Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)
Paris, France - August, 2025
Paper No. CIST 141
DOI: 10.11159/cist24.141

Improved BoW-BoC Indexing for Short Texts Using Large Language Models

Sara Bouzid[1], Loïs Piron[2]
[1] Cadi Ayyad University, UCA, ENSA Marrakesh, Complex Systems Modeling Laboratory,
BP 575 Abdelkrim Khattabi Avenue, Marrakesh, Morocco
sara.bouzid@uca.ac.ma;
[2]Independent Researcher
13120 Gardanne, France
lois.piron@gmail.com

Abstract **-** Document indexing is a critical component of effective information retrieval systems. However, when dealing with short texts containing limited context and domain-specific terminology, traditional indexing methods often fall short, resulting in low-quality retrieval performance. Traditional Bag-of-Words (BoW) representations that rely solely on document terms are insufficient in this case. This paper introduces an enhanced BoW-BoC indexing approach that leverages generative AI to extract high-level contextual concepts from short documents. These AI-generated concepts form a Bag-of-Concepts (BoC) representation, which is integrated with the BoW model to enrich document representations. During retrieval, the FastText embedding model is used to semantically match query terms with the combined BoW and BoC representations. Experimental results show that the proposed method significantly improves indexing and retrieval performance for semantically sparse short texts, without relying on static knowledge structures.

Keywords**:** Indexing, Short text, LLM, Generative AI, GPT, Document retrieval, Word embeddings, FastText

1. Introduction

The effectiveness of Information Retrieval Systems (IRS) largely depends on how well documents and user queries are represented. However, when documents contain sparse or rare terms with limited context, generating meaningful representations becomes challenging [1]. Common examples of documents containing figures and minimal text include financial report and statistical documents. In the literature [2] - [5], short documents also encompass microblogs (e.g., X posts), question-answer (QA) corpora, and abstracts. A widely used approach for representing documents in IRS is the Bag of Words (BoW) model, which extracts and indexes relevant terms from documents. However, short documents typically suffer from term sparsity and rarity, limiting their ability to capture semantic content.

In recent years, new techniques have emerged to create contextual representation of documents, improving hence the retrieval process in IRS. Notable among these are the Bag of Concepts (BoC) representation [6], [7] and word embeddings [8], [9]. BoC enriches term meaning by mapping terms to concepts from structured knowledge bases such as ontologies, thesauri, or dictionaries. Word embeddings, on the other hand, generate dense vector representations that capture semantic relationships based on pre-training on text corpora. Models like Word2Vec [10], GloVe [11], FastText [12] and BERT [13] are examples of word embeddings commonly used for this purpose. While these techniques improve retrieval performance for long, content-rich documents, they present challenges for short or domain-specific texts. BoC requires domain-specific knowledge structures, which are often complex to build and maintain [14], and word embeddings rely on pre-training with large, domain-relevant corpora that are often difficult to obtain. Moreover, both approaches assume a sufficiently rich initial text representation, which is often lacking in short documents.

Recent advances in Artificial Intelligence (AI) and machine learning have led to the development of powerful deep neural network models such as transformers [15], [16], which have shown remarkable performance when trained on large-scale text data. These developments have enabled the rise of large language models (LLMs) [17], including the widely known GPT series [18]. Generative AI tools like ChatGPT [19] and Claude [20], built on transformer architectures, demonstrate strong capabilities in understanding semantic meaning, even in texts with rare or sparse terms. These capabilities present new opportunities for enhancing document indexing and retrieval, particularly for short texts. In this paper, we propose an

improved indexing approach that combines the traditional BoW representation with a BoC representation, where the concepts are generated by a generative AI model. This approach builds upon a previous method [21], itself an extension of the BoW-BoC indexing technique presented in CIST'24 [1]. The original method employed a domain-specific lexicon to derive BoC concepts, chosen for its ease of implementation and maintenance. In contrast, our improved approach leverages generative AI, specifically the GPT-4o model, to extract relevant concepts without relying on handcrafted or domain-specific knowledge structures, which are often complex and costly to maintain.

The remainder of this paper is organized as follows. Section 2 reviews recent advances in document indexing. Section 3 introduces the concept of the improved BoW-BoC indexing approach. Section 4 presents preliminary results obtained from applying the approach to a short-text dataset. Conclusion and future work are presented in section 5.

2. Related Work

Numerous studies in the field of IRS have focused on improving query processing and document retrieval through techniques such as domain ontologies [22], [23], word embeddings [24] - [27], and, more recently, LLMs [21], [28]. However, relatively few have addressed the indexing stage, which is critical in IRS, as it directly impacts retrieval effectiveness.

Among the notable recent contributions to document indexing, Aliwy et al. [29] proposed a method combining word sense disambiguation (WSD) and named entity recognition (NER) to enhance indexing and retrieval tasks in digital library management systems. Their indexing system relies on Part-of-Speech (POS) tagging and NER techniques, with a voting mechanism employed to consolidate the outputs of these methods, enabling accurate annotation and indexing of the library content. Sharma and Kumar [30] introduced a hybrid semantic indexing approach for unstructured documents, integrating domain ontologies with the Skip-gram model from Word2Vec and a negative sampling-based machine learning approach to extract relevant concepts for indexing. In another study, Dai and Callan [31] explored the use of deep contextualized embeddings from BERT for indexing. The authors proposed a Deep Contextualized Term Weighting framework (DeepCT), a framework that computes context-aware term weights using a pre-trained BERT model. These weights are then incorporated into an inverted index to better capture term relevance, leading to improved retrieval accuracy. Recently, the use of LLMs has extended to generative retrieval [32], where models learn to associate queries directly with document identifiers (docids) during pre-training. While promising, this approach has limitations, including high training costs and poor adaptability to dynamic document collections. To overcome these issues, [33] introduced Few-Shot GR, a retrieval framework that uses LLM prompting to generate docids to both documents and queries, enabling efficient indexing and retrieval through a docid bank constructed in a few-shot manner.

Despite these advancements, most existing work focuses primarily on the retrieval phase and assumes that documents contain rich, diverse vocabulary. This overlooks a critical challenge - term scarcity - which is particularly problematic in short documents that often lack sufficient content for effective indexing. To address this gap, our work explores the use of generative AI for enhancing the indexing of short texts by enriching them with semantically meaningful concepts.

3. The BoW-BoC Indexing Approach Using Generative AI

To enhance the semantic representation of short documents, we propose a BoW-generated-AI-BoC indexing approach, in which the BoC component is generated using GPT-4o, building on the findings of our previous work [21]. Unlike the earlier method in [1], which relied on a domain-specific lexicon, our approach uses generative AI to extract contextual concepts directly from document content. The resulting terms from the indexing process are then used as an enriched representation for document retrieval, enabling the construction of semantic vectors through word embeddings.

3.1. Document Indexing

As illustrated in Fig. 1, the indexing process begins with standard text preprocessing of documents, including tokenization, lowercasing, and stopword removal, to produce the BoW representation. Simultaneously, each document is input to GPT-4o using zero-shot prompts to generate a set of high-level concepts that reflect the semantic themes of document content. These AI-generated concepts form the BoC representation. Since the generated concepts may include multi-word

expressions and overlap with existing BoW terms, a filtering step is applied to retain only single-word, unique concepts suitable for indexing.

The final outcome of this process is a BoW-T structure associated to an inverted index scheme. In this structure, each document's traditional BoW representation is paired with its AI-generated BoC counterpart. These combined representations are later used in the retrieval process to improve semantic matching and document relevance to user queries.

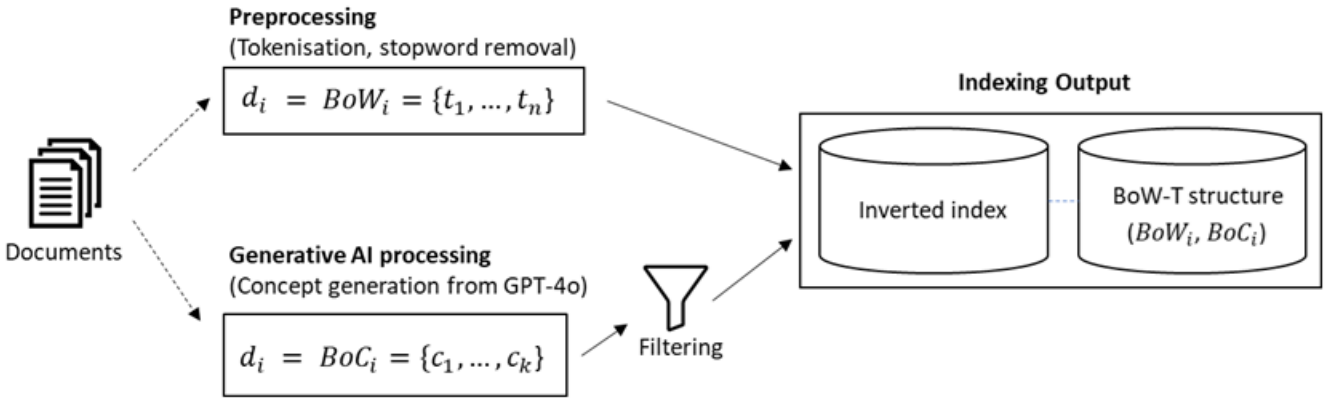

Fig. 1: BoW-generated-AI-BoC indexing approach.

3.2. Document Retrieval

During the retrieval phase (Fig. 2), we use FastText embeddings [12] to generate rich semantic representations of both documents and user queries, enabling more effective matching. FastText is a word representation technique developed by Facebook's AI Research (FAIR) lab that extends the Word2Vec model [10] by incorporating subword (character n-gram) information. Instead of learning vectors for entire words, FastText represents each word as a bag of character n-grams, enabling it to generate embeddings for rare or even out-of-vocabulary words. These features make FastText particularly well-suited for handling domain-specific terms, abbreviations, and morphological variants, which are common in this research context [1].

Fig. 2 depicts the retrieval process of short documents. The BoW and BoC representations of documents are initially retrieved from the index base and the BoW-T structure based on matching terms from the user query. The selected terms from each document are then input into FastText to generate dense vector representations. On the other hand, user queries undergo preprocessing (tokenization, stopword removal, and filtering to retain only single, unique words) before being processed by FastText. Finally, cosine similarity is used to compare the dense vectors of documents and the query, allowing for effective semantic retrieval and ranking of relevant documents.

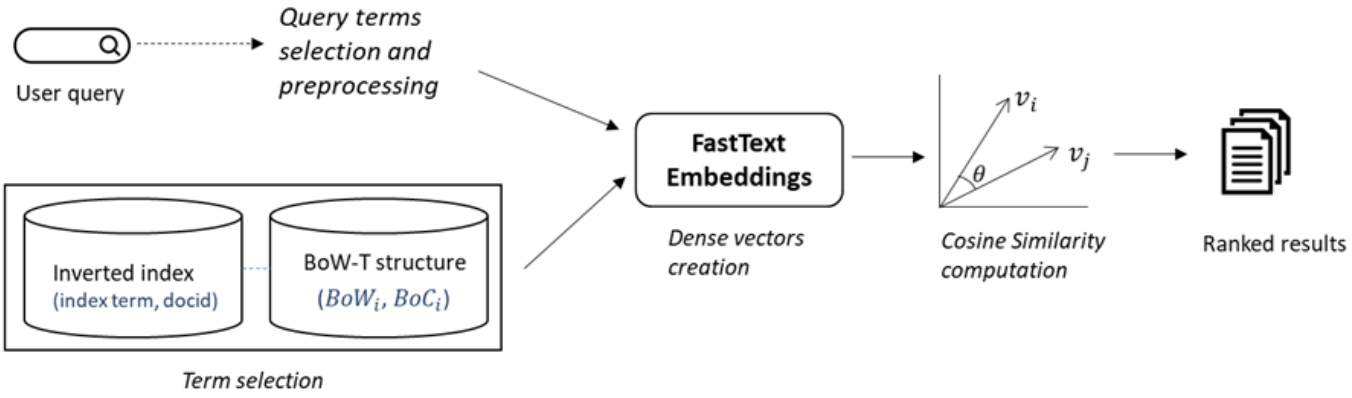

Fig. 2: Document retrieval process.

4. Experimentation

The BoW-generated-AI-BoC indexing approach was implemented using Apache Lucene (version 9.9) and tested on a machine with a Core i7 CPU, 16GB of RAM, and a 64-bit Windows operating system. To evaluate the new approach, we used the same open dataset from the Word Bank Group[1] and the same set of queries as in the previous version of the method [1] ensuring a fair comparison. For the retrieval phase using word embeddings, the FastText model was pre-trained on a custom corpus of 4,500 short documents extracted from Kaggle[2]. These documents, downloaded in CSV format, were selected using keywords related to economy and business, roughly corresponding to the domain of the test corpus.

Table 1 presents the results of the proposed method, evaluated using standard information retrieval metrics: precision, recall and F1-Score. For comparative analysis, we also implemented the retrieval process using the WScore formula (used in the earlier version of the approach) instead of FastText embeddings. Table 2 presents the F1-scores of the previous BoW-BoC indexing method and the new approach using the FastText and WScore retrieval strategies. Table 3 summarizes the performance gains achieved by our approach over other methods.

Table 1: Query processing results using BoW-generated-AI-BoC indexing and FastText retrieval.

Queries	Precision	Recall	F1-Score
Q1	0.929	0.867	0.897
Q2	1	0.938	0.968
Q3	1	0.858	0.924
Q4	0.938	0.883	0.91
Q5	1	0.819	0.901
Q6	0.858	0.924	0.89

Table 2: F1-Score results of three approaches: (1) BoW-BoC indexing, (2) BoW-generated-AI-BoC indexing with FastText retrieval strategy, and (3) BoW-generated-AI-BoC indexing with WScore retrieval strategy.

Queries	F1-Score of BoW-BoC indexing	F1-Score of BoW-generated-AI-BoC and FastText retrieval	F1-Score of BoW-generated-AI-BoC and WScore retrieval
Q1	0.858	**0.897**	0.871
Q2	0.934	**0.968**	0.942

[1] https://datatopics.worldbank.org/world-development-indicators/
[2] https://www.kaggle.com/datasets

Q3	0.881	**0.924**	0.901
Q4	0.824	**0.91**	0.883
Q5	0.816	**0.901**	0.897
Q6	0.751	**0.89**	0.816

Table 3: Comparison of improvement rates.

Approach	Average F1-Score	F1-Score improvement rate
BoW-BoC indexing	0.844	48.86%
BoW-generated-AI-BoC indexing and FastText retrieval	**0.915**	**61.38%**
BoW-generated-AI-BoC indexing and WScore retrieval	0.885	56.09%

These latest results clearly demonstrate that contextual information in document representation is essential for effectively addressing user queries. The proposed BoW-generated-AI-BoC indexing approach significantly enhanced short-document retrieval, achieving high precision and recall across all six queries, with an average F1-score of 0.915. This represents a 61.38% improvement over the traditional indexing approach, which does not apply enhancement techniques during indexing and retrieval. The proposed approach outperformed the previous BoW-BoC indexing method, showing a 25.62% increase in average F1-score.

The integration of FastText embeddings during retrieval also played a crucial role in boosting performance, as shown in Table 2 et Table 3. While the WScore formula, which combines BoW and BoC representations using predefined weights, improved the average F1-score to 0.885, a 56.09% increase over traditional indexing, it was still less effective than word embeddings, which provide richer semantic representations and, consequently, better retrieval results. Indeed, although word embeddings are widely recognized for their ability to enhance semantic understanding, they require a sufficiently representative initial term set to generate meaningful dense vectors. In our context - short texts featuring a few terms, including domain-specific ones -, document content is often sparse and lacks the depth needed for traditional semantic modeling. For this reason, our method combines BoW and AI-generated BoC representations, using LLMs like GPT-4o to extract high-level concepts and enrich the initial term space.

Ultimately, by leveraging generative AI to enhance document indexing, our approach demonstrates the importance of a semantically rich representation, especially in scenarios involving short texts. Even when advanced retrieval methods like word embeddings are applied, performance gains depend heavily on the quality of the document's initial semantic structure.

5. Conclusion

This paper presented the BoW-generated-AI-BoC indexing approach, an enhanced version of the BoW-BoC method introduced at CIST'24. The proposed approach leverages generative AI, specifically the GPT-4o model, to generate contextual concepts during the indexing phase, enriching the traditional BoW representation of documents with a complementary BoC layer. This integration enhances the semantic representation of short documents during their retrieval. In addition, the retrieval phase was improved by incorporating FastText embeddings, enabling more effective semantic matching between documents and user queries. Experimental results on short texts demonstrated significant performance gains over both the traditional indexing approach and the initial BoW-BoC method, highlighting the importance of incorporating semantically relevant terms to capture the full meaning of document content.

Notably, the findings also show that effective document indexing and retrieval can be achieved without relying on structured knowledge bases such as domain-specific lexicons and ontologies, which are often labor-intensive to build and maintain. By utilizing generative AI to identify high-level concepts, the approach offers a scalable and flexible alternative for improving short-text retrieval.

Future work will focus on optimizing the indexing structure by exploring vector store indexes, data structures that store documents as high-dimensional vectors derived from advanced embedding models. Planned experiments will evaluate the effectiveness of various embeddings, including those from OpenAI, S-BERT, and FastText, to further enhance semantic indexing and retrieval.

References

[1] S. Bouzid, "A BoW-BoC Indexing Method to Enhance Business-Related Document Representation and Retrieval," in *CIST'24: Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Science*, 2024, pp. 1–8. doi: 10.11159/cist24.162.

[2] M. Efron, P. Organisciak, and K. Fenlon, "Improving retrieval of short texts through document expansion," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, 2012, pp. 911–920. doi: 10.1145/2348283.2348405.

[3] M. Kozlowski and H. Rybinski, "Clustering of semantically enriched short texts," *J. Intell. Inf. Syst.*, vol. 53, no. 1, pp. 69–92, Aug. 2019, doi: 10.1007/s10844-018-0541-4.

[4] C. Li, S. Chen, and Y. Qi, "Filtering and Classifying Relevant Short Text with a Few Seed Words," *Data Inf. Manag.*, vol. 3, no. 3, pp. 165–186, 2019, doi: 10.2478/dim-2019-0011.

[5] J. Li, G. Huang, J. Chen, and Y. Wang, "Short Text Understanding Combining Text Conceptualization and Transformer Embedding," *IEEE Access*, vol. 7, pp. 122183–122191, 2019, doi: 10.1109/ACCESS.2019.2938303.

[6] W. Costa and G. V. Pedrosa, "A Textual Representation Based on Bag-of-Concepts and Thesaurus for Legal Information Retrieval," in *Symposium on Knowledge Discovery, Mining and Learning*, Brazil, Nov. 2022, pp. 114–121. doi: 10.5753/kdmile.2022.227779.

[7] P. Li, K. Mao, Y. Xu, Q. Li, and J. Zhang, "Bag-of-Concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base," *Knowledge-Based Syst.*, vol. 193, p. 105436, 2020, doi: https://doi.org/10.1016/j.knosys.2019.105436.

[8] X. Ye, H. Shen, X. Ma, R. Bunescu, and C. Liu, "From word embeddings to document similarities for improved information retrieval in software engineering," in *Proceedings of the 38th International Conference on Software Engineering*, Austin, 2016, pp. 404–415. doi: 10.1145/2884781.2884862.

[9] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," *Artif. Intell. Rev.*, vol. 56, no. 9, pp. 10345–10425, 2023, doi: 10.1007/s10462-023-10419-1.

[10] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient Estimation of Word Representations in Vector Space." pp. 1–12, 2013.

[11] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, pp. 1532–1543, 2014.

[12] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017, doi: 10.1162/tacl_a_00051.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, Minneapolis, 2019, pp. 4171–4186.

[14] S. Bouzid, "A Taxonomy Model to Enhance Short-Document Retrieval with Query Expansion," in *43rd IBIMA Computer Science Conference*, 2024, pp. 129–132. doi: https://doi.org/10.1007/978-3-031-79086-7_1.

[15] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the Transformer-based Models for NLP Tasks," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 2020, pp. 179–183. doi: 10.15439/2020F20.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems*, 2017, pp. 4752–4758. doi: 10.1145/3583780.3615497.

[17] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, Y. Wang, and E. Chen, "Large Language

Models for Generative Information Extraction: A Survey," *Front. Comput. Sci.*, vol. 18, no. 6, p. 186357, 2024, doi: https://doi.org/10.1007/s11704-024-40555-y.
[18] G. Yenduri, M. Ramalingam, G. C. Selvi, Y. Supriya, G. Srivastava, P. K. R. Maddikunta, G. D. Raj, R. H. Jhaveri, B. Prabadevi, W. Wang, A. V. Vasilakos, and T. R. Gadekallu, "GPT (Generative Pre-Trained Transformer) - A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions," *IEEE Access*, vol. 12, pp. 54608–54649, 2024, doi: 10.1109/ACCESS.2024.3389497.
[19] OpenAI, "OpenAI models." Accessed: Mar. 24, 2025. [Online]. Available: https://platform.openai.com/docs/models
[20] Anthropic, "Claude AI." Accessed: Mar. 31, 2025. [Online]. Available: https://www.anthropic.com/news/introducing-claude
[21] S. Bouzid and L. Piron, "Leveraging Generative AI in Short Document Indexing," *Electron.*, vol. 13, no. 17, p. 3563, 2024, doi: 10.3390/electronics13173563.
[22] K. Munir and M. Sheraz Anjum, "The use of ontologies for effective knowledge modelling and information retrieval," *Appl. Comput. Informatics*, vol. 14, no. 2, pp. 116–126, 2018, doi: 10.1016/j.aci.2017.07.003.
[23] M. N. Asim, M. Wasim, M. U. G. Khan, N. Mahmood, and W. Mahmood, "The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval," *IEEE Access*, vol. 7, pp. 21662–21686, 2019, doi: 10.1109/ACCESS.2019.2897849.
[24] W. X. Zhao, J. Liu, R. Ren, and J. R. Wen, "Dense Text Retrieval Based on Pretrained Language Models: A Survey," *ACM Trans. Inf. Syst.*, vol. 42, no. 4, pp. 1–41, 2024, doi: 10.1145/3637870.
[25] J. Wang, Z. Yang, and Z. Cheng, "Deep Pre-Training Transformers for Scientific Paper Representation," *Electronics*, vol. 13, no. 11, pp. 1–13, 2024, doi: https://www.mdpi.com/2079-9292/13/11/2123.
[26] W. Liu, J. Pang, Q. Du, N. Li, and S. Yang, "A Method of Short Text Representation Fusion with Weighted Word Embeddings and Extended Topic Information," *Sensors*, vol. 22, no. 3, pp. 1–15, 2022, doi: 10.3390/s22031066.
[27] F. C. Fernández-Reyes and S. Shinde, "CV Retrieval System based on job description matching using hybrid word embeddings," *Comput. Speech Lang.*, vol. 56, pp. 73–79, 2019, doi: https://doi.org/10.1016/j.csl.2019.01.003.
[28] H. Tan, S. Zhan, H. Lin, H.-T. Zheng, W. Kin, and V. Chan, "QAEA-DR: A Unified Text Augmentation Framework for Dense Retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 6, pp. 3669–3683, 2025, [Online]. Available: https://arxiv.org/abs/2407.20207v1
[29] A. Aliwy, A. Abbas, and A. Alkhayyat, "NERWS: Towards improving information retrieval of digital library management system using named entity recognition and word sense," *Big Data Cogn. Comput.*, vol. 5, no. 4, pp. 1–17, Dec. 2021, doi: 10.3390/bdcc5040059.
[30] A. Sharma and S. Kumar, "Machine learning and ontology-based novel semantic document indexing for information retrieval," *Comput. Ind. Eng.*, vol. 176, 2023, doi: 10.1016/j.cie.2022.108940.
[31] Z. Dai and J. Callan, "Context-Aware Term Weighting For First Stage Passage Retrieval," in *SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, China, 2020, pp. 1533–1536. doi: 10.1145/3397271.3401204.
[32] G. Budakoglu and H. Emekci, "Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning and Their Synergistic Fusion for Enhanced Performance," *IEEE Access*, vol. 11, pp. 30936–30951, 2023, doi: 10.1109/ACCESS.2025.3542334.
[33] A. Askari, C. Meng, M. Aliannejadi, Z. Ren, E. Kanoulas, and S. Verberne, "Generative Retrieval with Few-shot Indexing," *arXiv*, pp. 1–8, 2024, doi: https://doi.org/10.48550/arXiv.2408.02152.