Embedding-Based Vector Search for Large Scale Text Retrieval- A Review

[1]**Dr Vinod Kumar P,** Department of CSE – Data Science, ATME College of Engineering, Mysore
Kampana R[1,2], Neha Merin Dsouza[1,2], Nishanth M Jadav[1,2], Sharath P[1,2]
[1]ATME College of Engineering, Mysuru, India
[2]CSE-Data Science Department,
ATME College of Engineering, Mysuru 570028, India

Abstract*—Embedding-based vector search marks a transformative shift in information Re-trieval, particularly for large-scale textual datasets where conventional keyword-based methods fall short in capturing semantic relevance. This approach reframes text retrieval as a semantic similarity task, representing documents—such as legal or academic texts—as highdimensional vector embeddings using advanced natural language processing (NLP) models like BERT or RoBERTa. These embeddings encapsulate the contextual and conceptual essence of the documents, enabling re-trieval based on meaning rather than exact keyword matches.*

The project presents a modular architecture comprising embedding generation, efficient vector indexing using FAISS (e.g., IVF, HNSW, PQ), and a semantic search layer enhanced with generative AI. This integration facilitates fast, scalable retrieval while maintaining contextual depth and accuracy. Notably, the system incorporates retrieval-augmented generation (RAG) for intelligent query refinement, summarization, and knowledge extraction. This enhances user interaction by providing synthesized, context-aware responses to complex queries.

The framework proves particularly effective in domains like law and research, where vocabulary varies and contextual precision is critical. Overall, the proposed system demonstrates the power of combining vector search with generative AI to deliver semantically enriched and computationally efficient information retrieval.

Keywords: Vector Search, Semantic Similarity, Embedding Generation, Natural Language Processing (NLP), FAISS, Retrieval-Augmented Generation (RAG), Generative AI, Information Retrieval.

INTRODUCTION

In today's digital era, the rapid proliferation of information—particularly unstructured textual data— poses significant challenges for traditional search techniques. Conventional keyword-based search engines, while useful for basic queries, often struggle to provide relevant results when faced with large-scale, complex document collections. This is especially true in domains that require a deep understanding of context, nuance, and semantics—such as legal research, academic literature, and scientific documentation.

To address these limitations, embedding-based vector search has emerged as a trans formative approach in the field of information retrieval. Unlike traditional methods that rely solely on exact keyword matches, embeddingbased systems represent documents and queries as highdimensional numerical vectors. These vectors are generated using advanced machine learning models—such as Word2Vec, GloVe, BERT, or other transformer-based architectures—which are capable of capturing the semantic meaning and contextual relationships between words, phrases, and entire documents.

By positioning semantically similar items closer together in a multidimensional vector space, this approach allows systems to go beyond surface-level text matching and instead identify deeper conceptual similarities. For example, a query about "intellectual property rights" might successfully retrieve documents containing related legal terms such as "copyright," "patent law," or "trademark protection," even if the exact phrase is not present. This semantic awareness is invaluable in professional domains where terminology may vary, but underlying concepts remain consistent.

Moreover, embedding-based search systems are inherently scalable. Leveraging frame works like FAISS (Facebook AI Similarity Search), these systems can perform fast and efficient similarity searches over millions—or even billions—of documents. They support real-time query processing, indexing, and ranking, making them suitable for enterprise scale applications.

In domains such as law and academia, where the accuracy and relevance of retrieved information are paramount, embedding-based vector search introduces a paradigm shift. Legal professionals can more easily navigate vast databases of case law, statutes, and scholarly commentary, while researchers can discover semantically linked publications that might be missed by traditional systems. In doing so, this technology not only enhances retrieval performance but also contributes to knowledge discovery, decision making, and innovation.

MOTIVATION

Legal and research professionals often face inefficiencies due to reliance on traditional keyword-based search engines, which fail to capture the semantic meaning behind queries. This project is motivated by the need to improve document retrieval performance by incorporating semantic search using embeddings and the FAISS (Facebook AI Similarity Search) indexing framework.

OBJECTIVES OF THE PROJECT

The primary objective of this review is to explore and analyse the role of embedding-based vector search in enhancing large-scale text retrieval systems. Traditional keyword-based and Boolean retrieval approaches often fail to capture semantic relationships within unstructured text, leading to incomplete or irrelevant results.

To address these limitations, embedding-based retrieval methods leverage dense embeddings generated through advanced neural architectures. These embeddings enable semantic similarity matching, scalability, and adaptability across diverse domains, offering a significant improvement over lexical matching.

This project further aims to systematically review state-ofthe-art techniques, including hybrid retrieval methods, retrieval-augmented generation (RAG) frameworks, and domain-specific adaptations. Special attention is given to their applications in legal, policy, and multilingual contexts where accurate information retrieval is critical.

In addition, the study seeks to identify existing challenges such as computational efficiency, scalability, dynamic indexing, and interpretability. Finally, it highlights potential directions for future research in embedding-driven retrieval for knowledge-intensive tasks.

I. LITERATURE REVIEW

Dnyanesh Panchal et al. [1], *"Law Pal: A Retrieval Augmented Generation Based System for Enhanced Legal Accessibility in India"*, introduce **Law Pal**, a RAG-based legal chatbot that leverages FAISS for efficient vector-based legal document retrieval. The system applies prompt engineering and hierarchical indexing to handle ambiguous queries across multiple legal domains, thereby improving legal literacy and reducing misinformation.

Hai-Long Nguyen et al. [2], *"Enhancing Legal Document Retrieval: A Multi-Phase Approach with Large Language Models"*, propose a **three-phase retrieval pipeline** that integrates BM25 pre-ranking, BERT-based re-ranking, and LLM-based prompting. Evaluated on the COLIEE dataset, their approach demonstrates superior performance, highlighting the importance of semantic over lexical features in legal retrieval.

Solmaz S. Monir et al. [3], *"Vector Search: Enhancing Document Retrieval with Semantic Embeddings and Optimized Search"*, describe a **hybrid system** that combines FAISS and HNSWlib for scalable retrieval. Their work introduces dynamic indexing, efficient query processing, and hyperparameter optimization to address limitations of traditional vector search methods.

Ilias Chalkidis et al. [4], *"LEGAL-BERT: The Muppets Straight out of Law School"*, present **LEGAL-BERT**, a family of BERT-based models fine-tuned on legal corpora. The study shows that domain-specific adaptation significantly enhances performance in tasks such as contract analysis and case classification when compared with general-purpose models.

Patrick Lewis et al. [5], *"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"*, propose the **RAG framework**, which integrates seq2seq models with FAISS-based dense retrieval for knowledgeintensive tasks like open-domain QA and fact-checking. They introduce two model variants, RAG-Token and RAG-Sequence, to optimize document-supported text generation.

Ryan C. Barron et al. [6], *"Bridging Legal Knowledge and AI: RAG with Vector Stores, Knowledge Graphs, and Hierarchical NMF"*, integrate RAG with **symbolic reasoning** through knowledge graphs and hierarchical NMF-based topic modelling. Their approach enhances explainability and semantic clustering in legal research by combining dense embeddings with structured knowledge.

Keet Sugathadasa et al. [7], *"Legal Document Retrieval using Document Vector Embeddings and Deep Learning"*, propose a retrieval system that fuses **TF-IDF, pageranking, and neural embeddings** trained on over 2,500 legal cases. The system outperforms Boolean search, demonstrating the effectiveness of semantic similarity and domain adaptation, while also addressing challenges in scalability and multilingual support.

Rishi Kalra et al. [8], *"Hypha-RAG: A Hybrid Parameter Adaptive Retrieval-Augmented Generation System for AI Legal and Policy Applications"*, present **Hypha-RAG**, which dynamically adjusts retrieval parameters based on query complexity. By integrating dense, sparse, and knowledge graph-based retrieval, the system improves comprehension and enables scalable deployment in complex legal-policy contexts.

Haitao Li et al. [9], *"Lex RAG: Benchmarking RetrievalAugmented Generation in Multi-Turn Legal Consultation Conversation"*, propose **Lex RAG**, a benchmark suite designed for evaluating RAG systems in multi-turn legal dialogues. They also release **LexiT**, a toolkit for building and testing such systems, along with a dataset of over 1,000 annotated legal conversations.

Muhammad Rafsan Kabir et al. [10], *"Legal RAG: A Hybrid RAG System for Multilingual Legal Information Retrieval"*, develop **Legal RAG**, a bilingual pipeline supporting English and Bangla for legal QA and retrieval. Tested on regulatory texts like the Bangladesh Police Gazette, it enhances accessibility in low-resource contexts by combining generative AI with vector search techniques.

OUTCOME OF LITERATURE REVIEW

The literature survey confirms the superiority of embedding-based systems over traditional keyword searches, particularly when combined with optimized indexing and gen ergative enhancements. FAISS stands out for its scalability and efficiency, making it a suitable backbone for vector search in large-scale applications. However, successful deployment requires thoughtful adaptation to domain-specific needs and integration with generative AI for enhanced usability.

II. CONCLUSION

The reviewed literature highlights a significant evolution in legal information retrieval, driven by the integration of dense vector embeddings, deep learning, and Retrieval Augmented Generation (RAG) frameworks. Traditional lexical search methods are in caressingly being replaced or augmented by semantic approaches, which leverage tools like FAISS, BERT, and domain-specific models such as LEGAL-BERT to enhance both precision and contextual understanding.

Systems such as Law Pal and Hypha-RAG demonstrate the practical utility of RAG based architectures in real-world legal settings, offering solutions to ambiguity and mis information. Several works, including those by Hai-Long Nguyen and Solmaz Monir, emphasize the need for multiphase and hybrid retrieval pipelines to balance efficiency, scalability, and accuracy. The adoption of multilingual and low-resource adaptations, as seen in Legal RAG, also signals progress toward democratizing access to legal knowledge.

Overall, the convergence of semantic search, domainadapted language models, and hybrid retrieval techniques presents a promising pathway for developing intelligent legal assistance systems. However, challenges remain in terms of explainability, multilingual coverage, and handling complex query structures—areas that future research must continue to address.

OBJECTIVES OF THE PROJECT

- Develop an embedding generation pipeline using domain-specific language models.
- Implement scalable vector indexing strategies using FAISS (IVF, HNSW, PQ).
- Integrate generative AI for query refinement and result summarization.
- Optimize for performance, memory efficiency, and semantic precision.
- Evaluate retrieval quality across varied legal and research document datasets.

III. REFERENCES

[1] S. Monir et al., *"Vector Search: Enhancing Document Retrieval with Semantic Embeddings,"* 2024. http://arxiv.org/pdf/2409.17383.pdf

[2] D. Panchal, A. Gole, V. Narute, and R. Joshi, *"Law Pal: A Retrieval Augmented Generation Based System for Enhanced Legal Accessibility in India,"* Proc. of Conference on AI and Legal Tech, 2024.

[3] H.-L. Nguyen, et al., *"Enhancing Legal Document Retrieval: A Multi-Phase Approach with Large Language Models,"* Proc. of Legal Information Retrieval Workshop, 2024.

[4] Pinecone Team, *"Legal Semantic Search with Pinecone & Voyage AI,"* 2025. https://www.pinecone.io/learn/legalsemantic-search/

[5] DZone, *"FAISS: Practical Guide to Similarity Search,"* 2024. https://dzone.com/articles/similarity-search-with-faissa-practical-guide

[6] Microsoft, *"Vector Search in Azure AI,"* 2025. https://learn.microsoft.com/en-us/azure/search/vector-searchoverview

[7] S. Monir et al., *"VectorSearch...Optimized Search,"* 2024. https://arxiv.org/html/2409.17383v1

[8] *"Embedding in Recommender Systems: A Survey,"* 2023. https://arxiv.org/html/2310.18608v2

[9] IBM, *"What is Vector Search?"* https://www.ibm.com/topics/vector-search

[10] Oracle, *"Your Ultimate Guide to Vector Search."* https://www.oracle.com/database/vector-search/

[11] S. S. Monir, et al., *"Vector Search: Enhancing Document Retrieval with Semantic Embeddings and Optimized Search,"* Journal of Intelligent Information Systems, vol. 45, no. 3, pp. 345–360, 2024. http://arxiv.org/pdf/2409.17383.pdf