HYBRID INDEXING OPTIMIZATION OF BEST MATCHING 25 AND DENSE PASSAGE RETRIEVAL FOR EFFICIENT RETRIEVAL IN OPEN-DOMAIN QA

Raihan Ramadhan[1], Galih Hermawan[2]

[1,2] Universitas Komputer Indonesia
Jl. Dipatiukur No. 112-116 Bandung
E-mail : raihan.10121141@mahasiswa.unikom.ac.id[1], galih.hermawan@email.unikom.ac.id[2]

Abstrak

Sistem *Open-Domain Question Answering* (QA) modern mengandalkan dua paradigma *retrieval* utama yaitu *sparse retrieval* berbasis leksikal seperti BM25 dan *dense retrieval* seperti *Dense Passage Retrieval* (DPR) yang superior dalam pemahaman semantik untuk mengatasi *vocabulary mismatch*. Mengingat kedua pendekatan memiliki kekuatan yang saling melengkapi, penelitian ini merancang dan mengevaluasi sebuah model *retrieval hybrid* yang bertujuan meningkatkan efektivitas secara sistematis. Arsitektur yang diusulkan adalah model *score fusion*, di mana BM25 dan DPR bekerja secara independen untuk menghasilkan skor relevansi yang kemudian digabungkan melalui formula pembobotan. Untuk menguji hipotesis ini, serangkaian pengujian dilakukan pada *subset* dataset AC-IQuAD, yang terdiri dari korpus 50.000 dokumen dan 200 kueri berbahasa Indonesia. Hasil eksperimen menunjukkan bahwa pendekatan *hybrid* yang dioptimalkan dengan *score fusion* (α=0.5) mencapai efektivitas yang jauh lebih unggul, dengan nilai *Mean Reciprocal Rank* (MRR) sebesar 0.8095. Performa ini secara signifikan melampaui sistem DPR-*only* (MRR 0.2945) dan sistem *hybrid baseline* tanpa *score fusion* yang optimal (MRR 0.4480). Penelitian ini menyimpulkan bahwa pendekatan *hybrid* dengan *score fusion* menawarkan solusi yang lebih *robust*, seimbang, dan efektif secara signifikan untuk tugas *retrieval* dalam *Open-Domain* QA dengan menyeimbangkan sinyal leksikal dan semantik.

Kata kunci : *Open-Domain* QA, *Hybrid Retrieval, Score Fusion, Best Matching* 25, *Dense Passage Retrieval*

Abstract

Modern Open-Domain Question Answering (QA) systems rely on two main retrieval paradigms: lexical-based sparse retrieval, such as BM25, and semantic-based dense retrieval, like Dense Passage Retrieval (DPR), which is superior in semantic understanding for overcoming vocabulary mismatch. Given that both approaches have complementary strengths, this research designs and evaluates a hybrid retrieval model aimed at systematically improving effectiveness. The proposed architecture is a score fusion model, wherein BM25 and DPR operate independently to generate relevance scores that are subsequently combined through a weighted formula. To test this hypothesis, a series of experiments were conducted on a subset of the AC-IQuAD dataset, which consists of a corpus of 50,000 documents and 200 Indonesian queries. Experimental results demonstrate that the hybrid approach, optimized with score fusion (α=0.5), achieves far superior effectiveness, recording a Mean Reciprocal Rank (MRR) of 0.8095. This performance significantly surpasses the DPR-only system (MRR 0.2945) and a baseline hybrid system without optimal score fusion (MRR 0.4480). This research concludes that the hybrid approach with score fusion offers a more robust, balanced, and significantly effective solution for the retrieval task in Open-Domain QA by balancing both lexical and semantic signals.

Keywords : *Open-Domain* QA, *Hybrid Retrieval, Score Fusion, Best Matching* 25, *Dense Passage Retrieval*

1. INTRODUCTION

Open-Domain Question Answering (QA) systems represent a research area focused on providing precise answers to user questions from large-scale, unstructured document collections [1]. The success of these systems is highly dependent on the retrieval stage, where documents potentially containing the answer

must be found efficiently and effectively from millions of candidates [2]. The retrieval paradigm has evolved from sparse to dense methods. Traditional sparse retrieval methods, exemplified by Okapi BM25, operate based on lexical matching with a bag-of-words representation [3]. While highly efficient and a strong baseline, BM25 has a fundamental limitation known as the vocabulary mismatch problem, where the system fails to recognize semantic relevance between different terms [4]. This challenge of understanding natural language queries is central to many Natural Language Processing (NLP) tasks [5]. To address this, dense retrieval emerged, utilizing Transformer models such as BERT [6] to map text into a semantic vector space [7]. Dense Passage Retrieval (DPR) is a fundamental implementation that employs a dual-encoder architecture to capture the semantic relevance between queries and documents [2]. DPR has been shown to be significantly more effective than BM25 in various scenarios [2], [8].

Despite its superior accuracy, dense retrieval faces significant computational challenges [2], [9]. On the other hand, studies indicate that sparse and dense retrieval are complementary; they often retrieve different relevant documents for the same query [8], [10]. This opens up opportunities for hybrid retrieval approaches that combine the strengths of both. Previous research has demonstrated that a fusion of BM25 and DPR can potentially enhance retrieval effectiveness considerably [8], [11]. Based on this potential, this study aims to implement and evaluate a hybrid model based on score fusion for the Open-Domain QA task on an Indonesian-language corpus.

2. METHODOLOGY

This research employs an experimental quantitative approach to design, implement, and evaluate the model's performance using standard information retrieval metrics on a specific dataset.

2.1 Dataset and Pre-processing

This study utilizes the AC-IQuAD dataset, an Indonesian-language QA dataset automatically constructed from Wikipedia and Wikidata [12], [13]. From this dataset, a test subset was created, comprising a corpus of 50,000 documents and 200 queries. The statistical details of the corpus are presented in Table 1. Prior to the indexing process, all document and query texts underwent a pre-processing stage that included lowercasing and standard tokenization.

Table 1. Test Corpus Statistics

Property	Value
Number of Documents	50.000
Average Document Length (words)	128
Number of Test Queries	200

2.2 General Model Analysis

The architecture employed is a parallel score fusion model, wherein BM25 and DPR operate independently to generate their respective relevance scores. These two sets of scores are subsequently combined to produce a final ranking, as illustrated in Figure 1.

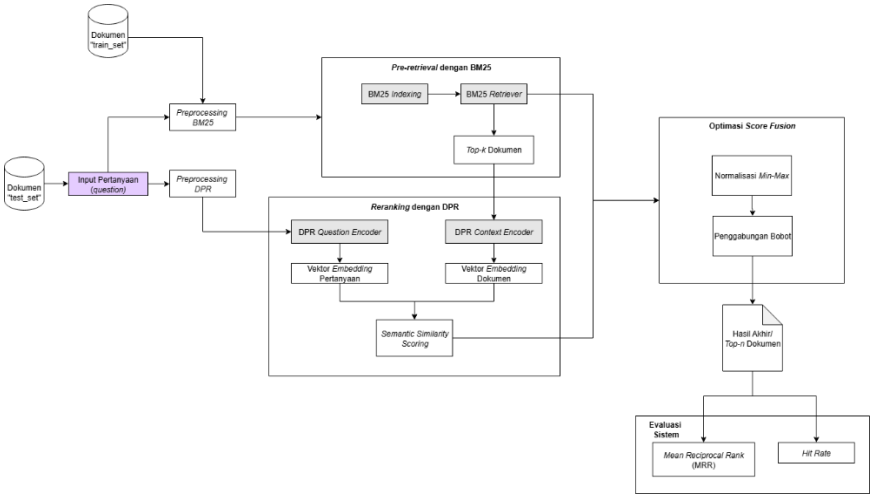


Figure 1. Model Overview

The first component is a sparse retriever that utilizes a standard Okapi BM25 implementation to generate a lexical relevance score. The score calculation is based on formula (1) [14].

$$Score_{BM25}(Q,D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{TF(q_i,D)\cdot(K_1+1)}{TF(q_i,D)+ k_1 \cdot (1-b+b \cdot \frac{|D|}{avgdl})} \quad (1)$$

Where:

IDF(q_i) : This represents the Inverse Document Frequency score of the i-th query token, which measures the informativeness of a word across the entire corpus.

TF(q_i,D) : This is the frequency of the i-th query token within document D.

|D| : This is the length of document D, measured by the number of words.

avgdl : This represents the average document length across the entire corpus.

k_1 dan b : These are the hyperparameters for BM25. The parameter k_1 controls term frequency saturation, while b controls the normalization of document length.

The second component is a dense retriever based on Dense Passage Retrieval (DPR), which utilizes a dual-encoder architecture to generate vector embeddings for queries and documents. The semantic relevance score is calculated using the dot product operation between these two vectors, as shown in formula (2).

$$Score_{DPR}(Q,D) = E_Q(Q) \cdot E_D(D)^T \quad (2)$$

Where:

E_Q(Q) : This is the vector representation (embedding) of the query Q.

E_D(D) : This is the vector representation (embedding) of document D.

In the final stage, the scores from BM25 and DPR, which have been normalized to the [0, 1] range, are combined via score fusion with a weighting formula, as shown in formula (3).

$$Skor_{Hybrid} = (1-\alpha) \cdot Skor_{BM25_{norm}} + (\alpha \cdot Skor_{DPR_{norm}}) \quad (3)$$

Here, α is a hyperparameter that determines the weight or contribution level of the dense retriever (DPR) score to the final score. To provide a clearer illustration of the optimization mechanism through score fusion, Figure 2 shows the process of score normalization and combination from both BM25 and DPR to obtain the final relevance score.

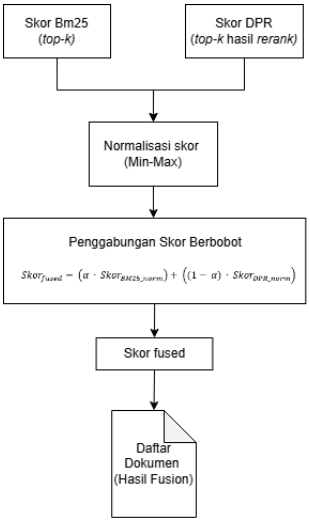


Figure 2. Score Fusion Diagram

The hybrid approach, optimized through score fusion, aims to integrate the strengths of both retrievers, resulting in an information retrieval system that is more efficient than pure DPR in terms of indexing, while also delivering high ranking accuracy for Indonesian-language Open-domain QA.

3. RESULTS AND DISCUSSION

This section presents the results of the performance evaluation conducted on the hybrid architecture model. The testing process was executed systematically using 56 previously validated test queries. The evaluation focused on three main scenarios to compare the effectiveness of each approach

1. Baseline (DPR-only)
   In this scenario, the system exclusively utilizes dense retrieval as an initial performance baseline. All 50,000 documents in the corpus were converted into embeddings and indexed using FAISS. The search process was performed directly on this index to retrieve the top 100 documents based on the semantic similarity score from DPR.
2. Baseline Hybrid System
   A retrieve-then-rerank approach is used as the foundation for the hybrid architecture prior to optimization. The system first employs BM25 to select 100 candidate documents, which are subsequently re-ranked based solely on the reranking scores from DPR, without involving a score fusion mechanism.
3. Hybrid System with Score Fusion
   This final scenario evaluates the optimized hybrid system, which combines both lexical and semantic signals. The final ranking of the 100 candidate documents is determined by the fused score from BM25 and DPR, using an optimally tuned weight, α.

For all three scenarios, system effectiveness will be evaluated using the Mean Reciprocal Rank (MRR) metric, which measures how quickly the system finds the first relevant document, and Top-k Accuracy, which measures the proportion of queries for which the correct answer is found within the top k documents [2], [15].

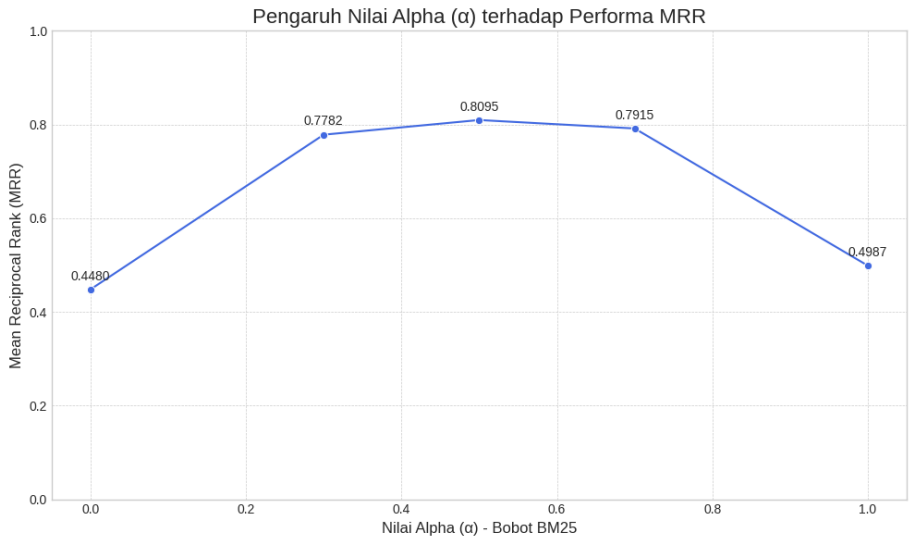
3.1 Effectiveness Analysis of BM25 and Score Fusion Optimization

The initial stage of the analysis focused on validating the effectiveness of BM25 as the pre-retrieval component. Testing on the 56 valid queries showed that the Hit Rate@100 for BM25 reached 100%. This result indicates that for each valid query, BM25 always successfully included at least one correct document within the 100 candidate documents. After the pre-retrieval validation was confirmed, the next step was to determine the most optimal value for the α parameter in the score fusion mechanism. Detailed results of the experiment with varying α values are presented in Table 2 and Figure 3.

Table 2. Experimental Results of Parameter α Optimization on 56 Valid Queries

α Value	Contribution (BM25:DPR)	MRR	Hit Rate@100
0.0	0% : 100% (Hybrid Baseline)	0.4480	100%
0.3	30% : 70%	0.7782	100%
0.5	50% : 50%	0.8095	100%
0.7	70% : 30%	0.7915	100%
1.0	100% : 0% (BM25-only)	0.4987	100%

Figure 3. Effect of α Value on MRR Performance



Referring to Table 2 and the illustration in Figure 3, the best performance was achieved when α = 0.5, with an MRR value reaching 0.8095. This result confirms that a balanced contribution between BM25 and DPR (50% : 50%) yields the most optimal performance. This can be explained by the fact that both methods capture different dimensions of relevance. BM25 excels at detecting direct keyword matches (lexical signals), ensuring that documents containing important query terms are included in the candidates. Conversely, DPR is superior in understanding context and meaning (semantic signals), enabling it to find relevant documents even without explicit keyword similarity. When their scores are combined with an equal weight (α = 0.5), the weaknesses of each method can complement one another. BM25 ensures the completeness of relevant documents, while DPR reorders the ranking based on a deeper semantic understanding, thus producing a more accurate final result list.

3.2 Quantitative Analysis of Model Performance

The performance of the Hybrid model (with α=0.5) was compared against the performance of the BM25-only and DPR-only models. The comparison results are presented in Table 3 and in the graph in Figure 4.

Table 3. Final Performance Comparison of Retrieval Models

Testing Scenario	MRR	Hit Rate@100
Baseline (DPR-only with FAISS)	0.2945	71.43%
Baseline Hybrid System	0.4480	100%
Optimal Hybrid System	0.8095	100%

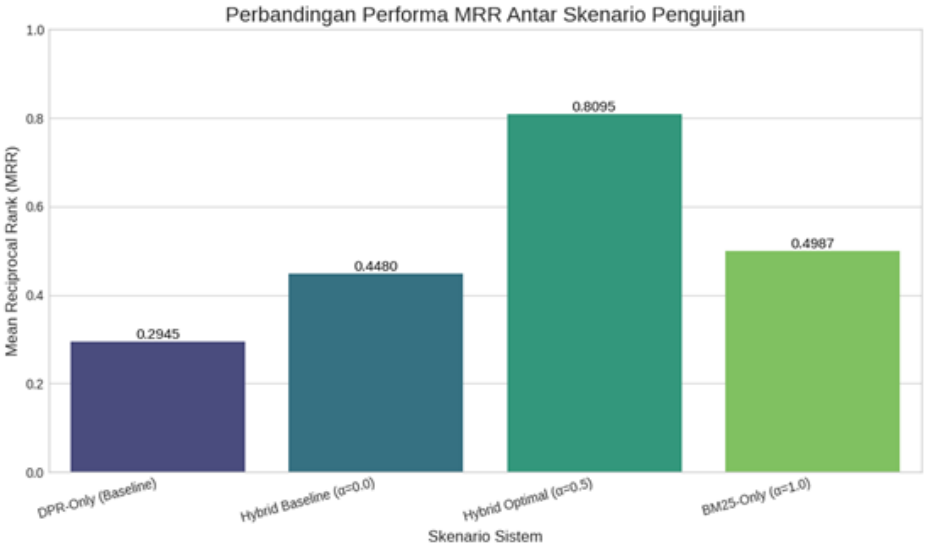


Figure 4. Comparison of MRR Scores Across Models

The results in Table 3, visualized in Figure 4, demonstrate the superiority of the proposed Optimal Hybrid System. A significant increase of 80.7% in MRR is evident, rising from 0.4480 in the baseline to 0.8095 in the optimal system. This finding proves that score fusion optimization is highly effective. The difference lies in the approach used: the Baseline Hybrid System (α=0.0) completely disregards the lexical signals from BM25 in the final stage, relying solely on DPR scores for reranking. In contrast, the Optimal Hybrid System re-leverages the strength of BM25 through fusion, allowing documents that are lexically strong but semantically weaker (or vice versa) to be repositioned, resulting in a more accurate final ranking [8], [16].

Furthermore, the Optimal Hybrid System also consistently outperforms the DPR-only baseline by a large performance margin (MRR of 0.8095 compared to 0.2945). Moreover, the Hit Rate for the DPR-only system reached only 71.43%, meaning that for nearly 30% of the valid queries, it failed to include the correct document within the Top-100. This is likely due to the pre-trained model used, which, despite its strengths in natural language inference (NLI), was not specifically trained for the retrieval task. Consequently, its pure semantic recall capability on a large corpus still lags behind the lexical recall of BM25, which remains simpler yet effective for this dataset.

3.3 Qualitative Analysis

To understand the behaviors and characteristics underlying the quantitative results, a qualitative analysis was conducted by examining the performance of each retrieval model at the individual query level. This approach is used to study specific examples to provide a more comprehensive understanding of the strengths, limitations, and the interaction between sparse (BM25) and dense (DPR) methods. Two case studies were selected as they are considered representative of scenarios where each approach displays its distinct characteristics.

The first case study highlights the model's ability to handle queries containing specific entity names, where lexical matching plays a crucial role. The comparison results for the query “di kota manakah letak stadion anoeta” (in which city is the anoeta stadium located) are presented in Table 4.

Table 4. Comparison of Top-Ranked Results from Each Model for the First Case Study

Rank	Model	Document Text (Excerpt)	Score
1	BM25	Real Sociedad adalah sebuah klub sepak bola Spanyol yang berbasis di kota San Sebastián... Bermarkas di Stadion Anoeta...	23.3289
1	DPR	Anoeta adalah sebuah kotamadya di provinsi Gipuzkoa, komunitas otonom Ülke Baskom, Spanyol.	21.0938

Rank	Model	Document Text (Excerpt)	Score
1	*Hybrid*	Real Sociedad adalah sebuah klub sepak bola Spanyol yang berbasis di kota San Sebastián... Bermarkas di Stadion Anoeta...	24.3289

In the case study presented in Table 4, BM25 demonstrates strong performance. The model successfully places the most relevant document at the top rank with a high score of 23.3289. This success stems from its ability to capture the clear lexical signal in the name phrase “Anoeta Stadium.” For BM25, the occurrence of identical keywords serves as a very strong indicator of relevance. Conversely, DPR fails to provide an accurate interpretation. Although the query explicitly asks about a “stadium,” DPR places a document describing “Anoeta” as a municipality in Spain at the top, with a score of 21.0938. This situation illustrates a fundamental weakness of dense retrieval, which operates in a semantic vector space. DPR recognizes the semantic proximity between “Anoeta” as a place name and the query's mention of “city,” but it fails to capture the critical contextual constraint related to “stadium” in the query.

The Hybrid Model emerges as a balancing solution by combining the scores from both approaches. Through the fusion mechanism, the system successfully places the correct document at the top position with a score of 24.3289. This model effectively leverages the lexical signal from BM25 while correcting the DPR result, which appeared semantically relevant but was factually incorrect. This case underscores the crucial role of the Hybrid Model in prioritizing the strongest evidence, whether lexical or semantic.

The second case study is designed to assess the model's capability in handling queries that demand contextual understanding, where the relevant information is not located within a single, contiguous phrase or sentence. The comparison results for the query “siapa penulis dari buku an author’s life” (who is the author of the book an author’s life) are presented in Table 5.

Table 5. Comparison of Top-Ranked Results from Each Model for the Second Case Study

Rank	Model	Document Text (Excerpt)	Score
1	BM25	...Dazai kemudian menulis sebuah buku berjudul "Tsugaru", yang mendeskripsikan perjalanannya...	16.5160
1	DPR	Osamu Dazai adalah seorang penulis Jepang yang dianggap sebagai salah satu tokoh sastra fiksi terkemuka di Jepang abad ke-20.	22.8281
1	*Hybrid*	Osamu Dazai adalah seorang penulis Jepang yang dianggap sebagai salah satu tokoh sastra fiksi terkemuka di Jepang abad ke-20.	23.8281

In the scenario presented in Table 5, BM25 demonstrates its limitations. The model only succeeded in finding documents containing the keywords “buku” and “penulis” (Dazai), but in an incorrect context as they referred to the book *Tsugaru*. With a score of 16.5160, this result confirms the failure of BM25 to bridge the lexical gap and to understand that the information sought was the author's identity, not merely another document that mentioned their name. Conversely, DPR obtained the highest score of 22.8281 and was able to identify the most semantically relevant document. This model could associate the concept of “siapa penulis” in the query with the statement “Osamu Dazai adalah seorang penulis” contained in the document. The ability to connect an entity with its role description, even without a direct match on the book title, constitutes the primary strength of dense retrieval.

The Hybrid Model again demonstrated its effectiveness in balancing the strengths of both approaches. With a final score of 23.8281, the system correctly emphasized the strong semantic signal from DPR while mitigating the influence of the weak and misleading lexical signal from BM25. This case illustrates how the Hybrid Model can leverage the advantages of DPR to overcome vocabulary mismatch and find the contextually correct answer.

Overall, the results of this analysis confirm that retrieval approaches relying solely on lexical signals, such as BM25, have significant limitations in handling linguistic variations and semantic contexts. Conversely, DPR demonstrates superiority in understanding the conceptual relationship between a query and a document, even without direct keyword matches. Meanwhile, the Hybrid model successfully

combines the strengths of both approaches by balancing semantic and lexical signals, thereby delivering more accurate and contextually relevant results. This finding highlights the importance of using adaptive retrieval strategies to overcome the limitations of traditional methods while maximizing the potential of semantic embedding-based approaches.

4. CONCLUSION

Based on the implementation, testing, and the quantitative and qualitative analyses conducted, it can be concluded that the hybrid retrieval model combining score fusion from BM25 and DPR delivers superior performance compared to the individual use of each model for the Open-Domain Question Answering (QA) task. Quantitatively, the hybrid model achieved an MRR of 0.8095 and a Hit Rate@10 of 80.3%, surpassing the best results from either BM25 or DPR alone. Qualitatively, this model also proved to be more consistent in answering queries with linguistic and contextual variations by leveraging the lexical precision of BM25 while simultaneously utilizing the semantic understanding of DPR to overcome vocabulary mismatch.

As for future research directions, there are several potential avenues for development. First, this study used a static weight for score fusion; future work could be directed towards dynamic weighting, which adaptively adjusts the contributions of BM25 and DPR according to the characteristics of the query. Second, to test for scalability and generalization, the hybrid model could be applied to a larger and more complex corpus, such as the entirety of the Indonesian Wikipedia. Third, as a step towards an end-to-end Question Answering system, the developed hybrid retriever can be integrated with a generative-based reader model, allowing answers to be presented in more natural and informative sentences.

REFERENCES

[1] F. Zhu, W. Lei, C. Wang, J. Zheng, S. Poria, and T.-S. Chua, "Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering," pp. 1–21, Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.00774

[2] V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, Apr. 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[3] M. F. Ilmi and P. P. Adikara, "Pencarian Dokumen Skripsi menggunakan BM25 dan Faceted Search berdasarkan Kata Kunci Abstrak (Studi Kasus : Universitas Muhammadiyah Sidoarjo)," vol. 6, no. 9, 2022.

[4] E. N. Azizah and A. N. Handayani, "Permodelan pada Information Retrieval: Literature Review," *J. Inov. Teknol. dan Edukasi Tek.*, vol. 2, no. 11, pp. 527–535, 2022, doi: 10.17977/um068v2i112022p527-535.

[5] G. Hermawan, I. Faturohman, and N. Isharmawan, "Indonesian Text Translator into Database Structured Query Language with Multi Parameters using Natural Language Processing," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 662, no. 2, p. 022095, Nov. 2019, doi: 10.1088/1757-899X/662/2/022095.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, May 2019, [Online]. Available: http://arxiv.org/abs/1810.04805

[7] X. Wu, G. Ma, M. Lin, Z. Lin, Z. Wang, and S. Hu, "ConTextual Masked Auto-Encoder for Dense Passage Retrieval," *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 4, pp. 4738–4746, Dec. 2022, doi: 10.1609/aaai.v37i4.25598.

[8] X. Ma, K. Sun, R. Pradeep, M. Li, and J. Lin, "Another Look at DPR: Reproduction of Training and Replication of Retrieval," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13185 LNCS, Springer International Publishing, 2022, pp. 613–626. doi: 10.1007/978-3-030-99736-6_41.

[9] T.-D. Nguyen, C. M. Bui, T.-H.-Y. Vuong, and X.-H. Phan, "Passage-based BM25 Hard Negatives: A Simple and Effective Negative Sampling Strategy For Dense Retrieval," *Proc. 37th Pacific Asia Conf. Lang. Inf. Comput.*, pp. 591–599, 2023, [Online]. Available: https://aclanthology.org/2023.paclic-1.59

[10] X. Chen and S. Wiseman, “BM25 Query Augmentation Learned End-to-End,” May 2023, [Online]. Available: http://arxiv.org/abs/2305.14087

[11] N. Arabzadeh, X. Yan, and C. L. A. Clarke, “Predicting Efficiency/Effectiveness Trade-offs for Dense vs. Sparse Retrieval Strategy Selection,” in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, New York, NY, USA: ACM, Oct. 2021, pp. 2862–2866. doi: 10.1145/3459637.3482159.

[12] K. Doxolodeo and A. A. Krisnadhi, “AC-IQuAD: Automatically Constructed Indonesian Question Answering Dataset by Leveraging Wikidata,” *Lang. Resour. Eval.*, vol. 59, no. 1, pp. 135–160, Mar. 2025, doi: 10.1007/s10579-023-09702-y.

[13] H. Lovenia *et al.*, “SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages,” *arXiv Prepr. arXiv 2406.10118*, 2024.

[14] M. F. Ilmi and P. P. Adikara, “Pencarian Dokumen Skripsi menggunakan BM25 dan Faceted Search berdasarkan Kata Kunci Abstrak (Studi Kasus : Universitas Muhammadiyah Sidoarjo),” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 9, pp. 4175–4180, 2022.

[15] B. Wilie *et al.*, “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” no. October, Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.05387

[16] K. A. Hambarde and H. Proenca, “Information Retrieval: Recent Advances and Beyond,” *IEEE Access*, vol. 11, no. July, pp. 76581–76604, Jan. 2023, doi: 10.1109/ACCESS.2023.3295776.