A Comparative Study of BERT Models for Semantic Retrieval of Brazilian Legal Precedents

Adrielson Ferreira Justino[1], Antônio Fernando Lavareda Jacob Junior[1], Fábio Manoel França Lobato[1,2]

[1] Universidade Estadual do Maranhão, Brazil
adrielferreir28@gmail.com
[2] Universidade Federal do Oeste do Pará, Brazil
fabio.lobato@ufopa.edu.br

Abstract. The growing volume of digital documents in Brazilian courts has intensified challenges related to judicial inefficiency, including the slow identification of established precedents and the risk of inconsistent rulings in similar cases. To address this, this study presents a systematic empirical study on the effectiveness of different classes of BERT-based embedding models for the semantic retrieval of legal documents, focusing on identifying relevant precedents for new complaints. A pipeline employing document chunking to handle long legal texts and using ElasticSearch for large-scale dense vector retrieval was implemented and evaluated. Using a corpus of legal complaints from the Maranhão State Court of Justice, three model categories were compared: (i) a general-purpose Portuguese model, (ii) three domain-specific models trained on Brazilian legal corpora, and (iii) a task-specific Sentence-BERT model fine-tuned for similarity tasks. Performance was assessed using a proxy-based protocol with precedent class labels as the ground truth, measured by Precision@k, MRR@15, and MAP@15. The results indicate that the task-specific SBERT-pt model achieved higher performance within the evaluated setting, consistently outperforming other tested models across the selected metrics. This model showed improvements in both immediate relevance (Precision@1 of 0.787) and overall ranking quality (MAP of 0.806). Although domain adaptation provided marginal benefits over the general-purpose baseline, task-specific fine-tuning for similarity appeared to be the most influential factor for retrieval quality in this scenario. These findings suggest that optimizing models for retrieval tasks can offer a promising direction for enhancing semantic searches in legal contexts. However, further studies using larger and more diverse datasets are needed to confirm the generalizability of these results and assess their impact on real-world judicial workflows.

CCS Concepts: • **Information systems** → **Similarity measures**; **Language models**; • **Computing methodologies** → **Natural language processing**; • **Applied computing** → **Law**.

Keywords: Dense Retrieval, Legal NLP, Sentence Embeddings, BERT, Judicial Decision Support

1. INTRODUÇÃO

A digitalização do setor jurídico brasileiro, impulsionada por sistemas como o Processo Judicial Eletrônico (PJe), otimizou o acesso à justiça, mas também intensificou o desafio de gerenciar número crescente de documentos judiciais [Maia Filho and Junquilho 2018]. Segundo o relatório "Justiça em Números" do Conselho Nacional de Justiça (CNJ)[1], o Brasil encerrou 2023 com 83,8 milhões de processos pendentes. Esse cenário compromete a confiança na Justiça e impacta negativamente o ambiente econômico [Magalhães and Freitas 2023]. Além do volume excessivo, a repetição de temas jurídicos entre diferentes tribunais gera decisões divergentes, prejudicando a segurança jurídica [Stemler 2019]. Para elucidar esse problema, o Sistema de Precedentes Judiciais, instituído pelo Código de Processo Civil de 2015 (Lei n.º 13.105/15), visa padronizar o julgamento de casos similares por meio de mecanismos como a Repercussão Geral (RG), os Recursos Repetitivos (RR), o Incidente de Assunção de Competência (IAC) e o Incidente de Resolução de Demandas Repetitivas (IRDR) [Brasil

[1] https://www.cnj.jus.br/wp-content/uploads/2024/05/justica-em-numeros-2024.pdf

2015]. Contudo, a sobrecarga processual dificulta sua aplicação manual, tornando a identificação de precedentes morosa e suscetível a erros [Toffoli and Gusmão 2019].

Iniciativas como o Programa Justiça 4.0 do CNJ impulsionam a modernização com Inteligência Artificial, destacando-se as plataformas *Codex* e Sinapses [de Souza and de Souza Salles 2022]. A cooperação Técnica n.º 002/2021, firmada entre a Universidade Estadual do Maranhão (UEMA) e o Tribunal de Justiça do Maranhão (TJMA) exemplifica essa tendência, desenvolvendo soluções de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM) como o "Robô Maria Firmina" para identificar precedentes jurídicos [Carmo et al. 2023; Marinato et al. 2022]. No entanto, em Recuperação de Informação (RI), as abordagens tradicionais de busca lexical falham em capturar significado implícito, essencial para análise de precedentes [Ruckdeschel 2020]. Como alternativa, a Recuperação de Informação Densa (RID) [Izacard et al. 2022] utiliza representações vetoriais de *Language Models (LM)*, como os da família *Bidirectional Encoder Representations from Transformers* (BERT) [Harispe et al. 2022]. A integração desses *embeddings* em motores de busca permite a implementação de sistemas de Busca Híbrida (*e.g., Elasticsearch*) [Ni et al. 2024].

A aplicação de RID em documentos longos, como as petições iniciais, apresenta seus próprios desafios. Modelos *Transformers* possuem limitação no comprimento da sequência de entrada (512 *tokens* no BERT) [Devlin et al. 2019]. Embora existam modelos de contexto longo (*e.g., Longformer*), seu alto custo computacional os torna menos acessíveis [Beltagy et al. 2020]. Consequentemente, adotou-se uma abordagem de segmentação textual, ou *chunking*, para viabilizar o uso de modelos pré-treinados eficientes [Karpukhin et al. 2020]. Neste contexto, este artigo apresenta um estudo comparativo de modelos de *embeddings* para recuperação de documentos jurídicos similares em português. As contribuições incluem: (i) análise comparativa entre modelos BERT de propósito geral, especializados no domínio jurídico e especializados em tarefa de similaridade; (ii) validação de um *pipeline* de *chunking-and-retrieval* como solução de baixo custo e melhor eficácia; (iii) desenvolvimento de uma linha de base para recuperação de precedentes em português. Dessa forma, espera-se contribuir para a triagem automática de petições e fortalecer a aplicação de precedentes judiciais, em alinhamento com os objetivos do Programa Justiça 4.0.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta uma discussão dos trabalhos relacionados. A Seção 3 detalha a metodologia empregada, incluindo o pipeline experimental, o conjunto de dados e o fluxo de avaliação. Na Seção 4, os resultados são apresentados e discutidos. A Seção 5 sintetiza as conclusões, contribuições e aponta direções para pesquisas futuras.

2. TRABALHOS RELACIONADOS

A aplicação de modelos de linguagem em textos jurídicos avançou com a arquitetura *Transformers* e mecanismos de busca vetorial têm ganhado destaque recente na literatura. O estudo de [Karpukhin et al. 2020] demonstrou que, partindo de um modelo BERT, é possível obter um recuperador de alto desempenho com uma função de aprendizagem contrastiva sem a necessidade de etapas de pré-treinamento adicionais e mais complexas. Esta seção sintetiza alguns dos estudos e modelos que servem como base e contraponto para este trabalho, partindo do modelo seminal BERT [Devlin et al. 2019], e seguindo para suas adaptações para a língua portuguesa, para o domínio jurídico específico e para a tarefa de similaridade semântica. Ao contrário de seus predecessores, que já haviam introduzido *embeddings* contextuais, o BERT foi o primeiro a alcançar uma bidirecionalidade profunda em todas as suas camadas [Devlin et al. 2019; Silva et al. 2024]. Gerados a partir de um pré-treinamento massivo em tarefas de linguagem auto-supervisionadas, como o *Masked Language Model* (MLM), esse avanço estabeleceu a base para especializações posteriores em diferentes idiomas e domínios. Para o português brasileiro, um dos trabalhos de maior impacto foi o BERTimbau [2], proposto por [Souza et al. 2020]. Ao ser pré-treinado do zero com um vasto *corpus* do português, o BERTimbau demonstrou

[2] `https://huggingface.co/neuralmind/bert-base-portuguese-cased`

um desempenho superior ao de modelos *multilíngues* em diversas tarefas, tornando-se uma base sólida para a pesquisa no Brasil [Silva et al. 2024].

A partir do BERTimbau, surgiram diversas especializações para o domínio jurídico. Modelos como o BERTikal[3] [Polo et al. 2021], o LegalBERT-pt[4] [Silveira et al. 2023] e o BumbaBert [Carmo et al. 2023] seguiram a abordagem de *continued pre-training* sobre um *checkpoint* do BERTimbau, utilizando documentos legais brasileiros, como petições iniciais, acórdãos e sentenças, para adaptar o vocabulário e a compreensão contextual às nuances da linguagem jurídica. Notavelmente, o BumbaBert incluiu dados do TJMA, alinhando-se ao domínio de aplicação deste trabalho. Os estudos supraditos demonstram um claro movimento da comunidade científica em direção à criação de modelos de linguagem cada vez mais especializados, reconhecendo que a compreensão das nuances de domínios específicos aumenta a eficácia de ferramentas de PLN. Paralelamente à especialização por domínio, outra frente de pesquisa adaptou a arquitetura BERT para tarefas específicas, como a RID por similaridade. Para a RI, a arquitetura original do BERT é computacionalmente ineficiente. A abordagem *Sentence-BERT* (SBERT), proposta por [Reimers and Gurevych], otimiza o modelo através de um *fine-tuning* com redes siamesas para gerar *embeddings* de sentenças diretamente comparáveis por similaridade de cosseno. Para o português, o SBERT-pt[5], é um exemplo dessa técnica, resultando em um modelo especialista na tarefa de produzir vetores para busca semântica, independentemente do domínio.

Esta seção discutiu avanços em classificação com vetorização de textos jurídicos, revelando uma lacuna na aplicação combinada de *embeddings* com mecanismos de recuperação vetorial escaláveis para busca de documentos similares. Embora estudos em RI tenham explorado abordagens híbridas lexicais-semânticas para documentos legais longos, como o trabalho de [Santos et al. 2024] que integra BM25 e BERT, este estudo propõe avaliar diferentes tipos de modelos densos para recuperação de documentos jurídicos por similaridade semântica, utilizando o *ElasticSearch* como infraestrutura de busca. A avaliação foi conduzida em cenário *zero-shot*, na qual não há ajuste fino dos modelos para a tarefa [Wortsman et al. 2022].

3. MATERIAIS E MÉTODOS

Este estudo foi guiado pela metodologia *CRoss-Industry Standard Process for Data Mining* (CRISP-DM), uma abordagem iterativa que estrutura projetos de ciência de dados em seis fases principais [Wirth and Hipp 2000]. As subseções a seguir detalham como cada fase foi aplicada para construir e avaliar um pipeline de recuperação de documentos jurídicos por similaridade semântica.

3.1 Entendimento do Negócio

A fase de Entendimento do Negócio, alinhada aos objetivos da cooperação técnica com o TJMA, consistiu na definição do problema central: a morosidade e a falta de escalabilidade no processo manual de identificação de precedentes judiciais, conforme detalhado na Seção 1. Diante disso, o objetivo deste estudo foi estabelecido como a investigação de técnicas de RID baseadas em similaridade semântica para desenvolver um *pipeline* de recomendação de precedentes judiciais, como apoio à decisão para analistas jurídicos e magistrados. Na Figura 1 é ilustrado o fluxo geral do pipeline.O *pipeline* deve ser capaz de, a partir de uma petição inicial submetida como consulta, recuperar uma lista ordenada de processos anteriores que sejam conceitualmente similares. O critério de sucesso dessa abordagem está, portanto, atrelado à sua capacidade de aumentar a eficiência na triagem processual e promover a consistência na aplicação da jurisprudência.

3.2 Entendimento dos dados

Os dados foram disponibilizados pelo departamento de Tecnologia da Informação (TI) do TJMA em arquivo *Comma-Separated Values* (CSV). A base contém 7.031 instâncias de petições iniciais, sendo

[3]`https://huggingface.co/felipemaiapolo/legalnlp-bert`
[4]`https://huggingface.co/raquelsilveira/legalbertpt_fp`
[5]`https://huggingface.co/rufimelo/bert-large-portuguese-cased-sts`

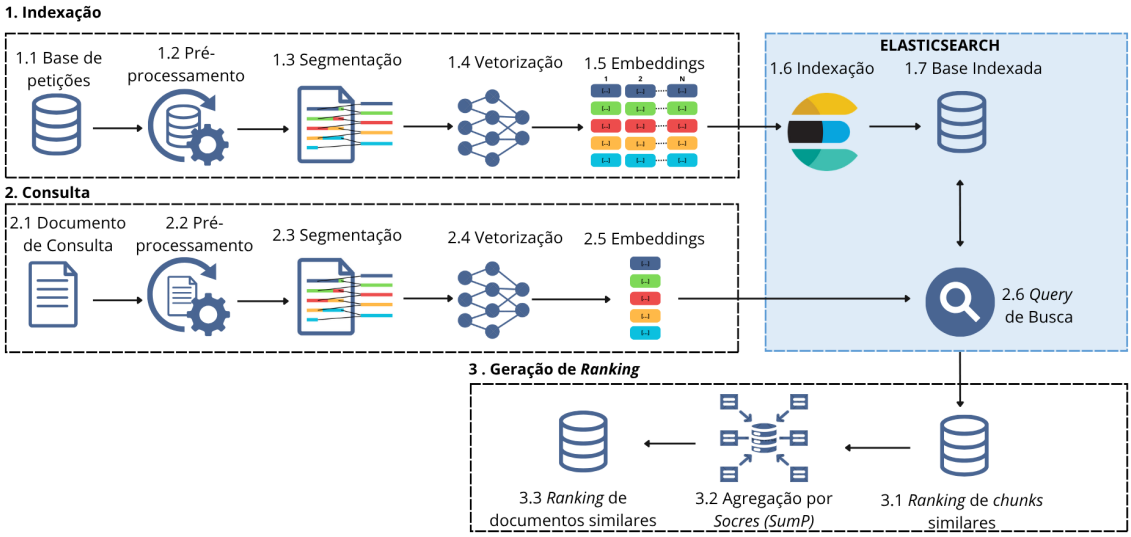

Fig. 1. Fluxograma do *Pipeline* de RID.
Fonte: Elaborado pelo autor (2025).

as colunas de maior relevância o texto completo das petições, o número do processo e o número do precedente que determina a classe de cada petição. Uma análise exploratória revelou desbalanceamento entre as 76 classes do campo tema, seguindo padrão de cauda longa (*long-tail*). Cinco classes concentravam a maioria das petições: IRDR1 (3.576), IRDR5 (1.058), RR986 (759), IRDR3 (470) e IRDR8 (248), enquanto as 71 restantes possuíam frequência muito baixa. Para reduzir os efeitos dessas classes no treinamento e avaliação, adotou-se agrupamento de categorias raras (*lumping*) [He and Garcia 2009]. As cinco classes mais frequentes foram mantidas, e as demais agrupadas na classe DEMAIS (920).

3.3 Preparação dos Dados

Nesta fase, os textos passaram por *pipeline* de preparação iniciado com limpeza textual padrão: remoção de caracteres especiais, ruídos (erros de codificação, textos HTML, quebras de linha desnecessárias, múltiplos espaços, URLs). Com os dados tratados, realizou-se divisão estratificada destinando 70% das petições para indexação e 30% para consultas de teste. Essa separação, feita antes da segmentação, evita *data leakage* e assegurando a validade da avaliação. Posteriormente, todos os documentos foram submetidos à *chunking*, etapa necessária para lidar com o limite de entrada dos modelos *Transformer* [Karpukhin et al. 2020]. Devido à limitação de 512 *tokens* dos modelos BERT, cada petição foi dividida em segmentos de até 480 *tokens*, com sobreposição de 100 *tokens* entre segmentos consecutivos para evitar perda de contexto semântico. Cada segmento manteve referência ao número do processo e tema da petição original para as etapas de indexação e avaliação.

3.4 Modelagem

A fase de Modelagem consistiu na aplicação dos modelos de linguagem pré-selecionados para converter os segmentos textuais em *embeddings*. Este processo foi executado de forma distinta para as diferentes categorias de modelos, a fim de respeitar suas arquiteturas específicas, e culminou na indexação desses vetores para a tarefa de busca. Para conduzir a análise comparativa, foram selecionados cinco modelos de linguagem pré-treinados para o português, todos baseados na arquitetura *BERT*. Estes modelos foram agrupados em três categorias funcionais, permitindo isolar e avaliar o impacto da especialização de domínio *versus* a especialização de tarefa na qualidade da busca semântica.

Modelo de Propósito Geral (*Baseline*): Como linha de base, foi utilizado o BERTimbau (*base*) [Souza et al. 2020], um treinado com corpus na língua portuguesa sem exposição prévia ao vocabulário

jurídico. **Modelos de Domínio Específico (Jurídico)**: Para avaliar o impacto do conhecimento de domínio, foram selecionados três modelos que passaram por pré-treinamento continuado em vastos *corpora* jurídicos brasileiros: o BERTikal [Polo et al. 2021], o LegalBERT-pt [Silveira et al. 2023] e o BumbaBert-small-SC [Carmo et al. 2023]. Cada um representa um esforço para adaptar o vocabulário e a compreensão contextual do modelo às nuances da linguagem legal. **Modelo de Tarefa Específica (Similaridade Semântica)**: Para avaliar o impacto da otimização para a tarefa de similaridade, foi selecionado o SBERT-pt [Reimers and Gurevych], modelo baseado na arquitetura Sentence-BERT para gerar *embeddings* otimizados para comparação direta via similaridade de cosseno, embora tenha sido treinado em um *corpus* de domínio geral.

Foi criado um índice dedicado no *ElasticSearch* para cada um dos modelos. Os *embeddings* gerados para todos os segmentos do *corpus* de indexação foram armazenados nestes índices em um campo do tipo *"dense vector"*, juntamente com seus metadados *"número de identificação do segmento", "número do processo" e "classe precedente"*. A dimensão desse campo em cada índice foi definida dinamicamente, inferindo-se o tamanho do vetor diretamente a partir dos *embeddings* gerados por cada modelo, assegurando a modularidade dos índices para as diferentes arquiteturas avaliadas (*e.g.,* BumbaBERT: 512; BERTimbau, BERTikal e LegalBERT-pt: 768 ; e SBERT-pt: 1024). Considerando que o objetivo do trabalho também incluía desenvolver uma solução de baixo custo, todos os experimentos foram conduzidos utilizando a plataforma *Google Colab Pro+*, em ambiente privado, com os experimentos sendo realizados em menos de um mês. As indexações podem ser utilizadas localmente, usando *Docker*, pois o maior custo computacional do *pipeline* reside na geração dos *embeddings*.

3.5 Avaliação

Dada a ausência de um *corpus* de teste com julgamentos de relevância manuais, ou *gold standard*, foi desenvolvido um desenho experimental baseado em *weak supervision*. Nesta abordagem, as etiquetas da coluna referente à classe do precedente são utilizadas como um indicador substituto (*proxy*), uma prática utilizada quando não disponibilizada anotação manual [Karpukhin et al. 2020].

O experimento foi conduzido de forma idêntica, para cada um dos modelos, para fins de comparação. O processo consistiu nas seguintes etapas: i) **Definição do Conjunto de Consultas (teste):** O conjunto de teste, composto por 30% das petições iniciais, foi utilizado como fonte de consultas. Cada petição neste conjunto serviu, uma a uma, como um documento de consulta ao sistema; ii) **Execução da Consulta:** Para cada petição de consulta do conjunto de teste, o mesmo processo de preparação de dados foi aplicado: o texto foi segmentado em *chunks* e os *embeddings* de cada *chunk* foram gerados utilizando o modelo correspondente ao índice que estava sendo avaliado; iii) **Agregação de Similaridade e Geração do *Ranking*:** A busca foi realizada passando cada *chunk* da petição de consulta pelo respectivo índice no *ElasticSearch*. Para a geração do *ranking* de documentos, adotou-se a estratégia de agregação *Sum of Passages* (SumP) [Ku et al. 2005]. Ela mitiga desbalanceamento em documentos longos, ao consolidar os *scores* de similaridade dos *chunks* em uma única pontuação por documento. O *score* final de cada documento corresponde à soma dos *scores* de todos os seus *chunks* presentes nos resultados da busca. A partir desse *score* agregado, os documentos são ordenados no *ranking*. Em particular a *SumP* valoriza múltiplas passagens relevantes dentro de um mesmo documento, superando limitações de outras estratégias, como o *Maximum Pooling (MaxP)*, que considera apenas o *chunk* mais relevante, e o *Average Pooling (AvgP)*, que tende a diluir a pontuação em documentos extensos. Para avaliar a qualidade dos *rankings* gerados por cada modelo, foram consideradas três métricas padrão da literatura de RI [Schütze et al. 2008], especialmente adequadas para avaliar a ordem dos resultados, mesmo em bases com desbalanceamento de classe [Ricardo and Berthier 2011]. **Precisão no Topo K (*Precision@k* ou *P@k*):** Mede a proporção de documentos relevantes entre os k primeiros resultados retornados pelo *pipeline*. Esta métrica avalia a capacidade do modelo de posicionar documentos relevantes nas primeiras posições do *ranking*. Neste trabalho, foi calculada para $k \in \{1, 5, 10, 15\}$, permitindo analisar o desempenho do *pipeline* em diferentes níveis de corte. ***Mean Reciprocal Rank* (MRR):** Avalia a capacidade do sistema de retornar o primeiro resultado correto o mais alto possível no *ranking*. O *MRR* é a média do inverso da posição do primeiro

acerto para todas as consultas. É particularmente útil em cenários onde o usuário pode se satisfazer ao encontrar rapidamente um único documento relevante. ***Mean Average Precision* (MAP):** É uma métrica mais completa que avalia a qualidade geral do *ranking*, considerando a ordem de **todos** os documentos relevantes encontrados. A *MAP* calcula a média das precisões em cada ponto em que um item relevante é recuperado, recompensando sistemas que não apenas encontram muitos itens corretos, mas que também os posicionam no topo da lista. É considerada uma das métricas mais estáveis para comparar o desempenho geral de sistemas de recuperação [Schütze et al. 2008].

3.6 Implantação

Embora a fase final do CRISP-DM, a implantação, esteja fora do escopo deste artigo, os resultados desta avaliação comparativa fornecem subsídios técnicos para uma futura integração desta solução de busca semântica ao fluxo de trabalho do TJMA, em alinhamento com os objetivos do Acordo de Cooperação. A entrega se deu por meio de relatório técnico e também disponibilização dos artefatos, em conformidade com o Acordo com o TJMA.

4. RESULTADOS E DISCUSSÕES

Esta seção apresenta os resultados dos experimentos de RID com os cinco modelos de *embedding* avaliados, analisando comparativamente o impacto da especialização por domínio e por tarefa na qualidade das buscas. As métricas de desempenho empregadas avaliam especificamente a qualidade do ordenamento dos documentos retornados, indicando a capacidade dos modelos de posicionar os resultados mais relevantes no topo da lista. N

a Tabela I estão sumarizados os resultados obtidos nesta avaliação, revelando um desempenho geral competitivo entre todos os modelos, mas com uma vantagem consistente para o SBERT-pt ($77,96\%$), que é especializado na tarefa de similaridade.

Table I. Resultados das avaliações dos modelos com média geral.

Modelo	**P@1**	**P@5**	**P@10**	**P@15**	**MRR**	**MAP**	**Média Geral (%)**
SBERT-pt (Tarefa)	**0,7872**	**0,7623**	0,7406	**0,7221**	**0,8592**	**0,8060**	**77,96**
BERTikal (Domínio)	0,7744	0,7573	0,7377	0,7154	0,8497	0,8018	77,27
LegalBERT-pt (Domínio)	0,7782	0,7600	**0,7411**	0,7214	0,8526	0,8048	77,64
BumbaBert (Domínio)	0,7801	0,7548	0,7366	0,7198	0,8523	0,8012	77,41
BERTimbau (Geral)	0,7777	0,7578	0,7369	0,7184	0,8514	0,8012	77,39

Conforme explicitado na Tabela I, o SBERT-PT alcançou o melhor desempenho em quase todas as métricas, incluindo as mais críticas para a RI, com uma superioridade de $0,91\%$ em P@1, $0,77\%$ em MRR e $0,15\%$ em MAP sobre o segundo colocado em cada métrica. Este achado sugere que, para a tarefa de busca semântica, a otimização da arquitetura do SBERT para gerar vetores de similaridade teve um impacto mais determinante do que a especialização do vocabulário no domínio jurídico. Ao comparar os modelos de domínio específico com a linha de base (BERTimbau), observa-se que todos apresentaram um desempenho muito próximo ou ligeiramente superior. O LegalBERT-pt e o BumbaBert, por exemplo, superaram o BERTimbau em P@1, confirmando a hipótese de que o conhecimento do vocabulário jurídico é benéfico. No entanto, a magnitude dessa melhoria foi marginal em métricas mais abrangentes como o MAP.

O resultado mais significativo refere-se à comparação entre a especialização de tarefa e a de domínio. Embora as variações nos valores de MAP sejam mínimas, o SBERT-pt apresentou melhor desempenho no topo do *ranking*, com destaque para P@1 ($0,7872$ contra $0,7801$ do BumbaBert) e MRR ($0,8592$ contra $0,8526$ do LegalBERT-pt). Essa capacidade de posicionar o documento mais relevante em primeiro lugar com maior frequência evidencia o impacto positivo da especialização de tarefa sobre a de domínio. Esse achado alinha-se com as conclusões de [Silva Junior et al. 2025], cujos autores também observaram que um modelo SBERT obteve a maior correlação com rótulos de similaridade atribuídos por especialistas, validando a eficácia da especialização de tarefa [Silva Junior et al. 2025]. No presente trabalho a única exceção notável foi na métrica P@10, onde o LegalBERT-pt obteve uma

ligeira vantagem, indicando sua robustez na recuperação de um conjunto um pouco mais amplo de documentos relevantes.

4.1 Ameaças à validade

É importante destacar que estes resultados foram obtidos utilizando as classes de precedentes como um indicador substituto (*proxy*) para a relevância. Estudos futuros podem incorporar uma validação qualitativa com especialistas da área para corroborar estes achados quantitativos e explorar o impacto desses modelos em um ambiente de produção. Outra limitação reside no agrupamento das classes minoritárias. Em trabalho futuros pretendesse explorar técnicas de *few-shot learning* ou métodos de reamostragem para aprimorar a capacidade do *pipeline* de distinguir entre precedentes com poucos exemplos de treinamento. Quanto etapa de agregação utilizou-se exclusivamente SumP. Embora essa abordagem favoreça documentos realmente relevantes, representa um *trade-off* metodológico. Trabalhos futuros avaliarão estratégias alternativas como MaxP e AvgP para identificar a abordagem mais eficaz para diferentes tipos de documentos jurídicos. Além disso, a avaliação foi limitada aos 15 melhores resultados retornados ($k = 15$), o que significa que as métricas de MRR e MAP reportadas referem-se, estritamente, MRR@15 e MAP@15. Essa escolha, entretanto, alinha-se com o objetivo prático de avaliar a capacidade do *pipeline* em apresentar, de forma eficiente, os resultados mais relevantes nas primeiras posições.

5. CONSIDERAÇÕES FINAIS

Este artigo apresentou um estudo empírico comparativo para avaliar a eficácia de diferentes classes de modelos de *embedding* na recuperação de documentos jurídicos. O objetivo foi analisar o desempenho desses modelos na busca por similaridade semântica, utilizando um *pipeline* composto por etapas de segmentação, indexação e busca vetorial. Foram testados modelos de propósito geral (BERTimbau), de domínio jurídico (BERTikal, BumbaBERT, LegalBERT-pt) e de tarefa específica (SBERT-pt), em um *corpus* de petições iniciais do TJMA. Os resultados indicaram que todos os modelos testados apresentaram desempenho competitivo, confirmando a viabilidade da busca semântica como uma abordagem eficaz para a identificação de precedentes. Destaca-se, em especial, a superioridade consistente do modelo SBERT-pt, especializado na tarefa de similaridade textual.

Esses achados têm implicações diretas para a implementação prática de sistemas de IA no Judiciário. Eles demonstram que, embora todos os modelos avaliados sejam tecnicamente capazes, a escolha de uma arquitetura otimizada para similaridade semântica, como o Sentence-BERT, representa a estratégia mais promissora para maximizar a precisão nos primeiros resultados retornados. Em um cenário prático, o *ranking* resultante traduz-se em uma lista ordenada dos precedentes mais prováveis, permitindo que magistrados e analistas jurídicos não apenas localizem documentos similares, mas também obtenham uma visão consolidada dos precedentes aplicáveis, agilizando a triagem de casos e apoiando análises orientadas pela jurisprudência. Para trabalhos futuros, destaca-se a ampliação da avaliação para um *corpus* mais diverso, incluindo outros tipos de peças processuais, como acórdãos, e documentos de diferentes tribunais brasileiros. Essa expansão visa testar a robustez e a capacidade de generalização dos modelos. Além disso, estudos qualitativos com especialistas jurídicos poderão validar a relevância prática dos resultados e seu impacto em fluxos reais de trabalho no Judiciário.

Agradecimentos e Uso de IA generativa

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-DT-303031/2023-9, POSDOC - 101057/2024-5; e Acordo de Cooperação Técnica N° 02/2021 (Processo N° 38328/2020 -TJ/MA).

Declara-se que os modelos de IA generativa *Gemini 2.5* e *GPT-4* foram utilizados como ferramentas de apoio, exclusivamente para a revisão gramatical e o aprimoramento do desenho da pesquisa. A autoria e a responsabilidade integral pelo conteúdo final, incluindo a verificação de plágio e correções, são de Adrielson Ferreira Justino.

REFERENCES

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer, 2020.

Brasil. Lei nº 13.105, de 16 de março de 2015. código de processo civil., 2015.

Carmo, F. A., Serejo, F., Junior, A. F. J., Santana, E. E., and Lobato, F. M. Embeddings jurídico: Representações orientadas à linguagem jurídica brasileira. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, 2023.

de Souza, C. M. F. and de Souza Salles, S. Acesso à justiça em tempos de pandemia: a experiência do núcleo permanente de métodos consensuais de tratamento de conflitos do tjrj. *Conhecimento & Diversidade*, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019.

Harispe, S., Ranwez, S., Montmain, J., et al. *Semantic similarity from natural language and ontology analysis*, 2022.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 2009.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. Unsupervised dense information retrieval with contrastive learning, 2022.

Karpukhin, V., Oguz, B., Min, S., Lewis, P. S., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, 2020.

Ku, L.-W., Wu, T.-H., Lee, L.-Y., and Chen, H.-H. Construction of an evaluation corpus for opinion extraction. In *NTCIR*, 2005.

Magalhães, R. A. and Freitas, F. O. A morosidade do poder judiciário e sua interferência nas relações contratuais. *Revista Jurídica Cesumar-Mestrado*, 2023.

Maia Filho, M. S. and Junquilho, T. A. Projeto victor: perspectivas de aplicação da inteligência artificial ao direito. *Revista de Direitos e Garantias Fundamentais*, 2018.

Marinato, M. S., Junior, A. F. J., Lobato, F. M., and Cortes, O. A. Classificação automática de petições iniciais usando classificadores combinados. In *Anais do XVI Brazilian e-Science Workshop*, 2022.

Ni, C., Wu, J., Wang, H., Lu, W., and Zhang, C. Enhancing cloud-based large language model processing with elasticsearch and transformer models. In *Proceedings of the International Conference on Image, Signal Processing, and Pattern Recognition (ISPP)*, 2024.

Polo, F. M., Mendonça, G. C. F., Parreira, K. C. J., Gianvechio, L., Cordeiro, P., Ferreira, J. B., de Lima, L. M. P., do Amaral Maia, A. C., and Vicente, R. Legalnlp-natural language processing methods for the brazilian legal language. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, 2021.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks.

Ricardo, B.-Y. and Berthier, R.-N. Modern information retrieval: the concepts and technology behind search, 2011.

Ruckdeschel, M. Term-based and embedding-based similarity search in large unknown text datasets, 2020.

Santos, J. A., Souza, E., Filho, C. J. B., Albuquerque, H. O., Vitório, D., de Lucena, D. C. G., Silva, N., and de Carvalho, A. Hirs: A hybrid information retrieval system for legislative documents, 2024.

Schütze, H., Manning, C. D., and Raghavan, P. *Introduction to information retrieval*, 2008.

Silva, E. C., de Medeiros, I. P., de Menezes, M. V., and Kamikawachi, D. S. L. Segmentation and summarization for extracting information about information technology equipment from government procurement notice. In *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, 2024.

Silva Junior, D. d., Oliveira, D. d., and Paes, A. Evaluating text representations for unsupervised legal semantic textual similarity in brazilian portuguese. *Discover Data*, 2025.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., and Furtado, V. Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Brazilian Conference on Intelligent Systems*, 2023.

Souza, F., Nogueira, R., and Lotufo, R. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, 2020.

Stemler, I. T. S. V. Identificação de precedentes judiciais por agrupamento utilizando processamento de linguagem natural, 2019.

Toffoli, J. A. D. and Gusmão, B. G. Inteligência artificial na justiça. *Brasília: CNJ*, 2019.

Wirth, R. and Hipp, J. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 2000.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models, 2022.