OPEN ACCESS
VOL. 8, NO. 2, PP. 75–89, MAY 2026

ISSN : 2622-8106 (ONLINE)
DOI: 10.20895/INISTA.V8I2.2116

Journal of Informatics, Information System, Software Engineering and Applications (INISTA)

Empirical Analysis of Query Expansion Strategies for IndoSBERT-Based Semantic Retrieval

Dela Puspita Lasminingrum[1], Eva Yulia Puspaningrum[2,*], Budi Mukhammad Mulyo[3]

Department of Informatics, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur, 60294, Indonesia

[1]22081010209@student.upnjatim.ac.id
[2]evapuspaningrum.if@upnjatim.ac.id
[3]budi.m.mulyo.fasilkom@upnjatim.ac.id

Received: 13-04-2026; Revised: 05-06-2026; Accepted: 10-06-2026.

Abstract

The advancement of information retrieval systems has shifted from keyword-based approaches to semantic retrieval using transformer-based models such as BERT and its variants. Despite their ability to capture contextual meaning, the vocabulary mismatch problem between queries and documents remains a key challenge. Query expansion (QE) is commonly used to address this issue, but its effectiveness in semantic retrieval is not always consistent. This study aims to analyze the impact of query expansion on a semantic retrieval system based on a fine-tuned IndoSBERT model using a dataset of undergraduate thesis titles and abstracts at repository UPN "Veteran" Jawa Timur. A hybrid QE approach is proposed by combining pretrained FastText and domain-specific Word2Vec embeddings, with and without filtering mechanisms. The system performance is evaluated using Precision@15, Recall@15, Mean Average Precision (MAP), and nDCG@15. The results show that QE can improve retrieval performance when properly controlled. Hybrid QE with filtering and Word2Vec with filtering produce the highest MAP score of 0.4249 and Recall@15 of 0.5366, while Hybrid QE without filtering achieves the highest nDCG@15 of 0.8747. In contrast, FastText-based QE without filtering results in performance degradation due to query drift, with MAP decreasing to 0.3378. It can be concluded that the effectiveness of QE in semantic retrieval is highly dependent on the quality of expansion terms and the application of filtering strategies. QE is not inherently beneficial, but requires careful design to improve retrieval performance.

Keywords: *Hybrid Embedding, IndoSBERT, Query Expansion, Semantic Retrieval, Word2Vec*

Corresponding Author:
*Eva Yulia Puspaningrum
Department of Informatics, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Email: evapuspaningrum.if@upnjatim.ac.id

I. INTRODUCTION

The development of information retrieval systems has shifted from keyword-based matching approaches toward semantic-based approaches by leveraging vector representations derived from transformer-based models. Models such as BERT and its variants are capable of representing queries and documents within a semantic space, thereby improving retrieval quality compared to traditional keyword-based methods [1]. This approach, known as semantic retrieval, focuses on understanding meaning rather than mere word matching.

In sentence-level retrieval, Sentence-BERT and its variants are designed to encode queries and documents into dense sentence embeddings that can be compared efficiently using similarity measures [2]. IndoSBERT is used in this study because the retrieval task requires sentence-level representations for Indonesian thesis titles and abstracts. Unlike IndoBERT, which is mainly a masked language model, IndoSBERT follows the Sentence-BERT architecture and produces sentence embeddings that are suitable for semantic similarity search. Therefore, the term IndoSBERT-based retrieval in this paper refers to retrieval using sentence-level embeddings generated by an IndoSBERT model.

Despite these advancements, the problem of vocabulary mismatch between queries and documents remains a major challenge in information retrieval systems [3]. User queries often do not use the same terms as those found in relevant documents, leading to decreased retrieval performance. To address this issue, query expansion (QE) techniques are employed by adding semantically related terms to enrich the query representation [4]. Essentially, QE aims to improve retrieval performance by enhancing the contextual representation of queries.

However, in the era of transformer-based models, the effectiveness of QE is not always consistent. A study by Min Pan *et al.* (2025) shows that while QE can improve performance by enriching query representations, it may also introduce irrelevant information that degrades retrieval quality, a phenomenon known as query drift [5]. The study further indicates that QE in dense retrieval systems still faces challenges in maintaining semantic consistency in expanded queries. This finding is reinforced by Allahim *et al.* (2025), who highlight that the selection of expansion sources is crucial to avoid disrupting semantic consistency in modern dense retrieval models [6].

Several recent QE studies have explored different expansion strategies. Aligned query expansion uses large language models to produce expansions that remain semantically close to the original query [4]. Gaussian kernel-based QE attempts to strengthen semantic associations in dense retrieval while controlling query drift [5]. A taxonomy of semantic QE further shows that the choice of expansion source, filtering strategy, and semantic control mechanism strongly affects retrieval performance [6].

On the other hand, the quality of query expansion is highly dependent on the methods used. Embedding-based approaches such as FastText can generate a wider variety of word expansions by leveraging subword (n-gram) representations, allowing them to capture similarities even for rare or unseen words. However, this approach tends to produce more general expansions, as it relies on surface-level similarity rather than domain-specific context [7]. In contrast, domain-specific Word2Vec can capture semantic relationships based on word co-occurrence within similar contexts, resulting in more domain-specific expansions, albeit with limited vocabulary coverage [8].

Therefore, combining these two approaches becomes an interesting direction, as it enables the integration of FastText's broad vocabulary coverage with Word2Vec's contextual semantic strength in domain-specific retrieval systems. For instance, Pertiwi *et al.* (2025), through the Fast2Vec model, demonstrate that Word2Vec excels at capturing contextual semantic relationships but struggles with out-of-vocabulary (OOV) words, whereas FastText effectively handles OOV through subword modeling but is less optimal in capturing complex semantic relationships [9].

Most studies have focused on using a single type of embedding or comparing them independently. There remains a research gap in evaluating the effectiveness of a hybrid combination of pretrained Indonesian FastText and locally fine-tuned Word2Vec as a query enrichment method for IndoSBERT-based retrieval systems. This study aims to address this gap by utilizing a dataset consisting of thesis titles and abstracts from a university repository, representing a specific academic domain.

The novelty of this research lies in empirically evaluating how word-level expansion terms generated by Word2Vec, FastText, and their hybrid combination affect sentence-level IndoSBERT retrieval, especially when filtering is applied to control semantic drift. FAISS is used as the vector indexing backend to retrieve dense document embeddings efficiently during similarity search [10]. Based on this background, the research questions are: (1) how does query expansion affect the performance of IndoSBERT-based semantic

retrieval, and (2) how do the characteristics of expansions generated by FastText and Word2Vec differ in influencing retrieval results? This study aims to analyze the impact of QE on a fine-tuned IndoSBERT model, both in terms of performance improvement and potential degradation, through evaluation on an academic repository dataset.

II. LITERATURE REVIEW

A. Semantic Retrieval with Transformer-Based Model

Information retrieval has evolved significantly from traditional bag-of-words models toward dense retrieval systems that leverage pre-trained language models. Iida and Okazaki (2021) demonstrate that incorporating semantic textual similarity alongside lexical matching substantially improves retrieval quality, as the model is able to represent both surface-form and contextual meaning simultaneously [1]. The shift to transformer-based approaches, such as BERT and its sentence-level variants (SBERT), enables the encoding of entire queries and documents into dense vector representations within a shared semantic space [2, 11].

Sentence-BERT (SBERT) and its multilingual variants have been widely adopted for semantic search tasks due to their efficiency in encoding sentence-level semantics using siamese and triplet network structures. Fine-tuning using ranking losses has been shown to be an effective training strategy for sentence embedding models, enabling the model to learn discriminative embeddings for retrieval tasks [12, 13]. The IndoSBERT model used in this study (`firqaaa/indo-sentence-bert-base`) is a sentence transformer specifically pre-trained on Indonesian text, providing a strong foundation for domain adaptation through fine-tuning [14].

B. Query Expansion Techniques

Query expansion (QE) is a well-established technique in information retrieval aimed at bridging the vocabulary mismatch between short user queries and more terminology-rich documents. Contextual embedding-based QE has been proposed to improve query representation by using contextualized semantic information [3]. More recently, Yang *et al.* (2025) introduced aligned query expansion using large language models (LLMs), showing that semantically aligned expansions reduce ambiguity while maintaining relevance [4].

Pan *et al.* (2025) further explored Gaussian kernel-based semantic enhancement for dense retrieval QE, highlighting that while QE can boost performance, uncontrolled expansion may cause query drift (the phenomenon where the expanded query deviates from the original user intent) [5]. Allahim *et al.* (2025) conducted a comprehensive taxonomy of semantic QE approaches, emphasizing that the selection of expansion sources critically affects semantic consistency in modern retrieval models [6].

C. Word Embedding Models: Word2Vec and FastText

Word embedding models form the backbone of embedding-based QE systems. Word2Vec, particularly the skip-gram variant, learns word representations by predicting surrounding context words, capturing semantic relationships through word co-occurrence patterns [8]. When trained on domain-specific corpora, Word2Vec produces embeddings that reflect the specialized vocabulary and contextual usage of that domain, making it a strong candidate for domain-adaptive QE. However, Word2Vec suffers from the out-of-vocabulary (OOV) problem, as it cannot represent words unseen during training.

FastText, developed by Facebook AI Research, addresses the OOV limitation through subword (character n-gram) representations, allowing the model to estimate embeddings for unseen words by composing subword units. Lumbantoruan *et al.* (2024) compared FastText and Word2Vec in Indonesian information retrieval tasks, finding that FastText achieves broader vocabulary coverage but may generate less domain-specific expansions [7]. Pertiwi *et al.* (2025) proposed the Fast2Vec hybrid model, demonstrating that Word2Vec excels in capturing contextual semantic relationships while FastText handles OOV words more effectively [9]. These complementary strengths motivate the hybrid embedding approach adopted in this study.

D. FAISS for Efficient Similarity Search

Facebook AI Similarity Search (FAISS) is a library designed for efficient similarity search over large collections of high-dimensional vectors [10]. FAISS supports both exact and approximate nearest neighbor search, enabling scalable dense retrieval over large document collections. Ramadhan *et al.* (2024) successfully applied FAISS in Indonesian-language passage retrieval for question answering, demonstrating its

effectiveness in combination with BERT-based encoders [15]. Wang *et al.* (2022) also used dense external expansion with FAISS-indexed document collections, showing improvements in zero-shot retrieval settings [16]. These works support the choice of FAISS as the indexing backend in the proposed system.

III. RESEARCH METHOD

This study adopts a structured methodology consisting of data collection, preprocessing, query expansion, semantic retrieval modeling, and evaluation. The entire framework is implemented in Python using Gensim for embedding models, Sastrawi for text preprocessing, `sentence-transformers` for IndoSBERT, and FAISS for efficient similarity search. Figure 1 shows the operational workflow of the proposed retrieval system during the search process. This figure illustrates how a user query is processed after the dataset, embedding models, and FAISS index have been prepared. The query is expanded using Word2Vec, FastText, or a hybrid strategy, encoded into a sentence-level representation using the fine-tuned IndoSBERT encoder, compared with indexed document embeddings through FAISS similarity search, and returned as ranked thesis documents.

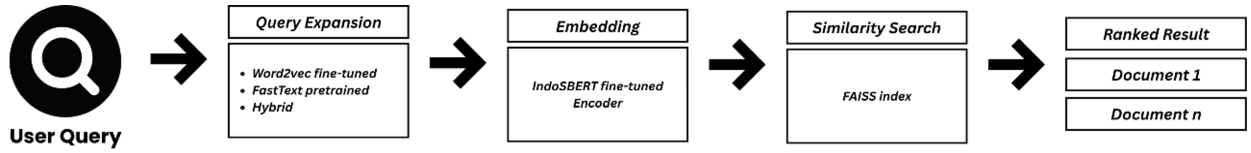

Fig. 1. System Workflow.

A. Data Collection Process

This study utilizes a dataset consisting of undergraduate thesis documents obtained from the repository of Universitas Pembangunan Nasional "Veteran" Jawa Timur. The collected data focus on thesis titles and abstracts, as these components represent the core research topics and serve as the primary elements in information retrieval systems. Data collection was conducted using a web scraping technique implemented in Python, as the repository does not provide a public API for large-scale data access. It is important to note that this process was carried out with permission from the repository administrators.

The data collection process was performed in two stages: (1) crawling to collect URLs of thesis detail pages from the repository index, and (2) scraping to extract title and abstract information from each document page. The collected data were stored in CSV format, where each row represents a thesis document. In total, the dataset consists of 15,326 thesis documents, including URL, title, abstract, study program, and faculty information. A representative sample of the final dataset used in this study is summarized in Table I below.

TABLE I. EXAMPLE OF DATASET SCRAPED FROM UPN JATIM REPOSITORY.

Url	Title	Abstract	Faculty	Department
https://repository.upnjatim.ac.id/1004/	Sistem penilaian usulan riset dan pengabdian kepada masyarakat menggunakan algoritma winnowing	Universitas berfungsi untuk memfasilitasi serta mewadahi …Kunci: sistem penilaian, algoritma menampi, plagiarisme, riset, pengabdian kepada masyarakat.	Faculty of computer science	Departemen of informatics
https://repository.upnjatim.ac.id/10052/	Pabrik asam phospat dari batuan phospat dan asam sulfat dengan proses wet	Pra rencana pabrik asam phosphat ini direncanakan untuk dapat berproduksi …selama 330 hari/tahun.	Faculty of engineering	Departement of chemical engineering

B. Data Preprocessing

The preprocessing stage aims to prepare thesis titles and abstracts for use in semantic retrieval and query expansion tasks. The raw data obtained from scraping contain variations in writing, symbols, and irrelevant terms, thus requiring several text cleaning steps. First, English abstracts were translated into

Indonesian to ensure language consistency with the IndoSBERT model. This step is essential to align the language of documents with the query representation, as semantic retrieval models operate within a shared embedding space. In the context of academic search, translation has been widely used to improve retrieval effectiveness by reducing language mismatches between queries and documents [17]. In this study, the translation process was performed automatically using spreadsheet functions (`=GOOGLETRANSLATE()`) after data collection.

Next, text normalization and case folding were applied by removing non-alphabetic characters, punctuation, and extra spaces, followed by converting all text to lowercase. This step ensures uniform text representation and prevents inconsistencies in semantic modeling.

The preprocessing pipeline is separated according to its purpose. For IndoSBERT-based semantic retrieval and fine-tuning, the system uses `text_semantic`, which retains normalized text without stopword removal or manual tokenization so that sentence context is preserved and processed by the IndoSBERT tokenizer. For embedding-based query expansion, the system uses `text_fasttext`, which applies tokenization and stopword removal using a combination of the Sastrawi stopword list and a custom domain-specific list. This separation ensures that word-level preprocessing is applied only to the QE pipeline, not globally to IndoSBERT input. This preprocessing pipeline aims to produce clean, consistent, and relevant text representations to improve embedding quality and retrieval performance.

C. Query Expansion

Query expansion (QE) is a technique used to enhance user queries by adding semantically related terms, thereby reducing the vocabulary mismatch between short queries and more terminology-rich documents. The core idea is to expand the original query with synonymous or contextually related words so that the retrieval system can find more relevant documents that use different terminology to express the same concept.

To illustrate how QE works, consider the following example. If a user submits the query “IoT”, the system identifies semantically similar terms from the embedding models. Word2Vec (trained on the thesis corpus) might suggest domain-relevant expansions such as “mikrokontroler” (microcontroller), “ESP32”, or “sensor”, because these words frequently co-occur with “IoT” in the thesis collection. FastText (pretrained on general Indonesian text) might additionally suggest “internet”, “jaringan” (network), or “smarthome”, leveraging broader language coverage. The expanded query then becomes: “IoT mikrokontroler ESP32 sensor internet jaringan”. This richer representation allows the retrieval model to retrieve documents mentioning, for example, “sistem pemantauan berbasis ESP32” (ESP32-based monitoring system), which would have been missed by the original single-word query “IoT”.

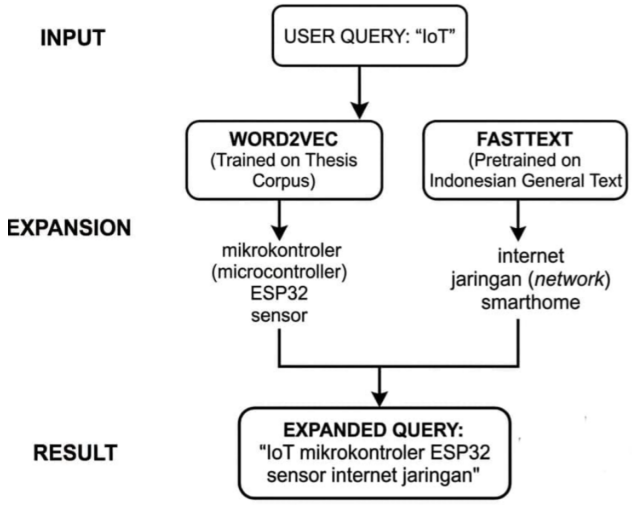

Fig. 2. Workflow of QE.

This study adopts a hybrid embedding approach that combines Word2Vec and FastText. The FastText model uses a pretrained Indonesian embedding (cc.id.300.bin) that leverages subword representations, allowing it to handle rare and out-of-vocabulary terms. In contrast, Word2Vec is trained on the thesis corpus

using the skip-gram architecture, enabling it to capture domain-specific semantic relationships. The illustration of how QE works shown at Figure 2.

The Word2Vec training configuration was determined based on empirical considerations and established practices in the literature. A vector size of 300 was chosen to match the dimensionality of the pretrained FastText model (cc.id.300.bin), enabling direct vector alignment between the two models during the hybrid initialization step. A window size of 8 was selected to capture longer-range contextual dependencies in academic texts, where related terms may appear several words apart within a sentence or paragraph, a larger window than typical (5) is appropriate for technical and scientific domains. A minimum word frequency of 5 was set to exclude very rare terms that appear fewer than 5 times in the corpus, filtering out typographical errors and overly specific terminology that would not generalize across queries. Training for 10 epochs was chosen as a balance between model convergence and computational efficiency on the 15,326-document corpus, following common practice for medium-sized domain corpora. To improve initialization quality, overlapping vocabulary between Word2Vec and FastText is aligned by transferring pretrained FastText vectors to the Word2Vec model before training, enabling the domain-specific training to refine rather than learn from scratch.

During query expansion, each token in the query is first matched against the Word2Vec vocabulary. If the token is not found, FastText is used as a fallback. For each token, the system retrieves a set of semantically similar words based on cosine similarity. Only candidates with similarity scores above 0.6 are considered. The candidate retrieval for hybrid query expansion process is summarized in algorithm at Figure 3.

To ensure the quality of expansion, a filtering mechanism is applied. This filtering removes stopwords, numerical tokens, low-frequency terms, and words that exceed a predefined document frequency ratio. Additionally, morphological variants of the original query terms are excluded to prevent redundant expansions. This filtering step plays a critical role in reducing noise and preventing semantic drift. Filtering process algorithm can be shown at Figure 4.

The output of the QE module is not directly used as word embeddings for retrieval. Instead, the selected expansion terms are concatenated with the original query to form an expanded textual query string. This reconstructed query is then passed to the IndoSBERT encoder, which produces a sentence-level embedding. Therefore, Word2Vec and FastText are used only to suggest expansion terms, while the final retrieval representation is still generated by IndoSBERT.

```
tokens = simple_tokenize(q)
exp = []
for t in tokens:
    if t in w2v:
        cands = w2v.most_similar(t, topn=10)
    elif t in ft:
        cands = ft.most_similar(t, topn=10)
    else:
        continue
    for c, s in cands:
        if s > 0.6 and c not in exp:
            exp.append(c)
expanded = " ".join(tokens + exp)
print("Expanded:", expanded)
```

Fig. 3. Pseudocode for hybrid candidate retrieval query expansion.

```
def is_valid(word, base_tokens, min_df=5, max_df_ratio=0.15):
    w = normalize(word)
    if w in [normalize(bt) for bt in base_tokens]:
        return False
    if w in STOPWORDS:
        return False
    if word.isdigit() or len(word) < 3:
        return False
    if doc_freq.get(word, 0) < min_df:
        return False
    if doc_freq.get(word, 0) / N_DOCS > max_df_ratio:
        return False
    for bt in base_tokens:
        if is_morphological_variant(normalize(bt), w):
            return False
    return True
```

Fig. 4. Pseudocode for filtering candidate query expansion.

D. Semantic Retrieval Model

The semantic retrieval system in this study is built using IndoSBERT, a Sentence-BERT based model implemented through the `sentence-transformers` library. Each input sentence is first encoded using a transformer-based encoder, and the final sentence embedding is obtained by applying mean pooling over token embeddings as follows:

$$\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i \tag{1}$$

where $\mathbf{h}_i$ represents the contextual embedding of the i-th token produced by the transformer, and n is the number of tokens in the sentence. The pretrained model (`firqaaa/indo-sentence-bert-base`) [14] with 768-dimensional embeddings is fine-tuned using thesis titles and abstracts to adapt to the academic domain. Fine-tuning IndoSBERT is necessary because while the pretrained model already captures general Indonesian sentence semantics, the academic thesis domain uses specialized terminology, abbreviations, and discourse patterns that differ substantially from the general pre-training corpus.

Fine-tuning adapts the embedding space to better align thesis titles (as queries) with their corresponding abstracts (as documents), enabling more precise semantic matching within this domain. To empirically validate this design choice, we compare the retrieval performance of the fine-tuned model against the baseline pre-trained IndoSBERT without fine-tuning in section 4.

Fine-tuning is performed using title-abstract pairs, where titles represent queries and abstracts represent documents. The model is trained using Multiple Negatives Ranking Loss (MNRL), which enables the model to maximize the similarity between relevant query-document pairs while simultaneously minimizing similarity with non-relevant pairs within the same batch [12, 13]. Document embeddings are stored using Facebook AI Similarity Search (FAISS), a library designed for efficient similarity search over high-dimensional vector representations [10]. The embedding matrix has dimensions $N \times d$, where N is the number of documents and d is the embedding dimension.

During retrieval, the expanded query is converted into an embedding and normalized. Similarity between the query vector $\mathbf{q}$ and document matrix $\mathbf{D}$ is computed using the inner product:

$$\text{score} = \mathbf{D} \cdot \mathbf{q} \tag{2}$$

Since vectors are normalized, this is equivalent to cosine similarity. FAISS then performs top-k selection to retrieve the most relevant documents efficiently. A conceptual illustration of the retrieval model process is shown in Figure 5.

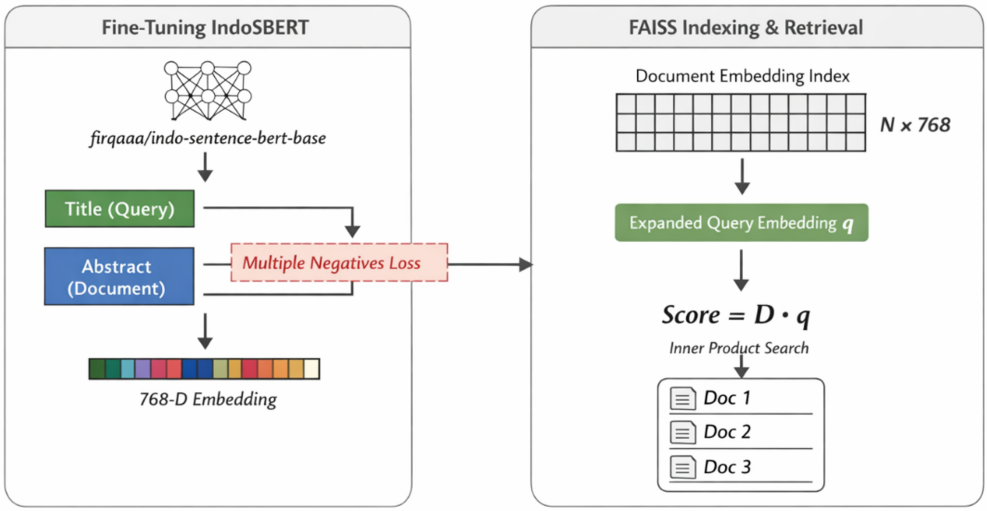

Fig. 5. Illustration model retrieval and indexing.

E. Experimental Setup

To evaluate the impact of query expansion, several experimental scenarios are defined, including a baseline without query expansion and multiple configurations using Word2Vec, FastText, and hybrid approaches, both with and without filtering (Figure 6). Each scenario is applied consistently across 22 representative queries selected from different departments to enable fair comparison. Although the number of queries is limited, each query was evaluated across seven retrieval scenarios, and each scenario returned the top-15 documents. As a result, the pooling process produced more than 2,000 retrieved results that required manual relevance assessment.

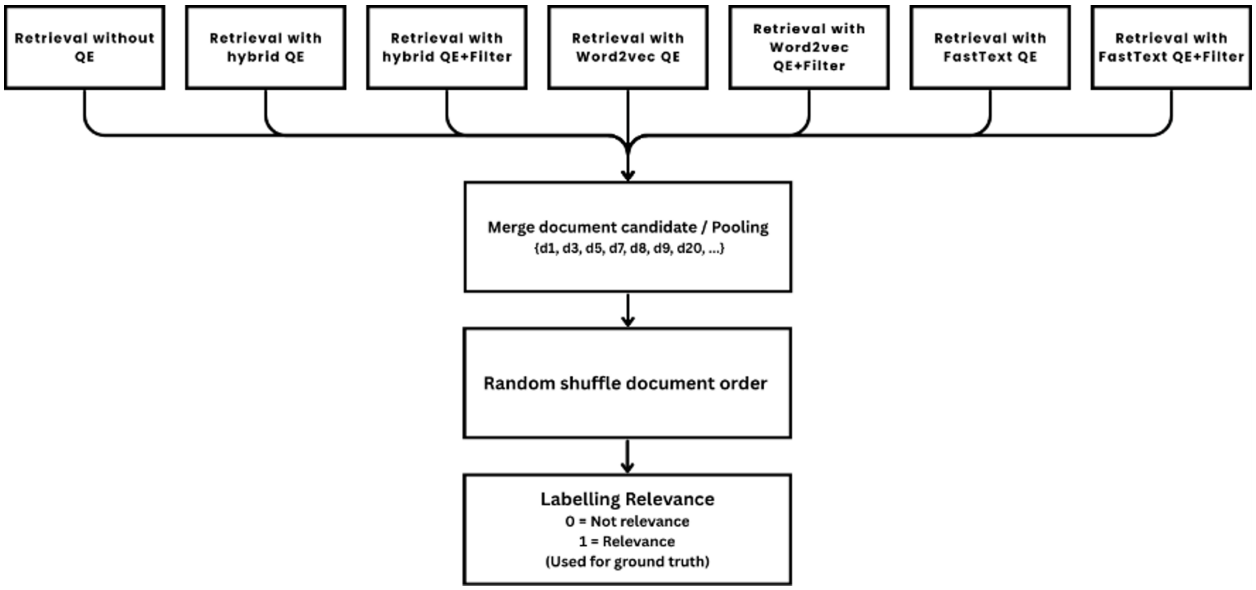

Fig. 6. Experimental Scenario.

Following the TREC-style evaluation protocol, the top-k ($k = 15$) retrieved documents from all systems were pooled and manually annotated. The relevance assessment was conducted by three raters, including the author and selected students from relevant study programs. The annotators were chosen based on their familiarity with the academic domain represented by the queries. Each annotator assessed the retrieved documents using binary relevance labels, where 1 indicates relevant and 0 indicates non-relevant. The assessment was based on topical relevance between the query and the retrieved thesis title. The final rel-

evance judgment was determined using majority voting across the three raters before being finalized into the `qrels` file.

To measure the consistency of the relevance judgments, an inter-rater agreement test was conducted using Fleiss' Kappa. The result showed a Fleiss' Kappa value of 0.39, which falls into the "Fair Agreement" category. Although the agreement level did not reach the "Substantial" category, the positive Kappa value indicates that the consistency among the three raters remains higher than agreement expected by chance. Therefore, the relevance assessment data used in the system evaluation has a statistically justifiable consistency foundation.

The use of 22 queries is acknowledged as a limitation of this study. However, increasing the number of test queries would also substantially increase the manual annotation workload. Given the limited research resources, a larger query set could potentially reduce annotation quality if raters were required to label too many retrieval results within a limited time. Therefore, this study prioritizes representative query selection and controlled manual annotation over a larger but potentially less consistent evaluation set. The query set was selected to represent different academic departments and realistic search needs in the repository. Nevertheless, future work should expand the number of queries, involve more trained annotators, and further validate the relevance judgments to improve generalizability.

F. Model Evaluation

The system is evaluated using standard information retrieval metrics: Precision, Recall, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (nDCG). These metrics assess accuracy, completeness, and ranking quality.

$$\text{Precision@k} = \frac{\text{number of document relevant in top-}k}{k} \tag{3}$$

$$\text{Recall@k} = \frac{\text{number of document relevant in top-}k}{\text{total relevant document}} \tag{4}$$

$$\text{MAP} = \frac{1}{Q}\sum_{q=1}^{Q} AP(q) \tag{5}$$

$$\text{nDCG@k} = \frac{DCG@k}{IDCG@k} \tag{6}$$

IV. RESULT AND DISCUSSION

A. Experimental Data Description

The evaluation in this study is conducted using 22 queries representing various academic domains, including informatics, economics, law, engineering, and other disciplines. The queries are manually selected to reflect diverse and realistic information needs within a thesis repository environment. The representative query used at this experiment presented at Table II. Each query is evaluated across all experimental scenarios, including semantic retrieval using fine-tuned IndoSBERT with various query expansion strategies. The system retrieves the top-15 documents for each query based on cosine similarity in the embedding space.

TABLE II. EXAMPLE OF QUERY SET USED IN THE EVALUATION.

qid	query
1	Audit Internal
2	Budaya Organisasi
...	...
22	Bahan bakar dan energi

B. Comparison: Baseline vs Fine-Tuned IndoSBERT

To justify the use of fine-tuning, we compare the retrieval performance of the baseline pre-trained IndoSBERT (`firqaaa/indo-sentence-bert-base`, without fine-tuning) against the fine-tuned variant, both evaluated without any query expansion on the thesis repository dataset (see Table III). The results show that fine-tuning improves Precision@15, Recall@15, and nDCG@15, while the improvement in MAP is relatively

small. This indicates that adapting the model to the academic thesis domain helps improve semantic alignment at higher ranks, although its effect on average precision remains limited.

TABLE III. PERFORMANCE COMPARISON BASELINE AND FINE-TUNED INDOSBERT.

system	P@15	R@15	MAP	nDCG@15
SEM_BASE	0.421212	0.171090	0.160492	0.765383
SEM_FT	0.448485	0.184724	0.161576	0.826517

C. *Overall System Performance*

The overall performance of the system across different query expansion methods is summarized in Table IV and visualized in Figure 7. The results show that query expansion generally improves retrieval performance compared to the baseline without expansion. The best performance in terms of MAP is achieved by Hybrid QE with filtering and Word2Vec with filtering, both reaching a MAP score of 0.4249. In contrast, the FastText-based QE produces the lowest MAP (0.3378), indicating that not all expansion strategies are beneficial and may even degrade retrieval performance. For clarity, the y-axis is adjusted to better highlight performance differences across methods.

Interestingly, Hybrid QE without filtering achieves the highest nDCG (0.8747), suggesting that unfiltered expansion may improve the ranking of top relevant documents. However, its MAP remains lower than the filtered variant, indicating that overall retrieval consistency is reduced due to the presence of noisy expansion terms.

TABLE IV. PERFORMANCE COMPARISON ACROSS QUERY EXPANSION METHODS.

system	P@15	R@15	MAP	nDCG@15
No QE	0.396970	0.421200	0.361947	0.861482
Hybrid QE	0.430303	0.506205	0.406123	0.874705
Hybrid QE+Filter	0.442424	0.536551	0.424915	0.858716
QE Word2Vec	0.430303	0.506205	0.405834	0.874568
QE Word2Vec+Filter	0.442424	0.536551	0.424915	0.858716
QE FastText	0.387879	0.426729	0.337783	0.744729
QE FastText+Filter	0.400000	0.425361	0.370015	0.854687

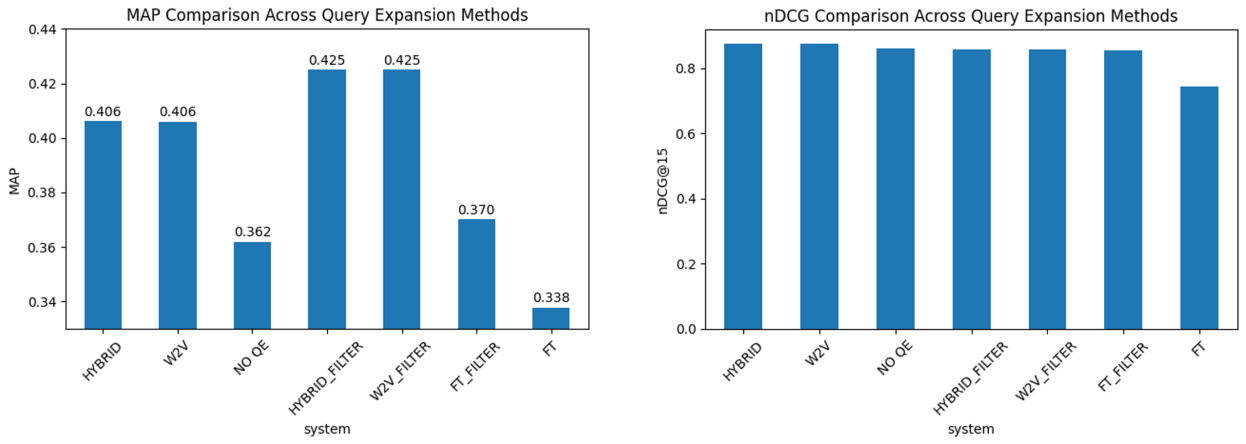

Fig. 7. Performance comparison using MAP and nDCG across methods.

D. *Precision and Recall Analysis*

To further analyze retrieval behavior, Precision@15 and Recall@15 are visualized in Figure 8. The results indicate that filtering consistently improves both precision and recall. The highest values are achieved by

Hybrid QE with filtering and Word2Vec with filtering, with Precision@15 reaching 0.6636 and Recall@15 reaching 0.5366. This is a significant finding, as improvements in both precision and recall simultaneously are not always guaranteed in retrieval systems. It indicates that the filtering mechanism successfully removes noisy expansion terms while preserving relevant semantic information.

In contrast, FastText-based methods show lower performance, particularly in recall. This suggests that FastText generates more general expansion terms that are less aligned with domain-specific queries, leading to suboptimal retrieval results.

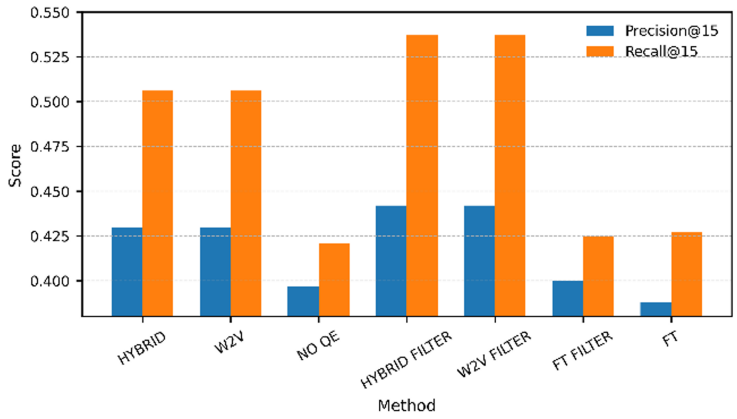

Fig. 8. Precision and Recall comparison across methods.

E. Case Study: Retrieval Example

To provide qualitative insight into system behavior, a case study is conducted using a representative query "iot" with top 3 document retrieval. Result can be shown at Table V.

The IoT case is presented as a representative detailed example rather than the sole evidence for system behavior. While complete retrieval outputs for all evaluated queries are not presented in detail, additional qualitative checks across different academic domains are summarized to support the discussion. The observed pattern remains consistent: domain-specific and filtered expansion generally produced more focused retrieval results, whereas uncontrolled FastText expansion was more likely to introduce broader terms that shifted retrieval away from the original intent.

F. Discussion and Implication

The results demonstrate that the effectiveness of query expansion in semantic retrieval depends not only on the expansion method but also on the quality of the generated terms.

1) Impact of Query Expansion Methods

Among the evaluated methods, Hybrid QE and Word2Vec-based QE produce nearly identical performance across all evaluation metrics. This indicates that domain-specific embeddings play a dominant role in improving retrieval performance, as they capture contextual relationships that are more aligned with the dataset.

The identical MAP and Recall values produced by Hybrid QE+Filter and Word2Vec+Filter suggest that, after filtering, the effective expansion terms retained by the hybrid method were dominated by Word2Vec-generated candidates. Many FastText candidates were likely removed because they were too general, too frequent across documents, or failed the filtering criteria. Therefore, the performance gain should not be interpreted as evidence that the hybrid method is strictly superior to Word2Vec. Instead, the result indicates that domain-specific Word2Vec expansion and filtering are the main contributors to retrieval improvement.

In contrast, FastText-based QE consistently shows lower performance. This suggests that FastText tends to generate broader and less specific terms, which may introduce noise into the query. As a result, the expanded query may deviate from the original intent, leading to reduced retrieval effectiveness. This phenomenon is commonly referred to as query drift and has been widely reported in previous studies [18].

The relatively low values of precision, recall, and MAP indicate that retrieving consistently relevant documents across all ranks remains challenging. Several factors may contribute to this result: the queries

TABLE V. EXAMPLE RETRIEVAL RESULTS ACROSS DIFFERENT METHODS.

Scenario	**Retrieval Result**
No QE	Pengembangan Platform iot Cerdas dengan Trigger Action Programming dan Canvas API berbasis Android Aplikasi indikator pergerakan harga pada cryptocurrency berbasis pivot point Aplikasi informasi arbitrasi cryptocurrency antar market berbasis api dan php
Hybrid QE	Model Predictive Control (MPC) pada Sistem Kendali Suhu itclab dan Pemantauannya Menggunakan Internet Of Things (IOT) Implementasi algoritma smote dan klasifikasi decision tree untuk mendeteksi kecurangan transaksi online Desain dan purwarupa gim edukasi teori pemrograman dasar menggunakan metode block-based programming
Hybrid QE+Filter	Pengembangan Sistem Pembaruan Firmware Over-the-Air untuk Perangkat iot Berbasis ESP32 dengan Integrasi MQTT dan HTTP Analisis perbandingan performasi antara metode pcc (per connection classifier) dengan ecmp (equal cost multi-path) pada jaringan akses internet menggunakan mikrotik router os Pengembangan Platform iot Cerdas dengan Trigger Action Programming dan Canvas API berbasis Android
Word2vec QE	Model Predictive Control (MPC) pada Sistem Kendali Suhu itclab dan Pemantauannya Menggunakan Internet Of Things (IOT) Implementasi algoritma smote dan klasifikasi decision tree untuk mendeteksi kecurangan transaksi online Desain dan purwarupa gim edukasi teori pemrograman dasar menggunakan metode block-based programming
Word2vec QE+Filter	Pengembangan Sistem Pembaruan Firmware Over-the-Air untuk Perangkat iot Berbasis ESP32 dengan Integrasi MQTT dan HTTP Analisis perbandingan performasi antara metode pcc (per connection classifier) dengan ecmp (equal cost multi-path) pada jaringan akses internet menggunakan mikrotik router os Pengembangan Platform iot Cerdas dengan Trigger Action Programming dan Canvas API berbasis Android
FastText QE	Analisis Penerimaan Audiens Terhadap Gerakan Boikot Produk Israel Pada Akun Instagram @gerakanbds Perancangan ui/ux aplikasi marketplace barbershop cutsplace menggunakan metode design thinking (studi kasus : pangkas rambut cutboss) Pengaruh Konsentrasi dan Waktu Aplikasi Zat Pengatur Tumbuh Paclobutrazol terhadap Pertumbuhan dan Hasil Kentang Hitam (Plectranthus Rotundifolius)
FastText QE+Filter	Pengembangan Platform iot Cerdas dengan Trigger Action Programming dan Canvas API berbasis Android Aplikasi indikator pergerakan harga pada cryptocurrency berbasis pivot point Aplikasi informasi arbitrasi cryptocurrency antar market berbasis api dan php

are short and sometimes ambiguous, the repository contains multiple academic domains with overlapping terminology, and title-abstract pairs do not always represent the same wording as real user queries. In addition, binary relevance labels may not fully capture partial relevance, which is common in academic thesis retrieval.

2) Effect of Filtering Mechanism

A clear performance difference is observed between filtered and non-filtered query expansion methods. Filtering consistently improves MAP and recall across all methods. For example, in the Hybrid approach, MAP increases from 0.4061 to 0.4249, while recall increases from 0.5062 to 0.5365 after applying filtering. This demonstrates that filtering effectively removes irrelevant or low-quality expansion terms, resulting in better semantic alignment between queries and documents.

However, a slight decrease in nDCG is observed after filtering. Rather than indicating the removal of relevant expansion terms, this behavior suggests that the filtering mechanism may reduce the diversity or coverage of expansion terms that help promote highly relevant documents to top ranks. In other words, stricter filtering improves overall consistency but may limit the ability of the model to exploit beneficial term variations for optimal ranking at the highest positions.

This trade-off highlights the importance of balancing expansion quality and coverage in semantic retrieval systems. Future improvements could explore more adaptive filtering strategies, such as parameter tuning or relevance-driven expansion methods (e.g., pseudo-relevance feedback), to better preserve useful variations while minimizing noise.

These findings highlight that query expansion is not universally beneficial but must be carefully controlled. The combination of hybrid query expansion and filtering provides a balanced approach, improving retrieval performance while maintaining semantic stability across different queries.

V. CONCLUSION

This study investigates the impact of query expansion on IndoSBERT-based semantic retrieval using a hybrid embedding approach combining Word2Vec and FastText, along with a filtering mechanism. The results show that query expansion generally improves retrieval performance by addressing vocabulary mismatch, as evidenced by higher MAP, precision, and recall compared to the baseline without expansion.

Among the evaluated methods, Hybrid QE+Filter and Word2Vec+Filter achieve identical highest MAP and Recall values, indicating that filtered domain-specific expansion plays an important role in producing more effective query representations. Furthermore, the findings reveal that domain-specific embeddings such as Word2Vec contribute more significantly to performance improvement than general-purpose embeddings like FastText, which tend to generate broader and less relevant terms.

The introduction of filtering plays a critical role in improving retrieval consistency by reducing noise and enhancing semantic alignment, although a trade-off is observed in slightly lower nDCG due to reduced expansion diversity. These results are consistent with prior studies on semantic query expansion that emphasize the importance of expansion source selection, term control, and semantic consistency in dense retrieval systems.

However, the identical performance of Hybrid QE+Filter and Word2Vec+Filter shows that the hybrid approach should not be overstated as strictly superior, instead, the findings suggest that the filtering mechanism and the domain-specific Word2Vec component are the main factors behind the observed improvement. Overall, this research contributes to the advancement of semantic retrieval by providing empirical evidence that the effectiveness of query expansion is conditional, depending on both the embedding method and the quality control mechanism.

This insight highlights the importance of balancing expansion coverage and semantic precision, and opens opportunities for future work in adaptive query expansion techniques, such as parameter optimization and relevance-driven expansion methods. The limited number of evaluation queries remains a limitation of this study and should be addressed in future work using a larger query set and more comprehensive relevance judgments.

ACKNOWLEDGMENT

The completion of this research and the preparation of this article would not have been possible without the support and contributions of various parties. Authors would like to express our sincere gratitude to our supervising lecturers, Mrs. Eva Yulia Puspaningrum, S.Kom., M.Kom and Mr. Budi Mukhamad Mulyo, S.Kom., M.T., for their guidance, valuable insights, and continuous support throughout the research process. Their constructive feedback and academic direction greatly contributed to the development of this study.

We also extend our appreciation to the administrators of the UPN "Veteran" Jawa Timur repository for granting permission and access to the data used in this research. Their support was essential in enabling the

data collection process. Finally, we would like to thank all individuals and colleagues who have provided assistance, directly or indirectly, during the completion of this research.

REFERENCES

[1] H. Iida and N. Okazaki, "Incorporating semantic textual similarity and lexical matching for information retrieval," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, (Shanghai, China), 2021.

[2] K. Diana and M. L. Khodra, "Indosbert: Enhancing indonesian sentence embeddings with siamese networks fine-tuning," in *2023 International Conference on Advanced Informatics for Computing Research (ICAICR)*, 2023. doi:10.1109/ICAICTA59291.2023.10390469.

[3] S. Naseri, J. Dalton, A. Yates, and J. Allan, "Ceqe: Contextualized embeddings for query expansion," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12656 LNCS, pp. 467–482, 2021. doi:10.1007/978-3-030-72113-8_31.

[4] A. Yang, M. Ai, G. Penha, and E. Palumbo, "Aligned query expansion : Efficient query expansion for information retrieval through llm alignment," *arXiv*, vol. 1, no. 1, 2025.

[5] M. Pan, W. Xiong, S. Zhou, M. Gao, and J. Chen, "Llm-based query expansion with gaussian kernel semantic enhancement for dense retrieval," *Electronics*, pp. 1–18, 2025.

[6] A. Allahim, A. Cherif, and A. Imine, "Semantic approaches for query expansion: taxonomy, challenges, and future research directions," *PeerJ Computer Science*, vol. 11, pp. 1–53, 2025. doi:10.7717/peerj-cs.2664.

[7] R. Lumbantoruan, M. Puspita, S. Nababan, and L. A. Saragih, "Analisis perbandingan fasttext dan word2vec pada sistem temu balik informasi," *Seminar Nasional Sains Data*, vol. 2024, no. Senada, pp. 1033–1041, 2024.

[8] S. Brandl, D. Lassner, A. Baillot, and S. Nakajima, "Domain-specific word embeddings with structure prediction," 2022.

[9] A. Pertiwi, A. Azhari, and S. Mulyana, "Fast2vec , a modified model of fasttext that enhances semantic analysis in topic evolution," *PeerJ Computer Science*, pp. 1–36, 2025. doi:10.7717/peerj-cs.2862.

[10] M. Douze *et al.*, "The faiss library." [Online]. Available: http://arxiv.org/abs/2401.08281, 2025.

[11] N. Fujishiro, Y. Otaki, and S. Kawachi, "Accuracy of the sentence-bert semantic search system for a japanese database of closed medical malpractice claims," *Applied Sciences*, vol. 13, no. 6, p. 4051, 2023. doi:10.3390/app13064051.

[12] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084.

[13] A. W. Anggoro, P. Corcoran, D. D. Widt, and Y. Li, "Harmonized system code classification using supervised contrastive learning with sentence bert and multiple negative ranking loss," *Data Technologies and Applications*, vol. 59, no. 2, pp. 276–301, 2024. doi:10.1108/DTA-01-2024-0052.

[14] F. Arasyi, "indo-sentence-bert: Sentence transformer for bahasa indonesia with multiple negative ranking loss." huggingface Repository. [Online]. Available: https://huggingface.co/firqaaa/indo-sentence-bert-base, 2022.

[15] T. I. Ramadhan, A. Supriatman, and T. R. Kurniawan, "Passage retrieval untuk question answering bahasa indonesia menggunakan bert dan faiss," *Jurnal Algoritma*, vol. 21, no. 2, pp. 156–163, 2024. doi:10.33364/algoritma/v.21-2.2100.

[16] X. Wang, C. Macdonald, and I. Ounis, "Improving zero-shot retrieval using dense external expansion," *Information Processing & Management*, vol. 59, no. 5, p. 103026, 2022. doi:10.1016/j.ipm.2022.103026.

[17] V. Petras, A. Lüschow, R. Ramthun, J. Stiller, C. España-Bonet, and S. Henning, "Query or document translation for academic search – what's the real difference?," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (A. Arampatzis *et al.*, eds.), pp. 28–42, Cham: Springer International Publishing, 2020.

[18] Y. S. N. E. P. and A. Musdholifah, "Pengembangan word embedding untuk domain spesifik ulasan hotel berbahasa indonesia," Master's thesis, Universitas Gajah Mada, 2020. [Online]. Available: https://etd.repository.ugm.ac.id/penelitian/detail/191233.