Evaluating BERT Models for Semantic Retrieval in Long Portuguese Legal Documents

Adrielson Ferreira Justino[1], **Antônio Fernando Lavareda Jacob Junior**[1], **Ricardo Marcondes Marcacini**[3], **Fábio Manoel França Lobato**[1,2,3]

[1]Universidade Estadual do Maranhão (UEMA) – São Luís – MA – Brazil

[2]Universidade Federal do Oeste do Pará (UFOPA) – Santarém – PA – Brazil

[3]Instituto De Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP) – São Carlos – SP – Brazil

adrielferreira28@gmail.com, fabio.lobato@ufopa.edu.br

Abstract. *The growing number of digital documents in the Brazilian judiciary creates new challenges to procedural efficiency. This study evaluated five BERT models for dense information retrieval from long court documents, utilizing segmentation and vector retrieval with Elasticsearch. General-purpose, domain-specific, and task-specific models were tested to measure the intra-cluster coherence. BumbaBERT (domain-specific) performed best, confirming that domain specialization is crucial for effective semantic retrieval in "zero-shot" scenarios in the Brazilian legal context.*

Resumo. *O crescente número de documentos digitais no Judiciário brasileiro cria novos desafios para a eficiência processual. Este estudo avaliou cinco modelos BERT na recuperação de informações densas para documentos judiciais longos, utilizando a segmentação e a recuperação de vetores com o Elasticsearch. Modelos de uso geral, específicos de domínio e específicos de tarefa foram testados para medir a coerência intra-cluster. O BumbaBERT (específico de domínio) teve o melhor desempenho, confirmando que a especialização de domínio é crucial para a recuperação semântica eficaz em cenários de "zero-shot" no contexto jurídico brasileiro.*

1. Introdução

A implementação do Processo Judicial Eletrônico (PJe) provocou uma digitalização massiva no sistema judiciário brasileiro, resultando em mais de 83,8 milhões de processos pendentes segundo o Conselho Nacional de Justiça (CNJ) [CNJ 2024]. Esse cenário desafia a eficiência judicial, contribuindo para morosidade e dificuldades na identificação de informações críticas [Magalhães and Freitas 2023]. Entre esses documentos, destacam-se os acórdãos judiciais, documentos jurídicos que relatam decisões de órgãos colegiados como as turmas recursais dos Tribunais de Justiça estaduais, Superior Tribunal de Justiça (STJ) e Supremo Tribunal Federal (STF), previstas no Art. 163 do Código de Processo Civil (CPC). Os acórdãos resultam da análise de recursos ou competência originária, tornando-se referências para casos similares [Guimarães 2004]. Embora fundamental para advogados e magistrados, sua extensão, linguagem técnica complexa e volume crescente tornam a busca manual ineficiente e propensa a erros [Toffoli and Gusmão 2019].

Diante desses desafios, o Programa Justiça 4.0 do CNJ impulsiona a modernização com Inteligência Artificial (IA).Abordagens de Processamento de Linguagem Natural (PLN) podem transformar grandes volumes de dados textuais não estruturados em representações computáveis, automatizando a classificação e organização por similaridade [Thakur et al. 2021]. Entretanto, o processamento de textos jurídicos apresenta desafios únicos: documentos longos, não estruturados e não rotulados dificultam a categorização manual [Costa and Dantas 2023]. Nesse contexto, algoritmos de aprendizado não supervisionado apresentam-se como alternativa para capturar padrões latentes [Oliveira and Sperandio Nascimento 2025]. Em especial, técnicas de agrupamento permitem descobrir grupos de documentos similares sem *ground truth* anotado, otimizando o fluxo de trabalho e promovendo consistência jurisprudencial [Scherrer et al. 2018].

Nesse cenário, a Recuperação de Informação Densa (RID) utiliza representações vetoriais densas de *Pre-trained Language Models* (PLM), como os da família *Bidirectional Encoder Representations from Transformers* (BERT) [Harispe et al. 2022, Devlin 2018]. Esses *embeddings* integrados em motores de busca modernos permitem Busca Híbrida, combinando precisão lexical com profundidade semântica em larga escala, destacando-se soluções baseadas em *Elasticsearch* [Ni et al. 2024]. A aplicação de RID em documentos longos enfrenta limitações dos modelos *Transformers*, como o BERT restrito a 512 *tokens* [Devlin 2018]. Embora existam alternativas para contextos longos como *Longformer*, seu alto custo computacional reduz a acessibilidade [Beltagy et al. 2020]. A segmentação de texto (*chunking*) apresenta-se como solução [Gao et al. 2023]. Esta técnica consiste em dividir os documentos em fragmentos menores, garantindo a compatibilidade com a janela de contexto fixa dos modelos e representando uma alternativa computacionalmente eficiente a arquiteturas mais custosas [Gao et al. 2023].

Apesar da proliferação de modelos de *embedding* para o português, sejam de propósito geral como BERTimbau[1] [Souza et al. 2020], de domínio específico como LegalBERT-pt[2] [Silveira et al. 2023] ou de tarefa específica como SBERT [Reimers and Gurevych 2019], percebe-se uma lacuna na literatura quanto avaliação de desempenho da recuperação de documentos jurídicos longos. A avaliação de modelos sob essa condição, na qual não há ajuste fino para a tarefa, caracteriza um cenário de avaliação *zero-shot* [Wortsman et al. 2022]. Diante disso, o objetivo deste estudo é investigar o desempenho de modelos de *embedding* baseados em BERT na RID de documentos longos, utilizando acórdãos judiciais como estudo de caso. Para tal, a avaliação por métricas de coerência semântica *intra-cluster*, quantificando proximidade semântica entre documentos recuperados sem dependência de *ground truth* manual.

As contribuições incluem: i) aplicação de fluxo metodológico experimental para avaliar coerência semântica em cenários não supervisionados; ii) investigação do desempenho de modelos BERT-like para documentos jurídicos longos (de propósito geral, o BERTimbau [Souza et al. 2020]; domínio específico, o BERTikal[3] [Polo et al. 2021], LegalBERT-pt [Silveira et al. 2023] e BumbaBert [do Carmo et al. 2023]; e tarefa específica o SBERT-pt[4]); e iii) demonstração de *pipeline* de *chunking* e RID em *Elastic-*

[1] https://huggingface.co/neuralmind/bert-base-portuguese-cased
[2] https://huggingface.co/raquelsilveira/legalbertpt_fp
[3] https://huggingface.co/felipemaiapolo/legalnlp-bert
[4] https://huggingface.co/rufimelo/bert-large-portuguese-cased-sts

search para exploração eficiente de jurisprudência não rotulada.

O restante do artigo está organizado como segue. Na Seção 2 são discutidos trabalhos relacionados. A Seção 3 detalha metodologia, dados e *framework* de avaliação. Os resultados são apresentados e discutidos na Seção 4. Por fim, as conclusões, contribuições e direções futuras são apresentados na Seção 5.

2. Trabalhos Relacionados

A recuperação de informação baseada em PLM evoluiu rapidamente, estabelecendo novos paradigmas para a busca semântica. Conforme revisado por [Zhao et al. 2024], o campo pode ser analisado sob quatro aspectos principais: arquitetura, treinamento, indexação e integração. Um dos avanços arquitetônicos centrais foi a popularização da abordagem *dual-encoder*, que mapeia consultas e documentos para representações vetoriais únicas [Karpukhin et al. 2020]. Alternativamente, a abordagem de "interação tardia" [Khattab and Zaharia 2020] calcula a relevância de forma mais granular, a nível de *token*. Embora estejam no estado da arte, esses métodos são frequentemente avaliados em cenários com dados para *fine-tuning* e para a recuperação de trechos curtos. Conforme destacado por [Wortsman et al. 2022], a aplicação desses modelos em novos domínios ou tarefas sem dados de treinamento específicos levanta o desafio da avaliação "*zero-shot*". Os autores exploraram soluções utilizando *Large Language Models* (LLMs) para fornecer *feedback* de relevância e refinar as consultas. Outro desafio central é o tratamento de documentos longos, que excedem o limite de contexto dos Transformers. A segmentação de texto (*chunking*) é a técnica prevalente para contornar essa limitação, e seu impacto na eficácia da recuperação é um tópico de pesquisa ativo, com estudos recentes analisando como o tamanho dos *chunks* influencia o desempenho [Gao et al. 2023].

No contexto da língua portuguesa, um dos trabalhos de maior impacto foi o BERTimbau [Souza et al. 2020]. Baseando-se no BERT [Devlin 2018], os autores treinaram o modelo usando um extenso *corpus* em português (pt-br), demonstrando desempenho superior ao de modelos multilíngues e tornando-se uma base sólida para a pesquisa no Brasil. A partir do dele surgiram especializações para o domínio jurídico, como BERTikal [Polo et al. 2021], LegalBERT-pt [Silveira et al. 2023] e BumbaBert [do Carmo et al. 2023], que seguiram a abordagem de *continued pre-training* em documentos legais para adaptar o vocabulário e a compreensão contextual às nuances do direito. Outra frente adaptou a arquitetura BERT para tarefas específicas, como a similaridade semântica. O *Sentence-BERT* (SBERT) [Reimers and Gurevych 2019] otimiza o modelo com redes siamesas para gerar *embeddings* diretamente comparáveis, e o SBERT-pt é um exemplo dessa técnica para o português.

Os trabalhos revisados demonstram o progresso da área e fornecem subsídios para construção do presente trabalho ao abordar decisões nos quatro pilares da RID [Zhao et al. 2024]. No pilar da arquitetura focou-se na abordagem *bi-encoder*, a mais comum para recuperação em larga escala [Karpukhin et al. 2020]; adotou-se um cenário de treinamento "*zero-shot*" [Wortsman et al. 2022], um desafio central na área; na indexação, a metodologia se apoia no uso do *Elasticsearch* como solução de busca vetorial de alta performance [Zhao et al. 2024, Ni et al. 2024]; e na integração foi implementado um *pipeline* que combina *chunking* e agregação de *score* para documentos longos [Gao et al. 2023]. Ao fazer isso, nosso estudo aborda uma lacuna específica na litera-

tura de PLN para o português. Enquanto esforços no domínio jurídico brasileiro têm se concentrado em tarefas de *clustering* [Costa and Dantas 2023], a avaliação da qualidade de *rankings* de recuperação para documentos longos em um cenário "*zero-shot*" permanece uma área menos explorada. Desta forma, este trabalho contribui ao oferecer uma investigação do desempenho de diferentes classes de modelos *BERT-like*, que dispensa a necessidade de anotação manual de relevância, um processo de alto custo, com trabalho intenso e propenso a erros.

3. Metodologia

O desenvolvimento deste trabalho foi guiado pela metodologia *CRoss-Industry Standard Process for Data Mining* (CRISP-DM), uma abordagem iterativa e flexível que estrutura projetos de ciência de dados em seis fases principais [Wirth and Hipp 2000]. Esta metodologia, reconhecida por sua aplicabilidade em diversos domínios, incluindo o jurídico [Costa and Dantas 2023], permitiu uma condução sistemática da pesquisa. O ciclo do CRISP-DM é composto por seis etapas, a saber: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem e Avaliação. Cada fase é detalhada de forma a se ter um *pipeline* de recuperação semântica de documentos judiciais nas subseções a seguir.

3.1. Fluxo metodológico

O **Entendimento de Negócio** envolveu a compreensão do domínio jurídico e auxiliou na delimitação do problema de pesquisa. O volume de documentos e complexidade na identificação de informações relevantes representam desafios significativos para a eficiência judicial brasileira [CNJ 2024]. A ausência de metadados estruturados limita métodos supervisionados de RID [Costa and Dantas 2023]. Conforme mencionado, o objetivo é aplicar *Dense Retrieval* com modelos *BERT-like* para recuperação semântica, superando limitações lexicais e da falta de rótulos. O critério de sucesso é demonstrar que, mesmo sem supervisão, é possível obter *rankings* com alta coesão semântica, promovendo segurança jurídica, uniformização da jurisprudência e agilidade processual.

Para a etapa de **Entendimento dos Dados** foram utilizados dados disponibilizados pelo Tribunal de Justiça do Maranhão (TJMA) em formato *JavaScript Object Notation* (JSON), totalizando aproximadamente 100 mil acórdãos. Mantiveram-se apenas as colunas essenciais, como o número do processo, ementa e inteiro teor com tamanho médio 1.846 *tokens* por documento, conforme ilustrado na Figura 1. O conjunto não contém rótulos de relevância, reforçando a necessidade de abordagem não supervisionada.

A **Preparação dos Dados** envolveu remoção de processos com menos de 100 caracteres no inteiro teor para eliminar registros incompletos, seguida de filtragem de duplicatas por número do processo. Após estas etapas, obtiveram-se 82.084 documentos válidos. Devido ao custo computacional com modelos *Transformer* [Beltagy et al. 2020], realizou-se amostragem aleatória estratificada. Com 99% de confiança e 2% de margem de erro, calculou-se amostra de 3.960 documentos, assegurando representatividade e minimizando viés de seleção. Posteriormente, os textos dos acórdãos passaram por *pipeline* de limpeza padrão, incluindo a remoção de caracteres especiais e de *stopwords*, remoção de ruídos (*e.g.,* erros de codificação, *tags* de marcação, quebras desnecessárias, múltiplos espaços e *hiperlinks*). Realizou-se divisão estratificada destinando 90% para

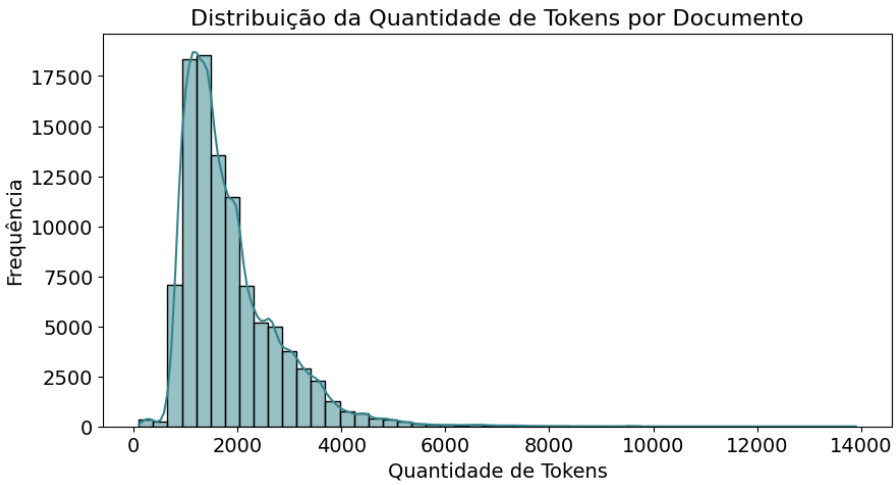

Figura 1. Histograma da quantidade de *tokens* por documento

base de indexação e 10% para consultas de teste, evitando vazamento de dados. Todos os documentos foram segmentados (*chunking*) devido ao limite de 512 *tokens* dos modelos BERT [Karpukhin et al. 2020]. Cada documento foi dividido em *chunks* de até 480 *tokens*, garantindo que a sequência final não excedesse o limite mesmo após a adição de *tokens* especiais pelos modelos BERT, como *Classification Token* (CLS) e *Separator Token* (SEP). Foi aplicada uma sobreposição de 100 *tokens* entre segmentos consecutivos para preservar o contexto semântico, mantendo-se a referência ao processo original para posterior agregação dos resultados.

A fase de **Modelagem** envolveu a conversão dos segmentos textuais em *embeddings* vetoriais densos utilizando modelos de linguagem pré-selecionados. O processo respeitou as arquiteturas específicas de cada modelo e culminou na indexação vetorial para busca semântica em larga escala. Foram selecionados cinco modelos pré-treinados para português, baseados na arquitetura *BERT* [Devlin 2018], organizados em três categorias funcionais para avaliar o impacto da especialização de domínio *versus* especialização de tarefa na recuperação semântica:

- **Modelo de Propósito Geral (*Baseline*)**: Como *baseline* foi utilizado o BERTimbau (*base*) [Souza et al. 2020]. Este modelo é robusto e possui profundo conhecimento da língua portuguesa geral, sendo pré-treinado em um vasto *corpus* sem exposição prévia ao vocabulário jurídico específico. Ele serve para estabelecer um ponto de referência para o desempenho dos modelos [Pires et al. 2024];
- **Modelos de Domínio Específico (Jurídico)**: Para avaliar o impacto do conhecimento de domínio, foram selecionados três modelos que passaram por pré-treinamento continuado (*continued pre-training*) em vastos *corpora* jurídicos brasileiros. São eles: o BERTikal [Polo et al. 2021], o LegalBERT-pt [Silveira et al. 2023] e o BumbaBert-small-SC [do Carmo et al. 2023]. Cada um representa um esforço da comunidade científica para adaptar o vocabulário e a compreensão contextual de um modelo pré-treinado às nuances do domínio, visando aprimorar a representação semântica de documentos jurídicos;
- **Modelo de Tarefa Específica (Similaridade Semântica)**: Para avaliar o impacto da otimização para a tarefa de similaridade textual, foi selecionado o

SBERT-pt [Reimers and Gurevych 2019], o qual utiliza a arquitetura *Sentence-BERT* (SBERT). O método otimiza o processo de geração de *embeddings* para que sejam diretamente comparáveis por similaridade de cosseno, mesmo que em um *corpus* de domínio geral.

Para cada modelo testado, foi estabelecido um índice específico no *Elasticsearch*. As representações vetoriais produzidas para todos os *chunks* do *corpus* de documentos foram armazenadas nesses índices em um campo do tipo *"dense_vector"*. Além das representações vetoriais, foram preservadas informações complementares necessárias para as etapas de busca e consolidação, incluindo numero do processo, ementa e o texto completo do acórdão original, para consulta e compilação. A dimensão desse campo vetorial em cada índice foi estabelecido dinamicamente, determinando o tamanho do vetor a partir dos *embeddings* gerados por cada arquitetura (512 para o BumbaBERT; 768 para os modelos BERTimbau, BERTikal e LegalBERT-pt; e 1024 para o SBERT-pt). Essa estratégia garantiu a flexibilidade dos índices para as diferentes arquiteturas avaliadas.

Na etapa de **Avaliação** da qualidade das recuperações de informação o desempenho dos diferentes modelos de *embeddings* foram medidos com base na coerência semântica *intra-cluster*. Considerando a ausência de um conjunto de referência manual (*gold standard*) com anotações de relevância entre acórdãos, essa abordagem é amplamente reconhecida e recomendada na literatura para cenários onde dados rotulados são escassos ou inexistentes [Harispe et al. 2022, Schütze et al. 2008]. Ao inferir a qualidade semântica dos *rankings* gerados a partir da similaridade interna entre os documentos recuperados, é possível validar a capacidade de representação dos modelos em uma tarefa de *downstream* sem a necessidade de supervisão humana [Harispe et al. 2022, Oliveira and Sperandio Nascimento 2025]. É fundamental ressaltar que, ao se comparar diferentes modelos de *embedding*, cada um opera em um espaço vetorial distinto. Consequentemente, a análise não se baseia na comparação de valores absolutos de similaridade entre os modelos, mas sim na consistência do desempenho relativo de cada um a partir das diferentes profundidades de *ranking* (K) e na magnitude das diferenças observadas dentro de cada espaço vetorial.

A última de etapa envolve a **Entrega** do modelo. Para o presente trabalho, os resultados da investigação de desempenho dos modelos de *embeddings* e o fluxo de avaliação de coerência semântica intrínseca fornecem subsídios técnicos fundamentais para o desenvolvimento e a futura integração de soluções de RID em fluxos de trabalho do Judiciário. Ao identificar os modelos que geram *rankings* mais coesos de documentos, este trabalho oferece diretrizes para sistemas de apoio à decisão, busca jurisprudencial e organização automática de documentos. Essa aplicação contribui diretamente para a otimização da eficiência e celeridade processual em um cenário real, alinhando-se aos objetivos de modernização e inovação do setor jurídico.

3.2. *Framework* experimental

O experimento foi realizado de forma idêntica para cada modelo avaliado, garantindo investigação consistente e controlada dos resultados. As etapas principais do *framework* foram sistematicamente executadas como descrito a seguir.

Preparação da Base de Indexação: Para cada modelo de linguagem avaliado, foi criado um índice dedicado no *Elasticsearch*, contendo os *embeddings* dos segmentos

textuais (*chunks*) dos documentos que compõem a *Base de Indexação*, correspondente a 90% da amostra total de acórdãos.

Definição da Base de Consultas: Os 10% restantes da amostra total foram utilizados para compor a *Base de Consultas*. Cada documento dessa base também foi segmentado em *chunks*, e os *embeddings* correspondentes foram utilizados como entradas nas buscas por similaridade, realizadas exclusivamente sobre a *Base de Indexação*.

Execução das Consultas de Similaridade: Para cada *chunk* do documento-consulta, foi realizada uma busca de similaridade utilizando a abordagem de k-Vizinhos Mais Próximos, em inglês, *k-Nearest Neighbors (k-NN)*, no índice correspondente do *Elasticsearch*. As consultas empregaram uma função de *Script Score* baseada na similaridade de cosseno entre o vetor do *chunk* de consulta e os vetores previamente indexados. Essa estratégia de recuperação por similaridade vetorial é uma prática amplamente adotada em sistemas de *Dense Retrieval*, especialmente em contextos de busca semântica, visando identificar documentos semanticamente relacionados [Karpukhin et al. 2020, Ni et al. 2024].

Agregação de *Scores* de Similaridade a Nível de Documento: Os resultados brutos da busca retornam *chunks* individuais, cada um com seu respectivo score de similaridade em relação ao *chunk* de consulta. Para calcular a relevância final de cada documento recuperado, foi adotada a técnica de Soma das Pontuações das Passagens, mais conhecido por *Sum of Passage Scores* (SumP) [Ku et al. 2005]. Nessa abordagem, o *score* final de um documento corresponde à soma dos *scores* de similaridade de todos os seus segmentos presentes nos resultados da busca, gerando assim um *ranking* de documentos mais representativos [Ku et al. 2005, Thakur et al. 2021]. A identificação dos documentos foi realizada com base no número do processo, assegurando que todos os *chunks* pertencentes a um mesmo acórdão fossem corretamente agrupados.

Profundidade de Avaliação (*Top-K*): Para cada consulta, foram considerados os documentos mais similares para três diferentes profundidades de *ranking*: $k = 5$, $k = 10$ e $k = 15$. A escolha desses múltiplos valores de k permite uma análise mais abrangente do desempenho dos modelos em diferentes granularidades de recuperação. Valores menores de k possibilitam avaliar a relevância imediata no topo do *ranking*, enquanto valores maiores permitem observar a capacidade do sistema em manter a coesão do *cluster* à medida que mais documentos são incluídos. A profundidade máxima de $k = 15$ é justificada por sua relevância prática, dado que usuários de sistemas de recuperação jurídica tipicamente inspecionam apenas os primeiros resultados para identificar e informações relevantes [Oliveira and Sperandio Nascimento 2025]. Adicionalmente, valores de k modestos em avaliações de coerência *intra-cluster* proporcionam análises concentradas na região mais crítica do *ranking*, onde se espera a maior concentração de documentos semanticamente similares.

A qualidade dos *rankings* de documentos gerados por cada modelo foi avaliada com base em duas métricas de coesão interna, a distância média *intra-cluster* e o desvio padrão dos *scores* de similaridade, os quais são detalhados a seguir.

Distância Média *intra-cluster*: Para cada consulta, calculou-se a *distância média de cosseno* entre todos os pares de documentos recuperados no Top-k. Diferentemente de métricas euclidianas, a distância de cosseno foca na orientação dos vetores, sendo in-

variante à magnitude. Isso é particularmente importante para modelos *Transformer*, nos quais a direção dos *embeddings* é semanticamente mais relevante [Singhal et al. 2001]. Para isso, cada documento foi representado por um único *embedding* agregado, gerado a partir de *Max Pooling* sobre os seus *chunks*. Essa técnica busca capturar os aspectos mais relevantes e distintivos do conteúdo, preservando informações significativas de forma mais eficaz do que a *média aritmética* [Reimers and Gurevych 2019]. Diferentemente da agregação de *scores* utilizada na etapa de recuperação, esta fase tem como objetivo consolidar os *embeddings* dos *chunks* em uma única representação vetorial, viabilizando o cálculo da *Distância Média Intra-cluster* [Moore et al. 2009].

Desvio Padrão dos *Scores* de Similaridade: Esta métrica avalia a consistência dos *scores* de recuperação atribuídos pelo *Elasticsearch* aos documentos retornados no Top-k. O desvio padrão foi calculado a partir dos *scores* finais de similaridade de cada documento recuperado, após a aplicação da estratégia de agregação *SumP*. Um desvio padrão reduzido sugere que o sistema produziu um conjunto de resultados com níveis de similaridade homogêneos em relação à consulta. Essa análise é útil para identificar modelos que geram *rankings* mais estáveis e menos dispersos, o que pode indicar maior confiança nos *scores* atribuídos, logo, na relevância percebida pelo usuário.

Vale ressaltar que, embora a distância média *intra-cluster* baseada em cosseno seja útil para avaliar a coesão, as diferenças entre os espaços vetoriais de cada modelo de *embedding*, e o espaço de características dos textos, podem afetar diretamente essa medida como critério único de comparação entre modelos. Por isso, neste trabalho, tal análise é sempre interpretada em conjunto com a avaliação do *ranking* gerado.

4. Resultados

O desempenho de cada modelo *Transformer* com base nas métricas de coerência semântica *intra-cluster* e na consistência dos *scores* de similaridade são apresentados de forma consolidada na Tabela 1, considerando as diferentes profundidades de *ranking*. As discussões subsequentes aprofundam as implicações desses achados para a RID em acórdãos jurídicos, podendo ser extensíveis para outros tipos de documentos.

Os resultados dispostos na Tabela 1 revelam um padrão consistente de desempenho entre os modelos avaliados para as diferentes profundidades de *ranking* ($K = 5, 10, 15$). As métricas Distância Média *Intra-cluster* e Desvio Padrão dos *Scores* buscam valores mais baixos, indicando maior coesão semântica e maior consistência dos *scores* de similaridade, respectivamente. O BumbaBERT [do Carmo et al. 2023], modelo especializado para o domínio, apresenta o melhor desempenho todas as profundidades de *ranking* avaliadas. Ele alcançou os menores valores para a Distância Média *Intra-cluster* em $K = 5$ (0,0190), $K = 10$ (0,0211) e $K = 15$ (0,0225), bem como para o Desvio Padrão dos *Scores* (0,8706 em $K = 5$, 1,0273 em $K = 10$ e 1,0531 em $K = 15$).

Este desempenho superior e consistente do BumbaBERT sugere que a sua especialização de domínio, possivelmente combinada com as características de seu pré-treinamento específico em dados jurídicos brasileiros (como acórdãos do TJMA [do Carmo et al. 2023]), é um fator decisivo para gerar *rankings* de documentos altamente coesos e com *scores* de similaridade estáveis em um contexto não supervisionado. Isso reforça a importância de modelos treinados em *corpora* específicos para o domínio em questão, como observado na literatura [Oliveira and Sperandio Nascimento 2025].

Tabela 1. Resultados das métricas *intra-cluster* para diferentes profundidades de *ranking* (valores em 10^{-2})

Modelo	Distância Média	Desvio dos *Scores*
Resultados para K=5		
BERTikal	2,77	91,22
LegalBERT-pt	2,70	96,30
BERTimbau	2,90	89,47
BumbaBERT	**1,90**	**87,06**
SBERT-pt	8,93	95,69
Resultados para K=10		
BERTikal	3,09	106,99
LegalBERT-pt	3,02	113,05
BERTimbau	3,21	105,29
BumbaBERT	**2,11**	**102,73**
SBERT-pt	9,88	112,67
Resultados para K=15		
BERTikal	3,29	112,57
LegalBERT-pt	3,21	119,33
BERTimbau	3,38	111,13
BumbaBERT	**2,25**	**105,31**
SBERT-pt	10,42	119,70

Os outros modelos de domínio específico, LegalBERT-pt e BERTikal, também apresentaram alta coesão *intra-cluster*, superando o BERTimbau (Geral) em todas as profundidades de *ranking* para a Distância Média *Intra-cluster*. Isso corrobora a hipótese de que o conhecimento de domínio é benéfico para a representação semântica de textos jurídicos, mesmo que o ganho em relação ao modelo de propósito geral seja marginal em alguns casos. O BERTimbau, como *baseline* geral, serviu como um ponto de referência. Embora seus valores de distância tenham sido consistentemente superiores aos dos modelos de domínio, seu Desvio Padrão dos Scores mostrou-se competitivo, indicando uma consistência na atribuição de *scores*.

O SBERT-pt, apesar de ser otimizado para a tarefa de similaridade de sentenças [Reimers and Gurevych 2019], apresentou os maiores valores para a Distância Média *Intra-cluster*. Este resultado, embora contraintuitivo, vai ao encontro de desafios conhecidos na literatura sobre o "*mismatch* de granularidade" em recuperação densa [Zhao 2012, Khattab and Zaharia 2020]. A otimização de um modelo para a comparação direta de textos curtos não garante que a agregação de seus *chunks* resulte em uma representação de documento coerente, especialmente quando a eficácia de estratégias de agregação de *scores*, como a *SumP*, varia entre modelos e tarefas [Gao et al. 2023, Thakur et al. 2021]. Esse achado sugere, portanto, que o processo de *fine-tuning* do SBERT, ao especializá-lo para a comparação direta, tenha diminuído sua robustez para operações de agregação, fazendo com que a ampla compreensão contextual dos modelos de domínio se mostrasse mais benéfica para esta tarefa do que a especialização para uma granularidade sentencial.

Observando a tendência geral entre as profundidades de *ranking*, todas as métricas, tanto Distância Média *Intra-cluster* quanto Desvio Padrão dos Scores, mostram um aumento gradual de $K = 5$ para $K = 15$. Isso é esperado, pois ao incluir mais documentos no *ranking*, é natural que a coesão e a consistência diminuam ligeiramente, já que os documentos adicionais tendem a ser progressivamente menos similares aos do topo. A

capacidade de um modelo de mitigar essa degradação, mantendo um aumento menor nas métricas, é um indicativo de sua robustez na recuperação de conjuntos mais amplos de documentos relevantes. Nesse sentido, o BumbaBERT demonstrou a menor degradação relativa em suas métricas de coesão e consistência, consolidando sua superioridade.

Em suma, os resultados obtidos demonstram a viabilidade da avaliação de sistemas de Recuperação Densa para documentos jurídicos em cenários não supervisionados, utilizando métricas de coerência *intra-cluster*. O desempenho superior e consistente do BumbaBERT destaca a importância da especialização de domínio em *embeddings* para tarefas de recuperação semântica em contextos altamente especializados, como o jurídico. A análise para diferentes valores de K oferece uma visão mais granular da performance, reforçando a aplicabilidade desses modelos para auxiliar na organização e exploração de vastos acervos de jurisprudência.

5. Considerações Finais

Este estudo apresentou uma análise comparativa, em um cenário "*zero-shot*", da eficácia de diferentes classes de modelos de *embedding* para a tarefa de RID em documentos jurídicos longos. Utilizando uma metodologia não supervisionada baseada em coesão semântica *intra-cluster*, os resultados demonstraram de forma consistente a superioridade dos modelos com especialização de domínio. O BumbaBERT [do Carmo et al. 2023], em particular, destacou-se ao gerar os rankings mais coesos em todas as profundidades avaliadas. Notavelmente, a especialização de domínio se mostrou um fator mais crítico para o desempenho do que a otimização para a tarefa de similaridade (SBERT-pt), um achado que evidencia o desafio do "*mismatch* de granularidade" ao aplicar modelos otimizados para sentenças em tarefas de agregação de documentos longos. A metodologia proposta se mostrou uma alternativa viável para aferir a qualidade de sistemas de recuperação em cenários com escassez de dados anotados, oferecendo *insights* para a implementação de ferramentas de IA no Judiciário [Zhao 2012, Thakur et al. 2021].

É salutar reconhecer as limitações deste estudo. A principal reside na própria natureza da avaliação não supervisionada: as métricas de coesão, medem a consistência interna do *ranking*, mas não substituem um julgamento de relevância em relação à consulta, que exigiria um gabarito anotado por especialistas. Adicionalmente, a comparabilidade direta dos valores numéricos de similaridade entre os diferentes espaços de *embedding* dos modelos é inerentemente limitada, devendo os resultados serem interpretados pelo seu ordenamento relativo e consistência. Ressalta-se que a avaliação foi conduzida sobre uma única divisão de dados, embora comum em estudos exploratórios, esta abordagem não garante que os resultados sejam generalizáveis a diferentes partições, representando uma ameaça à robustez estatística. Por fim, os achados são circunscritos ao *corpus* de acórdãos e à estratégia de agregação *SumP* utilizados, que pode introduzir um viés em favor de documentos mais longos.

Como trabalhos futuros, recomenda-se: i) a validação qualitativa dos *rankings* gerados com especialistas do domínio jurídico, para correlacionar a coesão semântica com a relevância percebida; ii) a implementação de validação cruzada para avaliação estatística e generalização dos resultados em diferentes partições dos dados; iii) a investigação de estratégias alternativas de agregação de *scores* (*e.g., MaxP, AverageP*) para mitigar possíveis vieses e avaliar o impacto na performance de cada classe de modelo; e iv) a replicação

deste estudo em outros corpora jurídicos (*e.g.,* de outros tribunais ou tipos de peça, como petições; v) e a avaliação em profundidades de *ranking* (K) mais elevadas.

Agradecimentos e Uso de IA generativa

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)- PQ-316507/2023-7, DT-303031/2023-9, POSDOC - 101057/2024-5; Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - 2023/10100-4; 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.

Referências

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

CNJ (2024). Relatório analítico anual da justiça em números 2023. `https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros`.

Costa, J. A. F. and Dantas, N. C. D. (2023). Análise comparativa de embeddings jurídicos aplicados a algoritmos de clustering. *Anais do Congresso Brasileiro de Computação Jurídica*.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

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

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Sun, J., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey.

Guimarães, J. A. C. (2004). *Elaboração de ementas jurisprudenciais: elementos teórico-metodológicos*.

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

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

Khattab, O. and Zaharia, M. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

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

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

Moore, D. S., McCabe, G. P., and Craig, B. A. (2009). *Introduction to the Practice of Statistics*.

Ni, C., Wu, J., Wang, H., Lu, W., and Zhang, C. (2024). Enhancing cloud-based large language model processing with elasticsearch and transformer models. In *ISPP*.

Oliveira, R. S. d. and Sperandio Nascimento, E. G. (2025). Analysing similarities between legal court documents using natural language processing approaches based on transformers. *PloS one*, 20(4):e0320244.

Pires, V. B., Guerreiro, D., et al. (2024). Portuguese fake news classification with bert models. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*. SBC.

Polo, F. M., Mendonça, G. C. F., Parreira, K. C. J., Gianvechio, L., Cordeiro, P., Ferreira, J. B., de Lima, L. M. P., Maia, A. C. d. A., and Vicente, R. (2021). Legalnlp–natural language processing methods for the brazilian legal language.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks.

Scherrer, L., Tomko, M., Ranacher, P., and Weibel, R. (2018). Travelers or locals? identifying meaningful sub-populations from human movement data in the absence of ground truth. *EPJ Data Science*, 7(1):1–21.

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

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V., and Furtado, V. (2023). Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Brazilian Conference on Intelligent Systems*, pages 268–282.

Singhal, A. et al. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

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

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.

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

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *4th Int. Conf. on Practical Applications of Knowledge Discovery and Data Mining*.

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

Zhao, L. (2012). *Modeling and solving term mismatch for full-text retrieval*. Carnegie Mellon University.

Zhao, W. X., Liu, J., Ren, R., and Wen, J.-R. (2024). Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*.