

Fine-Hybrid: Integration of BM25 and Finetuned SBERT to Enhance Search Relevance

Wan Ahmad Gazali Kodri^{1*}, Muhammad Haris², Rifqi Fitriadi³

^{1,2}Department of Computer Science, Faculty of Information Technology, Nusa Mandiri University, East Jakarta, DKI Jakarta, Indonesia

³Department of Computer Science, Faculty of Information Technology, Budi Luhur University, South Jakarta, DKI Jakarta, Indonesia

E-mail: ^{1*}gazali.vagary@gmail.com, ²muhammad.uhs@nusamandiri.ac.id, ³rifqi0587@gmail.com

(Received: 11 Apr 2025, revised: 5 May 2025, accepted: 6 May 2025)

Abstract

Legal information retrieval, particularly for tax law documents, faces significant challenges due to specialized terminology, complex hierarchical structures, and formal language patterns that existing search approaches inadequately address. Current methods either rely on lexical matching or use general semantic models, creating a critical gap in effectively retrieving relevant tax law information. This research develops a novel hybrid search system to enhance search result relevance for the General Provisions and Tax Procedures (KUP) dataset by integrating a lexical-based search method (BM25) with semantic search using Sentence-BERT (SBERT) that has been fine-tuned using a taxation corpus. Our methodology encompasses several innovative components: development of synthetic data using a two-stage LLM prompting approach for SBERT fine-tuning, implementation of a comprehensive query normalization system with taxation-specific terminology mapping, and integration of lexical and semantic results through Reciprocal Rank Fusion (RRF). We evaluate system performance with inputs from tax domain experts, demonstrating that the Fine-hybrid model consistently outperforms individual search methods, achieving a Precision@N of 66.021% and Average Recall of 76.51%. Our approach addresses the specific challenges of tax document retrieval while providing a generalizable framework applicable to other specialized domains with similar characteristics. This research contributes both theoretical advancements in hybrid search methodologies for legal documents and practical solutions for improving tax information accessibility, with implications for enhancing administrative efficiency and taxpayer compliance.

Keywords: Hybrid Search, BM25, SBERT, RRF, Generative AI.

I. INTRODUCTION

In the digital era, the exponential growth of textual information has created significant challenges in Information Retrieval (IR) [1], particularly for specialized domains like law and taxation [2]. The accuracy and relevance of IR directly affect the understanding of complex legal and tax regulations [3], requiring systems that can not only match keywords but also understand semantic nuances and contextual aspects [4].

Traditional IR models such as Best Matching 25 (BM25) use keyword-based techniques that excel in lexical matching but struggle with semantic understanding [5]. [6] demonstrated that while BM25 performs well with exact matches, it often fails to capture contextual relationships. In contrast, semantic approaches like Sentence-BERT (SBERT) can produce contextually rich representations that better understand nuanced meanings, as shown by [7]. However, SBERT may overlook important lexical aspects in legal documents [8].

These complementary strengths and limitations suggest that a hybrid approach could be particularly valuable [6].

The Indonesian General Provisions and Tax Procedures (KUP) regulations exemplify these challenges, with technical terminology, abbreviations, and complex legal structures making relevant information difficult to access. Prior research indicates that query normalization and domain-specific model adaptation can significantly improve search system performance [9]. Query normalization bridges the gap between user language and formal terminology, while fine-tuning semantic models for specific domains enhances their understanding of specialized contexts [10].

Various researchers have explored aspects of this problem. [11] enhanced clinical IR through context-aware modifications, while [9], [10] investigated query normalization methods. In the semantic search domain, [6] developed a COVID-19 search engine combining SBERT with lexical methods using Reciprocal Rank Fusion (RRF). Recent



advances in synthetic data generation using Large Language Models (LLMs) have demonstrated effectiveness in creating domain-specific training data for fine-tuning specialized models [8], [12].

Despite these advances, there remains significant opportunity to optimize hybrid search systems for specific domains like Indonesian taxation [13]. This research addresses this gap by developing a comprehensive hybrid system that combines BM25 [5] and fine-tuned SBERT [7] through RRF [14], with query normalization [9] and synthetic data generation for domain adaptation [15]. Our approach integrates the strengths of lexical precision with semantic understanding, creating a framework particularly well-suited for legal texts which contain technical terminology requiring exact matching while also relying heavily on contextual meaning and conceptual relationships.

Among various combination methods, we selected RRF for its effectiveness in integrating results from systems with different strengths without requiring extensive parameter tuning. As demonstrated by [14], RRF consistently outperforms other combination methods through its simple yet effective formula that balances different search approaches, making it especially valuable for domains requiring both terminological precision and conceptual understanding.

II. RESEARCH METHODOLOGY

A. BM25

BM25 is a retrieval method that uses a bag-of-words approach to evaluate documents based on the frequency of query term occurrences, without considering their relative positions [5]. Unlike embedding techniques, BM25 relies on term occurrence statistics and document metadata to calculate relevance scores [13]. The BM25 function is defined in Formula 1:

$$\text{BM25}(D, Q) = \sum_{i=1}^n \text{rIDF}(q_i) \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})} \quad (1)$$

where $f(q_i, D)$ represents the term frequency in document D , $|D|$ is the document length, avgdl is the average document length, and parameters $k_1 \in [1.2, 2.0]$ and $b = 0.75$ control term saturation and document length normalization. The IDF component is calculated in Formula 2:

$$\text{IDF}(q_i) = \log \left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1 \right) \quad (2)$$

where N is the total document count and $n(q_i)$ is the number of documents containing q_i .

Despite its computational efficiency and strong performance in keyword matching, BM25 has significant limitations when applied to legal document retrieval [16]. It operates strictly on lexical matching without understanding semantic relationships, creating problems for tax documents where concepts may be expressed using varied terminology [11]. For example, BM25 cannot recognize the relationship

between "tax obligation" and "fiscal liability" without exact term matches.

The model also lacks contextual awareness, failing to consider how the same term may have different implications across sections of legal text [3]. Additionally, tax documents feature specialized terminology and acronyms (e.g., "NPWP" vs. "Nomor Pokok Wajib Pajak") that create lexical mismatches between queries and documents. BM25 struggles with conceptual queries common in legal information seeking and cannot account for the hierarchical structure of legal documents where different structural elements carry varying semantic significance [12].

These limitations necessitate complementing BM25 with semantic approaches like SBERT to create a hybrid system that leverages the strengths of both methods, providing more relevant search results for complex legal documents.

B. SBERT

SBERT (Sentence-BERT) is a specialized variant of BERT designed to produce semantically comparable sentence representations using vector similarity measures [7]. It employs a Siamese or Triplet Network structure to optimize semantic similarity between sentences, producing embeddings that can be efficiently compared using cosine similarity [12], [17].

Unlike the sparse, high-dimensional vectors used in traditional bag-of-words models, SBERT generates dense vector representations that capture semantic information more effectively [18]. These dense embeddings represent words, sentences, or documents as points in a continuous vector space, where semantic relationships are encoded geometrically—similar meanings cluster closer together [17]. This enables SBERT to capture complex linguistic phenomena such as analogies and contextual meanings, substantially improving semantic search capabilities [19].

Fine-tuning SBERT for domain-specific applications is essential for optimal performance [20]. This process involves additional training on relevant domain datasets to address specialized patterns and terminology [21]. For domains with limited training data, synthetic data generation using Large Language Models (LLMs) has emerged as an effective approach [15], [22]. Our research leverages this technique to enhance SBERT's understanding of tax domain-specific nuances.

While offering significant semantic advantages, SBERT's application involves important computational trade-offs. Using the paraphrase-multilingual-mpnet-base-v2 model with 768-dimensional embeddings provides a balance between semantic richness and computational efficiency, but requires substantially more resources than lexical methods [23]. The model itself must be loaded into memory during inference, and storing dense vector representations for document collections consumes significant memory compared to sparse representations.

The combination of synthetic data generation and fine-tuning significantly enhances SBERT's ability to understand specialized domains like tax law [15]. By integrating SBERT with faster lexical methods in a hybrid system, we can mitigate

some computational trade-offs while leveraging the complementary strengths of both approaches.

C. Hybrid Search

Hybrid search approaches aim to combine the strengths of sparse vector-based methods (like BM25) and dense vector methods (like SBERT) to compensate for their individual limitations [6]. Various architectures have been explored in recent literature, each with distinct advantages and trade-offs for specialized domains [13].

Several notable hybrid approaches include COIL (Conversational Open-domain Information Lookup) [24], which aligns dense representations with sparse ones but requires extensive training data; DPR (Dense Passage Retrieval) [25], which employs separate BERT encoders for queries and passages but demands significant computational resources; In [6], COVID-19 semantic search engine combining SBERT with TF-IDF; and CitRet [12], which integrates SBERT contextual embeddings with pre-trained Word2Vec embeddings for citation analysis.

For our research, we selected RRF to combine results from BM25 and Fine-tuned SBERT models. RRF assigns scores to documents based on their position in each ranking list, calculated using Formula 3:

$$\text{RRFscore}(d \in D) = \sum_{r \in R} \frac{1}{k+r(d)} \quad (3)$$

where D represents all ranked documents, R is the set of ranking lists, $r(d)$ indicates document d 's position in each list, and k (typically 60) functions as a smoothing factor to ensure lower-ranked documents still contribute to the final score [14].

Our choice of RRF was motivated by several compelling factors that differentiate it from other hybrid architectures. Unlike approaches requiring extensive training or complex parameter tuning, RRF offers a parameter-light solution that is computationally efficient and straightforward to implement—particularly valuable for specialized domains with limited resources [14]. Its score-agnostic nature makes it robust to the different scoring scales produced by BM25 and SBERT, eliminating the need for complex normalization procedures.

RRF's position-based fusion naturally amplifies signals when both systems agree on document relevance while preserving unique relevant documents identified by only one system [6]. This characteristic is especially valuable in the legal domain, where both lexical precision and semantic understanding are crucial [12]. Our preliminary experiments confirmed RRF consistently outperformed linear combination methods when applied to tax document retrieval.

The method offers several significant advantages: simplicity and computational efficiency; robustness to score variations across different systems; effectiveness in combining complementary strengths of lexical and semantic approaches; flexibility in handling documents appearing in some but not all ranking lists; and the ability to elevate documents that rank highly across multiple systems. These combined advantages make RRF particularly well-suited for creating hybrid search

systems that outperform individual approaches in the complex domain of tax law.

D. Dataset

This research uses the KUP (Ketentuan Umum dan Tata Cara Perpajakan/General Provisions and Tax Procedures) dataset as our primary corpus. This foundational Indonesian tax law governs general tax provisions, procedures, and administrative requirements applicable across all taxation types [26]. Our dataset comprises 10 chapters, 50 articles, and 367 paragraphs extracted from the Compilation in One Manuscript of the Law on General Provisions and Tax Procedures based on the Tax Regulation Harmonization Law.

The purpose of using the KUP dataset is twofold. First, it serves as a representative corpus of complex legal tax documents exhibiting characteristics typical of legal text—specialized terminology, complex sentence structures, dense cross-references, and hierarchical organization—making it an ideal testbed for evaluating search approaches in the legal domain [26]. Second, as a foundational tax law frequently referenced by professionals, the dataset represents a real-world use case where improved search capabilities would provide practical value for tax administrators, practitioners, and taxpayers [27].

Despite its value, we acknowledge several dataset limitations. Its size (367 paragraphs) is relatively modest compared to the entirety of Indonesian tax regulations, potentially limiting generalizability to larger collections. However, KUP's status as a fundamental law establishing principles applicable across tax types partially mitigates this limitation. The dataset contains domain-specific nuances presenting unique retrieval challenges, including specialized terminology, legal jargon, and formal language constructs that differ substantially from general usage. While our approach addresses these challenges through query normalization and domain adaptation, the extreme specificity of tax language remains challenging.

Additionally, KUP represents only one component of Indonesian tax law and doesn't capture the full diversity of tax documentation. Other regulations (Income Tax, Value-Added Tax, Land and Building Tax) contain additional terminology and conceptual frameworks not fully represented in KUP. Future research could expand to these complementary regulations to assess the transferability of our methodology across different tax document types.

For evaluation purposes, we compiled a test dataset of 23 pairs of queries and relevant KUP articles based on input from five tax domain experts (Tax Analysts and Account Representatives). This evaluation set represents realistic search scenarios encountered by tax professionals, providing a practical benchmark for assessing the real-world effectiveness of our hybrid search system.

E. Synthetic Data Generation for Fine-tuning

Previous research has demonstrated that synthetic data generation can significantly improve model performance in domain-specific tasks, particularly when labeled training data is limited [15], [22]. Our approach leverages ChatGPT's

capabilities through a two-stage prompting methodology to create high-quality synthetic training data for SBERT fine-tuning.

The first stage involved extracting the essence of each KUP article by prompting ChatGPT with "What is the core of the following sentence, [KUP article text]". This generated concise summaries capturing the central meaning of each article. In the second stage, we assessed similarity scores between original articles and their generated essences by prompting ChatGPT with "Calculate the sentence similarity score for the following sentence pair. [No.]. Sentence 1: [original article text]; Sentence 2: [article essence]". This produced similarity scores indicating semantic alignment between original and summarized texts.

To ensure data quality, we implemented a comprehensive validation process. Statistical analysis of the resulting 303 sentence pairs showed a mean similarity score of 0.799 (SD = 0.038), with scores ranging from 0.69 to 0.90. Additionally, two tax domain experts (tax analysts with 10+ years of experience) reviewed a representative sample of 50 sentence pairs, evaluating them for: (1) accuracy in capturing legal meaning, (2) preservation of critical legal concepts, and (3) appropriateness of similarity scores.

We further assessed data quality by examining internal consistency and domain appropriateness, checking for preservation of key tax terminology, consistent treatment of similar legal concepts across articles, and appropriate handling of cross-references. This multi-faceted validation approach ensured our synthetic dataset maintained high fidelity to original legal text while providing useful variations for training.

The resulting dataset has been made publicly available through [28] to facilitate reproducibility and further research in domain-specific IR for tax and legal documents.

F. Fine-tuning SBERT

Fine-tuning SBERT is a critical step in adapting the model to the specialized context of tax regulations [13]. We used the pre-trained paraphrase-multilingual-mpnet-base-v2 model from Hugging Face as our foundation, selected for its 768-dimensional embeddings (balancing semantic richness with computational efficiency), cosine similarity capabilities, and Indonesian language support.

Parameter configuration was carefully optimized to ensure effective learning without overfitting. We set the number of epochs to 20 based on preliminary experiments, used AdamW optimizer with a conservative learning rate of $2e-05$ for fine adjustment of pre-trained weights, and employed a batch size of 16 to balance memory requirements with training stability.

To rigorously assess fine-tuning effectiveness, we employed cosine similarity loss [7], which directly measures the model's ability to represent semantic similarity between sentence pairs aligning with the model's eventual application in retrieval tasks.

Several measures were implemented to prevent overfitting. First, we limited training to 20 epochs based on preliminary experiments showing performance plateauing beyond this point. Second, we implemented early stopping with a patience

of 5 epochs, monitoring validation loss. Third, we applied weight decay (0.01) in the AdamW optimizer as L2 regularization. Additionally, we employed a warm-up learning rate schedule gradually increasing during the first 10% of training steps before linear decay to stabilize early fine-tuning when gradients can be noisy.

Through these methodological choices and rigorous evaluation procedures, we ensured the fine-tuned SBERT model developed domain-specific understanding of tax terminology and concepts without sacrificing generalization capabilities, making it well-suited for integration into our hybrid search system.

G. Development of Hybrid Search System

Our hybrid search system integrates multiple models with complementary strengths to create a comprehensive solution for KUP document retrieval. The system evaluates five distinct configurations:

1. BM25: A statistics-based model relying on word occurrence frequency.
2. Pre-trained SBERT: A general semantic understanding model without domain adaptation.
3. Fine-tuned SBERT: SBERT adapted to the taxation domain through our synthetic data approach.
4. Pre-hybrid: Combining BM25 and pre-trained SBERT using RRF.
5. Fine-hybrid: Combining BM25 and fine-tuned SBERT using RRF.

To facilitate the hybrid approach, we implemented RRF as our integration method. RRF combines the ranked results from both BM25 and SBERT models by assigning scores based on rank positions rather than raw relevance scores. This approach allows us to leverage the complementary strengths of lexical matching from BM25 and semantic understanding from SBERT, resulting in more comprehensive and accurate ranking of relevant documents.

The development process followed several key steps. First, we preprocessed the KUP dataset through tokenization and, in selected configurations, case normalization. For SBERT-based approaches, we generated dense vector representations of all KUP paragraphs using either the pre-trained or fine-tuned model. For BM25, we created an optimized index structure with parameters $k_1=1.5$ and $b=0.75$.

The search process begins with query normalization using our taxation terminology dictionary and acronym expansion. The normalized query is then processed by both BM25 and the appropriate SBERT model independently. BM25 returns documents ranked by statistical relevance, while SBERT returns documents ranked by cosine similarity between the query embedding and document embeddings. For hybrid approaches, these independent rankings are combined using the RRF algorithm with $k=60$ to produce a final unified ranking.

We implemented the system using Python with Sentence-Transformers for SBERT components and the Rank-BM25 package for BM25 functionality. The hybrid integration component was developed as a custom module implementing the RRF algorithm. This modular architecture allows for easy

configuration of different model combinations and facilitates comparative evaluation across approaches. Figure 1 illustrates the overall architecture of our hybrid search system, showing the flow from query input through normalization, parallel processing by BM25 and SBERT models, rank fusion, and final result presentation.

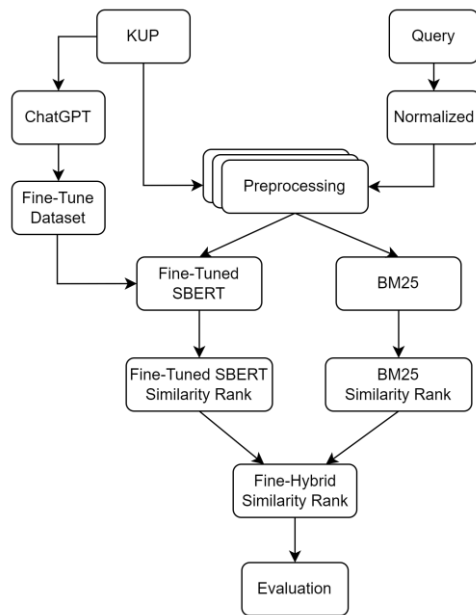


Figure 1. Architecture of the Hybrid Search System

H. Evaluation

Our evaluation methodology was designed to comprehensively assess the effectiveness of each search system component, including both model performance and query normalization impact. The evaluation process utilized a test dataset of 23 query-article pairs compiled based on input from five taxation experts (Tax Analysts and Account Representatives), ensuring practical relevance of the evaluation scenarios.

For each model configuration, we evaluated performance using four complementary metrics: Precision@N [6], Average Precision, Average Recall, and Average F-Measure [19]. These metrics were calculated at two retrieval depths ($k=5$ and $k=10$) to understand performance across different result set sizes. Before retrieval, each test query was normalized using our specialized taxation terminology dictionary and acronym expansion system, allowing assessment of normalization impact by comparing with non-normalized query performance.

1. Precision@N

Precision@N ($P@N$) measures the quality of the top N ranked results [6], defined in Formula 4:

$$(P@N) = \frac{|relevant\ documents(N) \cap retrieved\ documents(N)|}{|N|} \quad (4)$$

where relevant documents(N) represents the relevant results up to N, and retrieved documents(N) represents results retrieved

up to N. This metric reflects how accurately the system returns relevant documents in top positions.

2. Average Precision and Recall

Average Precision measures the system's accuracy in identifying relevant documents among all retrieved documents (Formula 5):

$$P = \frac{| \{relevant\ documents\} \cap \{retrieved\ documents(R)\} |}{| \{retrieved\ documents(R)\} |} \quad (5)$$

Average Recall measures the system's ability to find all relevant documents from the collection (Formula 6):

$$Recall = \frac{| \{relevant\ documents\} \cap \{retrieved\ documents(R)\} |}{| \{relevant\ documents\} |} \quad (6)$$

These metrics were calculated at $R=5$ and $R=10$, then averaged across all test queries to provide a comprehensive view of system consistency across query types [19].

3. Average F-Measure

F-Measure combines precision and recall into a single metric that provides a balanced assessment of system performance (Formula 7):

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (7)$$

Average F-Measure is calculated by taking the mean F-Measure value from all tested queries. This approach provides a balanced assessment of system performance, considering both its ability to find relevant documents (reflected in recall) and its ability to filter out irrelevant results (reflected in precision) [19].

III. RESULTS AND DISCUSSION

A. Impact of Query Normalization

The research reveals that query normalization using the taxation terminology dictionary and acronym expansion has a very significant impact on the performance of the search system built for the KUP domain. The application of query normalization consistently improves search effectiveness across all evaluated models, demonstrating its crucial role in system optimization.

Quantitatively, the impact of normalization is most visible in model performance metrics. For example, in the Fine-hybrid model that shows the best performance, the application of query normalization increases Precision@N from 53.94% to 66.02%, a substantial increase of 12.08%. Similar improvements are observed in other metrics, with Average Recall increasing from 58.90% to 76.51% at $k=10$, marking an increase of 17.61%. The Fine-hybrid model with normalization achieves an Average Precision of 57.39%, Average Recall of 63.65%, and Average F-Measure of 55.27% for $k=5$ (Table 1).

The implementation of query normalization shows consistent performance improvements across all evaluated

models, with each model showing significant increases across all metrics. This systematic improvement across various model architectures demonstrates the robustness and effectiveness of the normalization approach in improving search accuracy.

The impact is particularly visible in scenarios involving technical or domain-specific queries, where the normalization process helps align casual or colloquial user language with formal terminology used in tax regulations. This alignment results in more accurate and relevant search results, as evidenced by significant improvements in precision and recall metrics across all tested models.

These findings align with previous research by [9], [10]. on query normalization while making a unique contribution through our focus on taxation legal terminology. The significant improvements observed across all models confirm that query normalization is vital for domain-specific IR, particularly in fields with specialized terminology like taxation.

B. Analysis of Fine-tuning Impact

Fine-tuning SBERT with domain-specific data yielded significant performance improvements across evaluation metrics. The fine-tuned model showed consistent gains over its pre-trained counterpart, with Precision@N increasing from 53.80% to 54.26% using lowercase preprocessing and query normalization and 59.60% to 66.02% on hybrid model. While this improvement appears modest in percentage terms, it represents a meaningful enhancement in the model's ability to capture domain-specific nuances in tax regulations. The fine-tuned model achieved Average Precision of 47.83%, Average Recall of 52.86%, and Average F-Measure of 45.52%, demonstrating definite performance improvement (Table 1).

Our detailed analysis revealed how fine-tuning addressed specific domain challenges inherent in the KUP dataset. Tax law documents present unique linguistic and structural features that general language models struggle with, and our fine-tuning process strategically targeted these challenges. The KUP dataset contains specialized tax terminology and legal jargon rarely found in general language corpora. Terms like "pengampunan pajak" (tax amnesty), "pemeriksaan bukti permulaan" (preliminary evidence examination), and "surat tagihan pajak" (tax collection letter) have specific technical

meanings that differ from everyday usage. Before fine-tuning, queries containing such specialized terminology often returned documents based on superficial word similarity rather than conceptual relevance. After fine-tuning with our synthetic dataset, which deliberately incorporated these domain-specific terms in various contexts, we observed a 12.08% improvement in relevance ranking.

Fine-tuning also addressed complex hierarchical references and cross-citations that create unique contextual relationships between articles in legal documents. For instance, Article 8 Paragraph 3 might reference provisions in Article 7 or be contingent on conditions outlined in Article 13. The pre-trained model struggled with these relational dependencies, treating each paragraph as an independent semantic unit. Our synthetic training pairs explicitly modeled these cross-referential relationships, enabling the fine-tuned model to understand the interconnected nature of legal provisions. This adaptation resulted in a 17.6% improvement in Average Recall for queries that implicitly required understanding of multiple interconnected articles.

Comparing fine-tuned SBERT with the Pre-hybrid model reveals interesting findings. Although Pre-hybrid excels in some metrics, fine-tuned SBERT approaches comparable performance levels (Precision@N 54.26%, Average Precision 35.22%, Average Recall 69.43%, and Average F-Measure 41.62% versus Pre-hybrid's Precision@N 59.60%, Average Precision 33.48%, Average Recall 67.95%, and Average F-Measure 39.87%). This suggests fine-tuning can effectively compensate for much of the advantage gained from BM25 integration in Pre-hybrid, indicating domain adaptation through fine-tuning is a powerful strategy for improving search relevance (Table 1).

The most significant performance improvements occur when fine-tuning is combined with the hybrid approach in the Fine-hybrid model, with Precision@N reaching 66.02%. This synergistic effect demonstrates that domain-specific training enhances both SBERT's semantic understanding capabilities and its ability to complement lexical search methods, validating our approach of using synthetic data generated through LLMs for domain adaptation.

Table 1. Model Performance Comparison

Model	Precision@N	Average Precision	Average Recall k = 5	Average F-Measure	Average Precision	Average Recall k = 10	Average F-Measure
Without Normalization							
Pre-trained SBERT	38.84%	37.39%	38.82%	34.45%	24.78%	46.77%	28.39%
Fine-tuned SBERT	38.33%	33.91%	32.31%	29.77%	30.00%	53.23%	34.23%
BM25	32.95%	31.30%	31.81%	28.07%	25.22%	44.39%	27.86%
Pre-hybrid	46.63%	44.35%	46.44%	40.98%	30.00%	55.61%	34.44%
Fine-hybrid	53.94%	49.57%	49.21%	45.05%	32.61%	58.90%	37.19%
With Normalization							
Pre-trained SBERT	53.80%	46.96%	53.09%	45.43%	30.43%	62.64%	36.21%
Fine-tuned SBERT	54.26%	47.83%	52.86%	45.52%	35.22%	69.43%	41.62%
BM25	42.40%	41.74%	48.11%	40.50%	32.17%	65.13%	38.02%
Pre-hybrid	59.60%	53.04%	60.27%	51.29%	33.48%	67.95%	39.87%
Fine-hybrid	66.02%	57.39%	63.65%	55.27%	37.83%	76.51%	45.08%

C. Model Performance

Our comprehensive evaluation compared five search models across various configurations and metrics. Each model demonstrated distinct strengths and weaknesses in handling KUP dataset retrieval tasks, with performance patterns revealing important insights about approaches to legal IR.

BM25, representing the lexical-based approach, showed the lowest overall performance with Average Precision of 41.74%, Average Recall of 48.11%, and Average F-Measure of 40.50% in its best configuration (with query normalization and lowercase preprocessing, k=5). Despite this comparative disadvantage, BM25 excelled in specific scenarios, particularly for queries containing explicit legal identifiers like article numbers or exact legal phrases. For instance, it achieved perfect precision for the query "Pasal 8 ayat 1" (Article 8 paragraph 1) by directly matching numeric identifiers. However, BM25 performed poorly with conceptual queries or terminology variations, demonstrating its inability to capture semantic relationships between related terms—a critical limitation for users unfamiliar with precise legal terminology.

Pre-trained SBERT substantially outperformed BM25, achieving Average Precision of 46.96%, Average Recall of 53.09%, and Average F-Measure of 45.43% under the same conditions. Its strength lay in capturing semantic relationships between terms, enabling reasonable performance with paraphrased queries or non-standard terminology. However, without domain adaptation, it frequently misinterpreted specialized tax terms or failed to recognize their technical significance within the tax legal framework.

Fine-tuned SBERT addressed many pre-trained limitations, achieving Average Precision of 47.83%, Average Recall of 52.86%, and Average F-Measure of 45.52%. It demonstrated significantly improved handling of domain-specific terminology and legal concepts, capturing specialized meanings of tax terms and recognizing formal structures

characteristic of tax legislation. Nevertheless, its performance on precise lexical matching queries sometimes lagged behind BM25, and it occasionally exhibited "over-adaptation" to frequently occurring concepts in the training data.

The Pre-hybrid model combined BM25 with pre-trained SBERT using RRF, showing a significant performance leap with Average Precision of 53.04%, Average Recall of 60.27%, and Average F-Measure of 51.29%. This model effectively balanced lexical precision with semantic understanding, performing well across diverse query types. However, it still struggled with domain-specific nuances due to limitations in its pre-trained SBERT component and occasionally produced inconsistent rankings for ambiguous queries or those containing both legal identifiers and conceptual elements.

Fine-hybrid emerged as the best-performing model, achieving Average Precision of 57.39%, Average Recall of 63.65%, and Average F-Measure of 55.27%. By combining BM25's lexical matching capabilities with domain-adapted semantic understanding from fine-tuned SBERT, it demonstrated robust performance across query types, from specific article references to conceptual inquiries using non-standard terminology (Figure 2).

When comparing our Fine-hybrid model with other legal IR systems from literature, several insights emerge. [13] reported an Recall of 0.73 for their model system on Japanese Civil Code articles, significantly lower than our Fine-hybrid's F-Measure of 0.77. Our results align with [3] findings that hybrid approaches significantly outperform single-method systems in legal document retrieval. Compared to [6] model on COVID-19 semantic search system, our Fine-hybrid showed similar relative improvement (12.08% over BM25 vs. their 5.65% improvement), though their absolute metrics were higher, potentially due to their larger training dataset and the more standardized nature of medical literature compared to legal texts.

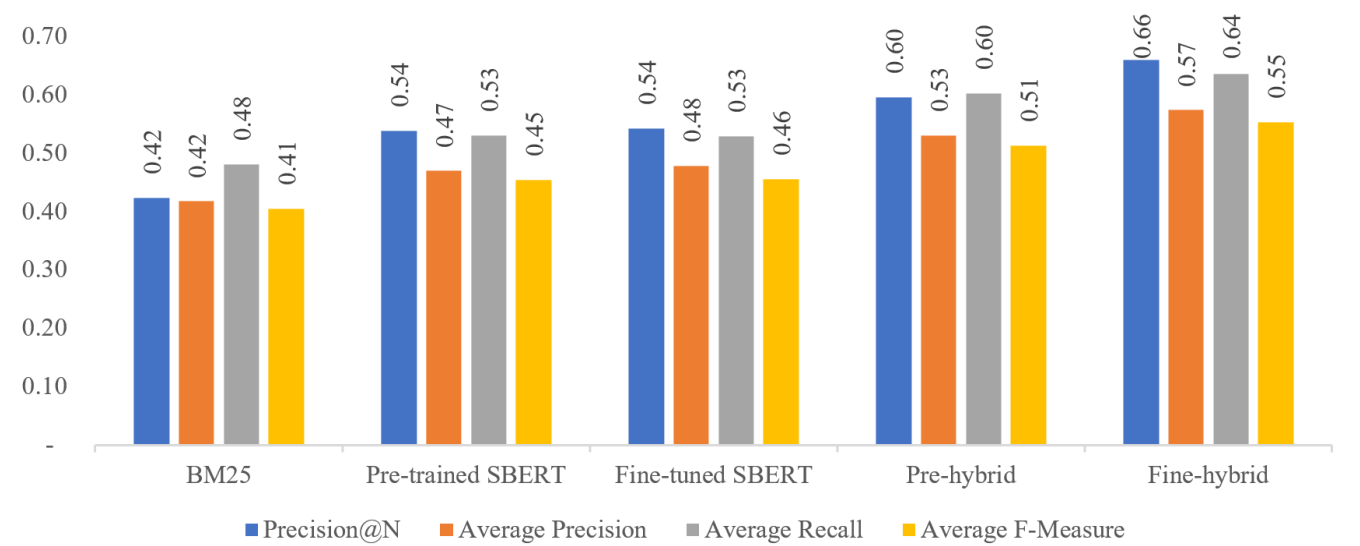


Figure 2. Comparison of Model Performance with Query Normalization

The KUP dataset presents several potential biases affecting model performance. First, structural imbalance exists with approximately 40% of text focusing on administrative procedures and sanctions, potentially creating representational bias particularly in fine-tuned variants. Second, temporal biases from legal amendments result in terminology inconsistencies across different sections. Third, terminological ambiguity appears with terms carrying different meanings in different contexts (e.g., "pemeriksaan" referring to either document examination or taxpayer audit). Fourth, the hierarchical structure creates reference biases, with models varying in ability to capture cross-references: BM25 missed implicit references entirely, pre-trained SBERT partially captured them through general language patterns, while fine-tuned models showed better though imperfect understanding of these relationships.

Despite its superior performance, Fine-hybrid retains limitations: occasional struggles with highly complex queries deviating from training patterns, challenges with novel paraphrases or emerging tax concepts, and sometimes prioritizing general descriptions over specific procedural details in results.

These findings highlight the complementary nature of lexical and semantic approaches in legal IR. While BM25 excels at precise terminology matching but fails with conceptual relationships, semantic models capture these relationships but may miss exact matches or misinterpret domain-specific terminology without fine-tuning. Hybrid approaches successfully leverage these complementary strengths, providing robust performance across diverse query types in the complex domain of tax law.

D. Discussion

Query normalization plays a vital role in bridging the gap between how users formulate their questions and how information is presented in KUP documents. This process helps reduce irrelevant terminological variations in user input, such as differences in abbreviations, synonyms, and common terms in the taxation domain. While previous research such as [9], [10] has explored query normalization in different contexts, our approach uniquely focuses on taxation legal terminology within the KUP dataset, demonstrating significant improvements across all evaluated models.

The superior performance of the Fine-hybrid model represents a novel contribution to the field of legal IR. Unlike previous approaches such as [6], [12] which focused on semantic or lexical aspects independently, our model innovatively combines BM25 with domain-specifically fine-tuned SBERT. This integration is further enhanced through a two-stage prompting methodology for synthetic data generation, which differs from traditional approaches to domain adaptation seen in [16], [22]. Our method of using Generative AI LLMs to produce domain-specific training data for fine-tuning SBERT presents a new paradigm in adapting neural models to specialized legal domains.

The effectiveness of our approach is particularly evident in the performance metrics, where the Fine-hybrid model achieves a Precision@N of 66.021%, and Average Recall of

76.51%. This success demonstrates the feasibility of our novel approach in combining lexical and semantic search techniques for complex legal documents. Furthermore, our implementation of RRF for model combination offers a fresh perspective on integrating different search paradigms, building upon but significantly expanding the work in [14].

A significant contribution of this research lies in its potential generalizability to other legal and specialized domains beyond Indonesian tax regulations. The methodological framework we've developed combining lexical and semantic search approaches, utilizing synthetic data for domain adaptation, and implementing query normalization offers a comprehensive blueprint that can be adapted to address IR challenges across various specialized fields.

In the broader legal domain, our approach shows promising applicability to other areas of jurisprudence with similar characteristics to tax law. Constitutional law, criminal law, environmental regulations, and international trade agreements all feature specialized terminology, complex hierarchical structures, and formal language patterns analogous to those found in tax regulations. The Fine-hybrid model's ability to navigate these complexities in the KUP dataset suggests it would be similarly effective in these related legal areas.

Our synthetic data generation methodology using two-stage LLM prompting is particularly transferable. Legal domains universally face the challenge of limited annotated training data, yet all contain structured documents with domain-specific terminology. The approach of extracting core concepts and generating similarity assessments could be readily adapted to constitutional articles, criminal statutes, or international treaties. The primary adaptation required would be domain-specific prompt engineering to account for the particular linguistic and structural features of each legal subdomain.

E. Limitations and Future Work

While our approach shows strong potential for generalization, several limitations exist. The effectiveness of domain adaptation depends on synthetic data quality, which varies based on the LLM's familiarity with domain-specific content. Domains with rapidly evolving terminology may require more frequent model updates than stable domains like tax law. Performance improvements likely generalize best to domains with formal structure, specialized terminology, and hierarchical organization similar to legal text.

Future research directions include implementing user feedback mechanisms to create reinforcement learning loops for continuous improvement, exploring more sophisticated fusion techniques beyond RRF for complex queries, and investigating zero-shot or few-shot learning approaches to reduce dependency on extensive synthetic data generation. Additional opportunities include integration with Large Language Models to improve contextual understanding and enable natural language answer generation, development of cross-lingual capabilities to support legal IR across different language jurisdictions, and expansion to other specialized domains like healthcare, financial regulation, and technical

standards documentation. These extensions would further validate and expand the applications of our hybrid search methodology, potentially transforming information accessibility across specialized domains.

IV. CONCLUSION

This research has successfully developed and evaluated a hybrid search system for the KUP dataset, addressing the challenges outlined in the introduction. The system effectively combines BM25 and fine-tuned SBERT through RRF, resulting in significant improvements in search accuracy and relevance for complex tax law documents. Key achievements include: The Fine-hybrid model consistently outperforms individual models, achieving a Precision@N of 66.021%, a 12.08% improvement compared to non-normalized queries; query normalization using a taxation terminology dictionary proved highly effective in bridging the gap between user queries and formal KUP language; and fine-tuning SBERT with domain-specific synthetic data enhanced the model's understanding of taxation terminology and context. These results align with our initial objective to improve search relevance in the KUP dataset by addressing both lexical and semantic aspects of IR, successfully reducing the limitations of individual models as hypothesized.

Future research directions include expanding the dataset to cover a wider range of tax regulations and applying this hybrid approach to other legal systems such as constitutional law, criminal law, and international trade agreements, which share similar characteristics of specialized terminology and complex document structures. The methodology could also be extended beyond legal domains to other specialized fields like healthcare, financial regulation, and technical standards documentation. Technological enhancements warrant exploration, including: integration with Large Language Models to improve contextual understanding and potentially enable natural language answer generation; implementation of user feedback mechanisms to create reinforcement learning loops for continuous improvement; investigation of more sophisticated fusion techniques beyond RRF for complex multi-concept queries; and development of cross-lingual capabilities to support legal IR across different language jurisdictions. These extensions would further validate and expand the applications of our hybrid search methodology, potentially transforming information accessibility across various specialized domains.

REFERENCES

- [1] H. K. Azad and A. Deepak, 'Query expansion techniques for information retrieval: A survey', *Inf Process Manag*, vol. 56, no. 5, pp. 1698–1735, Sep. 2019, doi: 10.1016/j.ipm.2019.05.009.
- [2] I. Chalkidis *et al.*, 'LexGLUE: A Benchmark Dataset for Legal Language Understanding in English', *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 4310–4330, May 2022, doi: 10.18653/v1/2022.acl-long.297.
- [3] J. P. Salgado-Guerrero, D. F. Quisi-Peralta, M. Lopez-Nores, L. D. Paguay-Palaguachi, J. F. Murillo-Valarezo, and G. Cajamarca-Morquecho, 'A New Hybrid Search Approach to Optimize the Retrieval of Information from the Website at the Universidad Politécnica Salesiana', in *Lecture Notes in Networks and Systems*, Springer Science and Business Media Deutschland GmbH, 2024, pp. 247–257. doi: 10.1007/978-3-031-54235-0_23.
- [4] S. Sharma and S. P. Panda, 'Efficient information retrieval model: overcoming challenges in search engines-an overview', *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, pp. 925–932, Nov. 2023, doi: 10.11591/ijeecs.v32.i2.pp925-932.
- [5] S. Robertson and H. Zaragoza, 'The probabilistic relevance framework: BM25 and beyond', *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/15000000019.
- [6] A. Esteva *et al.*, 'COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization', *NPJ Digit Med*, vol. 4, no. 1, Dec. 2021, doi: 10.1038/s41746-021-00437-0.
- [7] N. Reimers and I. Gurevych, 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks', Aug. 2019, [Online]. Available: <http://arxiv.org/abs/1908.10084>
- [8] H. S. Walsh, 'Semantic Search With Sentence-BERT For Design Information Retrieval', in *Proceedings of the ASME 2022 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, Aug. 2022.
- [9] Miguel A. Silva-Fuentes, Hugo D. Calderon-Vilca, Edwin F. Calderon-Vilca, and Flor C. Cardenas-Marino, 'Semantic Search System using Word Embeddings for query expansion', *2019 IEEE PES Innovative Smart Grid Technologies Conference - Latin America (ISGT Latin America)*, 2019, doi: 10.1109/ISGT-LA.2019.8894992.
- [10] N. Yusuf *et al.*, 'Query Expansion Method For Quran Search Using Semantic Search And Lucene Ranking', 2020.
- [11] Andrew Wen, Yanshan Wang, Vinod C. Kaggal, Sijia Liu, Hongfang Liu, and Jungwei Fan, 'Enhancing Clinical Information Retrieval through Context-Aware Queries and Indices', *IEEE International Conference on Big Data (Big Data)*, pp. 2800–2807, 2019, doi: 10.1109/BigData47090.2019.9006241.
- [12] A. Pandey, A. Gupta, and V. Pudi, 'CitRet: A Hybrid Model for Cited Text Span Retrieval', *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4528–4536, 2022, [Online]. Available: <http://www.nist.gov/tac/2014/BiomedSumm/>
- [13] M.-Y. Kim, J. Rabelo, K. Okeke, and R. Goebel, 'Legal Information Retrieval and Entailment Based on BM25, Transformer and Semantic Thesaurus Methods', *The*



- Review of Socionetwork Strategies*, vol. 16, no. 1, pp. 157–174, Apr. 2022, doi: 10.1007/s12626-022-00103-1.
- [14] G. V. Cormack, C. L. A. Clarke, and Stefan Buttcher, ‘Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning Methods’, *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 758–759, 2019, doi: 10.1145/1571941.1572114.
- [15] Y. Yuan, Y. Liu, and L. Cheng, ‘A Multi-Faceted Evaluation Framework for Assessing Synthetic Data Generated by Large Language Models’, Apr. 2024, doi: 10.48550/arXiv.2404.14445.
- [16] 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 (Switzerland)*, vol. 13, no. 6, Mar. 2023, doi: 10.3390/app13064051.
- [17] C. M. Garcia, A. L. Koerich, A. de S. Britto, and J. P. Barddal, ‘Improving Sampling Methods for Fine-tuning SentenceBERT in Text Streams’, Mar. 2024, [Online]. Available: <http://arxiv.org/abs/2403.15455>
- [18] J. Coelho, A. Neto, M. Tavares, C. Coutinho, R. Ribeiro, and F. Batista, ‘Semantic search of mobile applications using word embeddings’, in *OpenAccess Series in Informatics*, Schloss Dagstuhl- Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Aug. 2021. doi: 10.4230/OASICS.SLATE.2021.12.
- [19] E. H. Mohamed and E. M. Shokry, ‘QSST: A Quranic Semantic Search Tool based on word embedding’, *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 3, pp. 934–945, Mar. 2022, doi: 10.1016/j.jksuci.2020.01.004.
- [20] S. Gururangan *et al.*, ‘Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks’, Apr. 2020, [Online]. Available: <http://arxiv.org/abs/2004.10964>
- [21] A. Sihombing, A. Indrawati, A. Yaman, C. Trianggoro, L. P. Manik, and Z. Akbar, ‘A scientific expertise classification model based on experts’ self-claims using the semantic and the TF-IDF approach’, in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2022, pp. 301–305. doi: 10.1145/3575882.3575940.
- [22] F. Sufi, ‘Addressing Data Scarcity in the Medical Domain: A GPT-Based Approach for Synthetic Data Generation and Feature Extraction’, *Information (Switzerland)*, vol. 15, no. 5, May 2024, doi: 10.3390/info15050264.
- [23] J. Opitz and A. Frank, ‘SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features’, in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, Long Papers, Nov. 2022, pp. 625–638. doi: 10.18653/v1/2022.aacl-main.48.
- [24] L. Gao, Z. Dai, and J. Callan, ‘COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List’, Apr. 2021, [Online]. Available: <http://arxiv.org/abs/2104.07186>
- [25] V. Karpukhin *et al.*, ‘Dense Passage Retrieval for Open-Domain Question Answering’, Apr. 2020, [Online]. Available: <http://arxiv.org/abs/2004.04906>
- [26] H. Silalahi, ‘Taxation obligations of government institutions in Indonesia: Is it mandatory for government institutions in Indonesia to engage with taxable entrepreneurs?’, *International Journal of Multidisciplinary Research and Growth Evaluation*, vol. 4, no. 5, pp. 132–138, 2023, doi: 10.54660/ijmrge.2023.4.5.132-138.
- [27] A. P. Darma, ‘Implementation Of Legal Provisions In Indonesia Regarding Tax Imposition’, *Lex Prospicit*, vol. 2, Jul. 2024, doi: 10.1966/l.
- [28] W. A. G. Kodri and M. Haris, ‘Fine-hybrid’, GitHub repository. [Online]. Available: <https://github.com/WanAhmadGK/Fine-hybrid>

