Hybrid Summarization for Brazilian Judicial Decisions

Gabriele S. Araújo[1], **Ewaldo E. C. Santana**[1,2], **Fábio M. F. Lobato**[1,3,4]

[1]Programa de Pós-Graduação em Engenharia da Computação e Sistemas (PECS)
Universidade Estadual do Maranhão (UEMA)
São Luís – MA – Brasil

[2]Programa de Pós-graduação em Engenharia Elétrica (PPGEE)
Universidade Federal do Maranhão (UFMA)
São Luís – MA – Brasil

[3]Instituto de Engenharia e Geociências (IEG)
Universidade Federal do Oeste do Pará (UFOPA)
Santarém – PA – Brasil

[4]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
São Carlos – SP – Brasil

gabriele.ar4ujo@gmail.com, ewaldosantana@professor.uema.br
fabio.lobato@ufopa.edu.br

Abstract. ***Research Context:*** *Judicial decisions in Brazil have a complex textual structure, comprising reports, votes, and summaries, which hinders systematic analysis and affects information intelligibility. This also applies to many other countries in the Global South. This situation challenges not only legal professionals but also Information Systems (IS) that support the processing and organization of large volumes of documents.* ***Scientific and/or Practical Problem:*** *Many systems struggle with lengthy texts due to language model limitations and the need to preserve technical accuracy and cohesion. This restricts the development of reliable solutions for judicial activities and transparency.* ***Proposed Solution and/or Analysis:*** *This study proposes and evaluates a hybrid summarization pipeline for decisions from the Federal Supreme Court, integrating extractive and abstractive methods to handle long documents without information loss.* ***Related IS Theory:*** *The research is grounded in Task-Technology Fit (TTF), justifying the alignment between the pipeline and the legal summarization task to support information-processing objectives. Additionally, Design Science Research is employed as the methodology for artifact development and evaluation.* ***Research Method:*** *The pipeline was implemented and tested on the RulingBR corpus, which comprises Brazilian Supreme Court decisions. Five summarization methods were compared: TF-IDF (baseline), BumbaBERT, Gemini-2.5, and two hybrid variants, evaluated using standard metrics (ROUGE, BERTScore, and METEOR).* ***Summary of Results:*** *Hybrid approaches balanced technical accuracy and textual cohesion, while chunked processing expanded the applicability of specialized models to long documents. Results demonstrate practical impact through faster case review and improved jurisprudential research.* ***Contributions and Impact on IS area:*** *The study contributes to IS by proposing strategies that support information governance in organizational environments*

while considering processing limitations. Aligned with the Grand Challenge of IS in the Open World, it provides guidance for scalable, transparent, and interoperable decision-support systems, improving institutional transparency and supporting legal activities in complex digital ecosystems.

1. Introdução

As decisões judiciais apresentam uma estrutura textual singular, marcada por linguagem técnica, múltiplos votos e argumentações distribuídas em diferentes seções, característica comum de sistemas de *civil law*, predominantes em países do Sul Global [Glenn 2014]. Essa complexidade organizacional torna o processamento automático de documentos jurídicos distinto de outros domínios que envolvem análise de dados textuais, demandando soluções que preservem tanto a precisão técnica quanto a coesão argumentativa [Fama et al. 2024]. A análise dessas decisões envolve identificar fundamentos jurídicos, precedentes citados e dispositivos finais, informações que não estão concentradas em um único trecho do texto, mas distribuídas ao longo de relatórios, votos individuais e ementas [Luz de Araujo et al. 2023]. Essa dispersão semântica dificulta o trabalho de profissionais do direito e de pesquisadores interessados em compreender o raciocínio por trás das deliberações.

Em Sistemas de Informação (SI), esse cenário representa um desafio de organização e de transformação da informação em conhecimento útil e acessível [Casimiro and Teixeira 2024, Wang 2024]. Nos últimos anos, diferentes estratégias computacionais têm sido exploradas para apoiar a gestão de informações jurídicas, com mais de 140 projetos aplicados no judiciário brasileiro, com foco em classificação de documentos, triagem processual e automação de atividades repetitivas [Moreira and de Souza Moura 2023, Casimiro and Teixeira 2024, CNJ 2024]. Entre as atividades que mais demandam inovação estão a produção e utilização de sumários jurídicos, como as ementas. Elas sintetizam os fundamentos e o dispositivo de decisões judiciais, atuando como elemento central para a indexação em sistemas de busca, a classificação de casos semelhantes e a disseminação do entendimento jurisprudencial [CNJ 2021, Zhang et al. 2025].

Nesse contexto, a sumarização automática de textos jurídicos surge como um recurso estratégico para apoiar a transformação digital do setor ao condensar textos extensos em versões reduzidas que mantêm elementos relevantes da decisão [Bhattacharya et al. 2019, Fama et al. 2024]. Diferentes métodos têm sido investigados, incluindo: (i) abordagens extrativas, baseadas na seleção de sentenças relevantes; (ii) métodos abstrativos que reescrevem a informação em nova forma textual; e (iii) combinações híbridas que buscam equilibrar precisão e coesão de ambos os métodos [Kirmani et al. 2018, Liu and Lapata 2019, El-Kassas et al. 2021]. Apesar dos avanços, persistem desafios específicos na aplicação dessas técnicas ao domínio jurídico, incluindo textos longos que excedem a capacidade de processamento dos modelos tradicionais, linguagem altamente especializada e a necessidade de preservar terminologia técnica e relações argumentativas complexas [Bhattacharya et al. 2019, El-Kassas et al. 2021, Lai et al. 2024].

Neste estudo, buscou-se mitigar essa lacuna por meio da comparação de cinco métodos de sumarização aplicados a decisões do Supremo Tribunal Federal (STF): (i)

uma abordagem extrativa baseada em *Term Frequency-Inverse Document Frequency* (TF-IDF); (ii) uma abordagem extrativa baseada em modelo de língua especializado (o BumbaBERT, proposto por [Carmo et al. 2024], com processamento fundamentado em *chunks* [Alva Principe et al. 2025]; (iii) uma abordagem abstrativa com modelos de linguagem generativos como o Gemini-2.5; e (iv–v) duas variantes híbridas que combinam seleção extrativa e refinamento abstrativo. O artefato desenvolvido consiste em um *pipeline* de sumarização híbrida adaptado ao domínio jurídico brasileiro, capaz de processar documentos longos e gerar sumários consistentes, equilibrando a qualidade técnica e a viabilidade computacional.

O processamento híbrido reduz inicialmente o tamanho textual por meio de métodos extrativos antes de aplicar técnicas de IA generativa, otimizando o uso de recursos limitados e preservando a qualidade dos sumários [Jiang et al. 2024, Kuş and Acı 2024]. Embora estudos anteriores tenham investigado combinações semelhantes de modelos especializados com técnicas generativas, sua aplicação no contexto jurídico brasileiro permanece pouco explorada [Zhang et al. 2025, Araújo et al. 2025]. Para validação do trabalho, o *dataset* RulingBR foi utilizado, pois é um corpus bem estabelecido e amplamente referenciado na literatura de sumarização jurídica, que contém decisões do STF estruturadas em seções específicas, como relatório, fundamentação, votos e ementa, o que permite análise controlada e comparação com trabalhos anteriores [Feijó and Moreira 2018].

As principais contribuições deste estudo estão na articulação entre SI e gestão da informação jurídica em larga escala, fornecendo evidências experimentais sobre estratégias eficazes de sumarização no contexto brasileiro. Essa contribuição se alinha diretamente aos Grandes Desafios de SI no Mundo Aberto, sobretudo os relacionados à decisão escalável, transparente e interoperável, melhorando a transparência institucional e suportando atividades legais em ecossistemas digitais complexos e ricos em informação [Boscarioli et al. 2017]. Ademais, sua contribuição técnica consiste na implementação estratégica de síntese de conteúdo que permite ao BumbaBERT processar documentos longos, preservando suas representações mais relevantes e considerando também as limitações de recursos financeiros e computacionais enfrentadas por instituições ao lidar com a complexidade computacional desses sistemas.

O restante do artigo está organizado da seguinte forma: na Seção 2 os trabalhos relacionados são discutidos; a metodologia é descrita na Seção 3; na Seção 4 os resultados experimentais e suas implicações práticas são detalhados; por fim, na Seção 5 são apresentadas as considerações finais e direções futuras.

2. Trabalhos relacionados

A sumarização de textos jurídicos apresenta desafios únicos relacionados à complexidade linguística, à extensão documental e à necessidade de preservação da terminologia. Nesta seção, apresenta-se uma revisão de trabalhos relevantes sobre modelos de linguagem do domínio jurídico, sumarização extrativa, abordagens abstratas e métodos híbridos.

2.1. Modelos de linguagem especializados para o domínio jurídico

O ajuste de modelos de linguagem para o domínio jurídico tem gerado ganhos em diversas tarefas de Processamento de Linguagem Natural (PLN). Em [Chalkidis et al. 2020],

os autores desenvolveram Legal-BERT para textos jurídicos em inglês, demonstrando superioridade em relação ao BERT generalista, proposto por [Devlin et al. 2019], em tarefas de classificação de documentos jurídicos. Assim, evidenciando os benefícios da especialização de domínio, confirmados posteriormente por estudos que demonstraram vantagens do pré-treinamento em textos legais [Limsopatham 2021, Carmo et al. 2024].

No contexto brasileiro, [Souza et al. 2020] introduziram o BERTimbau, um modelo pré-treinado a partir do BERT em um corpus de domínio geral em português. Este motivou o ajuste fino em outros modelos, em especial no domínio jurídico, como o LegalBert-pt e o BumbaBERT. O LegalBert-pt foi pré-treinado em um corpus de 1,5 milhão de documentos legais de dez tribunais brasileiros [Silveira et al. 2023]; esses documentos consistiam em petições iniciais, petições, decisões e sentenças. O modelo apresentou vantagens em tarefas de PLN voltadas ao jurídico, em comparação ao BERTimbau-base. Já o BumbaBERT, proposto por [Carmo et al. 2024], é um modelo também pré-treinado para o domínio jurídico brasileiro, cujo treinamento foi realizado com um corpus de legislação e jurisprudência nacionais com mais de 5 milhões de documentos. Os resultados reportados por [Carmo et al. 2024] sugerem capacidades superiores em PLN jurídica, em comparação com modelos generalistas, confirmando evidências observadas em outros estudos.

Devido à arquitetura *Transformer*, os modelos baseados em BERT pecam ao lidar com processamento de documentos longos, pois o mecanismo de autoatenção, que exige que cada *token* na sequência de entrada atenda a todos os demais *tokens*, resulta em complexidade computacional quadrática em função do comprimento da sequência. Outras restrições de são do comprimento de entrada, onde modelos como o BERT-base processam sequências de até 512 *tokens* [Devlin et al. 2019]. Para lidar com essa limitação no contexto de entrada, técnicas de processamento baseadas em *chunks* têm sido exploradas, dividindo documentos longos em segmentos processáveis e agregando representações resultantes [Alva Principe et al. 2025], tornando-se viáveis para documentos jurídicos extensos e permitindo a aplicação de modelos especializados, mantendo representações semânticas ricas [Bhattacharya et al. 2019, de Castro and Ralha 2025].

Em [Silva Junior et al. 2025] foi conduzida a avaliação de 16 métodos de representação textual para similaridade semântica em português brasileiro no domínio legal, incluindo representações esparsas (*e.g.,* TF-IDF e LDA), *embeddings* estáticos (*e.g.,* word2vec, fastText e doc2vec) e contextualizados (*e.g.,* ELMo, BERT, Longformer, SBERT, SimCSE e DiffCSE). O objetivo foi identificar quais representações são mais eficazes para a recuperação de documentos semelhantes em cenários não supervisionados. Os autores compararam documentos do Superior Tribunal de Justiça (STJ) e do Tribunal de Contas da União (TCU), utilizando conjuntos rotulados heurísticos e por especialistas. Os resultados revelaram que representações simples, como TF-IDF e BM25, ainda produzem resultados competitivos, enquanto SBERT apresentou maior correlação com as anotações de especialistas. Adicionalmente, o estudo evidenciou que *Domain Adaptive Pre-training* (DAPT) modifica o espaço vetorial dos modelos *Transformer* de forma mais significativa do que as mudanças no mecanismo de atenção, observação relevante ao comparar BERT com Longformer.

Esses achados corroboram a necessidade de avaliar múltiplas representações textuais para tarefas de processamento jurídico, considerando *trade-offs* entre sofisticação

técnica, custo computacional e efetividade prática.

2.2. Sumarização de documentos

A sumarização de texto é uma tarefa de PLN que visa condensar um documento ou um conjunto de documentos em uma versão mais curta, mantendo suas informações mais relevantes e coerentes [Liu and Lapata 2019, Tsirmpas et al. 2024]. No contexto da classificação de documentos longos, a sumarização é uma estratégia para contornar as limitações de tamanho de entrada dos modelos de linguagem baseados em *Transformers* [Alva Principe et al. 2025], conforme discutido na Seção anterior. As técnicas de sumarização podem ser integradas de diferentes formas, como extrativas e abstrativas, as quais são detalhadas a seguir.

Sumarização extrativa. A sumarização extrativa tem como objetivo identificar as partes mais importantes do documento e apresentá-las em sua forma original, sem modificações linguísticas significativas [Liu and Lapata 2019]. [Polsley et al. 2016] desenvolveram o CaseSummarizer para a análise de decisões judiciais americanas, utilizando TF-IDF e análise estrutural. Seus resultados demonstraram eficácia na captura de informações relevantes, mas apresentaram limitações na geração de sumários coesos, um desafio comum em abordagens puramente extrativas [Bhattacharya et al. 2019].

Em [Jain et al. 2024], os autores propuseram DCESumm para a sumarização extrativa de documentos jurídicos longos, combinando um classificador supervisionado baseado em LegalBERT com o ajuste de pontuações por meio de *deep clustering* temático. A hipótese central é que sentenças relevantes tendem a se agrupar, permitindo o reforço ou a atenuação de sua relevância com base no contexto global. Avaliado no BillSum (legislação norte-americana) e no FIRE (Suprema Corte da Índia), o método superou abordagens clássicas, como o *TextRank* e *SummaRuNNer*, bem como métodos baseados em redes neurais artificiais, como a ferramenta Legal Pegasus. Os ganhos foram de 1-6 pontos de ROUGE no BillSum e de 6-12 pontos no FIRE, demonstrando efetividade em documentos extensos e não estruturados.

Os autores [Feijó and Moreira 2018] criaram o RulingBR, o primeiro corpus brasileiro de sumarização jurídica, contendo 10.623 decisões do STF, estruturadas em *ementa* (7%), *acórdão* (2%), *relatório* (22%) e *voto* (69%). Estabeleceram *baselines* extrativos (TextRank, LexRank e Luhn) com ROUGE-1 entre 0,21 e 0,31, o que indica que a dispersão semântica em textos jurídicos invalida estratégias posicionais típicas, como as utilizadas na sumarização de notícias. A análise revelou uma correlação fraca (r=0,39) entre o comprimento do documento e a ementa, evidenciando a complexidade na determinação do tamanho ideal do sumário. Destaca-se que este estudo de [Feijó and Moreira 2018] estabeleceu a base metodológica para pesquisas subsequentes, incluindo a exploração de abordagens abstratas no contexto brasileiro.

Sumarização abstrativa e modelos generativos. A sumarização abstrativa envolve a geração de novas frases e sentenças que podem não estar presentes no texto original, buscando compreender o conteúdo do documento e reescrevê-lo de forma concisa [Gupta and Gupta 2019]. Recentemente, modelos *sequence-to-sequence* pré-treinados, como BART [Lewis et al. 2020] e PEGASUS [Zhang et al. 2020], têm apresentado resultados superiores na sumarização abstrativa. BART utiliza uma arquitetura de *auto-encoder denoising* que aprende a reconstruir textos corrompidos, enquanto PEGASUS é

pré-treinado especificamente para a sumarização, com o objetivo de *gap-sentence generation*.

Além disso, os modelos generativos baseados em LLMs têm sido explorados para sumarização jurídica. Os autores [Preti et al. 2024] propuseram uma projeção extrativa que utiliza a capacidade generativa de LLMs para produzir resumos abstrativos, mas projeta o resultado em sentenças do documento original, eliminando a possibilidade de alucinação. O objetivo foi combinar a expressividade de modelos generativos com as garantias factuais de métodos extrativos. Para isso, foi utilizado o GPT-3.5-turbo para gerar um sumário abstrativo inicial, seguido de um algoritmo de projeção que identifica, no documento original, sentenças que melhor correspondem a cada sentença gerada, por meio de similaridade semântica. O método proposto retorna, então, uma sequência de sentenças originais que preserva o conteúdo do sumário gerado, sem introduzir texto novo.

Embora a sumarização abstrativa possa gerar sumários de alta qualidade e mais naturais, ela é computacionalmente mais intensiva e apresenta desafios maiores, como a alucinação de fatos ou a perda de fidelidade ao texto original, o que representa um risco considerável em aplicações sensíveis aos dados [Fabbri et al. 2019]. Atualmente, no contexto de lidar com as limitações dos *Transformers* em documentos longos, as técnicas abstrativas ainda não foram plenamente exploradas para esse propósito, o que constitui uma das motivações deste estudo [Alva Principe et al. 2025].

Abordagens híbridas para sumarização. A literatura sugere que a combinação de métodos extrativos e abstrativos oferece equilíbrio entre a cobertura de conteúdo e a coesão do sumário [Kirmani et al. 2018, El-Kassas et al. 2021], sendo relevante para a análise de decisões jurídicas em sistemas de informação [Liu and Lapata 2019, Bae et al. 2019]. No estudo de [Liu and Lapata 2019], os autores desenvolveram um *framework* híbrido que combina seleção extrativa hierárquica com geração abstrativa baseada em BERT, obtendo resultados superiores aos de métodos puros em *benchmarks* padrão. Com isso, os autores sugerem que o pré-treinamento extrativo inicial auxilia a geração abstrativa subsequente. Nos *datasets* CNN/DailyMail e BERTSUMEXTABS o método proposto por [Liu and Lapata 2019] alcançou ROUGE-1=42,13 *vs* 41,72 do modelo puramente abstrativo. A análise revelou que o modelo híbrido produz menos *n*-gramas novos do que a versão puramente abstrativa, indicando um viés extrativo herdado do pré-treinamento inicial, adequado a *datasets* com sumários parcialmente extrativos.

Já [Bae et al. 2019] investigaram uma integração específica para documentos legais, evidenciando a superioridade híbrida em métricas de coerência e de preservação técnica. Os autores relatam que a combinação de extração e abstração é particularmente eficaz em documentos jurídicos, devido à necessidade de preservar a terminologia técnica enquanto se mantém a coesão textual. Em [Janakiraman and Ghoraani 2025], foram comparados 17 modelos por meio de um *framework* multidimensional, com consistência factual de 35%, similaridade semântica de 25%, identificando DeepSeek-v3 como superior em acurácia factual, Gemini-1.5-Flash em custo-benefício ($0,00012/resumo) e na tensão entre consistência factual (ótima em 50 *tokens*) e qualidade percebida (ótima em 150 tokens). No contexto jurídico, [Arfat et al. 2024] aplicaram ChatGPT-3.5 e Gemini em decisões judiciais italianas, com Gemini indicando superioridade em BERTScore (0,67 *vs* 0,33), o que indica maior preservação semântica, embora ambos apresentassem *scores* ROUGE modestos (0,25 e 0,22, respectivamente) devido à natureza abstrativa.

2.3. Lacunas e contribuições do presente estudo

Embora estudos anteriores tenham investigado combinações de modelos especializados com técnicas generativas [Zhang et al. 2025], sua aplicação no contexto jurídico brasileiro permanece pouco explorada [Araújo et al. 2025]. Principalmente, a integração de modelos especializados, como BumbaBERT, com técnicas híbridas de sumarização para decisões judiciais brasileiras representa uma contribuição significativa, considerando as especificidades linguísticas, estruturais e jurídicas do sistema judiciário nacional.

Frente ao panorama apresentado, o presente trabalho se diferencia ao propor um *pipeline* híbrido que combina: (i) processamento baseado em *chunks* para superar limitações de tamanho de modelos especializados; (ii) o uso do BumbaBERT para representações semânticas ricas do domínio jurídico; (iii) refinamento por meio de modelos generativos; e (iv) avaliação com múltiplas métricas de qualidade. Esta abordagem busca equilibrar precisão técnica, viabilidade computacional e qualidade dos sumários gerados, considerando as limitações de recursos financeiros e computacionais frequentemente presentes em contextos de sistemas de informação no setor público brasileiro.

3. Metodologia

Esse estudo foi guiado pela abordagem metodológica de *Design Science Research* (DSR), que inclui etapas voltadas ao desenvolvimento e à avaliação de artefatos tecnológicos para resolver problemas organizacionais identificados [Hevner et al. 2004, Peffers et al. 2007]. A DSR é adequada para pesquisas em SI que buscam desenvolver soluções inovadoras, com relevância prática e rigor científico [Gregor and Hevner 2013]. O desenvolvimento deste trabalho seguiu as etapas estabelecidas por [Peffers et al. 2007], conforme detalhado a seguir.

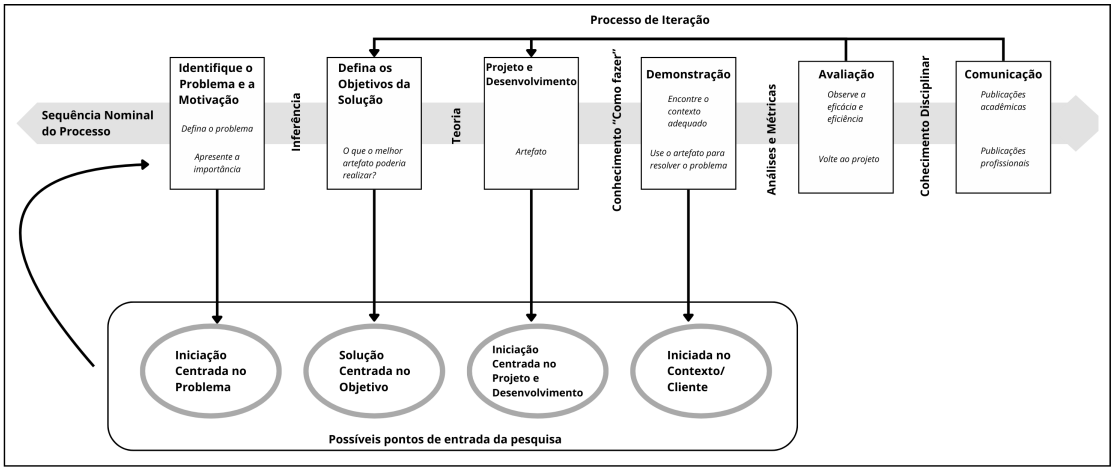

Figura 1. Etapas do DSR [Peffers et al. 2007]

Conforme a Figura 1, o DSR compreende seis etapas encadeadas. As duas primeiras, identificação do problema e definição dos objetivos, foram apresentadas na Seção 1, contextualizando as especificidades das decisões judiciais: extensão significativa, frequentemente de milhares de *tokens*, dispersão de informações relevantes ao longo de múltiplas seções e complexidade estrutural que dificulta tanto a análise manual por profissionais quanto o processamento automatizado. Essas características motivam a

exploração de estratégias de sumarização como solução viável, considerando as tecnologias existentes e as limitações computacionais dos modelos *Transformer* convencionais. As etapas subsequentes do DSR encontram-se descritas nas subseções a seguir.

3.1. Projeto e desenvolvimento

De acordo com [Hevner et al. 2004], essa etapa envolve o projeto e a construção do artefato tecnológico. O *pipeline* proposto foi desenvolvido considerando as propriedades específicas do domínio jurídico brasileiro e as limitações técnicas de processamento.

Dataset e caracterização dos dados. Neste estudo, foi utilizado o *dataset* RulingBR versão 1.2 [Feijó and Moreira 2018], um corpus especializado contendo 10.574 decisões judiciais do STF proferidas entre 2011 e 2018. Esse conjunto foi escolhido por ser o único corpus público brasileiro estruturado especificamente para sumarização jurídica, com ementas produzidas por especialistas que servem como referência de qualidade (*ground truth*). Cada documento está estruturado em oito campos distintos, que compreendem: (i) a *ementa*, um sumário de referência; (ii) o *relatório*, que descreve o caso; (iii) o *voto*, que apresenta a argumentação jurídica; (iv) a decisão final, chamada de *acórdão*; em continuidade têm-se os metadados, incluindo (v) *área*, que é o ramo do direito; (vi) *classe*, que é o instrumento processual; (vii) *relator*, que é o ministro responsável; e, por fim, (viii) o extrato, que apresentam as informações complementares. Na Tabela 1 são apresentadas as principais características estatísticas do corpus, calculadas utilizando o tokenizador do BumbaBERT [Carmo et al. 2024].

Tabela 1. Caracterização estatística do *dataset* RulingBR v1.2

Estatística	Textos Completos	Ementas
Número de documentos	10.574	10.574
Média (*tokens*)	5.458	540
Desvio padrão	7.864	434
Mediana	3.932	431
Mínimo	489	41
Máximo	182.675	6.862
Distribuições relevantes		
Docs $>$ 512 *tokens* (%)	99,98	38,74
Docs $>$ 1.500 *tokens* (%)	91,17	–
Docs $>$ 3.000 *tokens* (%)	64,81	–
Razão de compressão		
Média (%)	14,09	
Mediana (%)	11,43	

Devido às limitações computacionais do ambiente Google Colab, foram processados 5.792 documentos (54,8% do corpus), correspondendo ao maior volume viável de processamento integral sob restrições de memória e tempo de execução. A seleção foi realizada aleatoriamente (`random.seed(42)`), assegurando a reprodutibilidade. Essa amostra permanece, em volume, superior às utilizadas em estudos similares de sumarização jurídica [Arfat et al. 2024, Preti et al. 2024], proporcionando um poder estatístico adequado para a comparação entre métodos.

Os textos completos apresentaram média de 5.458 *tokens*, desvio padrão de 7.864 e mediana de 3.932, e 99,98% dos documentos excederam 512 *tokens* e 64,81% ultrapassaram 3.000 *tokens*. As ementas de referência apresentaram uma média de 540 *tokens*, com desvio padrão de 434 e mediana de 431. Essa extensão excede o limite arquitetural de 512 *tokens* dos modelos BERT convencionais [Devlin et al. 2019], o que evidencia a necessidade de estratégias para processar decisões judiciais completas sem comprometer o conteúdo relevante.

Pré-processamento textual. O pré-processamento foi implementado exclusivamente para a estratégia baseada em vetorização com TF-IDF, priorizando a preservação de características linguísticas relevantes ao domínio jurídico. Utilizou-se a biblioteca spaCy para segmentação de sentenças em português. Além do processo de normalização textual, que incluiu a preservação de acentos e de pontuação especializada relevantes para o domínio jurídico; normalização de espaços em branco e remoção de caracteres de controle; aplicação condicional de conversão para minúsculas, com base nos requisitos específicos de cada modelo; e filtragem de sentenças excessivamente curtas, estabelecendo um limite mínimo de 20 caracteres. As *stopwords* foram obtidas a partir da biblioteca spaCy para português, garantindo a remoção adequada de termos não informativos específicos do idioma, preservando terminologia técnica relevante.

Implementação dos métodos de sumarização. Complementando essa aderência metodológica, a escolha e o design dos métodos de sumarização baseiam-se na teoria de *Task-Technology Fit* (TTF), que postula que o desempenho de SI é otimizado quando há alinhamento entre as características da tecnologia e os requisitos da tarefa [Goodhue and Thompson 1995]. No contexto deste estudo, a produção de ementas jurídicas demanda a preservação da terminologia técnica, a síntese de fundamentos legais dispersos e a manutenção da coesão argumentativa. Os métodos propostos, tanto extrativos quanto abstrativos e híbridos, foram selecionados e configurados para maximizar esse ajuste tecnologia-tarefa, considerando *trade-offs* entre a qualidade da saída, a precisão jurídica e a viabilidade computacional. Isso justifica a avaliação multimétrica adotada e orienta as decisões de design do *pipeline*, assegurando o alinhamento entre as capacidades tecnológicas e os objetivos de processamento de informação jurídica [Zigurs and Khazanchi 2008, Pedroso et al. 2025].

Desse modo, o *pipeline* desenvolvido incluiu cinco diferentes abordagens de sumarização, permitindo avaliação comparativa de diferentes estratégias técnicas. Sendo composto tanto por métodos extrativos (*e.g.*, TF-IDF e BumBumBERT) quanto por métodos abstrativos (*e.g.*, Gemini-2.5). O método extrativo baseado em *TF-IDF* serve como *baseline*, implementando a técnica clássica de *Term Frequency-Inverse Document Frequency*, fundamentada no princípio de que termos frequentes no documento, mas raros no corpus, carregam maior carga informativa [Polsley et al. 2016]. As sentenças foram vetorizadas considerando *stopwords* em português, ranqueadas pela soma dos pesos de seus termos:

$$\text{score}(s) = \sum_{t \in s} \text{TFIDF}(t, d, D),$$

em que s representa uma sentença, t um termo pertencente a s, d o documento de origem da sentença e D o corpus de documentos utilizado para o cálculo do fator inverso de

frequência. As três sentenças de maior pontuação foram então selecionadas, mantendo a ordem original do documento.

No método extrativo com BumbaBERT, implementou-se o processo de decomposição-recomposição, dividindo os documentos longos em segmentos de 512 *tokens*, respeitando o limite arquitetural do modelo [Devlin et al. 2019]. Em seguida, esses *chunks* foram processados em lotes para otimização computacional, gerando *embeddings* contextualizados para cada segmento [Alva Principe et al. 2025]. Os *embeddings CLS* de cada bloco foram então agregados por média ponderada, resultando em uma representação semântica unificada do documento. Por fim, as sentenças foram ranqueadas pela similaridade cosseno

$$\mathrm{sim}(s_j, \mathbf{d}) = \frac{\mathbf{e}_{s_j} \cdot \mathbf{d}}{\|\mathbf{e}_{s_j}\| \cdot \|\mathbf{d}\|},$$

em que s_j representa a j-ésima sentença, $\mathbf{e}_{s_j}$ o vetor de *embedding* da sentença e $\mathbf{d}$ a representação semântica agregada do documento. Esse processo permite a seleção de sentenças com maior relevância semântica global, possibilitando o processamento de documentos extensos sem perda das representações contextuais ricas do modelo BumbaBERT [Polsley et al. 2016].

Para o método abstrativo, foi selecionado o modelo de linguagem de grande escala (`google/Gemini-2.5-flash-lite`), acessado via API. Este foi escolhido devido ao seu desempenho documentado em sumarização jurídica [Arfat et al. 2024], à relação custo-benefício favorável [Janakiraman and Ghoraani 2025] e à capacidade de contexto estendido adequada a documentos jurídicos longos. O *prompt* foi construído seguindo a abordagem *few-shot* [Brown et al. 2020], incluindo: (i) a estrutura padronizada de ementas com cabeçalho, tese central e conclusão; (ii) exemplos de ementas reais, como *in-context examples*, para orientação estilística; (iii) instruções explícitas para manutenção de linguagem formal e técnica, bem como para o comprimento adequado; e (iv) diretrizes específicas para preservação de informações jurídicas críticas, incluindo fundamentos legais e precedentes citados [CNJ 2021].

Os métodos híbridos combinam seleção extrativa inicial com refinamento abstrativo posterior, visando equilibrar a precisão na captura de conteúdo relevante com a coesão textual na apresentação final. Para isso, duas variantes foram implementadas: a primeira utiliza seleção TF-IDF de oito sentenças, seguida de refinamento por meio de um modelo generativo; a segunda emprega seleção com BumbaBERT de oito sentenças mais semelhantes ao documento completo, também seguida de refinamento abstrativo. A escolha deste número de sentenças (oito) para a etapa extrativa baseia-se em uma análise empírica preliminar que identificou esse volume como adequado para preservar informação suficiente, sem exceder os limites de contexto do modelo generativo.

3.2. Demonstração e avaliação do artefato

A avaliação foi conduzida por meio de uma abordagem multimétrica para capturar diferentes dimensões de qualidade dos sumários gerados [Koh et al. 2022].

Métricas de avaliação. A *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) nas variantes ROUGE-1, ROUGE-2 e ROUGE-L [Lin 2004] foi empregada para avaliar a similaridade lexical por meio da sobreposição de unigramas, bigramas e de

subsequências mais longas entre sumários gerados e ementas de referência. ROUGE-N mede a sobreposição de n-gramas entre o resumo gerado e a referência, enquanto ROUGE-L considera a subsequência comum mais longa - *Longest Common Subsequence* (LCS), capturando correspondências de ordem sequencial. O F1-*score* para ROUGE-N é calculado conforme Eq. 1.

$$\text{ROUGE-N}_{F_1} = \frac{2 \cdot P_N \cdot R_N}{P_N + R_N} \tag{1}$$

onde P_N representa a precisão (proporção de n-gramas do resumo gerado presentes na referência) e R_N a revocação (proporção de n-gramas da referência capturados no resumo gerado).

BERTScore [Zhang et al. 2019] foi utilizado como métrica semântica complementar, calculando a similaridade com base em *embeddings* contextualizados de modelos BERT. Esta métrica captura correspondências semânticas além da sobreposição lexical superficial, o que é particularmente relevante no domínio jurídico, em que diferentes formulações podem expressar conceitos equivalentes. Esta pontuação combina precisão e revocação por meio da média harmônica F1 (Eq. 2).

$$\text{BERTScore}_{F_1} = \frac{2 \cdot P \cdot R}{P + R} \tag{2}$$

onde P e R representam a precisão e a revocação das correspondências de *embeddings*, respectivamente.

Metric for Evaluation of Translation with Explicit ORdering (METEOR), proposta por [Banerjee and Lavie 2005], foi incluída por considerar correspondências exatas, radicais por meio de *stemming*, sinônimos por meio de WordNet e a ordem das palavras. Combina precisão (P) e revocação (R) na F-*score*, com penalização pela fragmentação (Pen), conforme Eq. 3.

$$\text{METEOR} = (1 - Pen) \cdot \frac{10 \cdot P \cdot R}{R + 9P} \tag{3}$$

onde $Pen = 0,5 \cdot \left(\frac{\text{n° fragmentos}}{\text{n° unigramas correspondidos}} \right)^3$.

Configuração experimental. Os experimentos foram conduzidos no ambiente Google Colab, utilizando GPU NVIDIA A100 para o processamento dos modelos. Para a implementação, utilizaram-se *Transformers*, PyTorch, spaCy (`pt_core_news_sm`), scikit-learn (TF-IDF) e o OpenAI SDK (Gemini via OpenRouter). Métricas calculadas com *rouge-score*, *bert-score*, nltk (METEOR) e sacrebleu (BLEU). A seleção de exemplos para composição dos *prompts* foi realizada por meio de amostragem aleatória com semente fixa (*seed=42*) a partir de ementas estruturadas do conjunto de dados, garantindo reprodutibilidade.

Os resumos gerados pelos diferentes métodos foram organizados em uma base de dados paralela ao corpus de referência, preservando a estrutura original dos documentos

e permitindo análises comparativas adicionais, podendo ser consultada no repositório do presente estudo[1], juntamente com os *prompts* utilizados no modelo generativo.

4. Resultados e discussão

Nesta seção são apresentados os resultados da avaliação comparativa dos cinco métodos de sumarização implementados: TF-IDF extrativo, BumbaBERT extrativo, abstrativo puro (Gemini-2.5), TF-IDF + Gemini e BumbaBERT + Gemini. A discussão contextualiza os achados em relação aos resultados do *baseline* reportados por [Feijó and Moreira 2018] no trabalho original do RulingBR, além de situá-los no panorama mais amplo da literatura de sumarização jurídica [Feijó and Moreira 2018].

4.1. Desempenho dos métodos implementados

Na Tabela 2, são apresentados os resultados quantitativos obtidos para cada método, por meio das métricas ROUGE (1, 2, L), BERTScore e METEOR. Todas as métricas foram calculadas com base nas ementas de referência produzidas por especialistas jurídicos, como *gold standard*. Para fins de legibilidade, o melhor resultado está destacado em fundo cinza enquanto o segundo melhor está sublinhado.

Tabela 2. Desempenho dos métodos de sumarização no *dataset* RulingBR v1.2

Método	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	METEOR
TF-IDF	0,3582	0,1480	0,2023	0,1377	0,2387
BumbaBERT	0,3579	0,1677	0,2070	0,1621	0,2730
Gemini-2.5	0,4524	0,2315	0,2964	0,1418	0,2726
TF-IDF + Gemini	0,4080	0,1871	0,2609	0,1236	0,2329
BumbaBERT + Gemini	0,4053	0,1841	0,2584	0,1257	0,2267

O método abstrativo baseado em Gemini-2.5 apresentou desempenho superior em todas as métricas ROUGE, alcançando ROUGE-1 de 0,4524, ROUGE-2 de 0,2315 e ROUGE-L de 0,2964. Este resultado representa ganhos de 26,3%, 38,0% e 43,2%, respectivamente, em relação ao método extrativo com métricas superiores, BumbaBERT, e um ganho de aproximadamente 45,9% em relação ao melhor resultado *baseline* (LexRank com 4 sentenças: 0,31) evidenciado por [Feijó and Moreira 2018]. Essa superioridade nas métricas ROUGE indica que o modelo generativo consegue capturar com maior efetividade o conteúdo lexical presente nas ementas de referência, provavelmente devido à sua capacidade de reformulação e paráfrase, além do *prompt* que manteve conhecimento estrutural sobre ementas jurídicas e exemplos *in-context*, funcionando como transferência de conhecimento implícita [Arfat et al. 2024]. Esta estratégia de *few-shot prompting* mostrou-se efetiva em domínios especializados e consistente com os achados de [Preti et al. 2024] na sumarização de decisões judiciais italianas.

Entre os métodos extrativos, o BumbaBERT demonstrou superioridade em métricas semânticas, com BERTScore: 0,1621 e METEOR: 0,2730, em comparação ao TF-IDF, com BERTScore: 0,1377 e METEOR: 0,2387, o que valida a hipótese de que representações contextualizadas especializadas para o domínio jurídico capturam relações semânticas mais ricas do que abordagens estatísticas clássicas. O Bumba-

[1] `https://github.com/GabrieleAraujo/hybrid_summarization.git`

BERT também apresentou ligeira vantagem em ROUGE-2 (0,1677 *vs* 0,1480) e ROUGE-L (0,2070 *vs* 0,2023), indicando melhor preservação de bigramas e estruturas sequenciais mais longas e melhoria de aproximadamente 15% em relação aos *baselines* de [Feijó and Moreira 2018]. É importante notar que, embora o BumbaBERT seja mais sofisticado tecnicamente, o TF-IDF manteve rendimento relativo, resultado consistente com achados [Silva Junior et al. 2025] que demonstraram a efetividade de representações simples baseadas em frequência para a similaridade textual em documentos jurídicos brasileiros, particularmente em cenários com dados não rotulados.

Por fim, os métodos híbridos apresentaram desempenho intermediário e consistente, com resultados superiores aos dos métodos extrativos puros, mas inferiores aos do método abstrativo. O TF-IDF + Gemini obteve ROUGE-1 de 0,4080 e ROUGE-2 de 0,1871, enquanto o BumbaBERT + Gemini alcançou 0,4053 e 0,1841, respectivamente. A proximidade dos resultados entre ambas as variantes híbridas sugere que o refinamento pelo modelo generativo tende a homogeneizar as diferenças decorrentes da etapa extrativa inicial. Isso aponta para uma possível redundância no *pipeline* híbrido, em que os modelos extrativos servem como mecanismo de redução prévia, em vez de uma seleção estratégica de conteúdo. Esta hipótese é explorada em maior profundidade na análise qualitativa a seguir.

4.2. Análise das limitações dos métodos híbridos

Na Figura 2, apresenta-se a distribuição do número de *tokens* dos sumários produzidos por cada método, em comparação com as ementas de referência, indicando padrões distintos de compressão que ajudam a contextualizar o desempenho observado nas métricas automáticas.

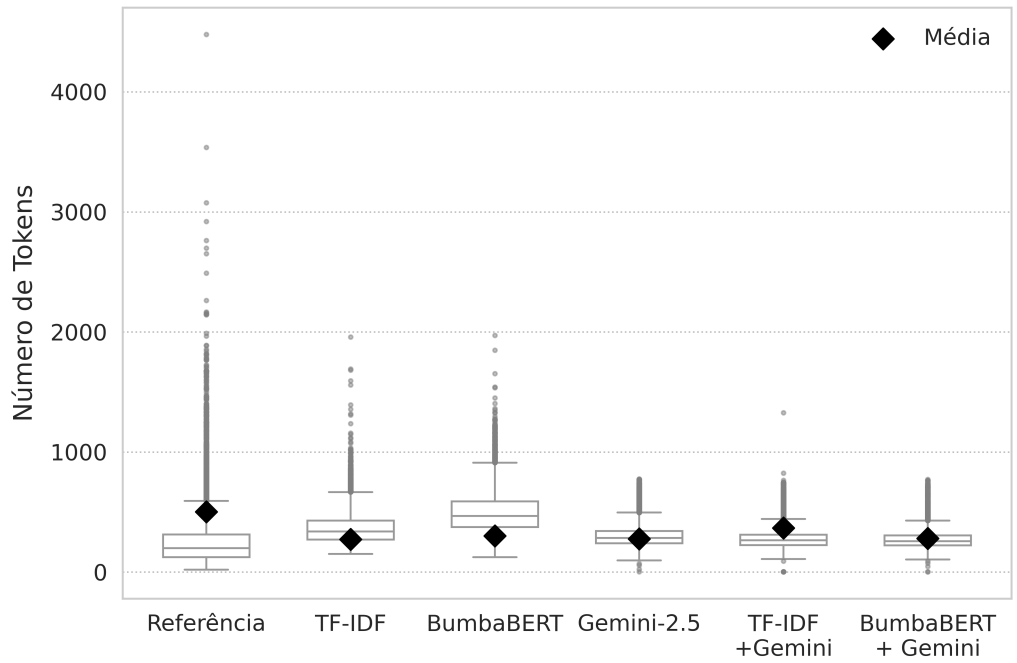

Figura 2. Comparação do número de *tokens* dos sumários gerados.

Os métodos híbridos, embora não tenham superado o método abstrativo nas métricas ROUGE, apresentam atributos favoráveis para aplicações com restrições com-

putacionais. Geraram resumos com médias de 300-350 *tokens*, o que representa uma redução de 35-40% em relação às ementas de referência (média: 540 *tokens*) e maior consistência do que o método abstrativo (média: 280 *tokens*, com alta variabilidade). Apesar dessa maior compressão, o TF-IDF + Gemini (ROUGE-1: 0,4080) ficou apenas 9,8% abaixo do Gemini-2.5, enquanto o BumbaBERT + Gemini (ROUGE-1: 0,4053) apresentou *déficit* de 10,4%, comportamento consistente com as observações de [Bae et al. 2019] sobre a efetividade de estratégias híbridas para balancear a fidelidade factual e a coesão textual em domínios especializados. A diferença qualitativa entre os métodos abstrativo e híbrido pode ser observada na Figura 3 por meio de um exemplo representativo de decisão sobre competência originária do STF em matéria previdenciária.

Referência: direito administrativo e processual civil. conversão de terço constitucional de férias em abono pecuniário e parcelamento de férias. incompetência originária do supremo tribunal federal. 1. a jurisprudência do supremo tribunal federal é firme no sentido de que não se caracteriza a competência prevista no art. 102, i, n, da constituição quando a verba, vantagem ou direito discutidos sejam estabelecidos, concomitantemente, em favor dos membros da magistratura e de outras categorias funcionais. precedentes. 2. agravo regimental a que se nega provimento.

Gemini-2.5: COMPETÊNCIA ORIGINÁRIA. AGRAVO REGIMENTAL. DIREITO ADMINISTRATIVO. SERVIDOR PÚBLICO. 1. A competência originária do Supremo Tribunal Federal, prevista no art. 102, I, 'n', da Constituição, não se caracteriza quando a verba, vantagem ou direito discutidos sejam estabelecidos, concomitantemente, em favor dos membros da magistratura e de outras categorias funcionais. 2. O direito à conversão do terço constitucional de férias em abono pecuniário e ao parcelamento de férias não é exclusivo da magistratura, sendo também titularizado por membros do Ministério Público e servidores públicos federais, o que afasta a incidência da norma constitucional. Agravo regimental não provido.

TF-IDF + Gemini: AGRAVO REGIMENTAL. COMPETÊNCIA ORIGINÁRIA DO STF. MAGISTRATURA E MINISTÉRIO PÚBLICO DA UNIÃO. 1. Direito de conversão do terço constitucional de férias em abono pecuniário e parcelamento de férias. 2. Inexistência de competência originária do STF, nos termos do art. 102, I, 'n', da Constituição Federal. 3. Vantagens e direitos que não são exclusivos da magistratura, sendo também titularizados por outras categorias funcionais. 4. Ação ajuizável em primeira instância. 5. Agravo regimental desprovido.

BumbaBERT + Gemini: COMPETÊNCIA ORIGINÁRIA. AGRAVO REGIMENTAL. MAGISTRATURA. MINISTÉRIO PÚBLICO DA UNIÃO. TERÇO CONSTITUCIONAL DE FÉRIAS. PARCELAMENTO DE FÉRIAS. 1. A competência originária do Supremo Tribunal Federal, prevista no art. 102, I, "n", da Constituição, não se caracteriza quando a verba, vantagem ou direito discutidos sejam estabelecidos concomitantemente em favor dos membros da magistratura e de outras categorias funcionais. 2. O direito à conversão do terço constitucional de férias em abono pecuniário e ao parcelamento de férias não constitui prerrogativa exclusiva da magistratura federal, sendo também titularizado por membros do Ministério Público e servidores públicos federais. 3. A controvérsia que envolve direitos comuns à magistratura e a outras categorias funcionais não atrai a competência originária do STF, devendo ser ajuizada em primeira instância. Agravo regimental desprovido.

Figura 3. Exemplo comparativo de resumos gerados pelos diferentes métodos.

A Figura 3 ilustra diferenças qualitativas relevantes entre as estratégias. A ementa de referência apresenta uma redação mais linear, com encadeamento em texto corrido. O Gemini-2.5, por sua vez, tende a reestruturar o conteúdo em tópicos e rótulos temáticos, reorganizando os fundamentos e explicitando relações de causa e consequência. Em contraste, as variantes híbridas produzem enunciados mais "ancorados" na seleção extrativa prévia, preservando termos e construções típicas do domínio, o que favorece fidelidade lexical e reduz risco de reformulações interpretativas.

Assim, enquanto o método abstrativo privilegia legibilidade e organização discursiva, os híbridos favorecem precisão terminológica e concisão, sendo especialmente úteis

em cenários que demandam citação literal, integração em *pipelines* com janelas de contexto restritas ou processamento em larga escala, nos quais a redução de *tokens* é decisiva para a viabilidade da análise [Jain et al. 2024].

4.3. Implicações práticas e viabilidade computacional

Além da efetividade medida por métricas automáticas, devem ser consideradas ponderações práticas de implementação. O Gemini-2.5, embora superior em desempenho, apresenta custos maiores, pois o processamento de contexto extenso implica latência de vários segundos e custos por requisição. Essa tensão entre qualidade e viabilidade econômica foi sistematicamente documentada por [Janakiraman and Ghoraani 2025], que identificaram que o Gemini-1.5-Flash apresenta melhor custo-benefício ($0,00012/resumo) entre 17 modelos avaliados, porém, ainda assim, é uma das opções de custo-benefício.

Os métodos extrativos, particularmente o TF-IDF, oferecem uma alternativa computacionalmente eficiente, executável em hardware convencional, sem dependência de serviços externos. O BumbaBERT, embora requeira GPU para um processamento eficiente, permite a implantação local. Esta autonomia é particularmente relevante para instituições públicas brasileiras que podem enfrentar restrições orçamentárias ou políticas de privacidade que desencorajam o compartilhamento de documentos jurídicos sensíveis com serviços de nuvem.

O *trade-off* entre qualidade e viabilidade prática sugere que diferentes métodos podem ser apropriados para contextos de aplicação distintos. Sistemas de produção com volumes elevados podem beneficiar-se de métodos extrativos rápidos, enquanto aplicações em que a qualidade máxima é prioritária podem justificar custos associados a métodos abstrativos.

4.4. Limitações e trabalhos futuros

Os resultados apresentados devem ser interpretados considerando limitações metodológicas importantes. A avaliação baseou-se exclusivamente em métricas automáticas, sem validação por especialistas jurídicos quanto à adequação das ementas geradas. Estudos futuros devem incluir uma avaliação humana estruturada, considerando a precisão jurídica, a completude dos fundamentos, a conformidade com os padrões estilísticos do CNJ e a adequação a diferentes perfis de uso, como os de advogados, juízes e pesquisadores.

Para além disso, o método abstrativo dependeu do modelo proprietário Gemini-2.5, cuja escolha foi fundamentada em critérios empíricos de desempenho, custo-benefício e capacidade de contexto estendido, mas constituiu uma decisão contingente às restrições computacionais e financeiras do estudo. Seus detalhes arquiteturais e de treinamento não são públicos, o que limita análises aprofundadas das características que contribuem para o desempenho observado. Investigações futuras devem explorar modelos abertos como LLaMA, Mistral e Sabiá, permitindo comparações sistemáticas de múltiplos LLMs, a análise de comportamentos distintos entre as arquiteturas e o ajuste fino com dados jurídicos brasileiros para validar a generalidade dos achados.

Quanto ao escopo, todos os métodos foram avaliados exclusivamente com base em decisões do STF, sem considerar variações estilísticas entre os relatores ou entre as

áreas do direito. A generalização para outros gêneros, como petições, contratos, pareceres, sentenças de instâncias inferiores ou tribunais, requer validação adicional, considerando que os aspectos estruturais variam entre os tipos documentais. Por fim, o corpus RulingBR foi coletado em 2018, antes da Recomendação CNJ n.º 154/2024, que estabelece novos padrões para a estruturação das ementas. Trabalhos futuros devem investigar a adaptação dos métodos a essas diretrizes recentes, potencialmente criando corpus alinhado às normas vigentes para validação comparativa.

5. Considerações finais

Neste estudo, foram avaliados cinco métodos de sumarização aplicados a decisões judiciais do STF, comparando abordagens extrativas (TF-IDF e BumbaBERT), abstrativa pura (Gemini-2.5) e híbridas que utilizam ambos os métodos. Os resultados indicam que o método abstrativo alcançou desempenho superior nas métricas ROUGE (ROUGE-1: 0,4524), representando um ganho de 45,9% em relação a *baselines* anteriores, enquanto os métodos híbridos ofereceram equilíbrio entre qualidade e eficiência ao produzir sumários 35-40% mais compactos, mantendo desempenho acentuado (ROUGE-1 $>$ 0,40).

A principal contribuição técnica consistiu na implementação de processamento baseado em *chunks* com o BumbaBERT, viabilizando a análise de documentos longos sem perda de representações semânticas especializadas. Metodologicamente, o estudo estabelece como equilibrar a qualidade técnica com a viabilidade computacional em contextos de recursos limitados, um desafio ainda inerente à infraestrutura do setor público. Os métodos híbridos mostraram-se adequados para aplicações como interfaces de busca jurisprudencial e sistemas de classificação, nas quais as restrições de contexto e de latência são determinantes.

Assim, o estudo alinha-se aos Grandes Desafios de SI no Mundo Aberto [Boscarioli et al. 2017], especificamente aos desafios relacionados à governança de dados no setor público, propondo estratégias de processamento que consideram limitações sistêmicas; transparência institucional, ao viabilizar acesso mais eficiente às decisões judiciais por meio de resumos de qualidade; e interoperabilidade, propondo *pipeline* adaptável a diferentes contextos de recursos computacionais e requisitos de qualidade.

Em termos de impacto prático, os achados contribuem em múltiplas dimensões, tanto para profissionais do direito, reduzindo o tempo de revisão de precedentes e de pesquisa jurisprudencial; para instituições judiciais, apoiando a produção de ementas e a classificação documental em larga escala; para cidadãos, melhorando o acesso à informação jurídica por meio de sistemas mais eficientes. Ao fornecer evidências empíricas sobre estratégias eficazes, considerando as limitações reais do contexto brasileiro e de países do Sul Global, o estudo contribui para o desenvolvimento de sistemas de apoio à decisão escaláveis, transparentes e contextualmente adequados.

Como mencionado anteriormente, o presente estudo ainda presenta algumas limitações, como a avaliação baseada em métricas automáticas, a dependência de um modelo proprietário e o escopo restrito às decisões do STF. Sendo assim, trabalhos futuros devem incorporar avaliação por especialistas jurídicos; explorar modelos abertos para ajuste fino com dados brasileiros; validar a generalização para outros gêneros jurídicos e instâncias; e investigar o alinhamento à Recomendação CNJ n.º 154/2024, que estabelece novos padrões de estruturação de ementas.

Agradecimentos

Este estudo foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - DT-303031/2023-9 e POSDOC-101057/2024-5; pela Financiadora de Estudos e Projetos - Pró-Amazônia - 2.373/24 - CTCCA-II; e pelo Acordo de Cooperação Técnica nº 02/2021 (processo nº 38328/2020-TJ/MA).

Referências

Alva Principe, R., Chiarini, N., and Viviani, M. (2025). Long document classification in the transformer era: A survey on challenges, advances, and open issues. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2):e70019.

Araújo, G. S., Jacob Junior, A. F. L., Santana, E. E. C., and Lobato, F. M. F. (2025). The artificial intelligence integration in the brazilian legal sector: A systematic review. In *Proceedings of the Brazilian Symposium on Information Systems (SBSI)*, Recife, PE, Brazil.

Arfat, Y., Colella, M., and Marello, E. (2024). Legal text analysis using large language models. In *International Conference on Applications of Natural Language to Information Systems*, pages 258–268. Springer.

Bae, S., Kim, T., Kim, J., and Lee, S.-g. (2019). Summary level training of sentence rewriting for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 10–20.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI, USA.

Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Ghosh, K., and Ghosh, S. (2019). A comparative study of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval*, pages 413–428. Springer.

Boscarioli, C., de Araujo, R. M., Maciel, R. S., Neto, V. V. G., Oquendo, F., Nakagawa, E. Y., Berrnardini, F. C., Viterbo, J., Vianna, D., Martins, C. B., et al. (2017). I grandsi-br: Grand research challenges in information systems in brazil 2016-2026.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Carmo, F. A., Serejo, F., Jacob Junior, A. F. L., Santana, E. E. C., and Lobato, F. M. F. (2024). Bumbabert: Um modelo de linguagem pré-treinado para o domínio jurídico brasileiro. *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 188–199.

Casimiro, J. S. C. and Teixeira, S. T. (2024). Artificial intelligence approaches within the brazilian judiciary's contemporary jurisdictional model. *Beijing L. Rev.*, 15:730.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., et al. (2020). Legal-bert: The muppets straight out of law school. In *Findings of EMNLP 2020*, pages 2898–2904.

CNJ, C. N. d. J. (2021). Diretrizes para elaboração de ementas e indexação de acórdãos. Technical report, CNJ, Brasília. https://www.cnj.jus.br/wp-content/uploads/2021/09/diretrizes-elaboracao-ementas-uerj-reg-cnj-v15122021.pdf.

CNJ, C. N. d. J. (2024). Justiça em números 2024. Technical report, CNJ, Brasília. https://www.cnj.jus.br/wp-content/uploads/2024/05/justica-em-numeros-2024.pdf.

de Castro, M. Q. and Ralha, C. G. (2025). Identificando divergências jurisprudenciais com técnicas de inteligência artificial para apoio de sistemas de informaçao judiciais. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 289–295. SBC.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Fabbri, A. R., Li, I., She, T., Li, S., and Radev, D. (2019). Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084.

Fama, I., Bueno, B., Alcoforado, A., Ferraz, T., Moya, A., and Costa, A. H. (2024). No argument left behind: Overlapping chunks for faster processing of arbitrarily long legal texts. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 129–138, Porto Alegre, RS, Brasil. SBC.

Feijó, D. V. and Moreira, V. P. (2018). Rulingbr: A summarization dataset for legal texts. In *International Conference on Computational Processing of the Portuguese Language*, pages 255–264. Springer.

Glenn, H. P. (2014). *Legal Traditions of the World: Sustainable Diversity in Law*. Oxford University Press, Oxford, UK, 5th edition.

Goodhue, D. L. and Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS quarterly*, pages 213–236.

Gregor, S. and Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2):337–355.

Gupta, S. and Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Hevner, A. R., March, S. T., Park, J., and Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1):75–105.

Jain, D., Borah, M. D., and Biswas, A. (2024). A sentence is known by the company it keeps: improving legal document summarization using deep clustering. *Artificial Intelligence and Law*, 32(1):165–200.

Janakiraman, A. and Ghoraani, B. (2025). An empirical comparison of text summarization: A multi-dimensional evaluation of large language models. *arXiv preprint arXiv:2504.04534*.

Jiang, Z., Yang, J., and Rao, D. (2024). An empirical study of leveraging plms and llms for long-text summarization. In *Pacific Rim International Conference on Artificial Intelligence*, pages 424–435. Springer.

Kirmani, M., Manzoor Hakak, N., Mohd, M., and Mohd, M. (2018). Hybrid text summarization: a survey. In *Soft Computing: Theories and Applications: Proceedings of SoCTA 2017*, pages 63–73. Springer.

Koh, H. Y., Ju, J., Liu, M., and Pan, S. (2022). An empirical survey on long document summarization: Datasets, models, and metrics. *ACM computing surveys*, 55(8):1–35.

Kuş, A. and Acı, Ç. İ. (2024). A hybrid approach to automatic text summarization of turkish texts: Integrating extractive methods with llms. In *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE.

Lai, J., Gan, W., Wu, J., Qi, Z., and Yu, P. S. (2024). Large language models in law: A survey. *AI Open*, 5:181–196.

Lewis, M., Liu, Y., Goyal, N., et al. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL 2020*, pages 7871–7880.

Limsopatham, N. (2021). Effectively leveraging bert for legal document classification. In *Proceedings of the natural legal language processing workshop 2021*, pages 210–216.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3730–3740.

Luz de Araujo, P. H., de Almeida, A. P. G., Ataides Braz, F., Correia da Silva, N., de Barros Vidal, F., and de Campos, T. E. (2023). Sequence-aware multimodal page classification of brazilian legal documents. *International Journal on Document Analysis and Recognition (IJDAR)*, 26(1):33–49.

Moreira, M. C. G. and de Souza Moura, P. N. (2023). A tecnologia como suporte para o judiciário e o acesso à justiça: uma proposta de aplicação no âmbito da violência doméstica. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 48–57. SBC.

Pedroso, B. C., Pereira, M. R., and Pereira, D. A. (2025). Performance evaluation of llms in the text-to-sql task in portuguese. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 260–269. SBC.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3):45–77.

Polsley, S. D., Jhaver, S., Mukerjee, S., et al. (2016). Casesummarizer: A system for automated summarization of legal texts. In *Proceedings of COLING 2016*, pages 258–268.

Preti, D., Giannone, C., Favalli, A., and Romagnoli, R. (2024). Automatic summarization of legal texts, extractive summarization using llms. In *Ital-IA 2024: 4th National Conference on Artificial Intelligence*, Naples, Italy.

Silva Junior, D. d., Oliveira, D. d., and Paes, A. (2025). Evaluating text representations for unsupervised legal semantic textual similarity in brazilian portuguese. *Discover Data*, 3(1):23.

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. Springer.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Tsirmpas, D., Gkionis, I., Papadopoulos, G. T., and Mademlis, I. (2024). Neural natural language processing for long texts: A survey on classification and summarization. *Eng. Appl. Artif. Intell.*, 133(PC).

Wang, Y. (2024). Design and application of legal information systems based on big data technology. *International Journal of Information Systems and Supply Chain Management (IJISSCM)*, 17(1):1–18.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, Y., Jin, H., Meng, D., Wang, J., and Tan, J. (2025). A comprehensive survey on automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Zigurs, I. and Khazanchi, D. (2008). From profiles to patterns: A new view of task-technology fit. *Information systems management*, 25(1):8–13.