Development and Evaluation of a Hybrid Information Retrieval System Applied to the Brazilian Legal Domain

Ana Carolina C. Bessa[1], Fábio M. F. Lobato[1,2], Antonio F. L. J. Junior[1]

[1]State University of Maranhão, Maranhão, Brazil,
[2]Federal University of Western Pará, Pará, Brazil

Correspondence: anabessa@aluno.uema.br,, fabio.lobato@ufopa.edu.br, antoniojunior@professor.uema.br

Abstract

The need for tools to manage processes, automate tasks, and speed up the judicial system justifies enhancing traditional Information Retrieval systems, which are often hindered by vocabulary mismatches and lengthy legal texts. Although Transformer-based models capture semantic nuances, they face input size limitations, making it challenging to process long texts without information loss. In this work, we introduce and evaluate a hybrid system for the legal domain that combines the BM25L algorithm with the BumbaLM language model. The system was evaluated using legal judgment summaries from TJMA. The experiments revealed that the standalone semantic model outperformed the hybrid approach. The lexical component struggled with natural-language and conceptual queries, resulting in false positives that degraded the hybrid system's overall performance.

1 Introduction

The volume of documents produced by judicial institutions continues to grow. According to the Brazilian National Council of Justice (CNJ), in its report *Justice in Numbers*[1], the year 2025 ended with 74,756,005 pending cases. This makes it difficult to organize and access information. With the current technological transformation from Industry 4.0, the legal sector has invested in intelligent tools to automate repetitive tasks and reduce procedural delays (Nascimento, 2024). In this sense, Information Retrieval (IR) has gained attention. IR can be defined as the process of finding relevant documents from large amounts of unstructured data. It can be used to locate documents that satisfy a user's information needs (Vitório et al., 2025). In the legal field, this is difficult because of the nature of the documents. They often contain technical language, lengthy texts, and mismatches in vocabulary. Terms in the user's query may not exactly match those in relevant documents (Moreira, 2024).

One technique to overcome this challenge is to use Transformer-based neural models, such as Sentence-BERT (SBERT). These models process entire text sequences at once to capture word relationships. However, using these architectures alone in the legal domain has limitations. Input size restrictions usually limit them to 512 tokens, where a token is roughly a word or character sequence the model can process at once. This requires truncating long texts, possibly losing relevant information in lengthy proceedings. Probabilistic algorithms like Best Matching 25 (BM25), which score documents by matching query terms, perform well in document retrieval and identifying technical terminology. Semantic models may underperform here due to noise or overgeneralization. Therefore, integrating these methods in hybrid systems can combine the semantic power of language models with the efficiency of lexical correspondence.

This research proposes the development of a hybrid legal information retrieval system. It combines the BM25 algorithm with the BumbaLM language model. The choice of BM25L is justified by its ability to handle lengthy documents. This matches the complex structure of legal proceedings. BumbaLM, a model trained on Portuguese legal data, serves as the semantic component. It addresses vocabulary differences and captures conceptual relationships.

2 Related Works

The legal domain presents unique challenges due to its specialized terminology and complex documents. Prior work by Vitório et al. (2025) compared 12 SBERT models to traditional Okapi BM25 and BM25L baselines in the Brazilian legislative context. Neural models improve semantic understanding. Selection of the appropriate BM25 variant is essential for baseline performance. In con-

[1]https://www.cnj.jus.br/wp-content/uploads/2025/11/justica-em-numeros-2025.pdf

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 186–190
April 13-16, 2026

trast, this research focuses on jurisprudential data, summaries of court rulings with different structures. It aims to find the best hybrid approach between BM25 and BERT-based models.

Kodri et al. (2025) introduce the Fine-Hybrid system, which combines BM25 with an SBERT model adapted to a tax corpus. Their results show that domain adaptation improves the model's ability to capture legal nuances. While Kodri et al. (2025) focuses on the tax domain, this study uses a 4-billion-parameter embedding model. It extends hybridization to general jurisprudence.

Another relevant work for the context is Fernandes et al. (2025). The authors introduced the JutisTCU dataset, with more than 16,000 documents from the Brazilian Federal Court of Accounts (TCU). Their experiments showed that integrating semantic methods based on OpenAI and BERT models improves case-law retrieval. Our work differs in that it uses an open-weight model, BumbaLM, which allows courts to run the system locally.

Baban Gain et al. (2019) and Kim et al. (2022b) describe the use of BM25 combined with BERT in Competition on Legal Information Extraction and Entailment (COLIEE) tasks for information retrieval in legal documents in other languages. The hybrid system developed focuses on adapting these architectures to the particularities of Brazilian Portuguese.

In summary, related work shows that combining BERT-based models with algorithms such as BM25 achieves better results for legal IR tasks than using either method alone.

3 Methodology

This section describes the experiments conducted to verify the efficiency of the hybrid model relative to individual information retrieval techniques, as well as the database and evaluation metrics used.

3.1 Data collection

The hybrid system experiments were conducted using a set of legal judgments from the Maranhão Court of Justice (TJMA). The dataset includes 100,000 records of final decisions on various cases, with details such as identification numbers (ID), unique case numbers, district, chamber, CNJ classification (ID and name), summary, and content. For the experiments, only the ID, a unique small number for each document, and the judgment summaries, written by the judges, were selected because their smaller size fit within the 512-token limit of the chosen language model.

The corpus underwent preprocessing to standardize words to lowercase, remove invalid symbols and special characters (e.g., "\n"), and filter out stopwords. This reduced text noise, leaving only the information necessary to improve the system's information retrieval performance.

3.2 BM25 Algorithm

The BM25L algorithm was chosen for exact-term-matching retrieval. This choice suits the legal area where document lengths vary greatly. BM25 is a widely used lexical search algorithm, known for its high efficiency and for requiring fewer computational resources than pre-trained large language models.

BM25L improves upon BM25, which overly penalizes long documents due to term-frequency saturation. BM25L adjusts for length normalization, so document relevance isn't underestimated by length (Kim et al., 2022a). Rare terms retain their discriminatory weight, even in lengthy legal documents.

3.3 BumbaLM

The vector representation was performed using the BumbaLM embedding language model (Carmo et al., 2023). BumbaLM was selected because it was trained on a collection of Portuguese legal documents from the Court of Justice, and it best matched the characteristics of the current dataset. It is important to note that BumbaLM is distributed solely as an open-weight model, enabling courts and legal institutions to run the system locally and ensuring data security during inference. However, the model's full training dataset and source code cannot be made open source due to strict privacy policies and confidentiality restrictions associated with the sensitive, real-world judicial documents used to train it.

Given the characteristics of the chosen model, the texts were truncated to a maximum of 512 tokens due to the model's context window. As a result, priority was given to using sentence summaries prepared by the judges, as they are shorter compared to the full text. The embeddings were created using the Mean Pooling method (averaging the last hidden layer), followed by L2 (Euclidean) normalization to ensure all vectors had unit length. These vectors were then stored in ChromaDB to eliminate the need to generate embeddings in sub-

sequent runs, with cosine distance defined as the chosen similarity metric.

3.4 Proposed system

The hybrid system combines lexical and semantic components for the IR task. The challenge at this stage was the incompatibility of the scoring scales: BM25L produces unlimited scores $[0, \infty]$, based on term frequency, while cosine similarity operates in the range $[-1, 1]$.

To address this issue, the approach used parallel retrieval, in which for each query q, the Top-K documents from each component are retrieved separately. The Min-Max Scaling technique was applied to the raw scores from each candidate list, normalizing them to the range [0, 1]. The final score S_f for each document d was computed using Equation 1.

$$S_f(d, q) = \alpha \cdot S'_1(d, q) + (1 - \alpha) \cdot S'_2(d, q) \quad (1)$$

where S' represents the normalized score, with S'_1 corresponding to BM25L and S'_2 for BumbaLM, and α is the control hyperparameter.

In the comparative experiments, α varied between 0.0 (purely semantic), 1.0 (purely lexical), and intermediate values (hybrid), with the value $\alpha = 0.5$ adopted as the baseline for evaluating the balance between the techniques.

3.5 Evaluation metrics

To assess the hybrid system and the individual techniques, the metrics used were Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and Mean R-Precision (MRP).

MAP is the main metric for overall quality because it considers the ranking order of all relevant documents retrieved and penalizes the system if an important document appears in lower positions (Zhang and Zhang, 2009). It is calculated by averaging the Average Precision (AP) scores for the set of queries Q. The AP of a query is given by Equation 2.

$$\text{MAP} = \frac{\sum_{k=1}^{n} (P(k) \times rel(k))}{\text{number of relevant documents}} \quad (2)$$

where $P(k)$ represents the precision at cutoff k, and $rel(k)$ is a binary function (1 if the document at position k is relevant, 0 otherwise.

MRR measures the system's ability to find the correct answer as quickly as possible (Caseli and Nunes, 2024). This metric assesses the position of the first relevant document in the list, as shown in Equation 3.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \quad (3)$$

where $rank_i$ indicates the position of the first relevant result for the query i.

The MRP metric measures precision at position R, where R is the total number of relevant documents for the current query (Beitzel et al., 2009), as in Equation 4.

$$\text{R-Precision} = \frac{\text{Relevant documents}}{R} \quad (4)$$

The experimental results considering the materials and methods presented are given in the following Section.

4 Results and Discussion

The experiment used 10 queries to retrieve specific results from the database, each linked to the document IDs it referenced. The queries were manually created by a person after analyzing various documents in the database, with each query serving as the main topic of the case, for example: "Property damage caused by a power grid failure". Manual query creation was chosen in this study to ensure that the queries were contextually meaningful and directly relevant to real-world scenarios observed in the dataset. Although this method generated precise, targeted queries, it limited the number of queries because of the time and expertise required for selection. It is also important to note that the dataset does not include annotations indicating the relevance of each judgment to its query at the time of this study. Table 1 shows the results from the models alone and with hybrid techniques.

	Metrics		
Model	MAP	MRP	MRR
BM25L	0,144	0,144	0,1
BumbaLM	**0,583**	**0,583**	**0,3**
Hybrid System	0,335	0,335	0,1

Table 1: Performance comparison between recovery strategies.

The results showed that using only BumbaLM was better than the other scenarios. The semantic model achieved a MAP and MRR of 0.583, meaning that, on average, the first relevant document was ranked between 1st and 2nd place. In comparison, BM25L had the lowest performance (0.144), likely

due to difficulty handling the queries' textual structure. The test queries, formulated in natural language and using conceptual terms, did not produce exact matches in the indexed summaries. Since BM25L relies on exact term frequency, it failed to retrieve documents that used synonyms or paraphrases. Conversely, BumbaLM showed strong generalization. Even without exact word matches, the generated embeddings effectively captured the search intent.

The queries used in the test, formulated in natural language and using conceptual terms, did not produce exact matches in the indexed summaries. BM25L relies on the frequency of exact terms, so it failed to retrieve documents that used synonyms or paraphrases. Even without exact word matches, the generated embeddings captured the search intent.

The hybrid system performed worse than the isolated semantic model, averaging the performance of BumbaLM and BM25L. This combination caused interference in the results by giving too much weight to a low-performing component. Using an average fusion coefficient (0.5), the system allowed BM25L false positives (irrelevant documents with high lexical scores) to overshadow the relevant documents identified by BumbaLM.

One point observed is the exact match between MAP and MRR values across all configurations. This can be explained because, for the set of queries tested, the retrieval functioned like searching for a known item, where the position of the first, and possibly only, relevant document in the template determined the average precision metric.

Therefore, the results showed that, for the evaluated corpus and queries, the BumbaLM language model performed best, and hybridization was harmful in this case. One solution to this issue is to work with a properly annotated dataset, including relevance notes for the judgments, and to formulate queries by legal professionals or LLMs, as presented in the works of Fernandes et al. (2025) and Vitório et al. (2025). Another option is to decrease the weight of the lexical component, using BM25L solely as a tiebreaker.

5 Conclusion

In this study, we presented a hybrid IR system that combines the BM25L algorithm and the BumbaLM model to assess its effectiveness for IR in legal documents. Compared to other work in this area, this work stands out for using an open-weight model with a large number of parameters. The results showed that the hybrid system performed worse than using only BumbaLM for the retrieval task. This is likely due to the structure of the search queries and the absence of relevance indicators in the judgments. Providing a set of annotated data and queries created by lawyers could improve the performance of the hybrid system compared to using each strategy alone.

The experimental results provide some clues into the performance of information retrieval using purely semantic models, purely lexical models, and a hybrid approach, and may guide broader experiments aimed at informing the development of efficient IR systems that meet the intrinsic needs of the legal domain.

For future work, the experiment will be conducted using an annotated database by legal experts and will include the development of an intuitive, accessible Graphical User Interface (GUI) that allows legal professionals to formulate queries in natural language and view the retrieved sentences, with highlights of the passages that contributed to lexical or semantic relevance. Usability tests and qualitative evaluations will be conducted with end users. The goal is to collect human feedback to verify whether the system retrieves the correct documents and is perceived as useful in real-world operational scenarios.

Limitations

As mentioned in the section 4, the limitation of this study is the lack of datasets with scores and answers for each document. In the experiments conducted, it was not possible to achieve the system's correct performance because the most relevant document to the query was not identified. Additionally, the queries were not formulated by legal experts, making it difficult to understand how searches are performed.

Acknowledgments

This study was supported by the National Council for Scientific and Technological Development (CNPq) - DT-303031/2023-9; by the Maranhão Foundation for Research and Scientific and Technological Development; the Financing Agency for Studies and Projects (FINEP) – (ProAmazonia - 2373/24 - CTCCA-II); and by Technical Cooperation Agreement Nº. 02/2021 (case Nº. 38328/2020-TJ/MA).

References

Baban Gain, Dibyanayan Bandyopadhyay, Tanik Saikh, and Asif Ekbal. 2019. Iitp in coliee@icail 2019: Legal information retrieval using bm25 and bert.

Steven M. Beitzel, Eric C. Jensen, and Ophir Frieder. 2009. *Average R-Precision*, pages 195–195. Springer US, Boston, MA.

Fabrício Carmo, Ferdinando Serejo, Antonio Jacob Junior, Ewaldo Santana, and Fábio Lobato. 2023. Embeddings jurídico: Representações orientadas à linguagem jurídica brasileira. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 188–199, Porto Alegre, RS, Brasil. SBC.

H. M. Caseli and M. G. V. Nunes, editors. 2024. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edition. BPLN.

Leandro Carísio Fernandes, Leandro dos Santos Ribeiro, Marcos Vinícius Borela de Castro, Leonardo Augusto da Silva Pacheco, and Edans Flávius de Oliveira Sandes. 2025. Juristcu: A brazilian portuguese information retrieval dataset with query relevance judgments. *arXiv preprint arXiv:2503.08379*.

Gyeongmin Kim, Minseok Kim, and Jaechoon Jo. 2022a. Enhancing code similarity with augmented data filtering and ensemble strategies. *JOIV: International Journal on Informatics Visualization*, 6(3):676–680.

Mi-Young Kim, Juliano Rabelo, Kingsley Okeke, and Randy Goebel. 2022b. Legal information retrieval and entailment based on bm25, transformer and semantic thesaurus methods. *The Review of Socionetwork Strategies*, 16(1):157–174.

Wan Ahmad Gazali Kodri, Muhammad Haris, and Rifqi Fitriadi. 2025. Fine-hybrid: Integration of bm25 and finetuned sbert to enhance search relevance. *Teknika*, 14(2):213–222.

Viviane P. Moreira. 2024. Recuperação de informação. In H. M. Caseli and M. G. V. Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3 edition, book chapter 21. BPLN.

Iury Gregory Chaves do Nascimento. 2024. A inteligência artificial no sistema judiciário trabalhista brasileiro.

Douglas Vitório, Ellen Souza, José Antônio dos Santos, André Carlos Ponce de Leon Ferreira, Adriano LI Oliveira, Nádia FF da Silva, and 1 others. 2025. Bm25 x vila sésamo: avaliando modelos sentence-bert para recuperaçao de informaçao no cenário legislativo brasileiro. *Linguamática*, 17(1):17–33.

Ethan Zhang and Yi Zhang. 2009. *Average Precision*, pages 192–193. Springer US, Boston, MA.