JSCE (Journal of System and Computer Engineering)
Vol.7, No.4, October 2026, pp. 432-442
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v7i4.2758

Research Article Open Access (CC–BY-SA) ■

Analysis Of A Textbook Recommendation System Based On Semantic Search And Full-Text Retrieval Based On Students' Learning Materials At The Labuang State Vocational High School Library

Muhammad Fadhil AY [1,a,*]; Supriadi Sahibu [2,a]; Respaty Namruddin [3,a]; Muhammad Yaumi [4,a]; Hazriani [5,a]; Abdul Latief Arda [6,a];

[1] Master Program of Computer Systems, Faculty of Computer Science, Universitas Handayani Makassar, Makassar 90231, Indonesia

[a] muhammadfadhilay22@gmail.com; [b] supriadi@handayani.ac.id; [c] respatynamruddin@handayani.ac.id; [d] muhammad.yaumi@uin-alauddin.ac.id; [e] hazriani@handayani.ac.id; [f] abdullatief@handayani.ac.id;

** Corresponding author*

Abstract

School libraries play an important role in providing relevant learning resources for students. However, conventional book search systems that rely primarily on titles or keywords may have limitations in identifying learning resources when the query and document content use different terms. This study aims to analyze a textbook recommendation system that combines semantic search and full-text retrieval based on students' learning materials at the Labuang State Vocational High School Library. This study applies Natural Language Processing (NLP) techniques for text preprocessing and represents textbook content using Term Frequency–Inverse Document Frequency (TF-IDF) and sentence embeddings. The similarity between learning-material queries and textbook content is calculated using cosine similarity. The lexical and semantic similarity scores are then combined using a Hybrid Search approach to generate document relevance scores. The dataset consists of 80 textbooks in PDF format, which were processed into 21,556 text paragraphs as the textual basis for the retrieval process. The extracted content was further divided into smaller chunks to facilitate semantic representation and retrieval.System performance was evaluated using Precision, Recall, F1-score, Average Precision (AP), and Mean Average Precision (MAP). The experimental results show that Hybrid Search achieved the highest MAP score of 0.361, compared with 0.344 for TF-IDF and 0.194 for Semantic Search. These results indicate that, within the experimental setting, combining lexical and semantic retrieval signals provided complementary information for ranking relevant textbooks. The findings suggest that the proposed approach can support more context-aware access to textbook resources based on students' learning materials. Further evaluation using more queries and broader datasets is required to assess the consistency of the observed retrieval performance.

Keywords— *Semantic Search, Full-Text Retrieval, Hybrid Search, TF-IDF, Sentence Embedding, Textbook Recommendation.*

1. Introduction

School libraries play an important role in providing learning resources, particularly textbooks that support students in understanding subject-specific materials. The development of

digital library services has encouraged the use of recommendation systems to help students identify relevant books. Content-based filtering has been implemented in school libraries to improve access to relevant collections [1]. Similar approaches use book descriptions and other textual attributes to identify books with related characteristics [2]. However, approaches based primarily on bibliographic information or predefined attributes may be less effective when students express their needs through learning topics or material descriptions rather than exact book titles or author names.

Information Retrieval (IR) provides mechanisms for matching user queries with document content. Lexical retrieval methods such as Term Frequency–Inverse Document Frequency (TF-IDF) represent terms according to their importance within documents and collections [3]. TF-IDF combined with cosine similarity has also been applied to book recommendation systems to identify books based on textual similarity [4]. However, lexical retrieval depends strongly on shared terms between queries and documents, which may limit its effectiveness when the same concept is expressed using different terminology.

Natural Language Processing (NLP) enables semantic retrieval by representing text as numerical vectors in a semantic space. Sentence-BERT and related SentenceTransformer models generate sentence-level embeddings that can be compared using cosine similarity [5]. Dense retrieval similarly uses vector representations to retrieve documents according to semantic similarity rather than exact lexical matching [6]. Although this approach can identify conceptually related content, semantic similarity does not always correspond to task-specific relevance, particularly when similar passages belong to different subjects or learning objectives.

Lexical and semantic retrieval can therefore provide complementary information. Hybrid retrieval systems combine different retrieval signals to improve the identification of relevant information [7]. Hybrid retrieval and reranking approaches have also integrated lexical and semantic information for passage retrieval [8]. In book recommendation, recent studies have continued to investigate content-based and hybrid approaches [9], while other studies have combined collaborative and content-based filtering to improve user–item representation [10]. Hybrid retrieval has also shown potential in broader textual retrieval applications [11]. Nevertheless, hybrid retrieval is not a new paradigm, and its effectiveness depends on factors such as weighting strategies, retrieval configurations, dataset characteristics, and application domains.

This issue is relevant to the Library of SMK Negeri Labuang, where textbook searches primarily rely on bibliographic information such as titles and author names. This mechanism may make it difficult for students to identify appropriate textbooks when their information needs are expressed as learning topics or descriptions of material, requiring them to examine multiple search results manually.

Therefore, this study investigates a hybrid retrieval approach for textbook recommendation based on students' learning materials. The proposed approach integrates TF-IDF-based Full-Text Retrieval with SentenceTransformer-based Semantic Search to utilize both lexical and semantic information from textbook content. The resulting similarity scores are combined using weighted hybrid scoring and document-level aggregation of chunk-level scores to generate textbook recommendations.

The research gap addressed in this study is not the introduction of hybrid retrieval as a new paradigm, but the implementation and empirical evaluation of a specific hybrid retrieval configuration for textbook recommendation in a vocational school library. Specifically, this study examines whether combining lexical and semantic similarity can improve textbook ranking compared with TF-IDF or Semantic Search independently under the same dataset, query set, and evaluation framework.

The contributions of this study are threefold: (1) implementing a hybrid textbook retrieval configuration that combines TF-IDF-based Full-Text Retrieval and SentenceTransformer-based Semantic Search; (2) empirically comparing TF-IDF, Semantic Search, and Hybrid Search using Precision, Recall, F1-Score, Average Precision (AP), and Mean Average Precision

(MAP); and (3) analyzing the complementary contribution of lexical and semantic retrieval signals to textbook recommendation in a vocational school library. Through this evaluation, the study provides empirical evidence regarding the applicability of hybrid retrieval and its effect on ranking learning resources within the specific context of the SMK Negeri Labuang library.

2. Related Work

2.1 Lexical Information Retrieval

Information Retrieval (IR) focuses on identifying and ranking information that is relevant to a user's information needs from a collection of documents. In a retrieval system, queries and documents are represented using specific representations, allowing their relevance to be measured and used to generate document rankings. Lexical retrieval represents textual information primarily through terms occurring in queries and documents, making explicit term matching an important component of the retrieval process.

One of the widely used approaches in lexical retrieval is Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF assigns a weight to a term based on its frequency within a document and its distribution across the document collection. Terms that occur frequently in a particular document but relatively infrequently across the collection receive higher weights. This characteristic makes TF-IDF suitable for identifying documents containing terms that explicitly correspond to a user's query.

In the context of textbook retrieval, lexical matching can provide strong evidence when subject-specific terms contained in a student's learning-material query also appear in the textbook content. For example, a query containing a specific concept or technical term can increase the retrieval score of textbooks containing the same term. However, lexical retrieval may encounter a vocabulary mismatch when the query and relevant document express the same concept using different words or linguistic structures.

BM25 is another widely used lexical retrieval method that considers term frequency, document length, and term distribution when calculating document relevance. Recent research has continued to investigate the integration of lexical retrieval signals with neural retrieval models, indicating that lexical evidence remains useful for improving retrieval and reranking performance [12]. This demonstrates that sparse retrieval remains relevant even as dense retrieval approaches become increasingly prevalent.

Therefore, lexical retrieval provides an interpretable mechanism for identifying explicit term matches, but it may have difficulty capturing semantic relationships when relevant textbook content does not use the same terminology as the user's query.

2.2 Semantic Search and Dense Retrieval

The development of Natural Language Processing (NLP) has enabled Information Retrieval systems to consider semantic relationships in addition to explicit word matching. Semantic Search represents queries and documents as vectors in an embedding space, allowing texts with related meanings or contextual characteristics to be identified based on vector similarity. This approach is commonly associated with dense retrieval because textual information is represented using dense numerical vectors.

SentenceTransformer-based models are widely used to generate semantic representations of sentences and document passages. Such representations allow queries and documents to be compared using similarity measures such as cosine similarity. This enables retrieval systems to identify textual content based on semantic relationships rather than requiring an exact match between query and document terms.

Dense retrieval approaches have also demonstrated the potential of learned vector representations for information retrieval. Contrastive learning has been used to improve dense representations so that semantically related queries and documents can be positioned closer together in the embedding space. This characteristic is particularly relevant to textbook

recommendation because students may describe their learning needs using terminology that differs from the wording used in textbook content.

For example, a student may formulate a query using a description of a learning concept while the relevant textbook discusses the same concept using different terminology. Semantic retrieval can potentially identify such content because the retrieval process considers the relationship between the meanings of the query and document rather than relying exclusively on shared terms.

However, semantic similarity does not necessarily correspond to task-specific relevance. Two textbook passages may be semantically related while addressing different subjects, learning objectives, or levels of conceptual detail. Consequently, a semantic retrieval system may retrieve content that is generally related to the query but is not necessarily the most relevant textbook for the student's specific information need.

The effectiveness of dense retrieval can also depend on the characteristics of the embedding model, language, document domain, and query formulation. Therefore, semantic retrieval provides an important alternative to lexical retrieval, but its performance should be evaluated within the specific domain and retrieval task in which it is applied.

2.3 Hybrid Retrieval

The different characteristics of lexical and semantic retrieval have motivated the development of Hybrid Retrieval approaches. Hybrid Retrieval combines sparse or lexical retrieval signals with dense or semantic retrieval signals to utilize both explicit term matching and contextual similarity.

Research on hybrid retrieval has demonstrated that sparse and dense representations can provide complementary retrieval information. Lee et al. investigated complementarity objectives for hybrid retrieval and highlighted the importance of exploiting the different information captured by sparse and dense retrieval representations. This complementarity is relevant because lexical retrieval can provide direct evidence of term matching, while semantic retrieval can provide additional evidence when relevant concepts are expressed using different terminology.

Hybrid retrieval has also been investigated in different information retrieval applications. Doan et al. developed a hybrid retrieval approach for Retrieval-Augmented Generation systems, demonstrating the use of multiple retrieval signals in a broader retrieval framework. Lu et al. proposed a hybrid-infused reranking approach for passage retrieval, further demonstrating the use of hybrid information for improving the ranking process.

These studies indicate that hybrid retrieval has been applied across several retrieval tasks, including passage retrieval and Retrieval-Augmented Generation. However, the retrieval configuration, scoring mechanism, dataset, and application domain differ across studies. Therefore, the effectiveness of hybrid retrieval cannot be assumed to be identical across different retrieval scenarios.

In the present study, Hybrid Retrieval is not positioned as a newly proposed retrieval paradigm. Instead, the contribution lies in the implementation and empirical evaluation of a specific hybrid configuration for textbook recommendation based on students' learning materials. The proposed system combines TF-IDF scores as lexical evidence with semantic similarity scores generated using a SentenceTransformer model. The two retrieval signals are normalized and combined using a weighted scoring strategy to produce a hybrid retrieval score.

This configuration enables the system to retain explicit lexical evidence while incorporating semantic information. A textbook containing terms that directly correspond to the student's query can therefore receive evidence from lexical retrieval, while a textbook using different terminology but expressing a related concept can receive additional evidence from semantic retrieval.

2.4 Book and Textbook Recommendation

Recommendation systems are designed to assist users in identifying items that correspond to their needs, preferences, or interests. In library environments, recommendation systems can assist users in identifying relevant books from available collections. Book recommendation approaches may utilize borrowing histories, user preferences, book metadata, textual descriptions, or the content of the books themselves.

Collaborative Filtering is a commonly used recommendation approach that utilizes patterns of interactions or preferences between users and items. This approach can be effective when sufficient interaction data, such as borrowing histories or ratings, are available. However, when recommendations are primarily determined by a student's learning-material query, a content-based approach is more directly aligned with the retrieval task because the system needs to match the information need with the characteristics or textual content of books.

Recent research has investigated content-based filtering for book recommendation by using textual or item-related characteristics to identify relevant book. Other studies have developed book recommendation systems using TF-IDF and cosine similarity, demonstrating the applicability of lexical text representation for matching user queries with book content.

Further research has explored book recommendation in school and library environments using content-based approaches. These studies demonstrate that textual characteristics of books can be utilized to identify items that are related to users' information needs. Recent work has also continued to develop content-based book recommendation systems for school e-libraries using TF-IDF and cosine similarity [13].

Hybrid recommendation approaches have also been investigated by combining multiple recommendation techniques. Such approaches aim to utilize complementary information from different recommendation mechanisms rather than relying on a single source of evidence [10]. However, recommendation systems based on user interaction data and conventional content attributes differ from the retrieval-oriented approach used in this study.

In this study, the recommendation process focuses specifically on content-based textbook retrieval. The student's query represents the required learning material, while the system searches the textual content extracted from the available textbooks. The retrieval results are subsequently aggregated at the textbook level and ranked to generate recommendations.

2.5 Research Positioning

Previous studies demonstrate that lexical retrieval and semantic retrieval have different but complementary characteristics. TF-IDF-based lexical retrieval provides explicit evidence when query terms occur in documents, while semantic retrieval can identify conceptually related content even when different terminology is used. Dense retrieval further provides a vector-based representation that can capture semantic relationships between queries and documents.

Research on hybrid retrieval indicates that combining sparse and dense retrieval signals can exploit complementary information from the two retrieval paradigms. Hybrid approaches have been investigated in passage retrieval, reranking, and Retrieval-Augmented Generation applications. These studies provide a foundation for combining lexical and semantic evidence, but they do not specifically address textbook recommendation based on students' learning-material queries in a vocational school library.

Research on book recommendation has primarily investigated content-based filtering, collaborative filtering, or combinations of recommendation techniques. Several studies have also applied TF-IDF and cosine similarity to book recommendation. However, the retrieval problem addressed in the present study differs from conventional book recommendation because the primary input is a student's learning-material query, while the target of retrieval is the textual content of textbooks.

Therefore, this study is positioned at the intersection of Information Retrieval, Semantic Search, Hybrid Retrieval, and textbook recommendation. The study does not claim that combining lexical and semantic retrieval represents a new retrieval paradigm. Instead, it focuses

on the implementation, configuration, and empirical evaluation of a hybrid retrieval approach for textbook recommendation in the specific context of the SMK Negeri Labuang library.

To examine the contribution of each retrieval approach, this study compares three configurations within the same experimental environment: TF-IDF, SentenceTransformer-based Semantic Search, and Hybrid Search. The systems are evaluated using Precision, Recall, F1-Score, Average Precision (AP), and Mean Average Precision (MAP). This evaluation allows the study to examine not only whether relevant textbooks are retrieved, but also how effectively the different approaches rank relevant textbooks.

The research positioning can therefore be summarized as follows: previous studies establish the usefulness of lexical retrieval, semantic retrieval, hybrid retrieval, and content-based book recommendation, while the present study brings these components together in a specific textbook retrieval scenario based on students' learning materials in a vocational school library. The empirical comparison of the three retrieval configurations provides the basis for analyzing the contribution of lexical and semantic information to textbook ranking within this particular application domain.

3. Method

This study employs a quantitative experimental approach to analyze the performance of a textbook recommendation system based on Semantic Search and Full-Text Retrieval. The research focuses on matching students' learning-material queries with the textual content of textbooks available in the Library of SMK Negeri Labuang.

The research process consists of dataset collection, text extraction, text preprocessing, chunking, semantic representation using SentenceTransformer embeddings, TF-IDF-based Full-Text Retrieval, similarity calculation using cosine similarity, integration of lexical and semantic scores through Hybrid Search, document-level score aggregation, and system evaluation. The retrieval approaches are evaluated under the same dataset, query set, and relevance judgments to enable a consistent comparison between TF-IDF, Semantic Search, and Hybrid Search.

3.1 Research Design

The research follows an experimental design in which three retrieval configurations are implemented and compared: TF-IDF-based Full-Text Retrieval, SentenceTransformer-based Semantic Search, and Hybrid Search. The first configuration represents lexical retrieval based on explicit term matching, while the second represents semantic retrieval based on vector representations. The third configuration combines the lexical and semantic similarity scores into a single retrieval score.

The research workflow consists of five main stages. First, textbooks are collected and their textual content is extracted from PDF documents. Second, the extracted text is preprocessed and divided into smaller chunks to facilitate retrieval and semantic representation. Third, each chunk is represented using both TF-IDF and sentence embeddings. Fourth, the similarity between the learning-material query and textbook content is calculated and combined using the proposed Hybrid Search scoring strategy. Finally, the retrieval performance of the three approaches is evaluated using Precision, Recall, F1-score, Average Precision (AP), and Mean Average Precision (MAP).

The experimental comparison is conducted using the same textbook collection and query set for all three retrieval configurations. This design allows the study to examine the relative contribution of lexical and semantic information to textbook ranking within the same experimental environment.

3.2 Research Stages

Figure 1 illustrates the overall stages involved in developing and evaluating the textbook recommendation system. The research consists of three main phases: system design, system development, and system testing and implementation.

The system design phase includes user-requirement identification, system architecture design, database design, and design of the textbook recommendation workflow. The purpose of this phase is to define how students' learning-material queries are processed and how textbook recommendations are generated.

The system development phase includes database construction, text extraction, Natural Language Processing (NLP) preprocessing, text chunking, semantic embedding, TF-IDF vectorization, and the integration of Semantic Search and Full-Text Retrieval through the Hybrid Search mechanism.

The system testing and implementation phase evaluates the retrieval performance using Precision, Recall, F1-score, Average Precision (AP), and Mean Average Precision (MAP). The developed system is subsequently implemented in the library environment of SMK Negeri Labuang.

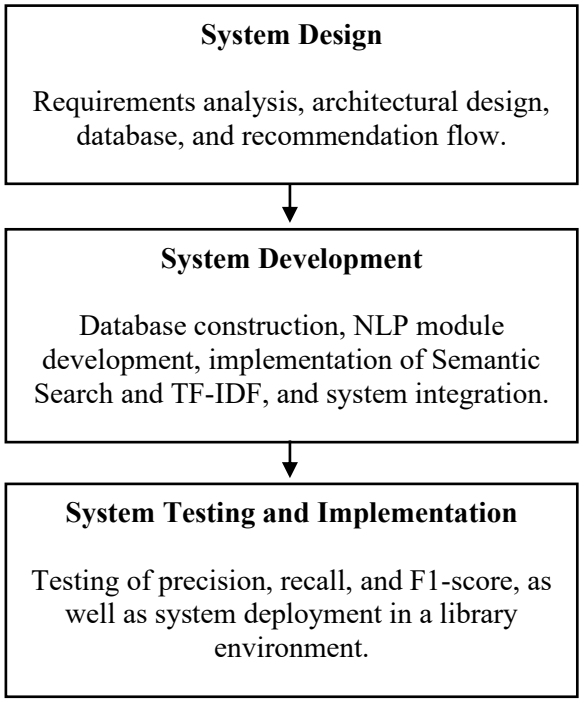


Figure 1. User Needs Identification Workflow

3.3 System Architecture

Figure 2 presents the architecture of the textbook recommendation system, consisting of three main components: users, the system application, and the database.

The user component includes students, teachers, and librarians who interact with the system. The primary input for the recommendation process is learning-material text entered by the user.

The system application consists of three main modules. The first is the learning-material input module, which receives and processes the user's learning-material query. The second is the book recommendation module, which performs preprocessing, text representation, similarity calculation, retrieval, and ranking to generate a list of relevant textbooks. The third is the data-management module, which manages textbook data, learning-material data, and other supporting information.

At the information-retrieval level, the system processes textbook content using NLP preprocessing and represents the text using two complementary approaches. TF-IDF is used for lexical representation, while SentenceTransformer is used for semantic representation. The

similarity between the query and textbook content is calculated using cosine similarity. The resulting lexical and semantic scores are subsequently combined through the Hybrid Search mechanism.

The database stores the textbook collection, extracted textual content, and supporting system data required for the recommendation process.

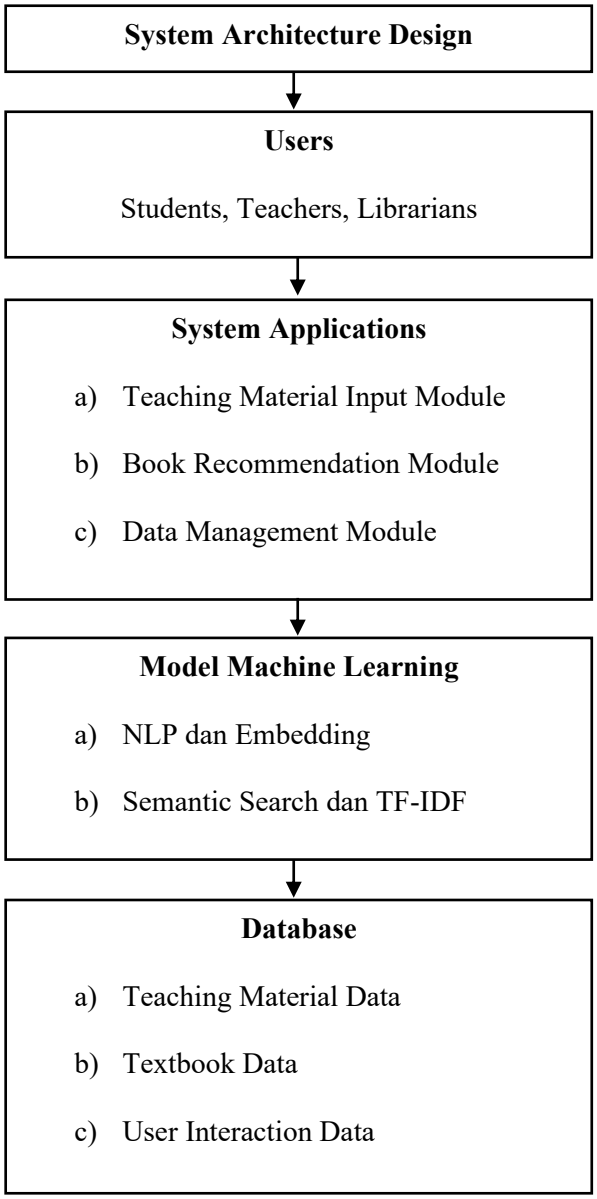


Figure 2. System Design Workflow.

3.4 Process Flow

The recommendation process begins when a user enters learning-material text into the system. The input text is subsequently processed through an NLP preprocessing pipeline consisting of case folding, tokenization, stopword removal, and stemming.

After preprocessing, the textual data are represented using two retrieval approaches. For semantic retrieval, the processed text is transformed into a numerical vector using a SentenceTransformer embedding model. For lexical retrieval, the text is represented using TF-IDF.

The similarity between the query representation and textbook-content representations is calculated using cosine similarity. Semantic similarity is used by the Semantic Search component, while TF-IDF similarity is used by the Full-Text Retrieval component.

The two similarity scores are then normalized and combined using the Hybrid Search scoring function. Because retrieval is initially performed at the chunk level, the resulting chunk-level scores are aggregated into document-level scores to obtain a relevance score for each textbook.

Finally, textbooks are ranked in descending order according to their document-level scores, and the system presents the Top-N textbooks as recommendations.

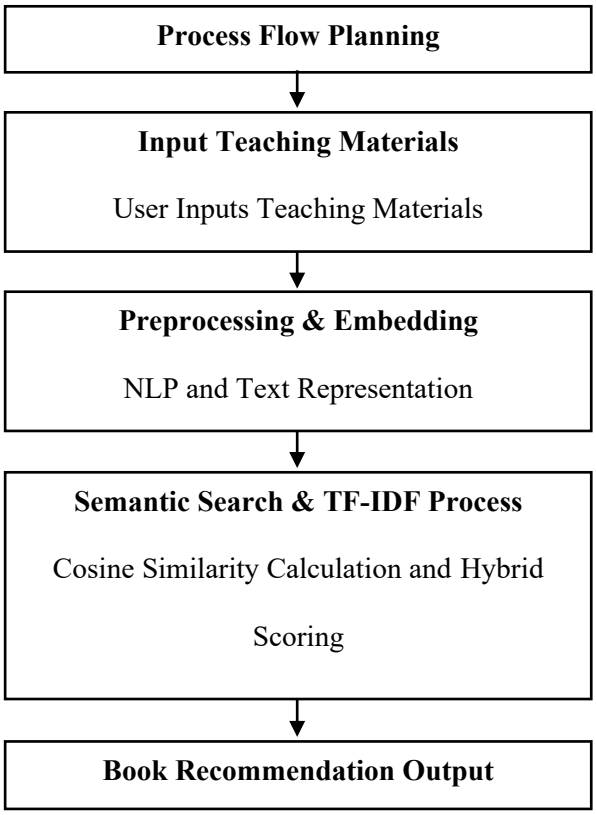


Figure 3. Process Flow Design.

3.5 Data Collection and Text Preprocessing

The dataset consists of 80 textbooks in PDF format obtained from the Library of SMK Negeri Labuang. The textbooks cover various vocational and general subject areas, including information technology, computer networking, entrepreneurship, and other learning materials.

The textual content of the textbooks is extracted from the PDF documents using a Python-based text extraction process. The extraction process produced 21,556 text paragraphs. These extracted paragraphs were subsequently used as the textual basis for the retrieval process.

The extracted text is preprocessed using several NLP operations. Case folding converts characters into a consistent letter case. Tokenization divides the text into individual tokens. Stopword removal removes words that provide limited discriminative information for retrieval. Stemming reduces words to their basic forms.

After preprocessing, the extracted content is divided into smaller text segments using word-based chunking with a maximum size of 500 words per chunk. The general number of chunks can be represented as:

$$n_{chunk} = [n/k] \quad (1)$$

where \(N\) represents the number of words in the processed text and \(k\) represents the maximum chunk size, which is set to 500 words.

Each chunk is assigned a unique identifier using the format:

BookTitle_PPage_CChunk

The identifier maintains the relationship between each chunk and its source textbook, page, and chunk position.

3.6 Document-Level Score Aggregation

Because the retrieval process operates at the chunk level, multiple chunks may originate from the same textbook. Therefore, the chunk-level scores are aggregated into a document-level score. The document-level score is calculated using:

$$Score_{document} = 0.7 \cdot max(score) + 0.3 \cdot mean(score) \quad (2)$$

Where *(max(score))* represents the highest-scoring chunk within a textbook and *(mean(score))* represents the average score of the retrieved chunks associated with that textbook.

The maximum score represents the strongest evidence of relevance within a textbook, while the mean score provides additional information regarding the overall relevance of the retrieved chunks. The resulting document-level score is used to rank textbooks in the recommendation output.

The weights of *0.7* and *0.3* are treated as configuration parameters in the current implementation. Their general optimality is not assumed beyond the experimental setting of this study.

3.7 Ground Truth and Evaluation Queries

System evaluation uses three representative queries corresponding to learning materials from different subject areas. Each query is associated with a set of textbooks considered relevant to the corresponding learning material.

The relevant textbooks constitute the ground truth used to evaluate the retrieval results. The relevance judgment should be based on the actual criteria used during the experiment. In the final manuscript, the following information should be explicitly stated:

The ground-truth relevance judgments were generated automatically by the system based on predefined relevance criteria. A textbook was classified as relevant if its textual content matched the subject or learning material specified in the query. These relevance judgments were used as the reference for evaluating TF-IDF, Semantic Search, and Hybrid Search.

For each query, the relevant textbooks are compared with the ranked retrieval results generated by TF-IDF, Semantic Search, and Hybrid Search.

Using the same queries and relevance judgments for all three retrieval configurations ensures that the comparison is conducted under the same evaluation conditions.

3.8 Research Time and Location

This research was conducted over four months, from December 2025 to March 2026, at the Library of SMK Negeri Labuang, located on Jl. Poros Majene, Laliko Village, Campalagian District, Polewali Mandar Regency, West Sulawesi.

SMK Negeri Labuang is a vocational secondary school whose library provides textbook resources to support students' learning activities. The research location was selected because the library provides the textbook collection required for the development and evaluation of the proposed textbook recommendation system.

4. Results And Discussion

This section presents the results of the textbook recommendation system and discusses the experimental findings based on the retrieval performance of TF-IDF, Semantic Search, and Hybrid Search. The evaluation considers Precision, Recall, F1-score, Average Precision (AP), and Mean Average Precision (MAP). The results are analyzed quantitatively to compare the retrieval and ranking performance of the three approaches.

4.1 Dataset and Text Extraction

The dataset used in this study consists of 80 textbooks in PDF format collected from various subject areas at the vocational secondary school (SMK) level, including information technology, computer networking, entrepreneurship, and other general subjects. Text extraction was performed using a Python-based text extraction process. The extraction process successfully processed all 80 textbooks and produced a total of 21,556 text paragraphs.

The extracted paragraphs were subsequently processed and used as the textual basis for the information retrieval process. Each extracted text segment was associated with its source document to maintain traceability between the retrieved content and the original textbook.

Important: if 21,556 is the number of paragraphs before chunking, keep the wording above. If 21,556 is actually the number of chunks after chunking, the sentence must be changed. This distinction should be consistent throughout the manuscript.

4.2 Data Preprocessing (Chunking)

To facilitate semantic embedding and retrieval at a more manageable textual granularity, the extracted textbook content was divided into smaller segments using word-based chunking with a maximum size of 500 words per chunk. The number of chunks can generally be represented as:

$$n_chunk = N / k$$

Where N represents the number of words in the processed text and \(k\) represents the maximum chunk size, which was set to 500 words.

Each resulting chunk was assigned a unique identifier using the format BookTitle_PPage_CChunk. This identifier was used to maintain traceability of each text segment with respect to its source textbook, page, and chunk position.

Analysis of three representative textbooks showed that the resulting chunk lengths were generally distributed within the range of approximately 100–400 words. The variation in chunk length resulted from the structure of the extracted text and the chunking process rather than from forcing every segment to have an identical length.

The chunking strategy allowed the retrieval process to operate on smaller textual units. Therefore, when a query was submitted, the system could identify specific portions of a textbook that were more closely related to the query rather than treating the entire textbook as a single textual unit.

4.3 Text Representation (Embedding)

Semantic representation was implemented using a SentenceTransformer embedding model to transform each text chunk into a numerical vector representation. The embedding process can be represented as:

$$V = f(x)$$

Where x represents the input text and \(f\) represents the embedding model that maps the text into a vector space.

The similarity between a query and a text chunk was subsequently calculated using cosine similarity. This representation enables the retrieval process to compare textual content based on the similarity of their vector representations rather than relying exclusively on exact keyword matching.

Principal Component Analysis (PCA) was also used to visualize the distribution of the generated embeddings in a reduced-dimensional space. The visualization showed that several chunks with similar textual contexts tended to appear relatively close to one another, although clearly separated clusters between textbooks were not observed.

The PCA visualization therefore provides a qualitative view of the structure of the generated representations. However, the visualization alone does not establish the effectiveness of the embedding model. Its retrieval effectiveness was instead evaluated quantitatively through Precision, Recall, F1-score, AP, and MAP.

4.4 TF-IDF Vectorization

In addition to semantic representation, this study implemented full-text retrieval using Term Frequency-Inverse Document Frequency (TF-IDF). The TF-IDF representation was generated using the TfidfVectorizer implementation provided by the scikit-learn library.

The resulting TF-IDF representation forms a matrix in which rows represent text chunks and columns represent terms contained in the corpus. Each matrix value represents the importance of a particular term within a text chunk relative to its occurrence across the corpus.

TF-IDF was used as the lexical retrieval baseline for comparison with Semantic Search and Hybrid Search. The approach is particularly suitable when query terms have direct lexical matches in the retrieved text. However, its representation primarily relies on term occurrence and therefore may be less effective when relevant information is expressed using different words or semantically related terms.

The TF-IDF results were subsequently compared with the semantic representation to determine whether combining lexical and semantic signals could provide complementary retrieval information.

4.5 Hybrid Search

To combine lexical and semantic information, this study implemented a Hybrid Search approach that integrates TF-IDF similarity and semantic similarity. Both similarity scores were normalized to a common scale using min-max normalization before being combined.

The hybrid score was calculated using:

$$Score_hybrid = \alpha \cdot (Sim_semantic \times 1.1) + (1 - \alpha) \cdot Sim_tfidf \quad (3)$$

Where $\alpha = 0.78$ represents the weight assigned to the semantic similarity component.

The factor 1.1 was applied as an amplification factor to the semantic similarity component. In the current implementation, the values \(\alpha=0.78\) and 1.1 were used as configuration parameters of the proposed hybrid retrieval system.

4.6 System Testing and Evaluation

System testing was conducted using three representative queries corresponding to learning materials from different subject areas. Each query was associated with a set of manually determined relevant documents that served as the ground truth for evaluation.

The retrieval performance of TF-IDF, Semantic Search, and Hybrid Search was evaluated using Precision, Recall, F1-score, Average Precision (AP), and Mean Average Precision (MAP).

Table 1. Evaluation results — TF-IDF (Full-Text Retrieval)

Query	Precision	Recall	F1-Score	AP
Q1	0.333	1.000	0.500	0.333
Q2	0.250	1.000	0.400	0.500
Q3	0.166	1.000	0.285	0.200

The TF-IDF results show that Recall reached 1.000 for all three queries, indicating that the relevant documents defined in the ground truth were retrieved. However, Precision remained relatively low, ranging from 0.166 to 0.333. This indicates that the retrieval results also contained non-relevant documents.

The AP values ranged from 0.200 to 0.500, with Q2 obtaining the highest AP. The MAP value of TF-IDF was 0.344. These results indicate that TF-IDF was able to retrieve relevant documents when query terms had lexical correspondence with the textbook content, but the ranking also included documents that were not considered relevant.

Table 2. Evaluation results — Semantic Search

Query	Precision	Recall	F1-Score	AP
Q1	0.000	0.000	0.000	0.000
Q2	0.166	1.000	0.285	0.250
Q3	0.200	1.000	0.333	0.333

Semantic Search produced different results across the three queries. For Q2 and Q3, Recall reached 1.000, while Q1 produced Precision, Recall, F1-score, and AP values of 0.000.

The MAP value of Semantic Search was 0.194, which was lower than both TF-IDF and Hybrid Search. The result indicates that semantic similarity alone did not consistently correspond to the manually defined relevance judgments in the evaluated queries.

The result for Q1 also shows that semantic similarity does not necessarily guarantee successful retrieval for every query. Differences in query formulation, terminology, subject-specific vocabulary, and the characteristics of the embedding representation may affect the ranking of relevant documents.

Table 3. Evaluation results — Hybrid Search

Query	Precision	Recall	F1-Score	AP
Q1	0.333	1.000	0.500	0.500
Q2	0.250	1.000	0.400	0.333
Q3	0.200	1.000	0.333	0.250

The Hybrid Search results show Recall values of 1.000 for all three queries, while Precision ranged from 0.200 to 0.333. The AP values ranged from 0.250 to 0.500.

The MAP value obtained by Hybrid Search was 0.361, which was higher than the MAP values of TF-IDF (0.344) and Semantic Search (0.194).

Table 4. Mean Average Precision (MAP) comparison

Method	MAP
Hybrid_Search	0.361
TF-IDF (Full-Text Retrieval	0.344
Semantic_Search	0.194

Based on Table 4, Hybrid Search obtained the highest MAP value among the three evaluated approaches. The MAP value increased by 0.017 compared with TF-IDF, from 0.344 to 0.361. Compared with Semantic Search, the difference was 0.167, from 0.194 to 0.361.

The difference between TF-IDF and Hybrid Search is relatively small compared with the difference between Hybrid Search and Semantic Search. This indicates that the contribution of the hybrid approach in this experiment was primarily reflected in the combination of lexical and semantic signals rather than in replacing lexical retrieval entirely.

4.7 Discussion

The experimental results show that TF-IDF, Semantic Search, and Hybrid Search produced different retrieval performance. TF-IDF achieved a MAP of 0.344, Semantic Search 0.194, and Hybrid Search the highest MAP of 0.361 within the evaluated setting.

TF-IDF remained useful when query terms occurred directly in textbook content, although its Precision values of 0.166–0.333 indicate that lexical matching could also retrieve documents outside the defined relevant set. Semantic Search obtained a lower MAP of 0.194, particularly due to Q1, where Precision, Recall, F1-score, and AP were 0.000. This indicates that semantic similarity alone did not consistently correspond to the ground-truth relevance judgments.

Hybrid Search achieved a MAP of 0.361, which was 0.017 higher than TF-IDF and 0.167 higher than Semantic Search. This result suggests that lexical and semantic signals provided complementary information in the evaluated dataset. TF-IDF captured explicit term matches, while semantic similarity provided an additional representation-based retrieval signal.

Recall alone was insufficient to describe ranking quality because it reached 1.000 in most cases, while Precision and AP varied across methods and queries. MAP therefore provided additional information about the ranking of relevant documents. The document-level aggregation also considered both the highest-scoring chunk and the average relevance of retrieved chunks, allowing the system to represent textbooks that contain only specific sections relevant to a student's learning material.

Overall, Hybrid Search obtained the highest MAP among the evaluated approaches. However, this result is limited to the current dataset of 80 textbooks and three evaluation queries. Further evaluation with more queries, broader subject coverage, and systematically tuned hybrid parameters is needed to assess the consistency of the results across different library collections and user information needs.

5. Conclusions

Based on the experimental results, the developed textbook recommendation system was able to retrieve textbook content relevant to students' learning materials by integrating Semantic Search and Full-Text Retrieval through a Hybrid Retrieval approach. The system processed a collection of 80 textbooks into 21,556 text paragraphs and applied natural language processing, TF-IDF representation, and semantic embedding to represent textual content for retrieval.

The evaluation using Precision, Recall, F1-score, Average Precision (AP), and Mean Average Precision (MAP) showed that the Hybrid Search configuration obtained a MAP value of 0.361, compared with 0.344 for TF-IDF and 0.194 for Semantic Search. Within the experimental setting of this study, the Hybrid Search configuration therefore produced the highest MAP among the three evaluated approaches. The results indicate that combining lexical matching from TF-IDF with semantic similarity can provide complementary retrieval signals, particularly when queries contain both explicit terms and concepts that may be expressed using different wording. However, the improvement over TF-IDF was relatively limited, indicating that the contribution of semantic retrieval should be interpreted in the context of the dataset, query set, and retrieval configuration used in this study.

For future research, the system should be evaluated using a larger and more diverse collection of textbooks and a broader set of queries to assess the robustness and generalizability of the retrieval results. Future studies may also investigate systematic parameter tuning for the Hybrid Retrieval configuration and compare additional lexical and dense retrieval methods. The recommendation component could further be extended with personalized or collaborative filtering approaches to incorporate user preferences and interaction history. In addition, usability testing involving students, teachers, and librarians should be conducted to evaluate the practical usefulness and user experience of the system. Integration with digital library services and further development of the user interface may also improve the applicability of the system in a real-world school library environment.

Acknowledgements

The author would like to express sincere gratitude to the Master of Computer Systems Program, Universitas Handayani Makassar, for its academic support throughout this research.

The author also thanks the thesis supervisors for their guidance, direction, and valuable suggestions during the completion of this study. Appreciation is further extended to the Library of SMK Negeri Labuang for granting access and providing the data required to support this research.

References

R. Ardiansyah, M. A. Bianto, and B. D. Saputra, "Sistem Rekomendasi Buku Perpustakaan Sekolah Menggunakan Metode Content-Based Filtering," Jurnal CoSciTech (Computer Science and Information Technology), vol. 4, no. 2, pp. 510–518, Oct. 2023, doi: 10.37859/coscitech.v4i2.5131.

L. Rosidah and P. Dellia, "Library Book Recommendation System Using Content-Based Filtering," Internet of Things and Artificial Intelligence Journal, vol. 4, no. 1, pp. 42–65, Feb. 2024, doi: 10.31763/iota.v4i1.693.

A. Askari, A. Abolghasemi, G. Pasi, W. Kraaij, and S. Verberne, "Injecting the BM25 Score as Text Improves BERT-Based Re-rankers," in Advances in Information Retrieval, 45th European Conference on Information Retrieval (ECIR 2023), Dublin, Ireland, Apr. 2023, vol. 13980, pp. 66–83, doi: 10.1007/978-3-031-28244-7_5.

C. G. Reswara, J. Nicolas, I. M. D. Widyatama, D. David, and P. Arisaputra, "Book Recommendation System Using TF-IDF and Cosine Similarity," AIP Conference Proceedings, vol. 3135, no. 1, p. 020003, May 2024, doi: 10.1063/5.0212477.

H. S. Walsh and S. R. Andrade, "Semantic Search With Sentence-BERT for Design Information Retrieval," in Volume 2: 42nd Computers and Information in Engineering Conference (CIE), St. Louis, Missouri, USA, Aug. 2022, p. V002T02A066, doi: 10.1115/DETC2022-89557.

G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave, "Unsupervised Dense Information Retrieval With Contrastive Learning," Transactions on Machine Learning Research, 2022.

D. Lee, S.-W. Hwang, K. Lee, S. Choi, and S. Park, "On Complementarity Objectives for Hybrid Retrieval," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, Jul. 2023, pp. 13357–13368, doi: 10.18653/v1/2023.acl-long.746.

J. Lu, K. Hall, J. Ma, and J. Ni, "HYRR: Hybrid Infused Reranking for Passage Retrieval," in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, May 2024, pp. 8528–8534.

P. Ouanlamai, T. Panyatip, P. Phothinam, A. Sumungkaset, N. Masusai, S. Phimphisan, and P. Thipsanthia, "Book Recommendation System Using Content-Based Filtering Technique," Journal of Engineering and Industrial Technology, Kalasin University, vol. 3, no. 3, pp. 36–53, Jun. 2025, doi: 10.14456/jeit.2025.13.

A. Bhajantri, N. K. Goudar, D. G. M., R. B. Kaliwal, V. Rathod, A. Kulkarni, and G. K., "Personalized Book Recommendations: A Hybrid Approach Leveraging Collaborative Filtering, Association Rule Mining, and Content-Based Filtering," EAI Endorsed Transactions on Internet of Things, vol. 10, no. 1, 2024, doi: 10.4108/eetiot.6996.

N. N. Doan, A. Härmä, R. Celebi, and V. Gottardo, "A Hybrid Retrieval Approach for Advancing Retrieval-Augmented Generation Systems," in Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024), Trento, Italy, Oct. 2024, pp. 397–409.

D. Ridhwanullah, Y. K. Kumarahadi, and B. D. Raharja, "Content-Based Filtering pada Sistem Rekomendasi Buku Informatika," Jurnal Ilmiah SINUS, vol. 22, no. 2, pp. 57–66, Jul. 2024, doi: 10.30646/sinus.v22i2.840.

E. Nikolaev, I. Bondarenko, I. Aushev, V. Krikunov, A. Glinskii, V. Konovalov, and J. Belikova, “FactDebug at SemEval-2025 Task 7: Hybrid Retrieval Pipeline for Identifying Previously Fact-Checked Claims Across Multiple Languages,” in Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), 2025, pp. 2190–2196.

M.-C. Lee, Q. Zhu, C. Mavromatis, Z. Han, S. Adeshina, V. N. Ioannidis, H. Rangwala, and C. Faloutsos, “HybGRAG: Hybrid Retrieval-Augmented Generation on Textual and Relational Knowledge Bases,” in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 879–893, doi: 10.18653/v1/2025.acl-long.43.

N. Saputri and P. Suryati, “Machine Learning-Based Book and Library Recommendation Application Using Content-Based Filtering Method,” Jurnal Teknologi Informasi dan Terapan, vol. 12, no. 1, pp. 25–33, Jun. 2025, doi: 10.25047/jtit.v12i1.445.

M. F. Roziq, R. N. Sarbini, H. Mukminna, and M. S. Anam, “Developing a Content-Based Book Recommendation System for School E-Libraries Using TF-IDF and Cosine Similarity,” Jurnal Ilmiah Informatika dan Komputer, vol. 3, no. 1, pp. 49–57, Jun. 2026, doi: 10.69533/informatech.volume3number1.543.