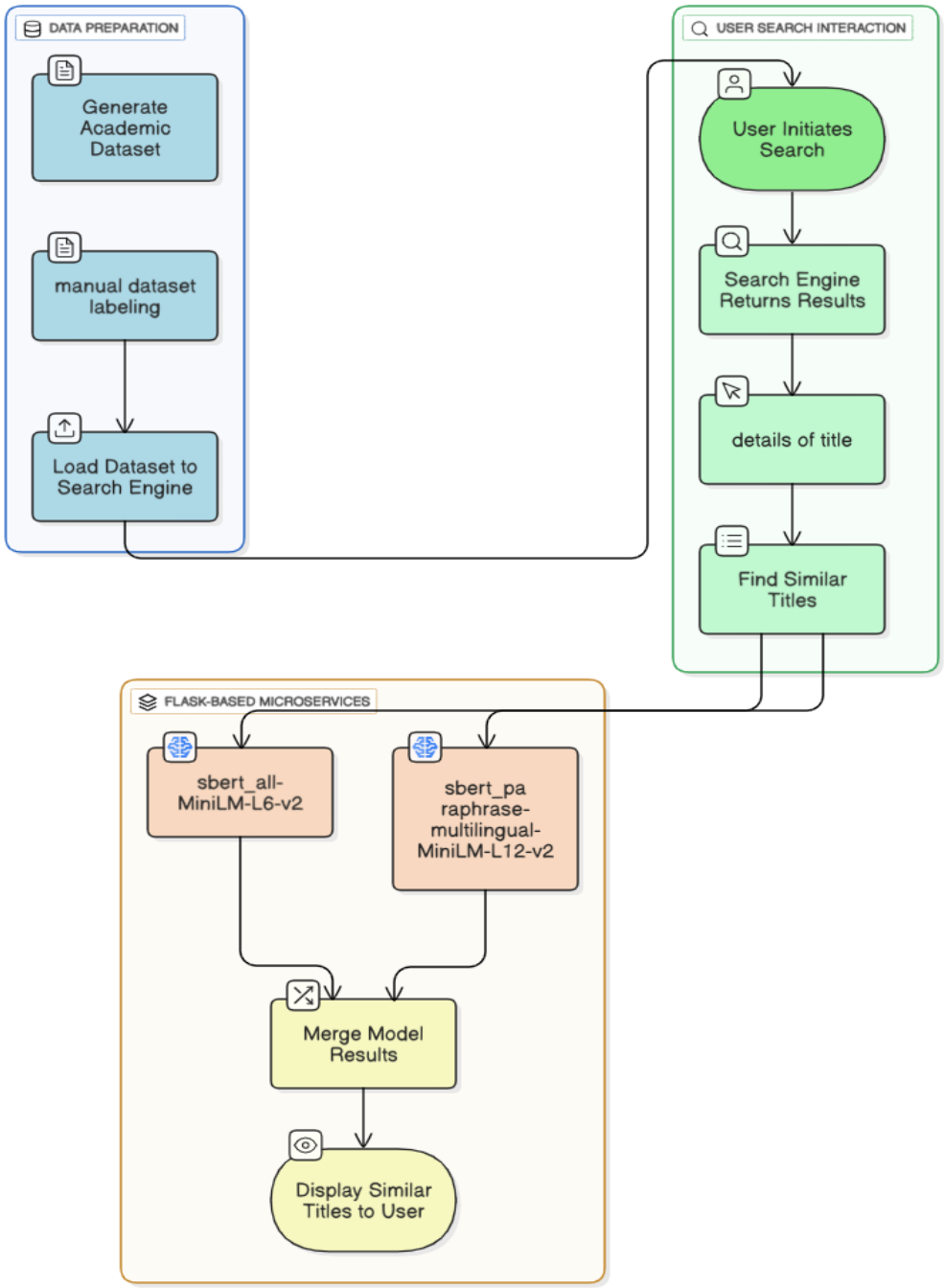

Figure 6: Workflow of the system.

3.7. Evaluation Protocol

This section outlines the evaluation protocol for a content-based recommendation system, designed specifically for a low-resource, title-only dataset. Due to the absence of abstracts or user logs, we used expert annotations for manual relevance judgments as ground truth. Using a representative set of 20 query titles from the dataset, these queries were consistently applied to baseline and fine-tuned models to ensure fair comparison.

3.8. Evaluation Metrics

Metrics have been used for evaluating recommended titles, as in equation 2. Precision at k measures the proportion of relevant recommended items in the top-k set. It is a measure of exactness. For this metric, a suggested title is considered 'relevant' if its graded relevance score meets or exceeds

a threshold of 0.6, signifying an intense match in at least two hierarchical categories (e.g., Domain and Subdomain).

$$\text{P@5} = \frac{|\{\text{Relevant Items}\} \cap \{\text{Top-5 Recommended Items}\}|}{5} \quad (2)$$

MRR evaluates the system's ability to return a relevant item at the highest possible rank. It is susceptible to the position of the *first* correct answer. The reciprocal rank for a single query is the multiplicative inverse of the rank of the first relevant item. The MRR is the average of these reciprocal ranks over all queries.

Like P@5, an item is considered relevant if its graded score is ≥0.6. The MRR is calculated as in equation (3).

$$\text{MRR} = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\frac{1}{\text{rank}_q} \quad (3)$$

Where |Q| is the total number of queries and rank_q is the rank of the first relevant recommendation for query $_q$. If no relevant item appears in the top-5 list, the reciprocal rank for that query is 0.

NDCG@5 is a more sophisticated metric that evaluates the ranking quality by considering both the position and the graded relevance of each item. It rewards placing highly relevant items at the top of the list and penalizes them for appearing lower. Unlike P@5 and MRR, NDCG utilizes the full granularity of graded relevance scores.

The metric, as in equation (4), derived from the discounted cumulative gain (DCG), is calculated as:

Where rel_i is the graded relevance score of the item at rank_i.

$$\text{DCG@5} = \sum_{i=1}^{5}\frac{rel_i}{\log_2(i+1)} \quad (4)$$

To enable fair comparison across queries, the DCG is normalized by the Ideal DCG (IDCG), which is the DCG of a perfectly sorted list of the recommendations.

$$\text{NDCG@5} = \frac{DCD@5}{IDCG@5} \quad (5)$$

The resulting NDCG@5 score is between 0.0 and 1.0, where 1.0 represents a perfect ranking. The final reported metric is the average NDCG@5 as in equation (5) across all test queries. Each recommendation could earn up to 0.9 points if fully matched. A threshold recommendation scoring ≥0.6 points (matching at least two criteria) is considered relevant to simplify scoring.

Computed ranking metrics such as Precision@5, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG) provided a straightforward, objective method to evaluate and compare recommendation quality before and after fine-tuning,

4. Results

This section presents the comprehensive evaluation results of SBERT-based recommendation models before and after fine-tuning, focusing on three key performance metrics:

4.1. Average Metrics by Models (Before vs. After)

The model performance evaluation was conducted using a 20-test set consisting of a diverse selection of academic research queries drawn from the SPU dataset. This test set was designed to represent various domains and subdomains, ensuring a robust assessment of each model's generalization capabilities. Table 3 and figure 5 comprehensively summarize the results, comparing the baseline and fine-tuned configurations of two Sentence-BERT models. The findings highlight that fine-tuning had a significant positive impact on retrieval quality, as both models exhibited measurable improvements across all key evaluation metrics—Precision@5, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain at rank 5 (NDCG@5).

For the sbert_all-MiniLM-L6-v2 model, fine-tuning resulted in a noticeable uplift in performance: Precision@5 increased from 0.790 to 0.820, marking a relative gain of 3.8%. At the same time, MRR improved from 0.863 to 0.867, indicating that relevant titles were more likely to be ranked higher in the top-5 results. Similarly, NDCG@5 rose from 0.895 to 0.922, reflecting a better balance between relevance and rank in the retrieved items. These improvements suggest that even lightweight English-only models like all-MiniLM-L6-v2 can benefit considerably from domain-specific fine-tuning. More notably, the sbert_paraphrase-multilingual-MiniLM-L12-v2 model outperformed all others before and after fine-tuning. Its baseline performance was already strong, but fine-tuning elevated its effectiveness even further. Precision@5 advanced from 0.890 to 0.940, indicating that 94% of the top-5 recommended research titles were considered relevant, while NDCG@5 improved from 0.974 to 0.991, showcasing near-optimal ranking quality. These metrics underscore the model's ability to prioritize highly relevant items and present them early in the recommendation list, which is essential for real-world usability. The multilingual nature of this model may have contributed to its robustness across a dataset that potentially includes titles in multiple languages.

In conclusion, the fine-tuned sbert_paraphrase-multilingual-MiniLM-L12-v2 model demonstrated superior performance across all metrics, making it the strongest candidate for deployment in academic recommendation systems. Its ability to achieve high accuracy, strong ranking fidelity, and effective multilingual handling positions it as a convenient and reliable choice for enhancing research discovery, academic matching, and other knowledge retrieval tasks in multilingual educational environments.

Table 3: Performance comparison of SBERT models before and after fine-tuning.

Model	Configuration	Precision@5	MRR	NDCG@5
sbert_all-MiniLM-L6-v2	Before Fine-tuning	0.790	0.863	0.885
sbert_all-MiniLM-L6-v2	After Fine-tuning	0.820	0.867	0.922
sbert_paraphrase-multilingual-MiniLM-L12-v2	Before Fine-tuning	0.890	1	0.974
sbert_paraphrase-multilingual-MiniLM-L12-v2	After Fine-tuning	0.940	1	0.991

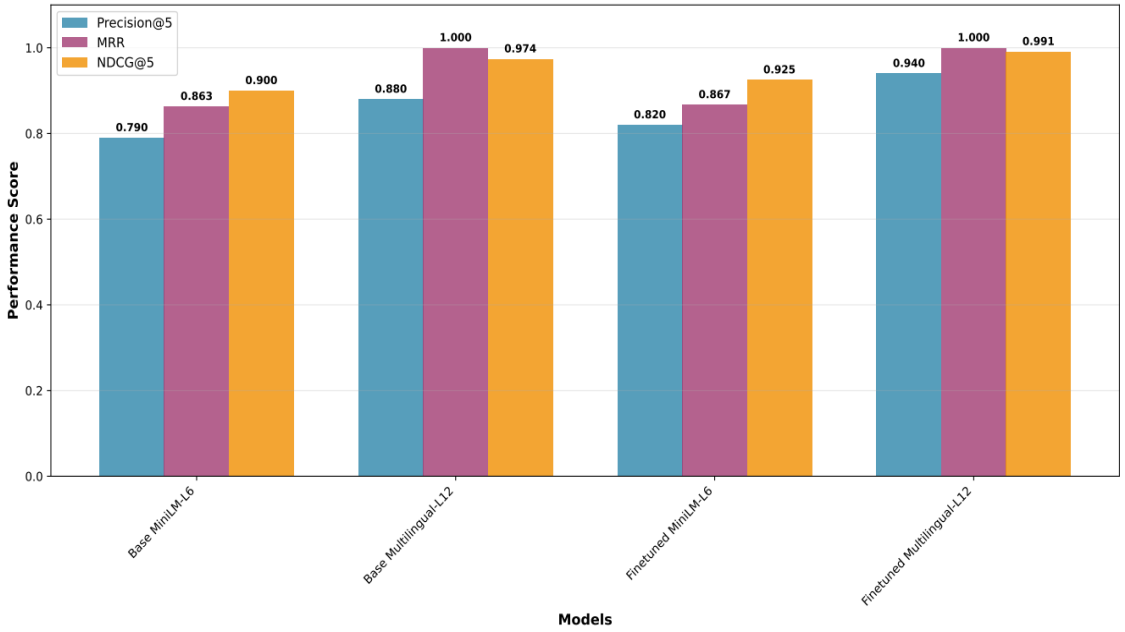

Figure 5: Model performance comparisons for precision@5, MRR, and NDCG@5 metrics.

5. Discussion

The results of this study clearly demonstrate that fine-tuning has a substantial positive impact on the performance of SBERT models in academic recommendation tasks. This improvement stems from several key factors, most notably domain adaptation, which enabled the models to better understand the terminology and language patterns common in academic research, as reflected in the significant 1 MRR gain for the sbert_all-MiniLM-L6-v2 model. Furthermore, the use of a three-level hierarchical taxonomy—comprising domain, subdomain, and specific topic—during fine-tuning allowed the models to learn structured semantic relationships between research topics, rather than relying on labelling with similar or dissimilar [20]. Integrating structured labelling with task-specific fine-tuning allows the system to generalize across domains, rather than remaining confined to one domain as in Juarto *et al.* [22] study.

A comparative analysis between the two architectures reveals why the sbert_paraphrase-multilingual-MiniLM-L12-v2 model outperformed its counterpart: its 12-layer transformer structure grants it superior representational capacity, enabling deeper semantic feature extraction and better generalization to complex academic domains. Its multilingual training on paraphrase data further equips it to understand semantically similar expressions across diverse linguistic styles and writing conventions, improving robustness across a broad spectrum of research titles. Regarding metric-specific performance, both models saw meaningful gains in Precision@5, with the multilingual variant achieving a remarkable 0.940, indicating that nearly all top-5 suggestions were relevant—a critical factor for real-world academic applications. MRR analysis also highlighted a notable ranking, best, showing that relevant recommendations consistently appeared at the top of the list. Similarly, high NDCG@5 scores of 0.922 and 0.991 confirm that both models effectively prioritize relevant content in a ranked setting, with the multilingual model nearing optimal performance.

Whereas Dina *et al.* [11] propose a keyword-driven recommender that returns papers based on user-typed interests, the system evaluated here infers the relevant domain from the queried title via domain-adapted embeddings and recommends semantically aligned academic titles. Both models achieved their strongest performance when using only research titles, whereas Juarto *et al.* [22] approach did not attain comparable results under the same setting.

These findings carry important implications for academic recommendation systems: first, they reinforce the necessity of domain-specific fine-tuning to achieve optimal outcomes; second, they provide

guidance on model selection, where the sbert_all-MiniLM-L6-v2 may serve well in resource-constrained settings, while the more powerful multilingual model is preferred for environments where accuracy is paramount; and third, they underscore the models' readiness for practical deployment, as evidenced by their high precision scores (0.820 and 0.940), which suggest a strong potential to enhance user satisfaction and streamline the discovery of relevant academic research.

all-MiniLM-L6-v2 is generally preferable for web deployment because it is a lightweight encoder that yields fast inference, modest memory use, and lower operational cost. By contrast, paraphrase-multilingual-MiniLM-L12-v2 employs a deeper, 12-layer architecture that typically increases latency, GPU/CPU utilization, and RAM footprint during inference. In resource-constrained environments (e.g., CPU-only servers or small VMs), these requirements can reduce throughput and limit batch sizes, which is a practical limitation for real-time recommendation. Although techniques such as quantization, caching, and distillation can mitigate overhead, the heavier model still poses stricter infrastructure demands in production.

A second limitation concerns language coverage. paraphrase-multilingual-MiniLM-L12-v2 supports cross-lingual semantic alignment, making it suitable for multilingual datasets and university settings where recommendations must bridge different languages (e.g., Kurdish, Arabic, and English). In contrast, all-MiniLM-L6-v2 is optimized for a single language and will generally recommend effectively only within the language of the input query, limiting its applicability in multilingual deployments. Consequently, results obtained with all-MiniLM-L6-v2 may not generalize to cross-language retrieval scenarios, whereas the multilingual model better accommodates such use cases—albeit at higher computational cost.

6. Conclusions

The experimental results conclusively demonstrate that fine-tuning SBERT models significantly enhances their performance in academic recommendation tasks. The fine-tuned sbert_paraphrase-multilingual-MiniLM-L12-v2 model achieved exceptional performance across all evaluation metrics (Precision@5: 0.940, MRR: 0.925, NDCG@5: 0.991), making it the optimal choice for deployment in the SPU research discovery system. The consistent improvements observed across both model variants validate the effectiveness of domain-specific fine-tuning using hierarchical academic taxonomy. These results prove that content-based recommendation systems can significantly enhance research discovery and facilitate interdisciplinary collaboration in university environments when properly adapted to academic domains. The high-quality performance metrics achieved by the fine-tuned models support the practical deployment of this recommendation system for academic research discovery, supervisor matching, and interdisciplinary collaboration facilitation at SPU and potentially other academic institutions and universities facing similar research discovery challenges.

Author contributions: Havan Wahid Rashid: Original article, Data curation, Writing – original draft. **Sarkar Hasan Ahmed:** Supervision, Validation, Visualization, Writing – review & editing.

Data availability: Data will be available upon reasonable request by the authors.

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

Funding: The authors did not receive support from any organization for conducting the study.

References

[1] 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, Dec. 2023, doi: 10.1016/j.isci.2023.108508.

[2] D. Dasri, A. Annisa, and T. Haryanto, "Two-way thesis supervisor recommendation system using MapReduce K-Skyband View Queries," *JOIV International Journal on Informatics Visualization*, 2025, Accessed: Jun. 10, 2025. [Online]. Available: www.joiv.org/index.php/joiv.

[3] H. Ko, S. Lee, Y. Park, and A. Choi, "A survey of recommendation systems: recommendation models, techniques, and application fields," *Electronics (Basel)*, vol. 11, no. 1, p. 141, Jan. 2022, doi: 10.3390/electronics11010141.

[4] W. M. Thackston, *Sorani Kurdish: A Reference Grammar with Selected Readings*. 2006, 250 pp. [Online]. Available: https://archive.org/details/thackston-2006-sorani-grammar-readings

[5] E. Öpengin and G. Haig, "Introduction to special issue - Kurdish: A critical research overview," *Kurdish Studies*, vol. 2, pp. 99–134, 2022, Accessed: Jun. 08, 2025. [Online]. Available: www.kurdishstudies.net.

[6] S. Ahmadi, "A tokenization system for the kurdish language," in *Proc. 7th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2020)*, Barcelona, Spain, 2020, pp. 96-101. [Online]. Available: https://aclanthology.org/2020.vardial-1.11

[7] S. Ahmadi, "KLPT – Kurdish language processing toolkit," in *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 72–84. doi: 10.18653/v1/2020.nlposs-1.11.

[8] M. Hafiz Ismail, T. Rosli Razak, M. Arif Hashim, and A. Faisal Ibrahim, "A simple recommender engine for matching final-year project student with supervisor," *CCMSE*, 2015, Accessed: Jun. 01, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.1908.03475.

[9] H. Amaad, N. Jhamat, K. Riaz, and Z. Arshad, "Context-aware and sequential pattern mining based recommendations for research papers: a hybrid approach," *Journal of Information Communication Technologies and Robotic Applications*, pp. 57–76, Dec. 2020, doi: 10.51239/jictra.v0i0.240.

[10] V. Stergiopoulos, M. Vassilakopoulos, E. Tousidou, and A. Corral, "An academic recommender system on large citation data based on clustering, graph modeling and deep learning," *Knowledge Information System*, vol. 66, no. 8, pp. 4463–4496, Aug. 2024, doi: 10.1007/s10115-024-02094-7.

[11] K. Church, O. Alonso, P. Vickers, J. Sun, A. Ebrahimi, and R. Chandrasekar, "Academic article recommendation using multiple perspectives," Jul. 2024, Accessed: Jun. 01, 2025. [Online]. Available: http://arxiv.org/abs/2407.05836.

[12] D. Mohamed, A. El-Kilany, and H. M. O. Mokhtar, "Academic articles recommendation using concept-based representation," in *Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 2*, Cham, Switzerland: Springer, 2021, pp. 733–744, doi: 10.1007/978-3-030-55187-2_52.

[13] C. Albusac, L. M. de Campos, J. M. Fernández-Luna, and J. F. Huete, "Content-based recommendation for academic expert finding," in *Proceedings of the 5th Spanish Conference on Information Retrieval*, New York, NY, USA: ACM, Jun. 2018, pp. 1–8. doi: 10.1145/3230599.3230607.

[14] S. Gheewala, S. Xu, and S. Yeom, "In-depth survey: deep learning in recommender systems—exploring prediction and ranking models, datasets, feature analysis, and emerging trends," *Neural Computing and Applications*, vol. 37, no. 17, pp. 10875–10947, Jun. 2025, doi: 10.1007/s00521-024-10866-z.

[15] A. Rodriguez and R. Vuppala, "A recommendation system for scientific papers through Bayesian nonparametric hybrid filtering," 2014, pp. 20–41. doi: 10.4018/978-1-4666-5063-3.ch002.

[16] R. Singh, G. Gaonkar, V. Bandre, N. Sarang, and S. Deshpande, "Scientific paper recommendation system," in *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, IEEE, Apr. 2023, pp. 1–4. doi: 10.1109/I2CT57861.2023.10126196.

[17] A. L. Lezama-Sánchez, M. Tovar Vidal, and J. A. Reyes-Ortiz, "An approach based on semantic relationship embeddings for text classification," *Mathematics*, vol. 10, no. 21, p. 4161, Nov. 2022, doi: 10.3390/math10214161.

[18] M. Fateen and T. Mine, "Using similarity learning with SBERT to optimize teacher report embeddings for academic performance prediction," in *Communications in Computer and Information Science*, vol. 1831, pp. 720–726, 2023, doi: 10.1007/978-3-031-36336-8_111.

[19] N. Yang, J. Jo, M. Jeon, W. Kim, and J. Kang, "Semantic and explainable research-related recommendation system based on semi-supervised methodology using BERT and LDA models," *Expert Systems with Applications*, vol. 190, p. 116209, Mar. 2022, doi: 10.1016/j.eswa.2021.116209.

[20] C. Yin and Z. Zhang, "A study of sentence similarity based on the all-minilm-l6-v2 model with 'same semantics, different structure' after fine tuning," in *Proc. 2024 2nd Int. Conference Image, Algorithms and Artificial Intelligence (ICIAAI)*, 2024, pp. 677–684, doi: 10.2991/978-94-6463-540-9_69.

[21] H. A. Mohamed, F. Gasparetti, and G. Sansonetti, "BERT, ELMo, USE and InferSent sentence encoders: the panacea for research-paper recommendation?" in *Proceedings 13th ACM Conf. Recommender Systems*," 2019. Accessed: Jun. 02, 2025. [Online]. Available: https://www.researchgate.net/publication/335555312.

[22] B. Juarto and A. Suganda Girsang, "Neural collaborative with sentence BERT for news recommender system," *JOIV : International Journal on Informatics Visualization*, vol. 5, no. 4, p. 448, Dec. 2021, doi: 10.30630/joiv.5.4.678.

[23] S. S. Roy, A. Kumar, and R. Suresh Kumar, "Metadata and review-based hybrid apparel recommendation system using cascaded large language models," *IEEE Access*, vol. 12, pp. 140053–140071, 2024, doi: 10.1109/ACCESS.2024.3462793.

[24] K. Sarode and S. R. Javaji, "Multi-BERT for embeddings for recommendation system," *arXiv preprint arXiv:2308.13050*, Aug. 2023. [Online]. Available: https://arxiv.org/abs/2308.13050

[25] X. Li, X. Wang, and H. Liu, "Research on fine-tuning strategy of sentiment analysis model based on BERT," in *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, IEEE, May 2021, pp. 798–802. doi: 10.1109/CISCE52179.2021.9445882.

[26] J. Zhang, W. Chang, H. Yu, and I. S. Dhillon, "Fast multi-resolution transformer fine-tuning for extreme multi-label text classification," *in Proc. 35th Conference Neural Information Processing Systems (NeurIPS)*, 2021. [Online]. Available: http://arxiv.org/abs/2110.00685.

[27] B. Nguyen and S. Ji, "Fine-tuning pretrained language models with label attention for biomedical text classification," arXiv, arXiv:2108.11809, 2022. [Online]. Available: http://arxiv.org/abs/2108.11809.

[28] J. Mücke, D. Waldow, L. Metzger, P. Schauz, M. Hoffman, N. Lell, and A. Scherp, "Fine-Tuning Language Models for Scientific Writing Support," in *Machine Learning and Knowledge Extraction (CD-MAKE 2023)*, Benevento, Italy, Aug. 29–Sep. 1, 2023, A. Holzinger *et al.*, Eds., Lecture Notes in Computer Science, vol. 14065. Cham, Switzerland: Springer, 2023, pp. 301–318, doi: 10.1007/978-3-031-40837-3_18

[29] T. Dhamecha, R. Murthy, S. Bharadwaj, K. Sankaranarayanan, and P. Bhattacharyya, "Role of language relatedness in multilingual fine-tuning of language models: A case study in Indo-Aryan languages," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for computational linguistics, 2021, pp. 8584–8595. doi: 10.18653/v1/2021.emnlp-main.675.

[30] Y. Ma, H. Chen, Q. Wang, and X. Zheng, "Text classification model based on CNN and BiGRU fusion attention mechanism," *ITM Web of Conferences*, vol. 47, no. 02040, 2022, doi: 10.1051/itmconf/20224702040.

[31] D. E. Cahyani and I. Patasik, "Performance comparison of TF-IDF and Word2Vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2878-2886, Oct. 2021, doi: 10.11591/EEI.V10I5.3157.

[32] J. Ye and H. Tian, "Learning to research: learning to ranking the similar papers via BERT fine-tuning," *Advances in Engineering Technology Research*, vol. 5, no. 1, p. 349, May 2023, doi: 10.56028/aetr.5.1.349.2023.

[33] P. Gao, J. Zhao, Y. Ma, A. Tanvir, and B. Jin, "HFT-ONLSTM: Hierarchical and Fine-Tuning Multi-label Text Classification," arXiv, arXiv:2204.08115, Apr. 2022. [Online]. Available: https://arxiv.org/abs/2204.08115

[34] N. Pal and O. Dahiya, "Analysis of educational recommender system techniques for enhancing student's learning outcomes," in *2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, IEEE, Feb. 2023, pp. 1–5. doi: 10.1109/ICIPTM57143.2023.10118132.

[35] M. Timmi, "Educational Video recommender system," *International Journal of Information and Education Technology*, vol. 14, no. 3, pp. 362–371, 2024, doi: 10.18178/ijiet.2024.14.3.2058.

[36] Alicia McNett, "Recommender systems research and theory in higher education: a systematic literature review," *Issues In Information Systems*, 2022, doi: 10.48009/3_iis_2022_113.

[37] G. Muzdybayeva, D. Khashimova, A. Amirzhanov, and S. Kadyrov, "A Matrix factorization-based collaborative filtering framework for course recommendations in higher education," in *2023 17th International Conference on Electronics Computer and Computation (ICECCO)*, IEEE, Jun. 2023, pp. 1–4. doi: 10.1109/ICECCO58239.2023.10147152.

[38] J. Kim, T. Kim, and B. Yun, "Development and application of an ai-based personalized research-paper recommendation system: an example from k university," *Korean Association for Educational Information and Media*, vol. 29, no. 3, pp. 705–730, Sep. 2023, doi: 10.15833/KAFEIAM.29.3.705.

[39] A. Zhao and Y. Ma, "Research on recommendation of big data for higher education based on deep learning," *Scientific Programming*, vol. 2022, pp. 1–8, May 2022, doi: 10.1155/2022/5448442.

[40] N. Reimers and I. Gurevych, "Sentence-BERT: sentence embeddings using Siamese BERT-Networks," *International Joint Conference on Natural Language Processing*, Aug. 2019. [Online]. Available: http://arxiv.org/abs/1908.10084.

[41] Sentence Transformers, "Pretrained models — sentence transformers documentation." Accessed: Jun. 07, 2025. [Online]. Available: https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

[42] D. Liao, "sentence embeddings using supervised contrastive learning," *arXiv preprint*, Jun. 2021, Accessed: Jun. 03, 2025. [Online]. Available: http://arxiv.org/abs/2106.04791.