Proposta recebida em Fevereiro 2025 e aceite para publicação em Junho 2025.

BM25 x Vila Sésamo: avaliando modelos Sentence-BERT para Recuperação de Informação no cenário legislativo brasileiro

BM25 vs. Sesame Street: assessing Sentence-BERT models for Information Retrieval within the Brazilian legislative scenario

Douglas Vitório
Centro de Informática,
Universidade Federal de Pernambuco

Ellen Souza
Universidade Federal Rural de Pernambuco

José Antônio dos Santos
Universidade de Pernambuco

André Carlos Ponce de Leon Ferreira de Carvalho
Universidade de São Paulo

Adriano L. I. Oliveira
Centro de Informática,
Universidade Federal de Pernambuco

Nádia F. F. da Silva
Universidade Federal de Goiás

Resumo

Modelos baseados em BERT vêm sendo largamente utilizados, tornando-se o estado da arte para muitas tarefas de Processamento de Linguagem Natural e também para Recuperação de Informação. A arquitetura Sentence-BERT permitiu que esses modelos fossem facilmente utilizados para a busca semântica de documentos, já que ela gera *embeddings* contextuais que podem ser comparados através de medidas de similaridade. Para melhor investigar a aplicação de modelos baseados em BERT para Recuperação de Informação, este trabalho avaliou 12 modelos Sentence-BERT, disponíveis publicamente, para a recuperação de documentos no cenário legislativo brasileiro. Duas variantes do algoritmo BM25 foram utilizadas como *baseline*: Okapi BM25 e BM25L. O BM25L alcançou melhores resultados, com significância estatística, mesmo no cenário em que os documentos não foram pré-processados, enquanto que apenas um dos modelos de linguagem, ajustado usando dados legislativos brasileiros, obteve um desempenho similar para uma das três bases de dados utilizadas.

Palavras chave

recuperação de informação; documentos legislativos; modelos de linguagem; BERT; BM25

Abstract

BERT-based models have been largely used, becoming the state-of-the-art for many Natural Language Processing tasks and for Information Retrieval. The Sentence-BERT architecture allowed these models to be easily used for the semantic search of documents, as it generates contextual embeddings that can be compared using similairty measures. To further investigate the application of BERT-based models for Information Retrieval, this work assessed 12 publicly available Sentence-BERT models for documents retrieval within the Brazilian legislative scenario. Two BM25 variants were used as baseline: Okapi BM25 and BM25L. BM25L achieved better results, considering statistical significance, even in the scenario in which the documents were not preprocessed, while only one language model, fine-tuned using Brazilian legislative data, could reach a similar performance for one of the three used datasets.

Keywords

information retrieval; legislative documents; language models; BERT; BM25

1. Introdução

A utilização de técnicas computacionais se tornou essencial para o funcionamento das instituições judiciais e legislativas, as quais necessitam lidar com um grande volume de dados que vem sendo constantemente produzido (Souza et al., 2021b). No cenário legislativo brasileiro, por exemplo, o Departamento de Consultoria Legislativa (Conle)[1] da Câmara dos Deputados realiza o processamento de uma enorme quantidade de documentos produzida durante o processo de elaboração e manutenção das leis. Desde 1930, ela já processou mais de 144 mil proposições legislativas (Brandt, 2020).

Durante o processo legislativo, matérias, comumente criadas por parlamentares, são submetidas à avaliação do Poder Legislativo — no caso

[1] https://www2.camara.leg.br/a-camara/estruturaadm/consultoria-geral/estrutura-1/conle/servicos-1

DOI: 10.21814/lm.17.1.474

do Brasil, composto pela Câmara dos Deputados e pelo Senado Federal — para que sejam analisadas e votadas, visando tornar-se legislação. Essas proposições legislativas, apontadas por Brandt (2020) como o elemento central do processo legislativo, têm como principais espécies, no âmbito brasileiro, os Projetos de Lei (PLs) e as Propostas de Emenda à Constituição (PECs). As proposições são documentos longos, geralmente contendo toda a legislação proposta, ou as mudanças à legislação já vigente, e uma justificação para as mudanças. O tamanho médio das proposições legislativas brasileiras se encontra em torno de 700 palavras, porém algumas podem possuir mais de 200 mil palavras (Vitório et al., 2025). A Figura 1 traz, como exemplo, a primeira página de uma proposição legislativa (PEC 221/2019), destacando suas principais partes.

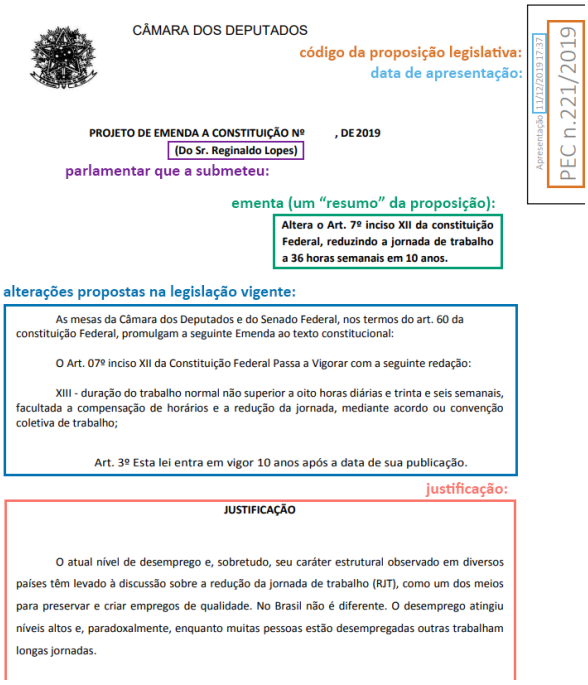

Figura 1: Exemplo de proposição legislativa submetida ao processo legislativo brasileiro.

Um dos papéis da Conle é encontrar documentos legislativos, tais como PLs, PECs e outras proposições legislativas, visando atender às requisições oriundas de parlamentares. Para que um parlamentar crie uma nova proposição legislativa, ele necessita, previamente, consultar a Conle para obter a lista de proposições submetidas anteriormente sobre um determinado tema. Essa é uma tarefa bastante dispendiosa caso seja realizada de forma manual.

A tarefa de buscar dados a partir de grandes bases é chamada de Recuperação de Informação (RI). Ela também é conhecida como "recuperação de documentos" quando se trata da obtenção de documentos textuais, como as proposições legislativas. Portanto, o objetivo principal da RI é encontrar material, dentro de uma coleção de dados geralmente não-estruturados, que satisfaça a necessidade de informação de um usuário, isto é, material que seja relevante para esse usuário (Manning et al., 2008).

Dentro do cenário legal, o interesse pelo uso de sistema de RI para recuperação de documentos se dá pelo fato de que todo o trabalho nessa área é baseado em documentos textuais, sendo o acesso e a posse desses documentos cruciais para o bom andamento do processo legal. Além disso, há um aumento no número de documentos legais sendo produzidos (Souza et al., 2021b), o que torna o trabalho dos profissionais da área ainda mais complexo. Por consequência, uma subárea específica da RI, chamada Recuperação de Informação Legal (RIL), desenvolveu-se para auxiliar tarefas dentro do domínio legal (Maxwell & Schafer, 2008). A subárea de RIL engloba tarefas como a análise de jurisprudência, bem como o suporte ao processo de confecção de leis. Dentro desse cenário legislativo, técnicas automatizadas de RI são necessárias para acompanhar o crescente número de documentos criados por parlamentares. Ademais, a organização, o acesso e a busca por esse tipo de dado traz desafios significativos devido à natureza não-estruturada desses documentos (Cantador & Sánchez, 2020).

Outro desafio no uso de técnicas de RI encontra-se na natureza das requisições dos usuários. Como os usuários expressam suas necessidades de informação através de consultas, os algoritmos de RI objetivam recuperar documentos relevantes cujo conteúdo corresponda ao conteúdo da consulta. Os termos utilizados na consulta, contudo, podem não ser os mesmos presentes no documento, surgindo um problema conhecido como "incompatibilidade de vocabulário" (Manning et al., 2008). Uma alternativa, então, é considerar a semântica e o contexto dos termos.

A aplicação de redes neurais para gerar *embeddings* contextuais e seu subsequente uso para o Processamento de Linguagem Natural (PLN) vêm aumentando substancialmente nos últimos anos, tornando-se o estado da arte para muitas tarefas ao melhorar o desempenho dos modelos, mesmo utilizando conjuntos de dados menores (Wolf et al., 2020; Caseli & Nunes, 2024). O mais conhecido e utilizado modelo ba-

seado nessas arquiteturas denominadas *Transformers* (Vaswani et al., 2017) é o Bidirectional Encoder Representations for Transformers (BERT) (Devlin et al., 2019), inclusive para RI e outras tarefas relacionadas ao ranqueamento de textos (Lin et al., 2022; Wang et al., 2024).

Devido ao alto custo computacional de modelos de linguagem para encontrar frases similares, Reimers & Gurevych (2019) propuseram uma modificação no BERT conhecida como Sentence-BERT (SBERT). O SBERT utiliza estruturas siamesas e triplas de redes neurais para derivar *embeddings* de frases que sejam semanticamente significantes, os quais podem ser comparados utilizando medidas de similaridade. Essa arquitetura tem permitido que modelos baseados no BERT sejam usados facilmente para tarefas como a Similaridade Semântica Textual (STS) e RI com busca semântica.

Embora os Grandes Modelos de Linguagem (LLMs) generativos venham ganhando cada vez mais atenção para tarefas de PLN nos últimos anos, devido à sua capacidade de gerar texto semelhante ao humano, ainda existem muitos desafios na sua utilização para problemas do mundo real, como para RI (Wang et al., 2024). Os LLMs necessitam de muitos recursos computacionais tanto para treinamento quanto para inferência, eles também só podem gerar respostas baseadas unicamente no conhecimento contido no seu pré-treinamento e há preocupações relacionadas à privacidade no uso das suas APIs. Por outro lado, o uso de modelos baseado em BERT não possui esses riscos de privacidade, também necessitando de significantemente menos recursos computacionais, sendo, dessa forma, mais viáveis para tarefas como a de recuperação de documentos. Além disso, o BERT pode ser adaptado para tarefas específicas através de pré-treinamento e *fine-tuning*.

O BERT tradicional, entretanto, possui uma limitação crucial para o ranqueamento de textos: sua incapacidade de processar textos longos (Lin et al., 2022). Esse modelo, e a maioria dos modelos baseados nele, possui um tamanho máximo de entrada de 512 tokens. Portanto, quando lidando com documentos cujo tamanho excede esse limite, o texto precisa ser truncado e uma grande parte da informação é perdida. Alternativas são as técnicas de *chunking* (Jaiswal & Milios, 2023), como a de janela deslizante, as quais, entretanto, necessitam de custo extra para processar os diferentes segmentos do texto e depois combiná-los. Esse é um problema grave ao realizar RI com documentos longos, como aqueles produzidos no cenário legislativo. Sendo assim, o tamanho médio de 700 termos das proposições legislativas brasileiras (Vitório et al., 2025) apresenta um obstáculo para a utilização desses modelos de linguagem. Apenas modelos mais recentes, como o ModernBERT (Warner et al., 2024), começaram a expandir a janela de contexto para utilizar um número maior de tokens como entrada.

Com isso, este estudo visa realizar uma avaliação abrangente e sistemática de modelos SBERT, disponíveis publicamente na plataforma HuggingFace[2], para RI no domínio legislativo brasileiro. Para isto, foram utilizadas duas variantes do algoritmo BM25 como *baseline*: Okapi BM25 e BM25L. O BM25 é um algoritmo probabilístico, até hoje bastante utilizado para o ranqueamento de textos, tanto para pesquisa acadêmica quanto para sistemas comerciais (Lin et al., 2022; Caseli & Nunes, 2024). O Okapi BM25 é a versão original desse algoritmo, enquanto o BM25L é uma variante criada para lidar com documentos mais longos, como os legislativos. O objetivo deste trabalho, portanto, é responder à questão: *tendo em vista os recentes avanços trazidos pelos modelos de linguagem, o algoritmo BM25 ainda os supera para Recuperação de Informação no cenário legislativo brasileiro, considerando a restrição do tamanho de entrada dos modelos?*

O restante deste artigo está organizado da seguinte forma: a Seção 2 traz os trabalhos relacionados; a Seção 3 apresenta a arquitetura SBERT, explicando sua utilização para RI, além de apresentar os modelos avaliados neste estudo. A configuração dos experimentos é detalhada na Seção 4; enquanto que os resultados são apresentados e discutidos na Seção 5. A Seção 6 traz as conclusões e trabalhos futuros.

2. Trabalhos relacionados

O uso de modelos baseados em BERT vem se tornando um tópico emergente no domínio legal, inclusive para RI. Na pesquisa realizada por Sansone & Sperlí (2022), foi relatado o uso de modelos BERT para tarefas como a recuperação de processos judiciais e a classificação de documentos legais.

Estudos realizando RI dentro do domínio legislativo, contudo, são escassos, principalmente utilizando dados em Português. Da mesma forma, não foram encontrados trabalhos cujo foco seja a avaliação e comparação de modelos de linguagem para RI dentro do domínio legal em Por-

[2] https://huggingface.co

tuguês. Sendo assim, os trabalhos apresentados nesta seção realizam a comparação do desempenho dos modelos com o desempenho dos seus respectivos sistemas propostos, à exceção do estudo realizado por Mandal et al. (2021), o qual avaliou técnicas da literatura, porém para documentos em Inglês.

dos Santos et al. (2024) propuseram um sistema híbrido combinando o algoritmo BM25L e um modelo baseado em BERT ajustado. Para escolher o melhor modelo a ser utilizado na abordagem híbrida, foram comparados cinco modelos SBERT para a recuperação de proposições legislativas brasileiras, três deles específicos para o domínio legal. O algoritmo BM25L também foi avaliado com e sem pré-processamento, mostrando que a versão com dados pré-processados superou todos os modelos baseados em BERT, mesmo quando eles foram ajustados com dados legislativos. O trabalho de dos Santos et al. (2024) é similar ao apresentado neste estudo, porém, neste trabalho, o foco se encontra numa comparação mais abrangente entre os modelos SBERT e o BM25. Um conjunto maior de modelos de linguagem foi avaliado, além de duas variantes do algoritmo BM25, também utilizando um número maior de bases de dados e de métricas de desempenho para comparar os resultados.

Visando calcular a similaridade entre relatórios de processos judiciais oriundos da Suprema Corte Indiana, Mandal et al. (2021) avaliaram métodos tradicionais e de *embeddings* para representação de textos: TF-IDF, LDA, Word2Vec, PScoreVect (Mandal et al., 2017), BERT e Law2Vec (Chalkidis & Kampas, 2019), o qual foi treinado com dados legais. Técnicas tradicionais como TF-IDF e LDA alcançaram os melhores resultados, ao passo que o Doc2Vec teve um bom desempenho para representação de documentos na presença de uma estrutura linguística. Os *embeddings* gerados pelo BERT alcançaram um desempenho muito inferior, o que, para Mandal et al. (2021), pode ser explicado pelo tamanho dos documentos legais e a limitação de entrada do BERT.

Outros estudos compararam o seu próprio sistema a algoritmos BM25 para o domínio legal utilizando documentos escritos em Português. Nesses casos, os sistemas propostos superaram o BM25 em cenários específicos. Melo et al. (2023) propuseram um sistema híbrido de busca semântica para auxiliar o trabalho do Supremo Tribunal de Justiça (STJ) de Portugal. O sistema combina o BM25 com o Legal-BERTimbau, uma versão ajustada do BERTimbau (Souza et al., 2020) criada pelos autores utilizando dados do próprio STJ. Os autores relataram uma grande melhoria nos resultados alcançados pelo sistema híbrido para o primeiro resultado da consulta, em comparação com a utilização somente do BM25.

Cordeiro et al. (2023), por sua vez, propuseram o Legal Semantic Search Engine (LeSSE), um sistema híbrido que combina o uso de informações semânticas e lexicais para RI. O trabalho objetivou a criação de um sistema para busca de legislação referente ao direito do consumidor de Portugal, substituindo o sistema predecessor, o qual era baseado apenas na busca por palavras-chave. Os autores utilizaram dois módulos no seu sistema híbrido: um que visa a realização de uma busca lexical, através do algoritmo Okapi BM25, e outro focado na busca semântica a partir de uma versão ajustada do BERTimbau. O processo de *fine-tuning* do BERTimbau utilizou uma base de dados anotada manualmente por especialistas contendo pares de consultas e trechos de leis do direito do consumidor que respondessem àquelas consultas. Os resultados mostraram um melhor desempenho do LeSSE em comparação à *baseline* composta pelo algoritmo BM25 utilizado de forma individual.

Os trabalhos apresentados nesta seção mostram que, mesmo quando modelos baseados em BERT são utilizados e obtêm resultados expressivos para RI no domínio legal, esse bom desempenho é decorrente da sua utilização dentro de sistemas híbridos, junto a algoritmos como o BM25.

3. Sentence-BERT e o seu uso para Recuperação de Informação

Como explicado anteriormente (Seção 1), a arquitetura SBERT permite o uso adequado de modelos baseados em BERT para tarefas como a de RI (Reimers & Gurevych, 2019).

A arquitetura SBERT gera *embeddings* semânticos de frases, os quais podem ser comparados através de medidas de similaridade. Portanto, a similaridade entre documentos, ou entre um documento e uma consulta, pode se calculada, resultando em um *ranking* de documentos similares a serem recuperados. Neste estudo, a similaridade do cosseno foi utilizada, a qual também foi usada pelos autores do SBERT.

Para realizar a avaliação, 12 modelos baseados em BERT foram selecionados. Após uma pesquisa na literatura, foram encontrados oito modelos, disponíveis publicamente, os quais foram treinados ou ajustados com dados escritos em Português para o domínio legal ou legislativo: Legal-BERTimbau, JurisBERT, BER-

TimbauLaw, LegalBert-pt, LegalBERTPT-Br, GovBERT-BR e versões do BERTimbau e LegalBert-pt ajustadas com dados legislativos. Ademais, o BERTimbau também foi selecionado, já que foi treinado para o Português Brasileiro, além de três modelos SBERT multilíngues: LaBSE, Paraphrase Multilingual MPNet e Paraphrase Multilingual MiniLM. Todos os modelos podem ser encontrados na plataforma HuggingFace e podem ser utilizados com a arquitetura SBERT. A Tabela 1 sumariza esses modelos.

BERTimbau (Souza et al., 2020) é a versão do BERT treinada para a Língua Portuguesa. Utilizando o brWac (Wagner Filho et al., 2018), um corpus grande e diverso de páginas web, o BERT foi pré-treinado para três tarefas de PLN: STS, Reconhecimento de Implicação Textual (RIT) e Reconhecimento de Entidades Nomeadas (REN). Dois modelos com tamanhos diferentes foram disponibilizados: Base[3] (com 110M parâmetros) e Large[4] (330M parâmetros), ambos com um tamanho máximo de entrada de 512 tokens. A versão Large foi utilizada neste estudo.

Legal-BERTimbau[5] (Melo et al., 2023) é uma versão ajustada do BERTimbau para o domínio legal português. Para realizar a adaptação de domínio, foram usados pares de sentenças legais oriundas do Supremo Tribunal de Justiça de Portugal. O seu tamanho máximo de entrada também é de 512 tokens.

JurisBERT[6] (Viegas et al., 2023) é um modelo baseado em BERT treinado do zero para a área jurídica brasileira. Primeiramente, ele foi pré-treinado utilizando documentos legais brasileiros disponíveis publicamente, tais como leis, decretos, decisões e acórdãos (decisões judiciais tomadas por colegiados), com uma predominância de leis. Como o JurisBERT foi criado especificamente para a realização de STS utilizando as ementas (resumos de textos legais) de acórdãos, os documentos foram divididos em parágrafos. Posteriormente, pares de acórdãos extraídos do Supremo Tribunal Federal (STF) brasileiro e de tribunais reginais foram utilizados para ajustar o modelo. Baseado no tamanho médio das ementas, Viegas et al. (2023) definiram o tamanho máximo de entrada do modelo como 384 tokens.

BERTimbauLaw Viegas et al. (2023) também disponibilizaram[7] uma versão ajustada do BERTimbau, para a qual o *fine-tuning* foi realizado da mesma forma que para o JurisBERT. Esse modelo, chamado BERTimbauLaw, também possui 384 tokens de entrada máxima.

LegalBert-pt (Silveira et al., 2023) foi pré-treinado usando 1,5 milhão de documentos oriundos de 10 tribunais de justiça brasileiros. O modelo foi criado para lidar com REN e tarefas de classificação de texto no domínio legal. Duas versões dele foram desenvolvidas e disponibilizadas: LegalBert-pt SC[8], a qual foi treinada do zero, e LegalBert-pt FP[9], a qual é uma adaptação do BERTimbau. Como o trabalho de Silveira et al. (2023) relata resultados superiores utilizando a versão LegalBert-pt FP, esse foi o modelo selecionado para este estudo. Por ter sido pré-treinado a partir do BERTimbau, seu tamanho máximo de entrada também é de 512 tokens.

LegalBERTPT-Br[10] (Silva et al., 2021) foi desenvolvido no cenário da Câmara dos Deputados brasileira visando ser utilizado para Modelagem de Tópicos. O treinamento do modelo foi realizado pela combinação do BERTimbau com o *framework* de Aprendizado Contrastivo SimCSE (Gao et al., 2021). Como conjunto de treinamento, foram utilizados discursos políticos, proposições legislativas e comentários de cidadãos acerca dessas proposições. Devido ao tamanho pequeno das frases utilizadas, o tamanho máximo de entrada do modelo foi definido como 32 tokens.

GovBERT-BR[11] (Silva et al., 2025) é um modelo focado em dados governamentais brasileiros, incluindo documentos administrativos e legais, com o objetivo de lidar com os desafios trazidos pela terminologia burocrática brasileira. Ele foi pré-treinado a partir do BERTimbau utilizando licitações públicas, trechos do Diário Oficial da União contendo informações sobre o processo de licitação e documentos legais do STF, oriundos do corpus VICTOR (Luz de Araujo et al., 2020). Assim como o BERTimbau, esse modelo também possui um limite de entrada de 512 tokens.

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

[4] https://huggingface.co/neuralmind/bert-large-portuguese-cased

[5] https://huggingface.co/rufimelo/Legal-BERTimbau-large

[6] https://huggingface.co/alfaneo/jurisbert-base-portuguese-sts

[7] https://huggingface.co/alfaneo/bertimbaulaw-base-portuguese-sts

[8] https://huggingface.co/raquelsilveira/legalbertpt_sc

[9] https://huggingface.co/raquelsilveira/legalbertpt_fp

[10] https://huggingface.co/ulysses-camara/legal-bert-pt-br

[11] https://huggingface.co/dccmpmgfinalisticas/GovBERT-BR

Modelo	Domínio	Língua	Entrada max.
BERTimbau	vários	Português Brasileiro	512
Legal-BERTimbau	legal	Português Europeu	512
JurisBERT	legal	Português Brasileiro	384
BERTimbauLaw	legal	Português Brasileiro	384
LegalBert-pt	legal	Português Brasileiro	512
LegalBERTPT-Br	legislativo	Português Brasileiro	32
GovBERT-BR	admin./legal	Português Brasileiro	512
Fine-tuned BERTimbau (FT BERTimbau)	legislativo	Português Brasileiro	512
Fine-tuned LegalBert-pt (FT LegalBert-pt)	legislativo	Português Brasileiro	512
LaBSE	vários	multilíngue	256
Paraphrase Multilingual MPNet	vários	multilíngue	128
Paraphrase Multilingual MiniLM	vários	multilíngue	128

Tabela 1: Sumário dos modelos baseados em BERT que foram avaliados.

Fine-tuned BERTimbau No trabalho de dos Santos et al. (2024), foi disponibilizada[12] uma versão do BERTimbau ajustada com dados legislativos brasileiros. O *fine-tuning* foi realizado usando pares de proposições legislativas relacionadas, oriundas de uma árvore de proposições contendo o relacionamento entre elas. O tamanho máximo de entrada para esse modelo é o mesmo do BERTimbau: 512 tokens. Neste trabalho, esse modelo será referenciado como "FT BERTimbau".

Fine-tuned LegalBert-pt: Os autores dos Santos et al. (2024) também disponibilizaram[13] uma versão do LegalBert-pt (Silveira et al., 2023), ajustada com a mesma técnica utilizada para a versão do BERTimbau (explicada no tópico anterior). Seu tamanho máximo de entrada também é o mesmo do modelo base: 512 tokens, e ele será referenciado como "FT LegalBert-pt".

LaBSE O modelo Language-agnostic BERT Sentence Encoder (LaBSE) (Feng et al., 2022) oferece suporte a 109 linguagens, tendo sido treinado usando dados monolíngues do CommonCrawl e Wikipédia, assim como pares bilíngues traduzidos de páginas web. Ele foi primeiramente avaliado para *Bitext Retrieval*, mas também foi utilizado para STS. Embora, no trabalho original, tenha sido relatado um tamanho de entrada de 512 tokens durante a etapa de pré-treinamento, o modelo disponível na plataforma HuggingFace[14], e utilizado neste estudo, possui um tamanho máximo de entrada de 256 tokens.

Paraphrase Multilingual MPNet[15] foi construído através do processo de Destilação do Conhecimento (Reimers & Gurevych, 2020) do modelo MPNet (Song et al., 2020), utilizando o XLM-RoBERTa (Conneau et al., 2020), o qual foi pré-treinado em 100 linguagens diferentes, como "modelo aluno". Portanto, esse modelo pode gerar *embeddings* a serem utilizados para uma variedade de tarefas de PLN, tais como clusterização e busca semântica, além de recuperação de documentos. Ele possui um tamanho máximo de entrada de 128 tokens.

Paraphrase Multilingual MiniLM, também oriundo do processo de Destilação do Conhecimento, o modelo Paraphrase Multilingual MiniLM[16] foi construído a partir do treinamento de uma versão multilíngue do MiniLM (Wang et al., 2021), o qual é baseado no XLM-RoBERTa, enquanto utilizando a versão monolíngue do MiniLM (Wang et al., 2020) como "modelo professor". Ele também possui uma entrada máxima de 128 tokens e pode ser utilizado para uma variedade de tarefas de PLN e para RI.

4. Configuração dos experimentos

Nesta seção, serão detalhados os experimentos realizados para comparação entre os modelos SBERT e as variantes do algoritmo BM25, bem como os corpora utilizados para execução dos experimentos e as métricas de avaliação.

[12] https://huggingface.co/josedossantos/bertimbau-tuned
[13] https://huggingface.co/josedossantos/legalbertpt-tuned
[14] https://huggingface.co/sentence-transformers/LaBSE
[15] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
[16] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

4.1. Corpora

Quatro corpora contendo dados legislativos foram utilizados para a avaliação dos modelos SBERT e das variantes do BM25 para RI: um contendo proposições legislativas a serem recuperadas e três contendo consultas e suas respectivas listas de documentos relevantes.

Devido à falta de bases de dados disponíveis na literatura na área de RI com documentos legislativos em Português — sendo o corpus construído por Vitório et al. (2025) a primeira base disponível publicamente —, todos os corpora utilizados neste trabalho foram obtidos em parceria com a Câmara dos Deputados brasileira, viabilizando a realização desta pesquisa.

4.1.1. Corpus de Proposições Legislativas

Para o processo de RI, um corpus contendo 105.669 proposições legislativas foi utilizado (Vitório et al., 2025). Portanto, o objetivo dos modelos de RI foi recuperar documentos a partir desse conjunto de dados. Essas proposições encontram-se disponíveis publicamente no portal da Câmara dos Deputados[17], bem como no GitHub[18], como parte do Ulysses-RFCorpus, o qual será detalhado posteriormente nesta seção.

Como os modelos baseados em BERT possuem um tamanho de entrada limitado, uma etapa de limpeza desse corpus foi realizada previamente à sua utilização neste estudo, fazendo uso do mesmo processamento usado e descrito por dos Santos et al. (2024), isto é, removendo símbolos inválidos e caracteres especiais (por exemplo,"\r", "\n") do conteúdo das proposições.

A média de palavras por proposição legislativa nesse corpus é de 706, com um desvio padrão de 1.799. A maioria do documentos possui de 100 a 1.000 palavras, porém quase 5.000 deles são compostos por mais de 2.000 palavras (Vitório et al., 2025).

4.1.2. Ulysses-RFCorpus

O Ulysses-RFCorpus (Vitório et al., 2025) foi construído em parceria com o departamento de Consultoria Legislativa (Conle) da Câmara dos Deputados brasileira. Ele contém um conjunto de 692 consultas e, para cada uma delas, uma lista de 12 proposições julgadas pelos consultores da Conle.

As consultas desse corpus foram criadas pelos consultores da Conle simulando consultas legislativas reais. Os mesmos consultores as utilizaram como entrada para o sistema de RI da Câmara dos Deputados, o qual é baseado em técnicas de NLP e no algoritmo BM25 (Souza et al., 2021b). Eles, então, julgaram os 12 resultados recuperados pelo sistema para cada consulta. Durante a construção do corpus, três categorias de relevância foram consideradas para julgar os documentos: "irrelevante", "pouco relevante" e "muito relevante". Em média, seis das 12 proposições foram julgadas como "irrelevantes", quatro como "muito relevantes" e duas como "pouco relevantes" para cada consulta. Entretanto, nenhuma proposição foi julgada como "pouco relevante" ou "muito relevante" para 46 das 692 consultas. Isto é, essas consultas apenas continham uma lista de documentos irrelevantes, portanto foram removidas da avaliação e apenas 646 consultas foram utilizadas neste estudo.

As consultas desse corpus possuem, em média, 20 palavras e um tamanho máximo de 121 palavras (Vitório et al., 2025), o que não constitui um problema para os modelos baseados em BERT.

4.1.3. Preliminary Search Corpus (PSC)

Outro corpus disponibilizado pela Câmara dos Deputados brasileira foi utilizado para este estudo. É trabalho do time de consultores da Conle a realização de uma busca por proposições e outros documentos legislativos para responder à requisição de um parlamentar, a qual é chamada "consulta legislativa". Até 2021, as consultas legislativas eram utilizadas como entrada para um sistema legado baseado em busca Booleana, durante um processo manual e extremamente custoso chamado de "Pesquisa Prévia".

O corpus utilizado neste estudo é o resultado da Pesquisa Prévia para 1.934 consultas legislativas reais, criadas por parlamentares. Portanto, cada consulta nesse corpus contém uma lista de proposições legislativas relevantes, selecionados por um consultor da Conle. O número de documentos relevantes para cada consulta varia de 1 a 20.

Essas consultas, contudo, não podem ser disponibilizadas publicamente devido a restrições de confidencialidade, já que contêm informações que identificam os parlamentares que as criaram. Além disso, serviços de consultoria e assessoramento institucional são confidenciais, de acordo com o artigo 13 da Resolução da Câmara dos Deputados nº 48, de 1993 (Câmara dos Deputados, 1993).

[17] https://www.camara.leg.br/busca-portal/proposicoes/pesquisa-simplificada

[18] https://github.com/ulysses-camara/Ulysses-RFCorpus

Das 1.934 consultas contidas nesse corpus, apenas 11 possuem um tamanho maior do que 512 palavras. No geral, elas são compostas por uma média de 80 palavras, também não representando um problema considerável para a utilização de modelos BERT.

4.1.4. Legislative Consultation Corpus (LCC)

Um parlamentar solicita à Conle, através de uma consulta legislativa, uma lista de documentos relevantes visando a criação de uma nova proposição legislativa.

O corpus utilizado por Souza et al. (2021b) e dos Santos et al. (2024), e também utilizado para este estudo, contém um conjunto de 295 consultas legislativas reais e sua respectiva proposição legislativa resultante. Assim sendo, cada consulta desse corpus contém apenas um documento relevante/relacionado. Esse corpus também não pode ser disponibilizado publicamente pelo mesmo motivo explicado na subseção anterior.

A maioria das consultas presentes nesse corpus possui entre 10 e 40 palavras (Souza et al., 2021b), com uma média de 66 palavras, sendo também possível para os modelos SBERT processarem a grande maioria delas.

Para sumarizar os corpora de consultas utilizados, a Tabela 2 apresenta a quantidade de consultas em cada corpus, o tamanho médio, em palavras, das consultas e a quantidade de proposições julgadas por consulta.

4.2. Análise experimental

Usando os três corpora de consultas, quatro aspectos diferentes foram levados em consideração para construção e execução dos experimentos: diferentes variantes do algoritmo BM25; o tamanho máximo de entrada para os modelos SBERT; a utilização, ou não, de técnicas de pré-processamento de texto para o BM25; e o tempo de processamento despendido pelos modelos avaliados.

4.2.1. Algoritmos BM25

Para que a comparação dos modelos não fosse restrita apenas a um algoritmo de *baseline*, optou-se pela utilização de duas variantes do BM25: Okapi BM25 e BM25L. Okapi é a versão original do BM25 (Robertson et al., 1994) e a mais difundida (Caseli & Nunes, 2024), enquanto que o BM25L (Lv & Zhai, 2011) é uma variante construída para lidar com documentos mais longos, sendo mais adequada para os domínios legal e legislativo. Souza et al. (2021b) concluíram que o BM25L é mais adequado para o cenário legislativo brasileiro, considerando o corpus LCC, portanto decidiu-se pela avaliação das duas variantes neste trabalho.

A função de pontuação do Okapi BM25 estima a relevância de um documento para uma consulta baseado nos termos da consulta que aparecem no documento. Já o BM25L tenta corrigir a preferência do Okapi BM25 por documentos mais curtos, modificando a função de pontuação para melhorar a pontuação de documentos mais longos, através de um novo parâmetro δ.

Para este estudo, os valores recomendados pelos artigos originais para os parâmetros presentes nas fórmulas do Okapi BM25 e BM25L foram utilizados: $k_1 = 1{,}5$, $b = 0{,}75$ e $\delta = 0{,}5$.

4.2.2. Tamanho máximo de entrada

O tamanho máximo de entrada dos modelos SBERT, isto é, a quantidade de tokens que podem ser utilizados como entrada para o modelo, pode ser um fator que impacta diretamente no seu desempenho. Para documentos maiores que esse limite, o texto é truncado e somente a quantidade limite de tokens é considerada para geração dos *embeddings*. Além disso, os modelos selecionados para este estudo possuem diferentes tamanhos máximos de entrada, variando de 32 a 512 tokens.

A arquitetura SBERT permite que o tamanho de entrada dos modelos seja modificado para o máximo disponível do modelo base — neste caso, para os 512 tokens do BERT. Por conta disso, e visando a realização de uma comparação mais justa entre os modelos baseados em BERT, eles foram avaliados em dois cenários: 1) mantendo o seu tamanho padrão de entrada (explicitados na Tabela 1); e 2) modificando-o para 512 tokens para todos os modelos. Portanto, cada modelo que possui um tamanho máximo de entrada menor que 512 foi avaliado duas vezes: JurisBERT, BERTimbauLaw, LegalBERTPT-Br, LaBSE, Multilingual MPNet e Multilingual MiniLM.

Dessa forma, os modelos SBERT podem ser avaliados diminuindo o tamanho do truncamento dos tokens das proposições legislativas para a geração dos *embeddings*, ao passo que também serão comparados com sua configuração original, isto é, da forma com que foram construídos.

Corpus	# consultas	Tam. médios das consultas	# proposições /consulta
Ulysses-RFCorpus	646	20	12
Preliminary Search Corpus (PSC)	1.934	80	1-20
Legislative Consultation Corpus (LCC)	295	66	1

Tabela 2: Sumário dos corpora de consultas utilizados, apontando a quantidade de consultas, o tamanho médio em número de palavras e a quantidade de proposições julgadas para cada consulta.

4.2.3. Pré-processamento

A etapa de pré-processamento do texto é uma etapa crucial para RI com algoritmos como o BM25. Técnicas como a de *stemming* podem melhorar consideravelmente o desempenho da recuperação de documentos, inclusive no domínio legislativo (Souza et al., 2021a,b), já que elas podem mitigar o problema de incompatibilidade de vocabulário (Manning et al., 2008; Caseli & Nunes, 2024). Por exemplo, Souza et al. (2021b) relataram um aumento de 0,6678 para 0,7356 na métrica de Revocação@20 para o corpus LCC ao utilizar o pré-processamento correto para o BM25L.

Dessa forma, optou-se por realizar os experimentos com as variantes do BM25 em dois cenários: com e sem técnicas de pré-processamento. A escolha pelo cenário sem pré-processamento visa equiparar os algoritmos BM25 aos modelos baseados em BERT, os quais são geralmente treinados e utilizados sem o pré-processamento do texto. Ao mesmo tempo, também se procurou avaliar as variantes do BM25 da forma que elas podem obter um melhor desempenho, isto é, com as técnicas de pré-processamento adequadas.

As técnicas de pré-processamento escolhidas foram aquelas que alcançaram o melhor desempenho no trabalho de Souza et al. (2021b): remoção de pontuação, acentuação e *stopwords*; *stemming* com o algoritmo Savoy (Savoy, 2006); e uma combinação de *unigram* e *bigram*.

4.2.4. Tempo de processamento

Os experimentos foram realizados utilizando o cluster Euler[19] do Centro de Ciências Matemáticas Aplicadas à Indústria (CeMEAI) da Universidade de São Paulo (USP). Esse cluster permite a execução de código utilizando processamento via GPU com as seguintes especificações: dois processadores Intel Xeon E5-2650v4 de 2.2 GHz com doze núcleos, 128 GB DDR3 1866MHz de memória e 1 GPU Nvidia Tesla P100 - 3584 Cuda cores - 16GB.

[19] https://euler.cemeai.icmc.usp.br

Para comparar os modelos SBERT com os algoritmos BM25 para além do desempenho na recuperação de documentos — o qual é computado através das métricas de avaliação —, calculou-se também o tempo de processamento gasto por essas técnicas para todas as etapas necessárias para o processo de RI. Assim, pode-se avaliar o custo computacional da escolha pela utilização de modelos de linguagem ou algoritmos BM25.

Para os modelos de linguagem, mediu-se o tempo despendido para o carregamento do modelo, para a geração dos *embeddings* e para a recuperação de documentos para uma consulta. Já para os algoritmos BM25, o tempo gasto para a etapa de pré-processamento e para a construção do modelo foi medido, bem como o tempo para a recuperação dos documentos para a mesma consulta. Cada experimento foi realizado cinco vezes, com a média das cinco execuções sendo reportada na seção de resultados.

4.3. Métricas de avaliação

Como métricas de avaliação para os modelos SBERT e para o BM25, foram utilizadas Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) e Mean R-Precision (MRP) para os três corpora. Entretanto, como o corpus LCC contém apenas um único documento relevante para cada consulta, os resultados de MRR seriam os mesmos que os de MAP. Além disso, como o Ulysses-RFCorpus contém diferentes níveis de relevância, os resultados em termos de Normalized Discounted Cumulative Gain (nDCG) também são apresentados para esse corpus.

MAP é uma métrica de avaliação bastante utilizada para avaliar sistemas de RI. Combinando Revocação (*Recall*) e Precisão, ela calcula a média dos valores de Precisão a cada documento relevante que foi retornado (Zhang & Zhang, 2009). A métrica RR aponta a posição do primeiro documento relevante retornado para cada consulta (Caseli & Nunes, 2024). R-Precision, ou a Precisão no ponto R, mede a Precisão considerando a quantidade de documentos que são

relevantes para aquela consulta (Beitzel et al., 2009). Por fim, nDCG é baseada na suposição de que documentos altamente relevantes são mais valiosos que documentos que possuam uma relevância apenas marginal. Assim, como, em certos corpora, os documentos não são julgados com o mesmo grau de relevância, os documentos com níveis mais altos devem ser identificados e apresentados aos usuários nas primeiras posições (Järvelin & Kekäläinen, 2009).

Para o cálculo de cada métrica, é necessário definir a quantidade de documentos a serem retornados pelos algoritmos. Para o Ulysses-RFCorpus, foi considerada a recuperação de 12 documentos, devido ao fato de que 12 foi o número máximo de proposições julgadas pelos consultores da Conle para cada consulta. Para os outros dois corpora, o número de documentos a serem recuperados foi 20, já que as consultas do corpus PSC contém 20 documentos relevantes no máximo, enquanto que Souza et al. (2021b) e dos Santos et al. (2024) usaram a métrica Revocação a 20 documentos (R@20) para o corpus LCC.

Por fim, para avaliar os métodos com significância estatística, o *Student's t-test* (Student, 1908) foi utilizado para a comparação entre o desempenho dos modelos baseados em BERT e o dos algoritmos BM25. Como apontado por Urbano et al. (2019), esse teste é o teste de significância mais robusto para RI. No seu trabalho, os autores utilizaram a métrica MAP para realizar a avaliação, sendo assim, o MAP também foi escolhido para a análise neste estudo.

5. Resultados e discussão

A Tabela 3 apresenta os resultados alcançados pelas variantes do BM25 e pelos modelos SBERT para os três corpora. Os resultados expõem o desempenho superior dos algoritmos BM25 pré-processados em comparação a todos os modelos de linguagem. O melhor desempenho alcançado devido ao uso de pré-processamento para o BM25 também reafirma a importância dessa etapa para esse algoritmo, a qual já havia sido demonstrada por Souza et al. (2021a,b); dos Santos et al. (2024). Porém, mesmo as versões sem pré-processamento do BM25L e Okapi BM25 alcançaram os melhores resultados na maioria dos casos, somente sendo superados pela versão do BERTimbau ajustada por dos Santos et al. (2024), e apenas para o corpus PSC.

Além disso, utilizando o *Student's t-test* para medir o grau de significância da diferença entre esses resultados, obteve-se valores-p de 0,093 e 0,003 comparando os valores de Average Precision (AP) alcançados pela versão com *fine-tuning* do BERTimbau e, respectivamente, as versões sem pré-processamento do BM25L e Okapi BM25, para o corpus PSC. Isso mostra que só existe uma diferença significativa na comparação com o Okapi BM25, considerando um nível de confiança de 95%. Sendo assim, não se pode afirmar que o FT BERTimbau supere, com significância estatística, a versão sem pré-processamento do BM25L para o corpus PSC, ao passo que essa versão do BM25L pode ser apontada como a mais adequada para os outros dois corpora, superando estatisticamente o FT BERTimbau. As Figuras 2, 3 e 4 apresentam todos os valores-p das comparações entre os modelos para, respectivamente, os corpora Ulysses-RFCorpus, PSC e LCC.

Utilizando novamente o *Student's t-test* para comparar o desempenho do FT BERTimbau com as versões pré-processadas dos algoritmos BM25, percebeu-se que também não há diferença estatística nos desempenhos dessas técnicas. Os valores-p obtidos para a comparação das versões pré-processadas do Okapi BM25 e BM25L com a versão ajustada do BERTimbau — de, respectivamente, 0,063 e 0,057 — mostram que não há diferença estatisticamente significativa na utilização dessas técnicas para o corpus PSC com um nível de confiança de 95%. Entretanto, para os demais corpora, tanto o BM25L pré-processado quanto o Okapi BM25 com pré-processamento alcançaram resultados superiores a todos os modelos SBERT.

Convém destacar que o corpus PSC pode ser considerado o mais difícil para realização de RI dentre os três, já que foi resultado de um processo manual de seleção dos documentos relevantes. Essa "dificuldade" está demonstrada pelo fato de o melhor algoritmo ter alcançado um MAP de apenas 0,1693, não apresentando uma variação expressiva em relação aos outros algoritmos, diferentemente do apresentado para os corpora Ulysses-RFCorpus e LCC. Para esses dois outros corpora, o algoritmo BM25L pré-processado obteve um desempenho de 17% a 57% superior em comparação ao melhor modelo de linguagem e considerando a métrica MAP.

Outro ponto encontrado ao analisar os resultados é que o tamanho de entrada do modelo de linguagem não impacta diretamente no seu desempenho para a totalidade dos casos analisados. Ao aumentar o tamanho de entrada de modelos que possuíam um limite inferior a 512, observou-se uma queda no desempenho na maioria dos casos. Isso mostra que convém utilizar os modelos da forma que eles foram treinados, considerando seu tamanho de entrada original.

Modelo	Ulysses-RFCorpus MAP	MRP	MRR	nDCG	PSC MAP	MRP	MRR	LCC MAP	MRP
Okapi BM25	0,3704	0,3834	0,6718	0,5399	0,1028	0,0840	0,1766	0,3762	0,2949
Okapi BM25 (pré-processado)	0,7171	0,6585	0,8598	0,7935	0,1681	**0,1369**	0,2816	**0,4826**	0,3898
BM25L	0,3801	0,3913	0,6811	0,5429	0,1107	0,0871	0,1843	0,4167	0,3322
BM25L (pré-processado)	**0,7677**	**0,6849**	**0,8596**	**0,8274**	**0,1693**	0,1322	**0,2858**	0,4809	**0,3932**
BERTimbau (512)	0,0113	0,0209	0,0542	0,0490	0,0188	0,0164	0,0438	0,0300	0,0203
Legal-BERTimbau (512)	0,1221	0,1469	0,3320	0,2871	0,0886	0,0720	0,1613	0,2188	0,1593
JurisBERT (384)	0,0744	0,0924	0,2482	0,1862	0,0425	0,0331	0,0849	0,1360	0,0983
JurisBERT (512)	0,0654	0,0822	0,2153	0,1752	0,0424	0,0340	0,0878	0,1095	0,0712
BERTimbauLaw (384)	0,1239	0,1571	0,3481	0,2871	0,0801	0,0659	0,1522	0,2211	0,1797
BERTimbauLaw (512)	0,1240	0,1574	0,3402	0,2857	0,0734	0,0599	0,1423	0,1947	0,1322
LegalBert-pt (512)	0,0482	0,0670	0,1792	0,1394	0,0444	0,0397	0,0924	0,1076	0,0678
LegalBERTPT-Br (32)	0,0118	0,0171	0,0618	0,0373	0,0087	0,0075	0,0230	0,0046	0,0000
LegalBERTPT-Br (512)	0,0440	0,0626	0,1654	0,1402	0,0440	0,0364	0,0916	0,1048	0,0678
GovBERT-BR (512)	0,0172	0,0260	0,0735	0,0597	0,0336	0,0286	0,0725	0,0627	0,0407
FT BERTimbau (512)	**0,1902**	**0,2163**	**0,4456**	**0,4070**	**0,1214**	**0,0941**	**0,2024**	**0,3117**	**0,2271**
FT LegalBert-pt (512)	0,1220	0,1523	0,2998	0,3050	0,0890	0,0639	0,1569	0,1991	0,1356
LaBSE (256)	0,0893	0,1080	0,2910	0,2268	0,0685	0,0549	0,1264	0,2018	0,1525
LaBSE (512)	0,0968	0,1178	0,2947	0,2239	0,0665	0,0530	0,1286	0,1962	0,1458
Multi MPNet (128)	0,1633	0,1891	0,4204	0,3491	0,0933	0,0752	0,1696	0,2159	0,1525
Multi MPNet (512)	0,0458	0,0656	0,1652	0,1602	0,0415	0,0372	0,0952	0,0805	0,0508
Multi MiniLM (128)	0,1208	0,1454	0,3376	0,2692	0,0695	0,0598	0,1378	0,1933	0,1424
Multi MiniLM (512)	0,0019	0,0028	0.,120	0,0064	0,0047	0,0041	0,0125	0,0060	0,0034

Tabela 3: Resultados alcançados pelas variantes do BM25 e pelos modelos de linguagem. O tamanho máximo de entrada de cada modelo está indicado entre parênteses, enquanto que os resultados obtidos para cada corpus pelo melhor modelo de cada categoria (modelos de linguagem e BM25) estão em negrito. "FT" é abreviação de "*fine-tuned*", para os modelos ajustados por dos Santos et al. (2024).

Modelos mais genéricos, ou seja, sem a definição de um domínio específico, como o Multilingual MPNet e o LaBSE, mesmo possuindo um tamanho de entrada menor, não apresentaram diferença significativa em comparação aos modelos baseados no BERTimbau e pré-treinados para o domínio legal. Isso, aliado ao fato de que os melhores resultados foram obtidos pelo FT BERTimbau, o qual foi ajustado especificamente para o domínio legislativo brasileiro, mostra uma diferença entre os documentos jurídicos e legislativos, a qual também foi apontada por Vitório et al. (2025). Além disso, demonstra a importância do *fine-tuning* para tarefas específicas (Chalkidis et al., 2020), já que dos Santos et al. (2024) realizaram um *fine-tuning* específico para RI.

Ao analisar o tempo de processamento despendido por cada algoritmo para diversas etapas do processo de RI, a Tabela 4 traz a informação em segundos. Dessa forma, pode-se notar que os algoritmos BM25, ao fazer uso de pouco pré-processamento, realizaram a recuperação dos documentos em um tempo menor que todos os modelos de linguagem. Além disso, mesmo com a adoção de diversas técnicas de pré-processamento, as variantes do BM25 também despenderam menos tempo computacional que a grande maioria dos modelos SBERT. Nesse caso, apenas o LegalBERTPT-Br, o qual possui um tamanho de entrada muito pequeno, e o Multilingual MiniLM demandaram menos tempo computacional que o BM25.

Ao combinar as análises do desempenho dos algoritmos/modelos com o tempo despendido por eles, pode-se afirmar que a utilização do BM25 é mais recomendada do que a dos modelos baseados em BERT para a tarefa de RI no domínio legislativo brasileiro. O melhor modelo de linguagem, embora tenha alcançado um desempenho que não apresentou diferença estatística significativa em comparação aos algoritmos BM25 para alguns cenários, despendeu um tempo cerca de 12 vezes maior que o BM25 com documentos pré-processados e cerca de 87 vezes maior que os algoritmos sem pré-processamento. Isso mostra que o BM25 ainda obtém melhores resultados com um menor custo computacional.

Pôde-se notar que o principal gargalo durante a execução de RI com os modelos de linguagem se encontra na geração dos *embeddings* (correspondendo a 99% do tempo despendido), enquanto que, para os algoritmos BM25, o maior gargalo está presente na etapa de pré-processamento (de 54% a 85% do tempo), a qual, mesmo assim, despende um tempo inferior que a etapa de *embeddings*. Porém, convém relatar que o tempo despendido na etapa da recuperação propriamente dita dos documentos é cerca de 20 vezes maior

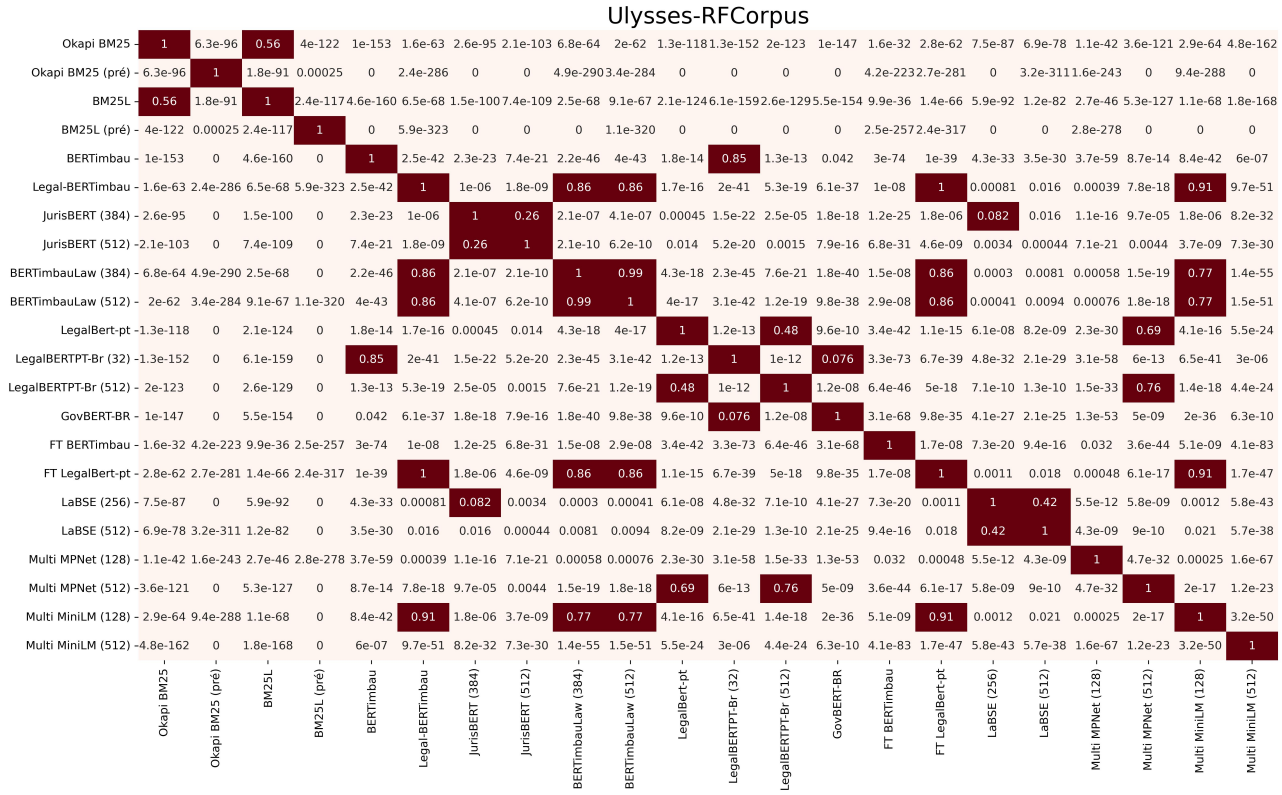

Figura 2: Valores-p obtidos com o *Student's t-test* para a comparação dos modelos utilizando o Ulysses-RFCorpus. Os valores em destaque indicam os casos nos quais não houve diferença significativa.

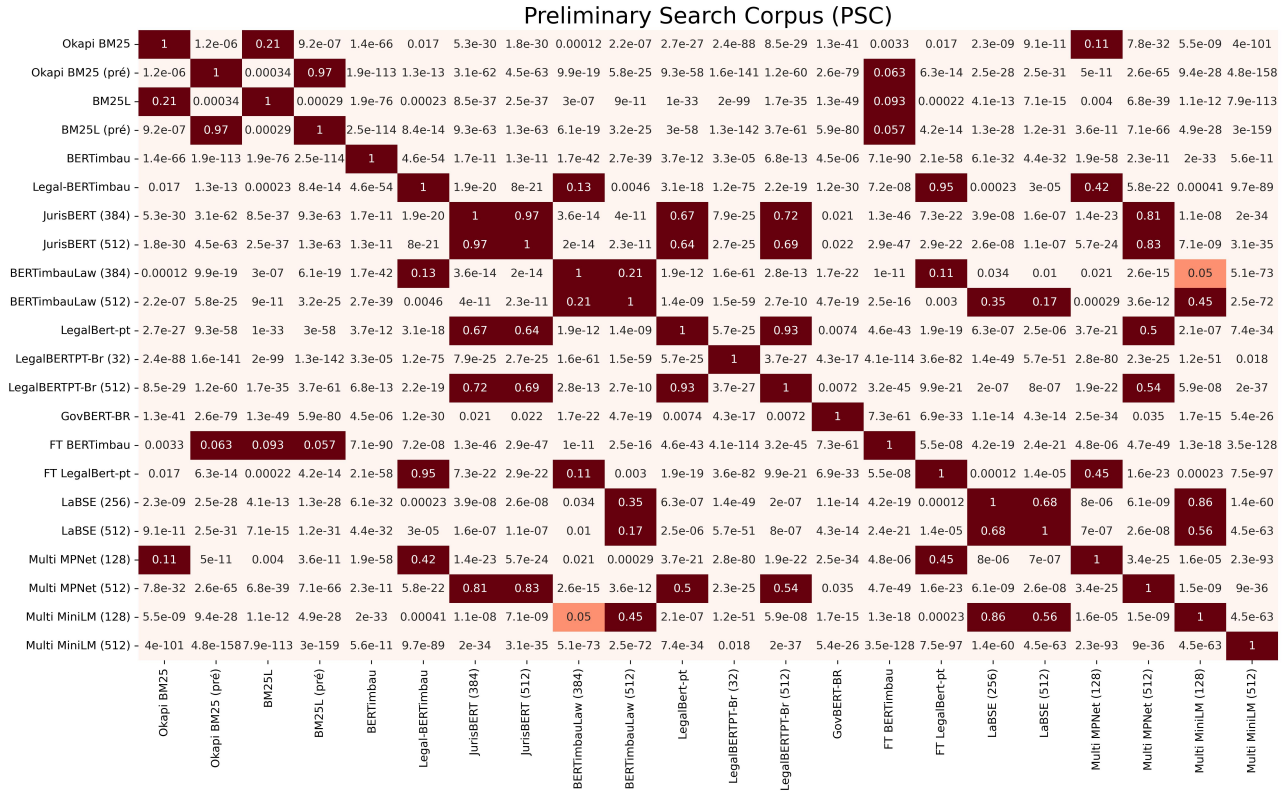

Figura 3: Valores-p obtidos com o *Student's t-test* para a comparação dos modelos utilizando o corpus PSC. Os valores em destaque indicam os casos nos quais não houve diferença significativa.

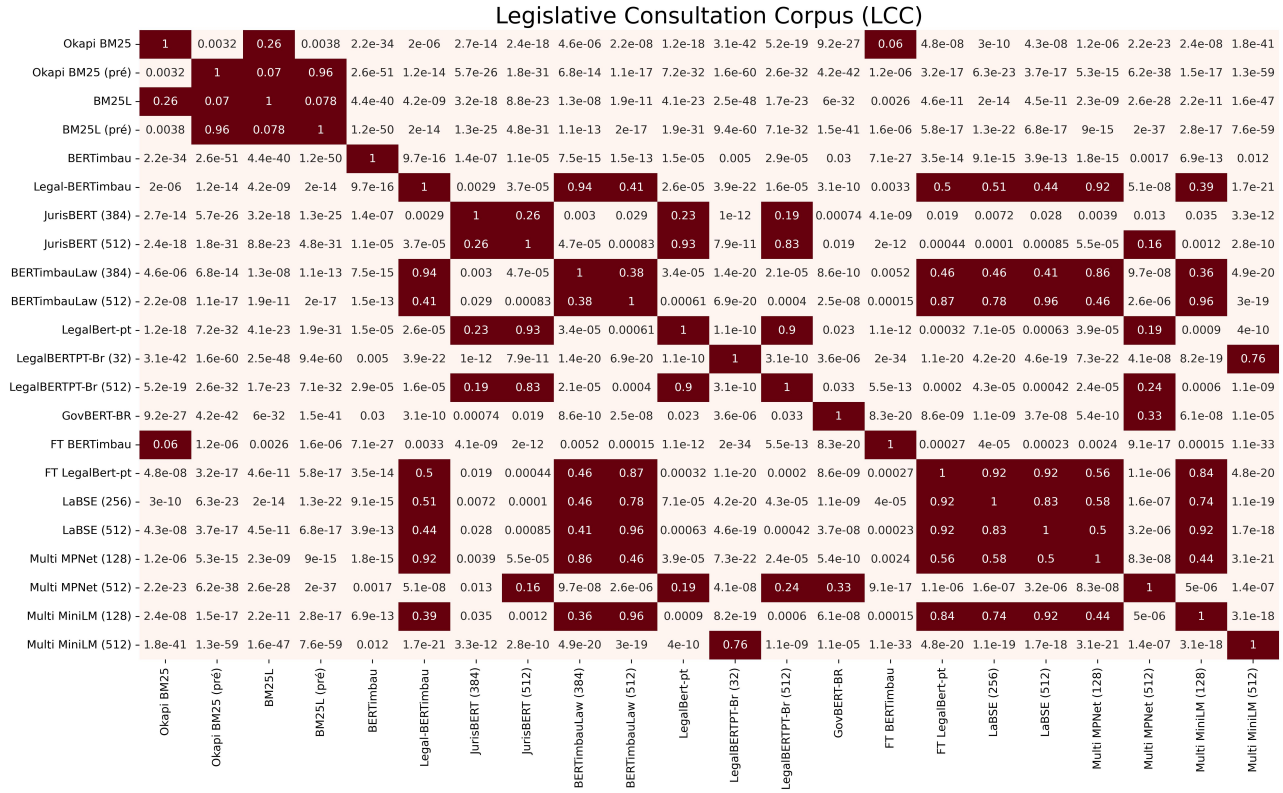

Figura 4: Valores-p obtidos com o *Student's t-test* para a comparação dos modelos utilizando o corpus LCC. Os valores em destaque indicam os casos nos quais não houve diferença significativa.

para o BM25 do que para os modelos SBERT, os quais utilizam uma medida de similaridade simples para realização dessa etapa, como a similaridade do cosseno utilizada neste estudo. Essa diferença, embora não impacte diretamente na recuperação de documentos para uma consulta individual (correspondendo a menos de um segundo), pode ter um impacto maior ao processar um número muito grande de consultas, enquanto que a etapa de geração dos *embeddings* é realizada apenas uma vez para cada base de documentos.

6. Conclusão

Neste trabalho, uma análise comparativa abrangente de modelos SBERT foi conduzida para a tarefa de recuperação de documentos no domínio legislativo brasileiro. Comparando o desempenho dos modelos de linguagem a *baselines* compostas por variantes do algoritmo BM25, objetivou-se responder à pergunta: *tendo em vista os recentes avanços trazidos pelos modelos de linguagem, o algoritmo BM25 ainda os supera para Recuperação de Informação no cenário legislativo brasileiro, considerando a restrição do tamanho de entrada dos modelos?*

Três pontos principais foram identificados a partir dessa análise, os quais constituem a principal contribuição deste trabalho. Em primeiro lugar, os resultados demonstraram que os algoritmos BM25 superaram os modelos baseados em BERT para a tarefa de RI com documentos legislativos, inclusive em cenários nos quais não foram utilizadas técnicas de pré-processamento. Vale ressaltar que o uso de pré-processamento pode melhorar o desempenho desse tipo de algoritmo. Dos 12 modelos SBERT avaliados, somente o modelo BERTimbau ajustado com dados legislativos por dos Santos et al. (2024) não apresentou diferença estatística significativa em comparação às variantes do BM25, e apenas para uma das três bases de dados avaliadas. Em comparação com os outros modelos, as variantes do BM25 foram superiores com significância estatística. Esse desempenho inferior, somado ao alto custo computacional de execução dos modelos de linguagem, apontam que o BM25 ainda é mais adequado para a realização de RI com documentos escritos em Português. Como visto também no trabalho de Mandal et al. (2021), esse desempenho inferior dos modelos de linguagem pode ser explicado pela limitação de tamanho de entrada do BERT.

Modelo	Pré-process.	Construção	*Embeddings*	Recuperação	Total
Okapi BM25	32,04	25,93	-	0,89	58,86
Okapi BM25 (pré-processado)	344,33	57,98	-	0,93	403,24
BM25L	**31,63**	25,52	-	0,76	**57,90**
BM25L (pré-processado)	344,65	57,72	-	0,95	403,31
BERTimbau (512)	-	5,95	5099,98	0,04	5105,97
Legal-BERTimbau (512)	-	10,56	5097,71	0,04	5108,31
JurisBERT (384)	-	8,63	1191,54	**0,03**	1200,20
JurisBERT (512)	-	6,21	1553,81	**0,03**	1560,04
BERTimbauLaw (384)	-	8,36	1210,01	**0,03**	1218,40
BERTimbauLaw (512)	-	7,15	1586,51	**0,03**	1593,70
LegalBert-pt (512)	-	4,93	1586,38	**0,03**	1591,34
LegalBERTPT-Br (32)	-	**4,14**	**147,08**	**0,03**	151,25
LegalBERTPT-Br (512)	-	5,13	1586,35	**0,03**	1591,52
GovBERT-BR (512)	-	**4,14**	1586,59	**0,03**	1590,77
FT BERTimbau (512)	-	8,11	5098,54	0,04	5106,69
FT LegalBert-pt (512)	-	7,13	5095,68	0,04	5102,85
LaBSE (256)	-	6,99	793,76	**0,03**	800,78
LaBSE (512)	-	7,23	1559,83	**0,03**	1567,09
Multi MPNet (128)	-	6,83	415,52	**0,03**	422,38
Multi MPNet (512)	-	7,01	1567,47	**0,03**	1574,51
Multi MiniLM (128)	-	6,18	181,71	**0,03**	187,93
Multi MiniLM (512)	-	5,84	710,63	**0,03**	716,50

Tabela 4: Tempo gasto, em segundos, pelas variantes do BM25 e pelos modelos baseados em BERT para cada etapa do processo de recuperação de documentos. O tamanho máximo de entrada de cada modelo está indicado entre parênteses, enquanto que os menores tempos para cada etapa estão em negrito. "FT" é abreviação de "*fine-tuned*", para os modelos ajustados por dos Santos et al. (2024).

Em segundo lugar, a importância de *fine-tuning* para tarefas específicas também foi confirmada, pois o modelo de linguagem o qual alcançou os melhores resultados foi ajustado com dados legislativos brasileiros para a tarefa de RI. Por fim, as diferenças nos desempenhos dos modelos avaliados também reafirmam a existência de diferenças consideráveis entre documentos legislativos e outros tipos de documentos legais, como os jurídicos. Esse ponto demonstra a necessidade da construção, treinamento ou adaptação de modelos específicos para esse tipo de documento, não se mostrando adequada a utilização de modelos treinados com outros tipos de dados, ainda que dentro do domínio legal.

Como trabalhos futuros, pretende-se ampliar essa avaliação incluindo sistemas híbridos, os quais combinam algoritmos BM25 com modelos de linguagem, como os apresentados por Melo et al. (2023), Cordeiro et al. (2023) e dos Santos et al. (2024). A utilização de técnicas de *chunking*, como a de janela deslizante, por exemplo, também pode ser avaliada, as quais podem mitigar o problema do tamanho de entrada do BERT. Além disso, pode-se executar a avaliação com outros tipos de documentos legais, bem como documentos legislativos escritos em outras línguas.

Agradecimentos

Pesquisa desenvolvida com utilização dos recursos computacionais do Centro de Ciências Matemáticas Aplicadas à Indústria (CeMEAI), financiados pela FAPESP (proc. 2013/07375-0).

Douglas Vitório, Adriano L. I. Oliveira e André Carlos Ponce de Leon Ferreira de Carvalho são financiados pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Agradecemos também à Câmara dos Deputados e ao Instituto Nacional de Inteligência Artificial (IAIA) pelo suporte e financiamento desta pesquisa.

Referências

Luz de Araujo, Pedro Henrique, Teófilo Emídio de Campos, Fabricio Ataides Braz & Nilton Correia da Silva. 2020. VICTOR: a dataset for Brazilian legal documents classification. Em *12th Language Resources and Evaluation Conference (LREC)*, 1449–1458.

Beitzel, Steven M., Eric C. Jensen & Ophir Frieder. 2009. Average R-Precision. Em *Ency-*

clopedia of Database Systems, 195. Springer. 10.1007/978-0-387-39940-9_491

Brandt, Mariana Baptista. 2020. *Modelagem da informação legislativa: arquitetura da informação para o processo legislativo brasileiro*: Universidade Estadual Paulista. Tese de Doutoramento.

Cantador, Iván & Lara Quijano Sánchez. 2020. Semantic annotation and retrieval of parliamentary content: A case study on the Spanish congress of deputies. Em *1st Joint Conference of the Information Retrieval Communities in Europe (CIRCLE)*,

Caseli, Helena M. & Maria Graça V. Nunes (eds.). 2024. *Processamento de linguagem natural: Conceitos, técnicas e aplicações em português*. BPLN 3rd edn.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras & Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. Em *Findings of the Association for Computational Linguistics*, 2898–2904. 10.18653/v1/2020.findings-emnlp.261

Chalkidis, Ilias & Dimitrios Kampas. 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law* 27. 171–198. 10.1007/s10506-018-9238-9

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. Em *58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. 10.18653/v1/2020.acl-main.747

Cordeiro, Nuno Pablo, João Dias & Pedro A. Santos. 2023. LeSSE—a semantic search engine applied to portuguese consumer law. Em *Portuguese Conference on Artificial Intelligence (EPIA)*, 118–130. 10.1007/978-3-031-49011-8_10

Câmara dos Deputados. 1993. Resolução da câmara dos deputados nº 48, de 1993.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186. 10.18653/v1/N19-1423

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan & Wei Wang. 2022. Language-agnostic BERT sentence embedding. Em *60th Annual Meeting of the Association for Computational Linguistics*, 878–891. 10.18653/v1/2022.acl-long.62

Gao, Tianyu, Xingcheng Yao & Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6894–6910. 10.18653/v1/2021.emnlp-main.552

Jaiswal, Aman & Evangelos Milios. 2023. Breaking the token barrier: Chunking and convolution for efficient long text classification with BERT. ArXiv [cs.CL]. 10.48550/arXiv.2310.20558

Järvelin, Kalervo & Jaana Kekäläinen. 2009. Discounted cumulated gain. Em *Encyclopedia of Database Systems*, 849–853. Springer. 10.1007/978-0-387-39940-9_478

Lin, Jimmy, Rodrigo Nogueira & Andrew Yates. 2022. *Pretrained transformers for text ranking: BERT and beyond*. Springer. 10.1007/978-3-031-02181-7

Lv, Yuanhua & ChengXiang Zhai. 2011. When documents are very long, BM25 fails! Em *34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1103–1104. 10.1145/2009916.2010070

Mandal, Arpan, Kripabandhu Ghosh, Saptarshi Ghosh & Sekhar Mandal. 2021. Unsupervised approaches for measuring textual similarity between legal court case reports. *Artificial Intelligence for Law* 29(3). 417–451. 10.1007/s10506-020-09280-2

Mandal, Arpan, Kripabandhu Ghosh, Arindam Pal & Saptarshi Ghosh. 2017. Automatic catchphrase identification from legal court case documents. Em *Conference on Information and Knowledge Management (CIKM)*, 2187––2190. 10.1145/3132847.3133102

Manning, Christopher D., Prabhakar Raghavan & Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press. 10.1017/CBO9780511809071

Maxwell, K. Tamsin & Burkhard Schafer. 2008. Concept and context in legal information retrieval. Em *Legal Knowledge and Information Systems (JURIX)*, 63–72. 10.3233/978-1-58603-952-3-63

Melo, Rui, Pedro A. Santos & João Dias. 2023. A semantic search system for the Supremo Tribunal de Justiça. Em *Portuguese Conference on Artificial Intelligence (EPIA)*, 142–154. 10.1007/978-3-031-49011-8_12

Reimers, Nils & Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Em *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. 10.18653/v1/D19-1410

Reimers, Nils & Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4512–4525. 10.18653/v1/2020.emnlp-main.365

Robertson, Stephen E., Steve Walker, Susan Jones, Micheline Hancock-Beaulieu & Mike Gatford. 1994. Okapi at TREC-3. Em *3rd Text REtrieval Conference (TREC)*, 109–126

Sansone, Carlo & Giancarlo Sperlí. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems* 106. 101967. 10.1016/j.is.2021.101967

dos Santos, José Antônio, Ellen Souza, Carmelo J. A. Bastos Filho, Hidelberg O. Albuquerque, Douglas Vitório, Danilo Carlos Gouveia de Lucena, Nádia Silva & André de Carvalho. 2024. HIRS: A hybrid information retrieval system for legislative documents. Em *Portuguese Conference on Artificial Intelligence (EPIA)*, 320–331. 10.1007/978-3-031-73497-7_26

Savoy, Jacques. 2006. Light stemming approaches for the French, Portuguese, German and Hungarian languages. Em *ACM Symposium on Applied Computing (SAC)*, 1031–1035. 10.1145/1141277.1141523

Silva, Mariana O., Gabriel P. Oliveira, Lucas G. L. Costa & Gisele L. Pappa. 2025. GovBERT-BR: A BERT-based language model for Brazilian Portuguese governmental data. Em *Brazilian Conference on Intelligent Systems (BRACIS)*, 19–32. 10.1007/978-3-031-79032-4_2

Silva, Nádia F. F. da, Marília Costa R. Silva, Fabíola S. F. Pereira, João Pedro M. Tarrega, João Vitor P. Beinotti, Márcio Fonseca, Francisco Edmundo de Andrade & André C. P. de L. F. de Carvalho. 2021. Evaluating topic models in portuguese political comments about bills from Brazil's chamber of deputies. Em *Brazilian Conference on Intelligent Systems (BRACIS)*, 104–120. 10.1007/978-3-030-91699-2_8

Silveira, Raquel, Caio Ponte, Vitor Almeida, Vládia Pinheiro & Vasco Furtado. 2023. LegalBert-pt: A pretrained language model for the Brazilian Portuguese legal domain. Em *Brazilian Conference on Intelligent Systems (BRACIS)*, 268–282. 10.1007/978-3-031-45392-2_18

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu & Tie-Yan Liu. 2020. MPNet: masked and permuted pre-training for language understanding. Em *34th International Conference on Neural Information Processing Systems (NIPS)*,

Souza, Ellen, Gyovana Moriyama, Douglas Vitório, André Carlos Ponce de Leon Ferreira de Carvalho, Nádia Félix, Hidelberg Albuquerque & Adriano L. I. Oliveira. 2021a. Assessing the impact of stemming algorithms applied to Brazilian legislative documents retrieval. Em *13th Brazilian Symposium in Information and Human Language Technology (STIL)*, 227–236. 10.5753/stil.2021.17802

Souza, Ellen, Douglas Vitório, Gyovana Moriyama, Luiz Santos, Lucas Martins, Mariana Souza, Márcio Fonseca, Nádia Félix, André Carlos Ponce de Leon Ferreira de Carvalho, Hidelberg O. Albuquerque & Adriano L. I. Oliveira. 2021b. An information retrieval pipeline for legislative documents from the Brazilian Chamber of Deputies. Em *Legal Knowledge and Information Systems*, 119–126. IOS Press. 10.3233/FAIA210326

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. Em *Brazilian Conference on Intelligent Systems (BRACIS)*, 403–417. 10.1007/978-3-030-61377-8_28

Student. 1908. The probable error of a mean. *Biometrika* 6(1). 1–25. 10.2307/2331554

Urbano, Julián, Harlley Lima & Alan Hanjalic. 2019. Statistical significance testing in information retrieval: An empirical analysis of type I, type II and type III errors. Em *42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 505–14. 10.1145/3331184.3331259

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. Em *31st International Conference on Neural Information Processing Systems (NIPS)*, 6000–6010.

Viegas, Charles F. O., Bruno C. Costa & Renato P. Ishii. 2023. JurisBERT: A new approach that converts a classification corpus into an STS one. Em *International Conference on Computational Science and Its Applications (ICCSA)*, 349–365. 10.1007/978-3-031-36805-9_24

Vitório, Douglas, Ellen Souza, Lucas Martins, Nádia F. F. da Silva, André Carlos Ponce de Leon de Carvalho, Adriano L. I. Oliveira & Francisco Edmundo de Andrade. 2025. Building a relevance feedback corpus for legal information retrieval in the real-case scenario of the Brazilian Chamber of Deputies. *Language Resources and Evaluation* 59. 1257–1277. 10.1007/s10579-024-09767-3

Wagner Filho, Jorge A., Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. Em *11th International Conference on Language Resources and Evaluation (LREC)*,

Wang, Jiajia, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md Tahmid Rahman Laskar & Amran Bhuiyan. 2024. Utilizing BERT for information retrieval: Survey, applications, resources, and challenges. *ACM Computing Surveys* 56(7). 1–33. 10.1145/3648471

Wang, Wenhui, Hangbo Bao, Shaohan Huang, Li Dong & Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. Em *Findings of the Association for Computational Linguistics*, 2140–2151. 10.18653/v1/2021.findings-acl.188

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang & Ming Zhou. 2020. MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. Em *34th International Conference on Neural Information Processing Systems (NIPS)*, 5776–5788.

Warner, Benjamin, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard & Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. ArXiv [cs.CL]. 10.48550/arXiv.2412.13663

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 38–45. 10.18653/v1/2020.emnlp-demos.6

Zhang, Ethan & Yi Zhang. 2009. Average precision. Em *Encyclopedia of Database Systems*, 192–193. Springer US. 10.1007/978-0-387-39940-9_482