UNIVERSIDADE ESTADUAL DO MARANHÃO

Programa de Pós-Graduação em Engenharia da Computação e Sistemas

Gabriele de Sousa Araújo

Processamento de documentos jurídicos longos:

comparação e avaliação de métodos baseados em Modelos de Linguagem

São Luís - MA

2025

Gabriele de Sousa Araújo

Processamento de documentos jurídicos longos: comparação e avaliação de métodos baseados em Modelos de Linguagem

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia da Computação e Sistemas da Universidade Estadual do Maranhão, como parte das exigências para obtenção do título de Mestra em Engenharia da Computação e Sistemas.

Universidade Estadual do Maranhão

Centro de Ciências e Tecnológicas

Programa de Pós-Graduação em Engenharia da Computação e Sistemas

Orientador: Prof. Dr. Fábio Manoel França Lobato

Coorientador: Prof. Dr. Ewaldo Eder Carvalho Santana

São Luís - MA

2025

Araújo, Gabriele de Sousa
Processamento de documentos jurídicos longos: comparação e avaliação de métodos baseados em modelos de linguagem. / Gabriele de Sousa Araújo. – São Luis, MA, 2025.
161 f

Dissertação (Mestrado em Engenharia da Computação e Sistemas) - Universidade Estadual do Maranhão, 2025.

Orientador: Prof. Dr. Fábio Manoel França Lobato.
Coorientador: Prof. Dr. Ewaldo Eder Carvalho Santana

1.Documentos Jurídicos Longos. 2.Modelos de Linguagem. 3.Justiça 4.0. 4.Inteligência Artificial. I.Titulo.

CDU: 004.434:004.8

Elaborado por Cássia Diniz - CRB 13/910

Gabriele de Sousa Araújo

Processamento de documentos jurídicos longos: comparação e avaliação de métodos baseados em Modelos de Linguagem

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia da Computação e Sistemas da Universidade Estadual do Maranhão, como parte das exigências para obtenção do título de Mestra em Engenharia da Computação e Sistemas.

São Luís - MA, 12 de novembro de 2025:

Documento assinado digitalmente
gov.br
FABIO MANOEL FRANCA LOBATO
Data: 15/12/2025 11:43:58-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Fábio Manoel França Lobato
Orientador - Universidade Federal do Oeste do Pará - UFOPA

Documento assinado digitalmente
gov.br
EWALDO EDER CARVALHO SANTANA
Data: 15/12/2025 17:41:21-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ewaldo Eder Carvalho Santana
Coorientador - Universidade Federal do Maranhão - UFMA

Documento assinado digitalmente
gov.br
DAVI VIANA DOS SANTOS
Data: 15/12/2025 17:21:56-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Davi Viana dos Santos
Examinador Externo - Universidade Federal do Maranhão - UFMA

Documento assinado digitalmente
gov.br
RICARDO MARCONDES MARCACINI
Data: 15/12/2025 12:04:49-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ricardo Marcondes Marcacini
Examinador Externo - Universidade de São Paulo - USP

São Luís - MA
2025

Agradecimentos

Aos meus familiares, em especial aos meus avós Rafael e Maria, e aos meus tios Lene e Elson, que contribuíram significativamente para minha formação pessoal e acadêmica, acolhendo-me como filha em momentos importantes da minha vida. Ao meu pai, Ronaldo, por sempre me encorajar a persistir e acreditar no meu potencial. Vocês são os pilares da pessoa que me tornei. Sou profundamente grata pelo amor incondicional, pelas valiosas lições e pelo apoio incansável ao longo desta jornada.

Ao Prof. Dr. Fábio Lobato, por ter me selecionado para integrar o grupo de pesquisa do LACA e, desde a graduação, me orientar com dedicação, ajudando-me a reencontrar meu caminho na computação e a acreditar no meu potencial. Agradeço também ao Prof. Dr. Antônio Fernando Lavareda Jacob Junior, por me receber na UEMA e assumir a coordenação deste projeto de mestrado. Suas orientações, confiança e expertise foram essenciais para o desenvolvimento e conclusão desta dissertação.

Aos meus amigos Adriel, Carlos e Domingas, pela parceria, incentivo e companheirismo ao longo dessa caminhada. Aos amigos que fiz durante o intercâmbio na Alemanha, que se tornaram minha família longe de casa — Tami, Amanda, Sarah, Júlia, Lara, Lívia, Carlos (Kaka) e Felipe, sou profundamente grata por tornarem os dias mais leves, pelas risadas, pelo apoio constante e por celebrarem comigo cada conquista, inclusive as pequenas. A presença de vocês fez toda a diferença.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização desta dissertação.

"É bom ter um destino final para a jornada, mas, no fim das contas, o que importa é a jornada em si."

(Ursula K. Le Guin)

Resumo

O sistema jurídico brasileiro enfrenta um cenário estrutural de acúmulo de demandas judiciais, o que compromete sua eficiência e o cumprimento do princípio constitucional da razoável duração do processo. Em agosto de 2024, o número de ações em tramitação ultrapassou 80 milhões, segundo dados da Base Nacional de Dados do Poder Judiciário. Esse volume expressivo de litígios afeta diretamente a celeridade e efetividade da prestação jurisdicional, demandando o desenvolvimento de instrumentos tecnológicos capazes de apoiar a gestão, triagem e compreensão de documentos legais, cuja extensão e complexidade desafiam as rotinas tradicionais de análise humana e automatizada. Nesse cenário, compreender como técnicas avançadas de processamento de texto podem contribuir para a racionalização da atividade jurisdicional constitui o cerne desta dissertação. Aplicações já existentes no domínio, como o BumbaBERT, apresentam resultados promissores para a otimização do fluxo processual, porém ainda limitadas pelas restrições estruturais da arquitetura *Transformer* que as compõe, sobretudo em razão de sua alta complexidade computacional. Visando contornar essa problemática, o presente projeto de dissertação propõe e avalia um conjunto de estratégias direcionadas ao processamento eficiente de documentos longos, tomando como estudo de caso as petições iniciais vinculadas a Incidentes de Resolução de Demandas Repetitivas (IRDR). Partindo das lacunas identificadas na literatura e da motivação prática advinda do acordo de cooperação técnica UEMA-TJMA no que tange à dificuldade de adaptação de modelos linguísticos a textos jurídicos extensos, o processo metodológico foi guiado pela *Data Science Trajectories* (DST). Essa abordagem forneceu base para a compreensão do domínio, planejamento das soluções, bem como a identificação de uma taxonomia de métodos capazes de organizar o campo de classificação automática de documentos longos em três vertentes: métodos de trucamento derivados dos *baselines* (*e.g.,* BumbaBERT, LegalBERT-PT); decomposição-recomposição (*e.g.,* ToBERT) e de síntese de conteúdo a partir de estratégias de seleção de sentenças (*e.g.,* TextRank, LexRank, SBERT, LlaMa). A partir dessa estrutura, procedeu-se à experimentação empírica e à validação estatística. Sendo assim, o estudo envolveu a implementação e a comparação de oito arquiteturas baseadas no ajuste fino do BumbaBERT, totalizando 40 experimentos que consideraram métricas de desempenho como acurácia, F1-*score*, precisão e revocação; indicadores de eficiência computacional como tempo, inferência e uso de memória; testes de significância estatística; e a viabilidade prática de implementação. Os resultados demonstraram que arquiteturas hierárquicas superam abordagens baseadas na síntese de conteúdo, alcançando um melhor equilíbrio entre precisão e estabilidade, ainda que com maior custo computacional. Tal constatação reforça a importância de preservar a estrutura argumentativa integral de textos jurídicos para garantir a consistência interpretativa e a confiabilidade das classificações automáticas. Assim, o trabalho contribui, em termos científicos, para o avanço do processamento de linguagem natural no domínio

jurídico, ao demonstrar como estratégias já consolidadas podem ser reinterpretadas e ajustadas para atender às especificidades linguísticas e estruturais dos textos jurídicos brasileiros. Do ponto de vista tecnológico e institucional, o estudo oferece um artefato reprodutível, passível de integração ao sistema de automação do TJMA, contribuindo para a redução do tempo de tramitação processual e para o fortalecimento de políticas de transformação digital no setor público. Por fim, em dimensão social, reafirma-se o papel da transformação digital como instrumento de democratização do acesso à justiça, promovendo uma inovação que alia precisão técnica, responsabilidade ética e compromisso com o interesse público.

Palavras-chave: Documentos Jurídicos Longos, Modelos de Linguagem, Justiça 4.0, Inteligência Artificial.

Abstract

The Brazilian legal system faces a structural scenario of case overload that undermines its efficiency and compliance with the constitutional principle of reasonable duration of proceedings. In August 2024, the number of pending lawsuits exceeded 80 million, according to data from the National Judiciary Database. This significant volume of litigation directly affects the speed and effectiveness of judicial services, requiring the development of technological tools capable of supporting the management, screening, and understanding of legal documents, whose length and complexity challenge traditional human and automated analysis routines. In this context, understanding how advanced text processing techniques can contribute to the rationalization of judicial activity is at the heart of this dissertation. Existing applications in the field, such as BumbaBERT, show promising results for optimizing procedural workflows but remain limited by the structural constraints of the Transformer architecture that underpins them, mainly due to its high computational complexity. To address this issue, this dissertation proposes and evaluates a set of strategies aimed at the efficient processing of long documents, using as a case study the initial petitions linked to Incidents of Resolution of Repetitive Demands (IRDR). Based on the gaps identified in the literature and the practical motivation arising from the UEMA–TJMA technical cooperation agreement regarding the difficulty of adapting language models to extensive legal texts, the methodological process was guided by the Data Science Trajectories (DST) framework. This approach provided a basis for understanding the domain, planning solutions, and identifying a taxonomy of methods capable of organizing the field of automatic long-document classification into three strands: truncation methods derived from baselines (e.g., BumbaBERT, LegalBERT-PT); decomposition–recomposition methods (e.g., ToBERT); and content synthesis methods based on sentence selection strategies (e.g., TextRank, LexRank, SBERT, LLaMa). Based on this structure, empirical experimentation and statistical validation were carried out. Thus, the study involved the implementation and comparison of eight architectures based on the fine-tuning of BumbaBERT, totaling 40 experiments that considered performance metrics (accuracy, F1-score, precision, and recall), computational efficiency indicators (time, inference, and memory usage), statistical significance tests, and practical implementation feasibility. The results demonstrated that hierarchical architectures outperform approaches based on content synthesis, achieving a better balance between precision and stability, albeit with higher computational costs (training time of 3h12min and total consumption of approximately 15 GB per epoch). This finding reinforces the importance of preserving the integral argumentative structure of legal texts to ensure interpretive consistency and reliability of automatic classifications. Thus, the work contributes scientifically to the advancement of natural language processing in the legal domain by demonstrating how established strategies can be reinterpreted and adjusted to meet the linguistic and

structural specificities of Brazilian legal texts. From a technological and institutional point of view, the study offers a reproducible artifact that can be integrated into the TJMA automation system, contributing to the reduction of procedural processing time and the strengthening of digital transformation policies in the public sector. Finally, in the social dimension, it reaffirms the role of digital transformation as an instrument for democratizing access to justice, promoting innovation that combines technical precision, ethical responsibility, and commitment to the public interest.

Keywords: Long Legal Documents, Language Models, Justice 4.0, Artificial Intelligence

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ADC	*Automatic Document Classification*
AENN	*Autoencoder Neural Network*
ALDC	*Automatic Long Document Classification*
ANN	*Artificial Neural Networks*
ANOVA	Análise de Variância
API	*Application Programming Interface*
AUC-ROC	*Area Under the Receiver Operating Characteristic Curve*
BART	*Bidirectional & Autoregressive Transformer*
BERT	*Bidirectional Encoder Representations from Transformers*
BiLSTM	*Long Short-Term Memory Bidirecional*
BPE	*Byte Pair Encoding*
brWaC	*Brazilian Web as Corpus*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CIAPJ	Centro de Inovação, Administração e Pesquisa do Judiciário
CNJ	Conselho Nacional de Justiça
CNN	*Convolutional Neural Network*
CogLTX	*Cognize Long TeXts*
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
CSAT	*Customer Support Analysis Tasks*
CPC	Código de Processo Civil
CPU	*Central Processing Unit*
CSV	*Comma-Separated Values*
DHSN	*Duale Hochschule Sachsen*

DSR	*Design Science Research*
DST	*Data Science Trajectories*
ECHR	*European Court of Human Rights*
F1-M	*Macro-F1*
F1-W	*Weighted-F1*
FAVOR+	*Fast Attention por meio de positive Orthogonal Random features*
FedCSIS	*Conference on Computer Science and Intelligence Systems*
FIRE	*Forum for Information Retrieval Evaluation*
FWER	*Familywise error rate*
FGV	Fundação Getúlio Vargas
FN	*False Negatives*
FP	*False Positives*
GNN	*Graph Neural Network*
GRADE	*Grading of Recommendations, Assessment, Development, and Evaluation*
GPT	*Generative Pre-trained Transformer*
GPU	*Graphics Processing Unit*
HPT	*Hierarchy-aware Prompt Tuning*
HSD	*Honestly Significant Difference*
IA	Inteligência Artificial
INPI	Instituto Nacional da Propriedade Industrial
IRDR	Incidente de Resolução de Demandas Repetitivas
LDC	*Long document classification*
LLaMA	*Large Language Model Meta AI*
LLMs	*Large Language Models*
LSH	*Locality-Sensitive Hashing*
LSTM	*Long Short-Term Memory*

LTs	*Long Transformers*
MLM	*Masked Language Modeling*
MLP	*Multilayer Perceptron*
NER	*Named-Entity Recognition*
NUGEPNAC	Núcleo de Gerenciamento de Precedentes
NSP	*Next Sentence Prediction*
OCR	*Optical Character Recognition*
PDF	*Portable Document Format*
PJe	Sistema Nacional de Processos
PLMs	*Pretrained Language Models*
PLN	Processamento de Linguagem Natural
PNUD	Programa das Nações Unidas para o Desenvolvimento
PRISMA	*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*
QUADAS-2	*Quality Assessment of Diagnostic Accuracy Studies*
RAG	*Retrieval-Augmented Generation*
RNN	Redes Neurais Recorrentes
RoBERT	*Recurrence over BERT*
RoBERTa	*Robustly Optimized BERT Approach*
SBERT	*Sentence-BERT*
SBSI	Simpósio Brasileiro de Sistemas de Informação
SCOTUS	*Supreme Court of the United States*
STF	Supremo Tribunal Federal
STJ	Superior Tribunal de Justiça
STM	Superior Tribunal Militar
STS	Similaridade Textual Semântica
SVC	*Support Vector Classifier*

SVM	*Support Vector Machine*
T5	*Text-to-Text Transfer Transformer*
TFF	*Task-Technology Fit*
TF-IDF	*Term Frequency-Inverse Document Frequency*
TJMA	Tribunal de Justiça do Maranhão
TN	*True Negatives*
TP	*True Positives*
TPU	*Tensor Processing Unit*
TSE	Tribunal Superior Eleitoral
TST	Tribunal Superior do Trabalho
ToBERT	*Transformer over BERT*
UEMA	Universidade Estadual do Maranhão
UFPA	Universidade Federal do Pará
UFOPA	Universidade Federal do Oeste do Pará
XAI	*Explainable AI*

Sumário

1 Introdução

Neste Capítulo, são abordadas as considerações iniciais sobre o presente trabalho, contextualizando-o e destacando as razões e justificativas que o fundamentam, a fim de compreender sua relevância e suas motivações. Em seguida, são apresentados os objetivos alcançados, as principais contribuições e os resultados obtidos. Por fim, é descrita a sequência em que o documento está organizado, o que fornece uma visão geral da estrutura do trabalho.

1.1 Contextualização e Motivação

O cenário jurídico brasileiro tem sido marcado por um crescimento expressivo no volume de processos judiciais ano após ano. De acordo com as estatísticas fornecidas pela Base Nacional de Dados do Poder Judiciário, gerenciada pelo Conselho Nacional de Justiça (CNJ), o número de ações em andamento nos diversos Tribunais e Varas do país ultrapassou a marca de 80 milhões em julho de 2024 (CNJ, 2024a). Essa sobrecarga processual reflete-se em todas as instâncias do sistema judiciário, desde as varas de primeira instância até os tribunais superiores, impactando significativamente a celeridade e a eficiência da prestação jurisdicional, acarretando extrapolação de prazos, morosidade do serviço, ineficácia dos comandos judiciais e até mesmo o descrédito do sistema judiciário (ALMEIDA; PINTO, 2022).

A complexidade e extensão dos documentos jurídicos exacerbam os desafios de processamento e análise, uma vez que petições, sentenças, acórdãos e outros documentos legais frequentemente ultrapassam centenas ou milhares de páginas, contendo informações complexas e inter-relacionadas (KALAMKAR et al., 2022). A análise manual desses documentos é um processo demorado e propenso a erros, demandando tempo considerável dos profissionais do direito e impactando diretamente a celeridade da justiça. Outro ponto diz respeito à natureza subjetiva da interpretação humana, aliada à possibilidade de erros na análise manual, o que pode resultar em decisões judiciais inconsistentes, comprometendo a equidade e a eficácia do sistema jurídico (KALAMKAR et al., 2022).

Diante desse cenário, o Poder Judiciário brasileiro tem buscado soluções para modernizar e otimizar seus processos. O CNJ instituiu a Resolução N^o 332, de 21 de agosto de 2020, que “dispõe sobre a ética, a transparência e a governança na produção e no uso de Inteligência Artificial no Poder Judiciário” (CNJ, 2020). Essa resolução reconhece que a aplicação da Inteligência Artificial (IA), “no âmbito do Poder Judiciário, visa promover o bem-estar dos jurisdicionados e a prestação equitativa da jurisdição, bem como descobrir métodos e práticas que possibilitem a consecução desses objetivos” (CNJ, 2020). Do

ponto de vista técnico, a implementação de IA no sistema judiciário visa aumentar a produtividade, reduzir o tempo de tramitação processual e minimizar os custos operacionais por meio da automação inteligente de tarefas (TAUK; SALOMÃO, 2023).

Como desdobramento dessa regulação, surgiu o programa Justiça 4.0 em 2021 (CNJ; PNUD, 2021), uma parceria entre o CNJ e o Programa das Nações Unidas para o Desenvolvimento (PNUD) que busca promover o acesso à justiça por meio do uso de novas tecnologias e IA (CNJ, 2024b), representando um marco na transformação digital do sistema judiciário e alinhando-se às tendências globais de inovação tecnológica no setor público. Essa iniciativa catalisou o desenvolvimento de aplicações de IA no Poder Judiciário, com foco em avanços recentes no campo do Processamento de Linguagem Natural (PLN) e na aplicação de modelos de linguagem de grande escala (do inglês, *Large Language Models*, LLMs) ao domínio jurídico (CARMO, 2024; POLO et al., 2021).

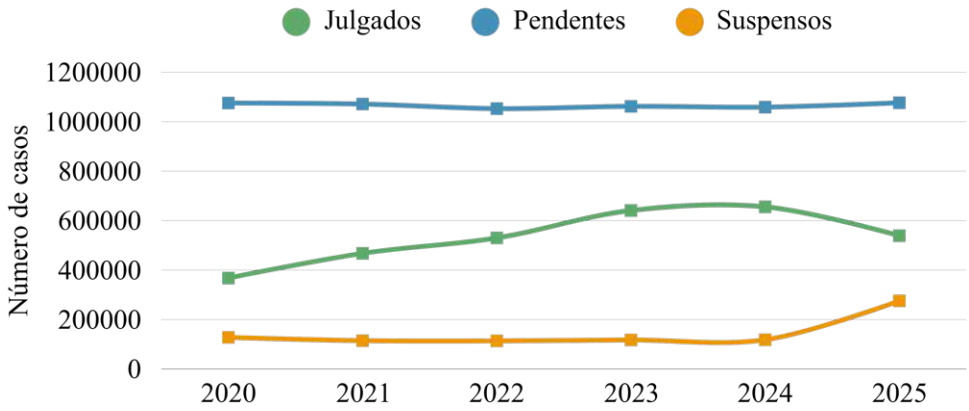
Segundo o relatório "Pesquisa uso de Inteligência Artificial no Poder Judiciário – 2023" publicado pelo CNJ (2024), há uma adoção heterogênea de IA pelos tribunais brasileiros, com 140 projetos identificados em 62 tribunais, distribuídos entre as justiças estadual (68), eleitoral (23), trabalhista (20), e em tribunais superiores e federais (29), que embora em menor número de instituições, concentram uma maior proporção de projetos por tribunal.

Vale destacar que esse mapeamento revelou que as soluções de IA nos tribunais concentram-se principalmente em tarefas como a classificação de documentos, a similaridade de texto e a busca semântica. Esses achados foram confirmados posteriormente no estudo de (ARAÚJO et al., 2025b). A classificação de texto, presente em 49,3% dos projetos, exemplifica o uso de IA para categorizar conteúdos de documentos jurídicos, o que é relevante em um sistema que lida com um número extraordinário de processos. Adicionalmente, 47,1% dos projetos focam no aumento da precisão e da consistência de tarefas repetitivas, tendo em vista a necessidade de análise de um grande conjunto de processos jurídicos (CNJ, 2024).

Essa realidade nacional se reflete de forma acentuada no contexto específico do Maranhão. O Tribunal de Justiça do Maranhão (TJMA) administra um acervo de aproximadamente 1,76 milhão de processos em tramitação em 2025, atendendo a uma população de cerca de 7 milhões de habitantes distribuídos em 217 municípios. Com apenas 228 magistrados em atividade[1]. A Figura 1 ilustra a evolução desse cenário, evidenciando o crescimento constante do acervo processual e a necessidade de soluções tecnológicas.

[1] <https://www.tjma.jus.br/transparencia/portal/pessoal/quantitativo-cargos/situacao-funcional-magistrados-ativos>

Figura 1 – Processos em tramitação no TJMA (09/2025)



Fonte: Adaptado de CNJ (2024).

Diante desse cenário de sobrecarga processual demonstrado na Figura 1, e reconhecendo a necessidade de modernização tecnológica para enfrentar tal desafio, destaca-se o Acordo de Cooperação Técnica Nº 002/2021 celebrado entre o TJMA e a Universidade Estadual do Maranhão (UEMA), voltado para o desenvolvimento conjunto de soluções tecnológicas para otimizar o processamento de documentos jurídicos, alinhando-se com a crescente demanda por eficiência e celeridade processual no âmbito do poder judiciário estadual (CNJ, 2024). Assim, o acordo prevê a aplicação de IA e automação de rotinas em sistemas de processo judicial, bem como o estudo de formas de integrar bases de dados por meio de técnicas de IA para o desenvolvimento de aplicações no sistema de informações do judiciário[2], com destaque em tecnologias como o BumbaBERT (CARMO, 2024), modelo pré-treinado para tarefas de PLN e a Robô Maria Firmina[3], ferramenta que analisa automaticamente o texto da petição inicial, identificando os precedentes qualificados aplicáveis (BEZERRA et al., 2025).

O modelo BumbaBERT, pré-treinado especificamente para o domínio jurídico brasileiro, baseia-se na arquitetura *Bidirectional Encoder Representations from Transformers* (BERT) (CARMO, 2024; DEVLIN, 2018). Este foi treinado com um *corpus* jurídico composto por legislações, jurisprudências e doutrinas brasileiras, visando capturar as nuances e especificidades da linguagem jurídica nacional, demonstrando capacidades notáveis em tarefas de PLN, como a classificação de petições iniciais (CARMO, 2024). Apesar de seu desempenho expressivo, o modelo está sujeito à limitação estrutural dos modelos baseados em BERT, cujo tamanho máximo de entrada é restrito a 512 *tokens*. Essa restrição inviabiliza o processamento direto de documentos jurídicos extensos, como as petições iniciais que apresentam uma média geral de 10.234 *tokens*, evidenciando um obstáculo para a aplicação de modelos convencionais em contextos jurídicos reais (KALAMKAR et al., 2022; PAPPAGARI et al., 2019).

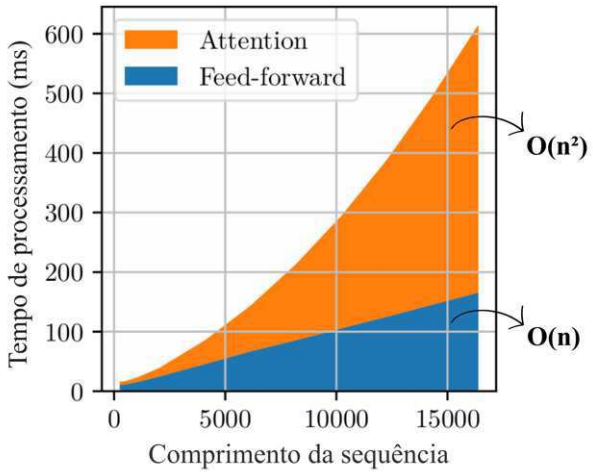
[2] <https://www.uema.br/2022/03/uema-assina-acordo-de-cooperacao-tecnica-com-tj-ma/>
[3] <https://bit.ly/4o1tSuh>

Frente ao exposto, o presente trabalho investiga e compara diferentes métodos baseados em modelos de linguagem para o processamento de documentos longos, com especial atenção à classificação de documentos jurídicos, uma das principais tarefas a que o BumbaBERT foi direcionado (CARMO, 2024). Dessa maneira, buscou-se avançar na automação do fluxo processual, possibilitando a extração precisa de informações relevantes e a classificação eficiente de documentos jurídicos. No âmbito do acordo de cooperação com o Tribunal de Justiça do Maranhão, antecipa-se que a implementação destas soluções tecnológicas contribuirá para a redução do tempo de tramitação processual, o aprimoramento da qualidade das decisões judiciais e o aumento da produtividade, alinhando-se aos objetivos do programa Justiça 4.0.

1.2 Justificativa

Apesar dos avanços consideráveis, os modelos, como o BERT, possuem restrição de tamanho da entrada, uma característica inerente à arquitetura *Transformers*. Essas limitações estão relacionadas com a estrutura interna desses modelos, que possuem um tamanho máximo de entrada fixado em 512 *tokens*, o que equivale a 500 palavras aproximadamente (DEVLIN, 2018). Isso se deve ao fato de que o mecanismo de atenção, fundamental para o funcionamento desses modelos, escala quadraticamente com o comprimento da sequência de entrada (Figura 2), o que aumenta significativamente o custo computacional para documentos longos (VASWANI, 2017).

Figura 2 – Custo computacional do mecanismo de atenção em função do comprimento da sequência



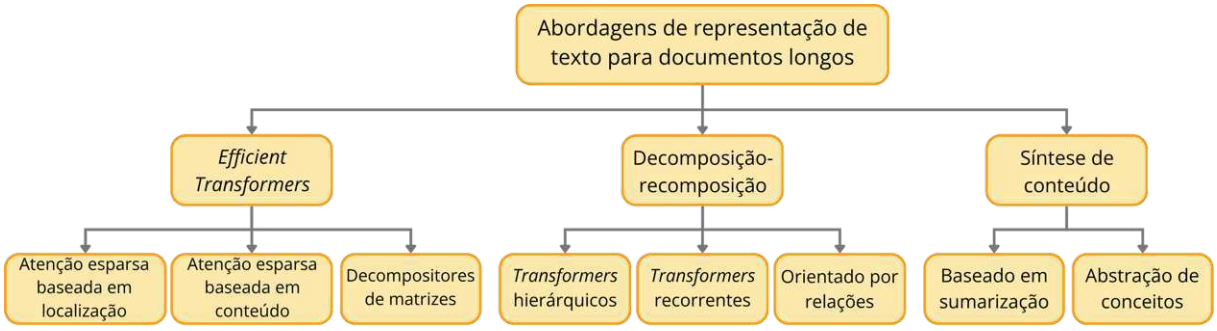
Fonte: Adaptado de Pagliardini et al. (2023). O componente de atenção ("*attention*") cresce quadraticamente ($O(n^2)$), enquanto o das camadas *feed-forward* aumenta linearmente ($O(n)$).

Conforme demonstrado, o componente de atenção domina progressivamente o tempo de processamento à medida que o comprimento da sequência aumenta, enquanto

outras operações, como *feed-forward*, mantêm crescimento linear (Figura 2). Esse gargalo computacional, decorrente da complexidade quadrática $O(n^2)$ do mecanismo de autoatenção, torna impraticável o processamento completo de textos longos sem técnicas de otimização específicas. Adicionalmente, a extensão de documentos amplifica o problema de esparsidade de dados, manifestando-se no crescimento exponencial do número de combinações possíveis de palavras à medida que o vocabulário aumenta, enquanto o conjunto de treinamento permanece fixo, elevando a probabilidade de encontrar sentenças não vistas durante o treinamento (BENGIO et al., 2003).

Contornando essas limitações, surgiram abordagens conhecidas como *Long Transformers* (LTs), projetadas especificamente para lidar com os desafios da Classificação Automática de Documentos Longos, do inglês *Automatic Long Document Classification* (ALDC) (PRINCIPE; CHIARINI; VIVIANI, 2025), conforme ilustrado na Figura 3.

Figura 3 – Abordagens em representação de texto para documentos longos



Fonte: Adaptado de Principe, Chiarini e Viviani (2025).

A taxonomia estruturada apresenta três categorias principais com suas respectivas subcategorias:

- *Efficient Transformers*: modificam o mecanismo de autoatenção por meio de técnicas como atenção esparsa baseada em localização, utilizando padrões fixos como janelas deslizantes; atenção esparsa baseada em conteúdo, onde padrões de esparsidade são aprendidos; ou aproximação de *low-rank* por meio de decomposição de matrizes (BELTAGY; PETERS; COHAN, 2020; PRINCIPE; CHIARINI; VIVIANI, 2025);
- Modelos de decomposição-recomposição: dividem documentos longos em segmentos menores para processamento sequencial, incluindo abordagens hierárquicas ao processar documentos em múltiplos níveis (PAPPAGARI et al., 2019); as recorrentes, conectando segmentos por meio de componentes de Redes Neurais Recorrentes - do inglês, *Recurrent Neural Network* (RNN), por fim, as orientadas por relações de grafos para modelar relacionamentos entre entidades (LI et al., 2023);

- Modelos de síntese de conteúdo: simplificam o documento original por meio de sumarização, selecionando segmentos-chave, ou por meio da abstração de conceitos, transformando-o em representações de alto nível, mantendo, assim, informações suficientes para gerar uma representação significativa (DING et al., 2020; PARK; VYAS; SHAH, 2022).

Entretanto, ainda persistem obstáculos adicionais na aplicação desses modelos ao domínio jurídico, como a linguagem técnica específica, a estrutura complexa dos documentos legais e a necessidade de compreensão contextual extensa (HUA et al., 2022; CHALKIDIS et al., 2020; PRINCIPE; CHIARINI; VIVIANI, 2025). Soma-se a isso a preocupação com questões de privacidade e a necessidade de profissionais qualificados para rotular dados jurídicos, o que, assim, limita a disponibilidade de conjuntos de dados diversificados e abrangentes para treinamento (KALAMKAR et al., 2022).

Embora os LLMs mais avançados sejam capazes de abranger o tamanho desses documentos jurídicos, ainda persistem preocupações quanto à transparência dos dados de treinamento desses modelos, sobretudo quando se referem aos "caixas-pretas" (do inglês, *black-boxes*) (BENDER et al., 2021; BOMMASANI et al., 2021) em questões relacionadas à privacidade dos dados processados por esses sistemas (CARLINI et al., 2021; KEARNS; ROTH, 2019).

Nesse contexto, o uso de modelos pré-treinados ou treinados do zero, especificamente para o domínio jurídico brasileiro, ainda apresenta vantagens consideráveis (CHALKIDIS et al., 2020; ZHONG et al., 2020). A aplicação desses modelos, em sua maioria, é direcionada ao contexto de Classificação Automática de Documentos (do inglês *Automatic Document Classification* - ADC), no qual são treinados em grandes *corpora* cuidadosamente curados e controlados, que permitem a transferência flexível de conhecimento para tarefas subsequentes (RIBEIRO et al., 2020). Podem também ser otimizados para capturar nuances específicas da linguagem jurídica brasileira e do sistema jurídico nacional, algo que modelos generalistas podem não conseguir fazer com a mesma precisão (CHALKIDIS et al., 2020).

O modelo pré-treinado BumbaBERT insere-se nesse cenário, como um *baseline* ainda promissor para a classificação de petições iniciais, por preservar a integridade dos textos jurídicos sem processamento externo prévio (CARMO, 2024). Assim, a otimização de métodos que estendam a capacidade do BumbaBERT para documentos longos justifica-se pela necessidade de combinar especialização no domínio, fidelidade textual e transparência metodológica, aspectos determinantes em aplicações jurídicas.

Diante do exposto, constata-se a necessidade premente de explorar e desenvolver soluções computacionais eficientes para o processamento de textos jurídicos longos, adaptadas às particularidades do sistema jurídico brasileiro. Tais soluções devem não apenas superar as limitações técnicas dos modelos atuais, mas também lidar com os

desafios inerentes ao domínio, contribuindo para uma prestação jurisdicional mais célere, consistente e equitativa.

Sendo assim, o trabalho contribui, em termos científicos, para o avanço do processamento de linguagem natural no domínio jurídico, ao demonstrar empiricamente como estratégias já consolidadas podem ser reinterpretadas e ajustadas para atender às especificidades linguísticas e estruturais dos textos jurídicos brasileiros. Do ponto de vista tecnológico e institucional, o estudo oferece um artefato reprodutível, passível de integração ao sistema de automação do TJMA, contribuindo para a redução do tempo de tramitação processual e para o fortalecimento de políticas de transformação digital no setor público. Por fim, em dimensão social, reafirma o papel da ciência de dados como instrumento de democratização do acesso à justiça, promovendo uma inovação que alia precisão técnica, responsabilidade ética e compromisso com o interesse público.

1.3 Objetivos

Diante do contexto apresentado, estabeleceram-se os seguintes objetivos:

1.3.1 Objetivo Geral

O objetivo geral deste trabalho foi comparar e avaliar métodos baseados em modelos de linguagem para o processamento eficiente de documentos jurídicos longos, buscando superar as limitações inerentes aos modelos de arquitetura *Transformer*.

1.3.2 Objetivos Específicos

À luz do objetivo geral, destacam-se os seguintes objetivos específicos:

- Identificar e analisar sistematicamente os projetos de inteligência artificial implementados no setor jurídico brasileiro, por meio de uma revisão sistemática e meta-análise de estudos existentes, visando mapear o panorama atual e quantificar o impacto desses projetos no setor público;
- Mapear e selecionar os principais métodos existentes na literatura para o processamento de textos longos baseados em Modelos de Linguagem, se possível, com ênfase em abordagens aplicáveis a documentos jurídicos;
- Implementar e adaptar os métodos selecionados para o contexto específico de documentos jurídicos longos, focando em tarefas de classificação de petições iniciais;
- Realizar a análise comparativa dos métodos implementados, considerando métricas de desempenho, como acurácia, precisão, revocação e F1-*score*, bem como eficiên-

cia computacional e escalabilidade, e validação estatística por meio de testes de significância.

1.4 Principais contribuições e resultados

Esta dissertação gerou impactos relevantes tanto na esfera acadêmica quanto na prática, contribuindo para o avanço do estado da arte no processamento de documentos jurídicos longos e oferecendo benefícios tangíveis para problemas reais.

1. **Contribuições para a comunidade científica e acadêmica.** O estudo contribui com avanços no processamento de textos longos, onde a comparação e avaliação dos métodos implementados evidenciam estratégias mais eficientes para lidar com as limitações dos modelos baseados em *Transformers* (*e.g.*, BERT, BumbaBERT). Os resultados demonstram a viabilidade prática de soluções adaptadas às especificidades do domínio jurídico brasileiro, estimulando novas pesquisas e colaborações nesta área;

2. **Contribuições para o domínio jurídico.** O presente estudo focou em melhorias na eficiência e na transparência do sistema judiciário, oferecendo uma solução tecnológica implementável aos desafios inerentes às arquiteturas dos modelos. No contexto do acordo UEMA-TJMA, a dissertação oferece subsídios técnicos para a integração e a otimização dos sistemas de identificação de precedentes vinculantes, com aumento da produtividade de magistrados e servidores. Com isso, alinhando-se diretamente aos objetivos do programa Justiça 4.0 ao aprimorar a capacidade de processar e analisar documentos jurídicos em sua totalidade, promovendo um sistema judiciário mais ágil, eficaz e, principalmente, transparente;

3. **Contribuições metodológicas e mapeamento do panorama nacional.** A revisão sistemática conduzida (ARAÚJO et al., 2025b) trouxe uma compreensão ampla da aplicação de IA no contexto jurídico brasileiro. Por meio do mapeamento do estado-da-arte, foram identificados sistematicamente 112 projetos de IA distribuídos em 90 tribunais brasileiros, mapeando suas características técnicas, metodologias empregadas e resultados alcançados em estudos científicos. Essa análise preencheu uma lacuna importante, estabelecendo uma compreensão sobre a adoção de IA no judiciário brasileiro e servindo de referência para estudos futuros. A abordagem metodológica desenvolvida contribuiu para o avanço na interseção entre IA, sistemas de informação e o setor jurídico, destacando questões que necessitam de maior atenção e fornecendo evidências para impulsionar a transformação digital do sistema judiciário;

4. **Produções científicas e disseminação do conhecimento.** Ao longo do desenvolvimento desta dissertação, foram geradas seis produções científicas ,diretas e indiretas, que contribuíram para a validação e o amadurecimento metodológico do estudo, das quais duas foram publicadas e quatro estão submetidas. Destes, as produções diretas incluem: (i) artigo resultante da revisão sistemática sobre o uso de IA no sistema judicial brasileiro, publicado nos anais do Simpósio Brasileiro de Sistemas da Informação (SBSI 2025), citado como Araújo et al. (2025b) e (ii) estudo sobre técnicas híbridas de sumarização em decisões judiciais, submetido ao SBSI 2026, cuja versão integral encontra-se no Apêndice A. Também foram desenvolvidas produções indiretas decorrentes da missão de estágio internacional realizada na Alemanha, no âmbito do Programa Abdias Nascimento, aprovado pelo Edital N^o 16/2023 da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), envolvendo temas correlatos à governança de dados, à tradução automática aplicada à acessibilidade e a um *pipeline* híbrido para detecção de pneumonia (ARAÚJO et al., 2025a). Tais itens foram descritos com mais detalhes no Capítulo 7.

5. **Contribuições técnicas** Foi desenvolvido um *pipeline* computacional reprodutível que integra o BumbaBERT a dois métodos comparativos para o processamento de documentos longos. O primeiro, baseado em decomposição-recomposição hierárquica, inclui a segmentação adaptativa em *chunks* de 200 *tokens* com sobreposição de 50 *tokens*, a agregação hierárquica por *Transformer Encoder* (2 cabeças de atenção) e *mean pooling* para obter uma representação unificada do documento. O segundo baseia-se na síntese de conteúdo, por meio da seleção extrativa de sentenças com base na relevância semântica e abstrativa, utilizando *few-shot prompting* neste último caso. O código-fonte completo e documentado encontra-se em um repositório privado do TJMA e está em processo de validação pelo stakeholder para posterior registro junto ao Instituto Nacional da Propriedade Industrial (INPI) e transferência de tecnologia, com acompanhamento da implantação em ambiente de produção do TJMA.

1.5 Organização do trabalho

Este documento encontra-se estruturado como segue. Nos Capítulos 2 e 3 são fundamentados os principais conceitos jurídicos e computacionais, abordando desde sistemas de precedentes brasileiros, modelo métodológico *Data Science Trajectories* (DST), conceitos de PLN e variantes baseadas em *Transformers*, incluindos os modelos de linguagem de domínio geral e específico. No Capítulo 3, esses conceitos são aprofundados no contexto de documentos longos, fundamentando os métodos para lidar com essa extensão e avaliá-los. Por seguinte, o Capítulo 4 compõe a parte mais aplicada dos fundamentos introduzidos anteriormente, detalhando por meio de trabalhos relacionados estudos que abordam o processamento de textos longos utilizando modelos de linguagem por meio de diferentes

estratégias, e com isso identificando as lacunas e posicionando o presente estudo frente ao estado-da-arte. No Capítulo 5, são descritos os materiais e métodos empregados na condução do estudo, incluindo o planejamento da trajetória DST para a comparação e avaliação de métodos de processamento de documentos longos, o entendimento do domínio, a exploração e a preparação dos dados, a configuração experimental e os procedimentos de avaliação e validação. O Capítulo 6 apresenta a análise e a interpretação dos resultados, incluindo a comparação de desempenho e de eficiência computacional dos métodos propostos, a análise das diferenças estatísticas observadas e a discussão das implicações práticas dos achados. Finalmente, o Capítulo 7 sintetiza os principais resultados, alinhando-os às contribuições científicas, tecnológicas e institucionais, levando em conta os impactos sociais, além das limitações e suas perspectivas futuras. Complementam este documento o Apêndice A, contendo o artigo submetido ao SBSI 2026 sobre sumarização híbrida de decisões judiciais, e o Anexo A, com o comprovante de submissão do referido artigo.

2 Fundamentos jurídicos-computacionais

Neste Capítulo, é apresentado o referencial teórico que sustenta o desenvolvimento da presente dissertação, delineando bases conceituais, jurídicas, tecnológicas e metodológicas que o orientam. O conteúdo foi elaborado a partir de uma revisão da literatura do tipo *ad hoc*, selecionando fontes pertinentes para proporcionar uma compreensão do tema. A organização segue uma progressão lógica do tema geral desde fundamentos jurídicos até computacionais relacionados à PLN. Na Seção 2.1 é contextualizado o domínio jurídico, descrevendo a natureza dos documentos processuais e o funcionamento do sistema de precedentes brasileiro, com ênfase nos IRDRs do TJMA. Em seguida, nas Seções 2.3 e 2.5 são apresentados os fundamentos de PLN, desde a classificação de textos até a arquitetura *Transformer* e os modelos de linguagem que constituem a base tecnológica deste estudo.

2.1 Fundamentos jurídicos

A discussão jurídica insere-se no contexto da iniciativa Justiça 4.0 conduzida pelo CNJ, que visa à transformação digital do Poder Judiciário por meio de soluções baseadas em dados e IA. Nesse cenário, compreender a estrutura e o funcionamento do processo judicial é importante para desenvolver tecnologias capazes de auxiliar magistrados e servidores na triagem, análise e classificação de documentos processuais.

2.1.1 Processo judicial

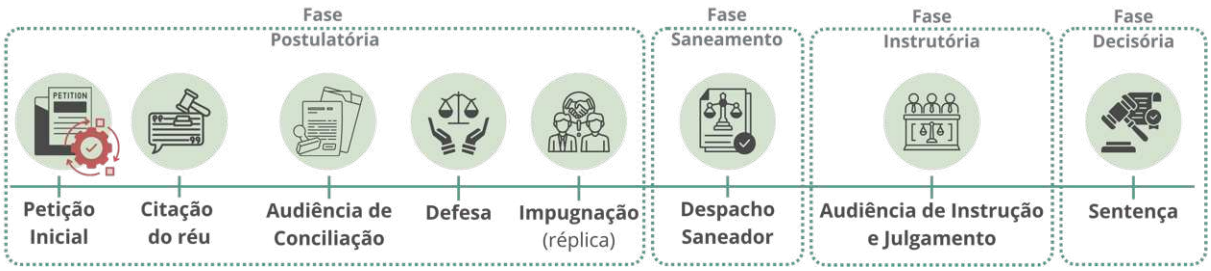
O domínio jurídico brasileiro é composto por um ecossistema complexo de atores institucionais, incluindo magistrados, advogados, membros do Ministério Público, defensores públicos, servidores e cidadãos que interagem por meio de documentos formais e atos processuais regulados pela legislação vigente (RUSSO, 2009). Entre esses instrumentos, a petição inicial representa o ponto de entrada do cidadão no sistema de Justiça e o marco inaugural de qualquer processo judicial.

Por meio dela, o jurisdicionado formula sua pretensão ao Estado-Juiz, expondo os fatos e os fundamentos jurídicos que justificam o pedido de tutela jurisdicional. Sua estrutura típica inclui a qualificação das partes, a narração dos fatos, a fundamentação jurídica e os pedidos, elementos que devem ser adequadamente identificados e processados por sistemas automatizados (NEVES, 2015; Brasil, 2015). Assim, a petição inicial concretiza o direito fundamental de acesso à Justiça, previsto no art. 5º, inciso XXXV, da Constituição

Federal de 1988, segundo o qual "a lei não excluirá da apreciação do Poder Judiciário lesão ou ameaça a direito" (BRASIL, 1988)

O procedimento comum no processo judicial brasileiro pode ser sintetizado em quatro fases principais, conforme ilustrado na Figura 4.

Figura 4 – Fluxo simplificado das fases do procedimento comum no processo judicial brasileiro.



Fonte: Produzido com base em (JR; BRAGA; OLIVEIRA, 2010).

A fase postulatória define a controvérsia a partir da exposição dos fatos, fundamentos jurídicos e manifestação das partes. Na fase de saneamento, o magistrado verifica os requisitos processuais e delimita os pontos controvertidos relevantes para a instrução. A fase instrutória é voltada à produção de provas e esclarecimentos necessários para o julgamento. Por fim, a fase decisória encerra o processo com a sentença, na qual o juiz aprecia o mérito e entrega a tutela jurisdicional (JR; BRAGA; OLIVEIRA, 2010; Brasil, 2015).

Do ponto de vista procedimental, a petição inicial desencadeia um fluxo documental que compreende o protocolo eletrônico, a autuação e distribuição do processo, a triagem inicial por servidores e a subsequente análise judicial. Esse fluxo é responsável por movimentar grande parte da carga de trabalho do Poder Judiciário, o que torna sua automação estratégica para a celeridade processual e a efetividade do princípio constitucional da duração razoável do processo (art. 5º, LXXVIII, CF/88) (BRASIL, 1988).

A análise dessas petições pelo Poder Judiciário resulta em decisões que, quando reiteradas e uniformes, constituem a jurisprudência, tradicionalmente entendida como o "conjunto de decisões uniformes e constantes dos tribunais resultante da aplicação de normas a casos semelhantes" (TELLA, 2011). Quando tais decisões adquirem caráter vinculante e passam a orientar casos futuros análogos, configuram-se como precedentes judiciais Mendes (2015), instituto detalhado na próxima Subseção no contexto do sistema processual civil brasileiro contemporâneo.

2.1.2 Sistema de precedentes e demandas repetitivas

A partir das decisões judiciais reiteradas, formam-se os precedentes. Segundo Jr, Braga e Oliveira (2010), precedente,

> "é a decisão judicial tomada à luz de um caso concreto, cujo núcleo essencial pode servir como diretriz para o julgamento posterior de casos análogos".

A importância dos precedentes foi significativamente ampliada com o advento do Código de Processo Civil (CPC) de 2015, que estabeleceu um sistema estruturado para seu tratamento, representando uma aproximação do sistema jurídico brasileiro, tradicionalmente fundamentado no *common law* (MENDES; TEMER, 2015).

O CPC/2015 instituiu mecanismos processuais específicos para promover a uniformização da jurisprudência e a segurança jurídica, introduzindo, assim, o Incidente de Resolução de Demandas Repetitivas (IRDR). Este instrumento processual visa estabelecer uma tese jurídica única para questões idênticas, otimizando o julgamento de processos semelhantes (MENDES; TEMER, 2015; Brasil, 2015). Os incs. I e II do art. 976 do CPC/2015 mencionam que:

> "É cabível a instauração do incidente de resolução de demandas repetitivas quando houver, simultaneamente: I - efetiva repetição de processos que contenham controvérsia sobre a mesma questão unicamente de direito; II - risco de ofensa à isonomia e à segurança jurídica (Brasil, 2015)."

A exigência de "efetiva repetição" não pressupõe número mínimo absoluto de processos, mas demanda quantidade suficiente para caracterizar contencioso de massa que justifique intervenção uniformizadora do tribunal (TEMER et al., 2019). O requisito de "controvérsia sobre a mesma questão unicamente de direito" delimita que o IRDR não se presta a uniformizar questões fáticas, mas sim interpretações divergentes sobre matéria jurídica aplicável a situações análogas. Já o "risco de ofensa à isonomia e à segurança jurídica" manifesta-se quando decisões contraditórias sobre a mesma questão jurídica geram tratamento desigual para jurisdicionados em situações equivalentes, violando o princípio constitucional da isonomia (art. 5º, *caput*, CF/88) (BRASIL, 1988) e gerando instabilidade na aplicação do direito (Brasil, 2015). Adiciona-se a esses requisitos um terceiro, de natureza negativa, consistente na inexistência de afetação pelos Tribunais Superiores de recurso para definição de tese sobre questão de direito repetitiva (art. 976, §4º) (ZUFELATO; OLIVEIRA, 2024).

Após a instauração do IRDR, o tribunal competente (2ª instância) suspende todos os processos pendentes em sua jurisdição que versem sobre a mesma questão jurídica, até o julgamento do incidente (TEMER et al., 2019). Uma vez julgado, a tese firmada deve ser aplicada a todos os processos suspensos e aos casos futuros que versem sobre idêntica questão de direito, no âmbito de competência do respectivo tribunal (TEMER et al., 2019).

2.1.3 IRDR no TJMA

No âmbito do TJMA, objeto de análise específica neste estudo, o Núcleo de Gerenciamento de Precedentes (NUGEPNAC) instituiu um sistema de catalogação e gestão dos IRDRs julgados em sua jurisdição (TJMA, 2025). Até a conclusão desta pesquisa, o TJMA havia catalogado 11 temas de IRDR[1], distribuídos conforme apresentado na Tabela 1.

Tabela 1 – Temas de IRDR catalogados pelo TJMA

Tema	**Processo**	**Objeto da Controvérsia**	**Área do Direito**
01	17.015/2016	Direito de servidores estaduais à diferença de reajuste de 21.7% em razão da concessão de índices diferenciados pela Lei nº 8.369/2006	Direito Administrativo
02	22.965/2016	Revisão de reajuste do percentual de 6.1% aos servidores públicos estaduais	Direito Administrativo
03	48.732/2016	Nomeação de candidatos excedentes em concurso público para professor do Estado	Direito Administrativo
04	3.043/2017	Legalidade de descontos de tarifas bancárias em conta de beneficiários do INSS	Direito do Consumidor / Bancário
05	53.983/2016	Validade de cláusulas contratuais em empréstimos consignados	Direito do Consumidor / Bancário
07	54.699/2017	Cabimento e cálculo de honorários sucumbenciais em execução individual de sentença coletiva	Direito Processual Civil
08	0801095-52.2018.8.10.0000	Termo inicial da prescrição em ações de promoção de militares estaduais	Direito Administrativo / Militar
09	0819580-95.2021.8.10.0000	Procedimento de revisão de tese do IRDR 07 (honorários sucumbenciais)	Direito Processual Civil
10	0817757-23.2020.8.10.0000	Cabimento de ações rescisórias ajuizadas pelo Estado do Maranhão sobre reajustes de 21.7% e 6.1%	Direito Administrativo
11	0823994-05.2022.8.10.0000	Termo inicial do prazo prescricional da sentença proferida em ação coletiva nº 6.542/2005	Direito Processual Civil
12	0827453-44.2024.8.10.0000	Revisão de teses do IRDR nº 5 (empréstimos consignados)	Direito do Consumidor / Bancário

Fonte: TJMA (2025).

A distribuição temática dos IRDRs do TJMA reflete o perfil do contencioso judicial maranhense, com predominância de questões relacionadas ao Direito Administrativo (temas 01, 02, 03, 08 e 10), notadamente envolvendo servidores públicos estaduais e militares, seguida de matérias de Direito do Consumidor e Bancário (temas 04, 05 e 12), concentradas em relações de consumo de serviços financeiros. Questões processuais (temas 07, 09 e 11) também figuram entre os IRDRs, evidenciando a necessidade de uniformização não apenas do direito material, mas também do procedimento. Essa concentração temática alinha-se

[1] O TEMA 06 foi cancelado, razão pela qual a numeração apresenta descontinuidade.

aos achados da pesquisa nacional, que identificou Direito Administrativo (48 IRDRs) e Direito Processual (31 IRDRs) como as áreas com maior número de incidentes admitidos até junho de 2018 (ZUFELATO; OLIVEIRA, 2024), sugerindo que matérias envolvendo entes públicos constituem núcleo preferencial de utilização do instituto.

As petições iniciais que se enquadram em temas de IRDR apresentam características estruturais específicas que influenciam seu processamento automatizado. Embora mantenham a estrutura convencional de qualificação das partes, narrativa fática, fundamentação jurídica e pedidos, essas petições frequentemente apresentam peculiaridades relevantes (NEVES, 2015). A fundamentação jurídica tende a ser mais extensa, com citações de precedentes, doutrina e legislação pertinente, uma vez que os advogados buscam demonstrar a relevância e a repetitividade da questão (ARIOZO; DOMINGOS, 2025).

A narrativa fática, por sua vez, apresenta padrões de similaridade entre casos enquadrados no mesmo tema IRDR, com estrutura padrão e variações pontuais (valores, datas, nomes), o que pode facilitar ou dificultar a identificação do tema central dependendo da estratégia de processamento adotada (ZUFELATO; OLIVEIRA, 2024). Adicionalmente, é comum a reprodução literal de trechos extensos de leis, regulamentos e normas infralegais, bem como referências a decisões judiciais anteriores, incluindo cópias integrais de acórdãos ou sentença paradigma[2], resultando em documentos com dezenas de milhares de palavras (Jusbrasil, 2025; KALAMKAR et al., 2022).

Essas características estruturais impõem desafios metodológicos à classificação automatizada de petições de IRDR. A identificação do tema IRDR possui implicações para a eficiência judiciária, viabilizando a suspensão processual tempestiva de demandas correlatas (art. 982, CPC/2015), a aplicação uniforme de teses vinculantes (art. 985) e a governança baseada em evidências do acervo processual (CNJ, 2024; Brasil, 2015). Nesse contexto, os projetos de IA desenvolvidos no âmbito do programa Justiça 4.0 têm buscado aprimorar a gestão processual por meio do uso de dados estruturados, da interoperabilidade entre sistemas e da criação de artefatos reprodutíveis aplicáveis a diferentes tribunais (CNJ; PNUD, 2021; ARAÚJO et al., 2025b). Tais exigências reforçam a necessidade de empregar abordagens metodológicas que integrem o domínio jurídico às etapas de ciência de dados de modo iterativo e transparente, favorecendo o desenvolvimento de soluções alinhadas às especificidades do sistema de justiça. Nesse contexto, a Seção a seguir apresenta esse tipo de abordagem, fundamentada em variantes metodológicas que evoluíram ao longo do tempo.

[2] Decisões judiciais utilizadas como referência para comparação ou uniformização de entendimentos em casos análogos, cuja autoridade depende do trânsito em julgado (art. 502, CPC/2015).

2.2 Modelo metodológico DST

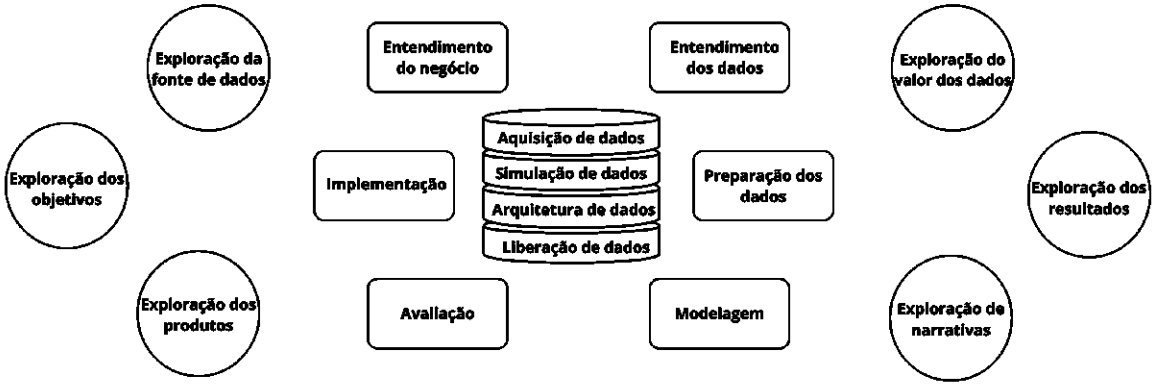
O modelo *Cross Industry Standard Process for Data Mining* (CRISP-DM), desenvolvido na segunda metade da década de 1990 (WIRTH; HIPP, 2000), consolidou-se como metodologia *de facto* para projetos de mineração de dados e descoberta de conhecimento, permanecendo amplamente utilizado em estudo relacionados à predição de risco de doenças crônicas (SISKA; OKTAVIA, 2024), sistemas de detecção de *fake news* em redes sociais (CAVUS; GOKSU; OKTEKIN, 2024) e classificação automatizada de textos (CHI et al., 2024), por exemplo. Contudo, após mais de duas décadas, o campo evoluiu consideravelmente. A ubiquidade de dispositivos eletrônicos e sensores, o uso massivo de redes sociais e a capacidade de armazenar e processar volumes de dados exponencialmente maiores transformaram fundamentalmente a natureza dos dados disponíveis, bem como os próprios processos para extrair valor deles (MARTÍNEZ-PLUMED et al., 2019). Mais importante, enquanto o CRISP-DM pressupõe projetos orientados por metas (*goal-directed*) com objetivos de negócio bem definidos desde o início, a ciência de dados contemporânea caracteriza-se por uma abordagem frequentemente exploratória, em que os dados assumem papel central e os objetivos podem ser descobertos iterativamente ao longo do projeto (MARTÍNEZ-PLUMED et al., 2019).

Reconhecendo essas limitações, Martínez-Plumed et al. (2019) propõem o modelo DST, que generaliza CRISP-DM integrando atividades exploratórias e de gerenciamento de dados ausentes no modelo original, mantendo compatibilidade retroativa com trajetórias tradicionais de mineração de dados. Diferentemente da prescritividade sequencial do CRISP-DM, o DST oferece um mapa flexível de atividades possíveis sem ordem predeterminada, permitindo que cientistas de dados componham trajetórias específicas adaptadas ao contexto de cada projeto (MARTÍNEZ-PLUMED et al., 2019).

2.2.1 Estrutura do modelo DST

O mapa DST, ilustrado na Figura 5, organiza-se em três camadas concêntricas de atividades complementares que capturam a diversidade de projetos contemporâneos de ciência de dados (MARTÍNEZ-PLUMED et al., 2019).

Figura 5 – Mapa DST com três camadas de atividades: exploratórias (círculo externo), direcionadas por metas (círculo interno) e gerenciamento de dados (núcleo central)



Fonte: Adaptado de Martínez-Plumed et al. (2019)

O círculo externo do mapa DST integra atividades exploratórias, ausentes no CRISP-DM original, que caracterizam projetos em que objetivos, dados ou produtos são descobertos iterativamente (MARTÍNEZ-PLUMED et al., 2019). Na Tabela 2 são sumarizadas essas atividades.

Tabela 2 – Atividades exploratórias do modelo DST

Atividade	**Descrição**
Exploração dos objetivos	Identificação de objetivos de negócio que podem ser alcançados de forma orientada por dados, frequentemente descobertos por meio da análise exploratória inicial dos dados disponíveis.
Exploração das fontes de dados	Descoberta e avaliação de novas fontes de dados potencialmente valiosas, internas ou externas à organização, incluindo APIs, sensores IoT, redes sociais e repositórios públicos.
Exploração do valor dos dados	Investigação sistemática do valor potencial contido nos dados, identificando padrões, anomalias e oportunidades não evidentes *a priori*.
Exploração dos resultados	Relacionamento iterativo dos resultados obtidos de análises e modelos com os objetivos de negócio, frequentemente levando ao refinamento ou à redefinição de objetivos.
Exploração da narrativa	Extração e estruturação de narrativas valiosas (visuais ou textuais) a partir dos dados, comunicando *insights* de forma acessível a *stakeholders* não técnicos.
Exploração do produto	Identificação de formas de transformar o valor extraído dos dados em serviços, aplicativos ou produtos que entreguem valor novo e tangível a usuários e clientes.

Fonte: Elaborada com base nos conceitos de Martínez-Plumed et al. (2019)

Essas atividades são tipicamente mais abertas do que as fases estruturadas do CRISP-DM, requerendo especialistas com conhecimento profundo do domínio e habilidades tanto técnicas quanto de comunicação (MARTÍNEZ-PLUMED et al., 2019). A ordem de execução dessas atividades depende do domínio, das descobertas realizadas e das decisões do cientista de dados, podendo ocorrer múltiplas vezes ao longo do projeto. Já o círculo interno preserva as seis fases do CRISP-DM, que permanecem válidas e relevantes para projetos com objetivos bem definidos desde o início (MARTÍNEZ-PLUMED et al., 2019). Contudo, no contexto do DST, essas fases não necessariamente ocorrem na ordem canônica, podendo ser interrompidas por atividades exploratórias ou executadas parcialmente quando os produtos intermediários, como dados preparados para publicação, constituem o resultado final do projeto (MARTÍNEZ-PLUMED et al., 2019).

As fases CRISP-DM incorporadas ao DST incluem desde a compreensão do negócio, estabelecendo objetivos e requisitos do projeto; compreensão dos dados, coletando e explorando dados iniciais para identificar problemas de qualidade e formular hipóteses; preparação dos dados, construindo *dataset* final por meio de limpeza, transformação, integração e formatação; modelagem, selecionando e aplicando técnicas de modelagem com calibração de parâmetros; avaliação, revisando modelos para garantir cumprimento de objetivos de negócio; e implantação, organizando e apresentando conhecimento adquirido para uso pelo cliente (WIRTH; HIPP, 2000).

Por fim, o núcleo central do mapa DST reconhece que os dados não constituem apenas um insumo estático do processo, mas também um objeto de trabalho técnico que pode constituir um produto final do projeto (MARTÍNEZ-PLUMED et al., 2019). A Tabela 3 detalha essas atividades.

Tabela 3 – Atividades de gerenciamento de dados do modelo DST

Atividade	**Descrição**
Aquisição de dados	Obtenção ou criação de dados relevantes por meio de instalação de sensores, desenvolvimento de aplicativos, APIs, *web scraping* ou aquisição de *datasets* de terceiros.
Simulação de dados	Geração de dados sintéticos por meio de simulações de sistemas complexos para produzir dados úteis, avaliar cenários hipotéticos ou responder perguntas causais do tipo "e se" (*what-if*).
Arquitetura de dados	Projeto do *layout* lógico e físico dos dados, incluindo a integração de múltiplas fontes, a definição de esquemas e a construção de *data warehouses* e *data lakes*.
Liberação de dados	Disponibilização de dados por meio de bancos de dados, interfaces (APIs), visualizações ou repositórios públicos, potencialmente como produto final do projeto.

Fonte: Elaborada com base nos conceitos de Martínez-Plumed et al. (2019)

Essas atividades refletem mudanças fundamentais na prática de ciência de dados em relação à mineração de dados tradicional, em que os dados frequentemente possuem

múltiplos usos além do contexto original de coleta, com complexidade e volume que requerem infraestrutura especializada e dados curados que podem constituir produtos comercializáveis independentes (MARTÍNEZ-PLUMED et al., 2019).

2.2.2 Trajetórias e categorização de projetos

Uma trajetória no modelo DST é definida como um grafo acíclico direcionado sobre atividades, que representa a sequência de passos executados em um projeto específico (MARTÍNEZ-PLUMED et al., 2019). Diferentemente do CRISP-DM, que prescreve um processo relativamente fixo, embora com iterações permitidas, o DST é exemplar, em vez de prescritivo, oferecendo um catálogo de trajetórias comuns que podem servir como *templates*, mas projetos podem compor trajetórias originais combinando atividades conforme necessário (MARTÍNEZ-PLUMED et al., 2019).

Martínez-Plumed et al. (2019) propõem categorização de projetos de ciência de dados baseada na predominância das três camadas de atividades, resultando em sete regiões possíveis no diagrama de Venn com projetos tradicionais de mineração de dados com objetivos bem definidos, descoberta sem objetivo predefinido, focados em infraestrutura e curadoria, projetos onde descoberta precede ou intercala modelagem, com forte componente de engenharia de dados, de descoberta e publicação de dados e complexos envolvendo todas as três dimensões (MARTÍNEZ-PLUMED et al., 2019).

Essa categorização auxilia o planejamento de projetos quanto a recursos humanos, tempo e custos, permitindo estimativas mais precisas por meio da comparação com projetos similares, em vez de forçar a adequação a um único processo (MARTÍNEZ-PLUMED et al., 2019). Essa flexibilidade do DST permite que pesquisadores projetem seu próprio fluxo de trabalho, removendo as setas rígidas do CRISP-DM e adotando uma abordagem mais adaptativa. Essa característica é particularmente importante para o presente estudo, partindo do pressuposto levantado por Martínez-Plumed et al. (2019) quanto à ciência de dados aplicada, em que não apenas a natureza dos dados mudou, mas também os processos de extração de valor a partir deles. Esses dados podem assumir diferentes formas e estruturas e são utilizados para a criação de modelos, a projeção de artefatos e, em geral, para ampliar a compreensão do assunto. Nesse sentido, o DST não se limita a dados tabulares, mas abrange múltiplos tipos de informação, incluindo dados textuais e linguagem natural, tema da próxima Seção.

2.3 Processamento de linguagem natural

O PLN é um campo interdisciplinar que combina linguística computacional, IA e ciência da computação para desenvolver sistemas capazes de processar e compreender a linguagem humana (HIRSCHBERG; MANNING, 2015). O PLN abrange uma ampla

gama de tarefas, como a tradução automática (HIRSCHBERG; MANNING, 2015), a sumarização de texto (LIU; LAPATA, 2019) e a análise de sentimentos (ZHANG; WANG; LIU, 2018). No domínio jurídico, como reportado por CNJ (2024), a classificação de textos e o Reconhecimento de Entidades Nomeadas - do inglês *Named-Entity Recognition* (NER) - são os mais explorados pelos tribunais brasileiros.

A classificação de textos, em particular, é uma tarefa que envolve a atribuição automática de categorias predefinidas a documentos de texto (AGGARWAL; ZHAI, 2012; KOWSARI et al., 2019). O processo de classificação de textos, embora possa variar dependendo da aplicação, segue um padrão estruturado em seis etapas principais: (i) coleta de dados; (ii) análise para anotação de classes; (iii) construção e ponderação de características; (iv) seleção e projeção de características; (v) treinamento do modelo de classificação (vi) e, finalmente, avaliação da solução (MIROŃCZUK; PROTASIEWICZ, 2018). O processo se inicia com um conjunto de dados textuais brutos, geralmente estruturado como $D = \{X_1, X_2, \ldots, X_N\}$, em que cada X_i representa um ponto de dados (por exemplo, um documento ou segmento de texto) composto por s sentenças, cada uma contendo w_s palavras com l_w letras. Cada ponto de dados é rotulado com um valor de classe proveniente de um conjunto de k índices de valores discretos distintos (AGGARWAL; ZHAI, 2012; KOWSARI et al., 2019).

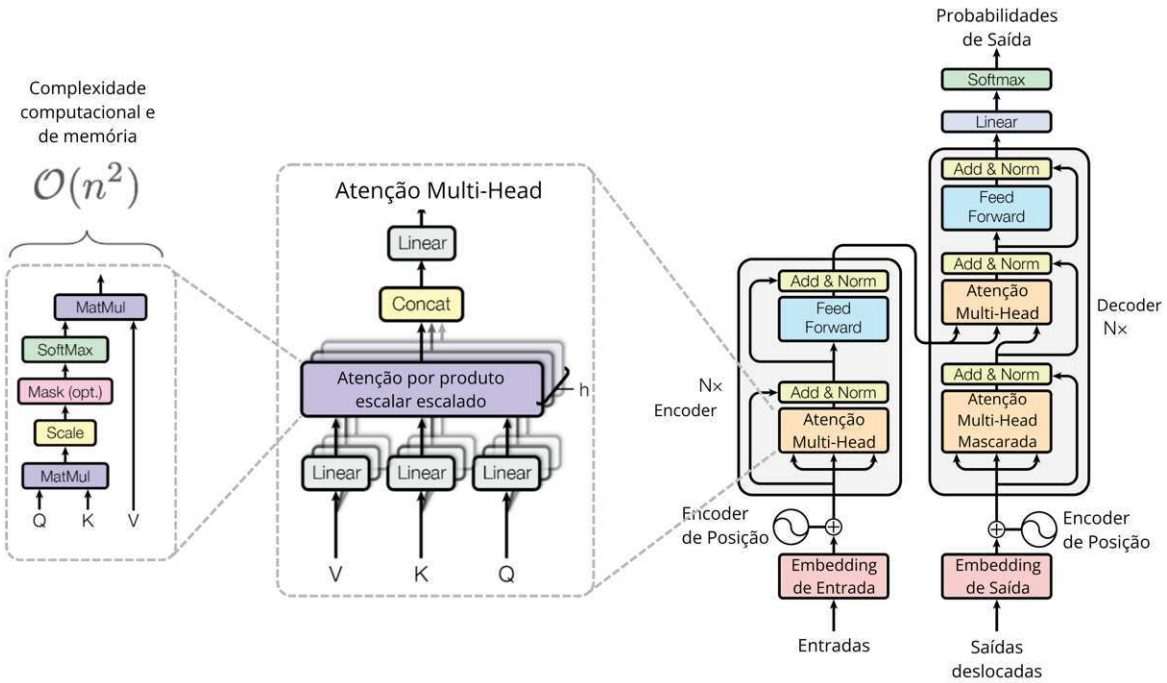
Ao longo dos anos, múltiplas abordagens têm sido propostas para a classificação de textos. Partindo de técnicas clássicas, como *Naïve Bayes*, árvores de decisão e Redes Neurais Artificiais - do inglês, *Artificial Neural Networks* (ANN) (SEBASTIANI, 2002). O advento do *Deep Learning* trouxe uma nova dimensão à área, com arquiteturas como RNNs, Redes Neurais Autoassociativas - do inglês, *Autoencoder Neural Network* (AENN), Redes Neurais Convolucionais - do inglês, *Convolutional Neural Networks* (CNN) e *Long Short-Term Memory* (LSTM) ganhando proeminência (MINAEE et al., 2021; KOWSARI et al., 2019). Complementando esses avanços, a combinação de diferentes algoritmos de classificação por meio de métodos de *Ensemble* tem frequentemente superado o desempenho de técnicas individuais (ANDERLUCCI; GUASTADISEGNI; VIROLI, 2019; DONG et al., 2020).

Além da notória revolução introduzida por esses métodos, é importante notar que a escolha da abordagem de classificação mais adequada depende de diversos fatores, incluindo o prazo, *know-how* da equipe e os recursos computacionais à disposição. A eficácia da solução está intrinsecamente ligada à sua capacidade de entregar resultados precisos em tempo adequado (KOWSARI et al., 2019). Neste contexto de busca contínua por melhorias, há modelos baseados na arquitetura *Transformer* introduzida por Vaswani (2017), marcando um ponto de inflexão na forma como a linguagem natural é processada e compreendida. Esta arquitetura é abordada a seguir.

2.4 Arquitetura *Transformers*

A arquitetura *Transformer* revolucionou o campo do PLN, em que, diferentemente das arquiteturas recorrentes clássicas como RNNs, que processam dados sequencialmente, ou das CNNs, que operam em janelas fixas, os *Transformers* utilizam um mecanismo de atenção que permite o processamento paralelo e a captura de dependências de longo alcance em sequências de texto (VASWANI, 2017; WOLF et al., 2020). O componente central dos *Transformers* é o mecanismo de autoatenção (*self-attention*), que permite que cada elemento da sequência de entrada interaja com todos os demais, ponderando sua importância relativa. A arquitetura *Transformer*, ilustrada na Figura 6, consiste em dois componentes principais: o Codificador (*Encoder*) e o Decodificador (*Decoder*), onde cada um desses é composto por várias camadas idênticas, permitindo um processamento profundo e refinado da informação (VASWANI, 2017).

Figura 6 – Modelo da arquitetura *Transformer* com *Encoder* (à esquerda) e *Decoder* (à direita).



Fonte: Adaptado de Vaswani (2017)

O módulo Codificador, formado pela camada *Positional Encoding*, é responsável por processar a sequência de entrada; cada camada dele possui dois subcomponentes principais. O *Multi-Head Attention* permite que o modelo atenda simultaneamente a diferentes posições da sequência de entrada, capturando diferentes tipos de relações. O *Feed Forward Network*, por sua vez, é uma rede totalmente conectada, aplicada a cada posição de forma independente e idêntica, permitindo um processamento não linear adicional (VASWANI, 2017). Já o Decodificador, por outro lado, é encarregado de gerar a

sequência de saída, onde cada camada contém três subcomponentes principais, o *Masked Multi-Head Attention*, similar à atenção do Codificador, mas mascara as posições futuras para preservar a auto-regressividade durante a geração; o *Multi-Head Attention*, que atende à saída do Codificador, permitindo que o Decodificador integre informações da entrada; por fim o *Feed Forward Network*, similar ao do Codificador (VASWANI, 2017). Essa arquitetura modular permite o processamento paralelo de sequências inteiras, ao mesmo tempo em que captura dependências de longo alcance por meio do mecanismo de atenção.

Matematicamente, o mecanismo de autoatenção é implementado por meio de três matrizes: *Query* (Q), *Key* (K) e *Value* (V), derivadas da entrada por meio de projeções lineares (VASWANI, 2017). O mecanismo de atenção é expresso pela Eq. 2.1:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

onde d_k é a dimensionalidade das chaves.

Uma característica distintiva dos *Transformers* é o uso de múltiplas "cabeças" de atenção em paralelo, um conceito conhecido como *multi-head attention*, que permite ao modelo focar simultaneamente em diferentes aspectos da entrada, enriquecendo significativamente sua capacidade de representação (VASWANI, 2017). Essa abordagem é representada pela Eq. 2.2.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{2.2}$$

onde $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$

Nas etapas finais do Decodificador, a informação processada passa por uma série de transformações. Inicialmente, uma rede *feed-forward* refina as representações. Em seguida, uma camada de projeção linear mapeia estas representações para o espaço do vocabulário. A distribuição resultante é normalizada por meio da função *softmax*, o que resulta em probabilidades para cada *token* do vocabulário. O modelo seleciona iterativamente o *token* mais provável, repetindo o processo até gerar um marcador especial de fim de sequência. Esta abordagem permite ao *Transformer* gerar texto de forma autorregressiva, palavra por palavra (VASWANI, 2017).

A arquitetura *Transformer* tem servido de base para muitos modelos de linguagem de última geração, incluindo BERT (DEVLIN, 2018), *Generative Pre-trained Transformer* (GPT) (RADFORD et al., 2018) e suas variantes, como RoBERTa (LIU et al., 2019), XLNet (YANG et al., 2019) e *Text-to-Text Transfer Transformer* (T5) (RAFFEL et al., 2020), devido à sua capacidade de processar eficientemente grandes volumes de dados textuais e de capturar relações complexas entre palavras e frases (QIU et al., 2020). Mais detalhes sobre alguns desses modelos são apresentados na Seção seguinte.

2.5 Modelos de linguagem

Os modelos de linguagem são sistemas de aprendizado de máquina projetados para compreender, gerar ou manipular linguagem natural. Eles são treinados em grandes *corpora* de texto para aprender padrões e estruturas linguísticas (JURAFSKY; MARTIN, 2000). Como mencionado na Seção anterior, há muitos modelos já desenvolvidos. Assim, serão definidos os modelos relevantes para a compreensão do presente estudo.

2.5.1 Modelos do domínio geral

O BERT, introduzido por Devlin (2018), consiste em um *Transformer Encoder* multicamada, que integra a arquitetura *Transformer*. Esse modelo utiliza aprendizado bidirecional para gerar representações contextualizadas de palavras, superando as limitações dos modelos unidirecionais anteriores. Sua arquitetura é composta por várias camadas de autoatenção com múltiplas cabeças (DEVLIN, 2018).

Antes que um texto possa ser processado pelo modelo, ele precisa ser convertido em uma sequência de unidades discretas denominadas *tokens*. Esse processo, chamado *tokenização*, é realizado por um componente denominado *tokenizador*, que segmenta o texto em palavras ou subpalavras e as converte em identificadores numéricos correspondentes ao vocabulário do modelo. Em geral, o BERT e suas variantes utilizam *tokenizadores* baseados em *WordPiece* (DEVLIN et al., 2019), que combinam eficiência de representação e cobertura lexical. Essa etapa define o comprimento efetivo da sequência que o modelo pode processar, nesse caso, limitado a 512 *tokens*.

O BERT utiliza *tokens* especiais, o *token [CLS]* e o *token [SEP]*. Os *tokens [CLS]* são adicionados no início de cada sequência de entrada, sendo sua representação final na última camada usada como agregação de toda a sequência para tarefas de classificação, e o *token [SEP]* é usado para separar pares de sentenças em tarefas que envolvem duas sequências de texto (DEVLIN, 2018). Conforme apresentado por Devlin (2018), cada *token* de entrada x_i é representado por um vetor que combina três tipos de *embeddings*, conforme definido na Eq. 2.3.

$$h_i = E_{w_i} + SE_{w_i} + PE_{w_i} \tag{2.3}$$

onde E_{w_i} é o *embedding* do *token*, SE_{w_i} é o *embedding* do segmento, e PE_{w_i} é o *embedding* posicional (WANG; CUI; ZHANG, 1911).

Adicionalmente, o modelo é pré-treinado em duas tarefas, *Masked Language Modeling* (MLM) e *Next Sentence Prediction* (NSP). No MLM, uma fração de *tokens* de entrada é mascarada aleatoriamente e substituída pelo *token* especial *[MASK]*, cabendo ao modelo predizer os *tokens* mascarados. Matematicamente, a função de perda do MLM é apresentada na Eq. 2.4.

$$\mathcal{L}_{MLM}(x^{corrupt}, \theta) = \frac{1}{k} \sum_{i \in m} -\log p_{M_i}(x_i^* | x^{corrupt}) \tag{2.4}$$

onde $p_{M_i}(x_i^* | x^{corrupt})$ é a probabilidade estimada pelo modelo para o *token* original x_i^* na posição i, $x^{corrupt}$ é a sequência de entrada com *tokens* mascarados, m é o conjunto de posições mascaradas, e k é o número de *tokens* mascarados(DEVLIN, 2018; WANG; CUI; ZHANG, 1911; SOUZA, 2020). Já o NSP treina o modelo para entender a relação entre pares de sentenças, melhorando sua capacidade de lidar com tarefas que envolvem múltiplas sentenças (DEVLIN, 2018). A função de perda para o NSP é uma classificação binária, conforme descrito na Eq. 2.5.

$$\mathcal{L}_{NSP}(x_{corrupt}, \theta) = -(y_{NSP} \log p_N(\hat{y}_{NSP} = 1) + (1 - y_{NSP}) \log(1 - p_N(\hat{y}_{NSP} = 0))) \tag{2.5}$$

onde y_{NSP} é o rótulo verdadeiro (1 se a segunda sentença segue a primeira, 0 caso contrário), e $p_N(\hat{y}_{NSP} = 1)$ é a probabilidade estimada pelo modelo de que a segunda sentença segue a primeira (DEVLIN, 2018; SOUZA, 2020). A perda total de pré-treinamento é a soma das perdas de MLM e de NSP. Com isso, a combinação dessas duas tarefas de pré-treinamento permite ao BERT capturar eficientemente dependências bidirecionais e de longo alcance em sequências de texto, tornando-o altamente eficaz em várias tarefas de PLN (DEVLIN, 2018).

Embora o BERT tenha vantagens em várias tarefas de PLN, é importante notar que o modelo original foi treinado principalmente em textos em inglês, o que limita sua aplicabilidade direta em outros idiomas (DEVLIN, 2018). A natureza não multilíngue do BERT apresenta desafios em outros contextos linguísticos. Assim como a simples tradução de textos para treinamento ou aplicação do modelo pode introduzir vieses e perder nuances específicas do idioma-alvo (SOUZA; NOGUEIRA; LOTUFO, 2020; SOUZA, 2020).

Reconhecendo essas limitações e a necessidade de um modelo específico para o português brasileiro, Souza, Nogueira e Lotufo (2020) propuseram o BERTimbau. Esse modelo consiste na adaptação do BERT para o português brasileiro, mantendo a arquitetura original do BERT, mas pré-treinado em um *corpus* denominado *Brazilian Web as Corpus* (brWaC), que contém 2,68 bilhões de *tokens* provenientes de 3,53 milhões de documentos (SOUZA; NOGUEIRA; LOTUFO, 2020). Este modelo utiliza um vocabulário próprio de 30.000 unidades de subpalavras, gerado pelo algoritmo *Byte-Pair Encoding* (BPE) e convertido para o formato *WordPiece*. O pré-treinamento foi realizado com as mesmas tarefas do BERT (MLM e NSP), utilizando exemplos gerados com um fator de duplicação de 10. O BERTimbau demonstrou desempenho superior em várias tarefas de PLN em português, estabelecendo-se como um modelo de referência para o idioma e servindo de base para novas adaptações em contextos específicos, como o jurídico, a ser discutido a seguir.

2.5.2 Modelos do domínio jurídico brasileiro

No contexto jurídico, foram desenvolvidos modelos especializados para atender às particularidades da linguagem e às tarefas específicas desse domínio. Dentre estes modelos como BERTikal (POLO et al., 2021), JurisBERT (VIEGAS; COSTA; ISHII, 2023), LegalBERT-PT (SILVEIRA et al., 2023), BumbaBERT (CARMO, 2024) e RoBERTaLexPT (GARCIA et al., 2024).

O BERTikal é um modelo de linguagem especializado para o domínio jurídico brasileiro, baseado no BERTimbau e proposto por Polo et al. (2021). Este modelo foi treinado utilizando um *corpus* de aproximadamente 6 milhões de documentos jurídicos, incluindo publicações, decisões judiciais e movimentações processuais de vários tribunais brasileiros entre os anos de 2019 e 2020. O BERTikal mantém a arquitetura do BERT, mas com um vocabulário adaptado para incluir termos jurídicos específicos. Em avaliações comparativas, o BERTikal demonstrou desempenho superior ao BERTimbau em tarefas como a classificação de documentos jurídicos e o reconhecimento de entidades nomeadas no contexto legal (POLO et al., 2021).

O JurisBERT é baseado na arquitetura BERT treinada do zero (*from scratch*) com base em *corpus* de textos jurídicos brasileiros, com aproximadamente 1,5 milhão de sentenças extraídas de leis, decretos federais, súmulas, decisões judiciais, acórdãos e tratados de diversos ramos do direito brasileiro (VIEGAS; COSTA; ISHII, 2023). O vocabulário específico do JurisBERT foi gerado com 30.000 unidades de subpalavras usando o tokenizador WordPiece, mantendo acentuação e distinção entre maiúsculas e minúsculas (*case-sensitive*) devido às particularidades do português jurídico (VIEGAS; COSTA; ISHII, 2023). O modelo foi pré-treinado exclusivamente com o mascaramento MLM (descartando *NSP* por evidências de baixa efetividade), com 20 épocas, *batch size* de 128 e sequências de comprimento máximo de 384 *tokens*, totalizando aproximadamente 7 dias de treinamento em duas GPUs NVIDIA GeForce RTX 3080 (VIEGAS; COSTA; ISHII, 2023). Avaliações em tarefa de Similaridade Textual Semântica (STS) usando ementas de acórdãos demonstraram que o JurisBERT obteve um ganho de 1,6% no F1-Score em comparação ao BERTimbau (VIEGAS; COSTA; ISHII, 2023).

O LegalBERT-PT foi desenvolvido em duas variantes, LegalBERT-PT SC (*from scratch*) e LegalBERT-PT FP (*further pretraining*), propostas por Silveira et al. (2023). O LegalBERT-PT SC (*from scratch*) foi treinado exclusivamente em *corpus* jurídico, com vocabulário especializado de 36.345 subpalavras, incluindo 5.977 identificadores de legislação brasileira. Enquanto o LegalBERT-PT FP (*further pretraining*) foi inicializado com pesos do BERTimbau-Base e submetido a um pré-treinamento adicional de 2,4 milhões de passos em *corpus* jurídico (SILVEIRA et al., 2023). O *corpus* de pré-treinamento foi obtido do sistema Codex do CNJ, contendo 1,5 milhão de documentos jurídicos (*e.g.,* petições iniciais, decisões e sentenças) de 10 tribunais brasileiros, totalizando aproximadamente 12

milhões de sentenças após limpeza e pré-processamento. As avaliações demonstraram que o LegalBERT-PT FP apresentou menor perplexidade (3.700) do que o LegalBERT-PT SC (3.822) e o BERTimbau-Base, indicando maior compreensão da linguagem jurídica (SILVEIRA et al., 2023).

Por sua vez, o BumbaBERT representa uma evolução ainda mais especializada no processamento de linguagem jurídica brasileira desenvolvido por (CARMO, 2024). Esse modelo é uma adaptação do BERT, com versões de arquitetura *base* (12 camadas, 768 unidades de *hidden state*, 12 cabeças de atenção) e *small* (uma versão com arquitetura reduzida, com 6 camadas, 512 unidades de *hidden state*, 8 cabeças de atenção). O BumbaBERT foi treinado em um *corpus* extenso e diversificado de aproximadamente 5,5 milhões de documentos jurídicos em português, abrangendo legislações, jurisprudências e petições, provenientes de diversas esferas do sistema de justiça brasileiro, como o TJMA, além de tribunais superiores como o Supremo Tribunal Federal (STF), Superior Tribunal de Justiça (STJ), Tribunal Superior do Trabalho (TST), Superior Tribunal Militar (STM) e Tribunal Superior Eleitoral (TSE). O modelo inclui uma versão BumbaBERT-*base* FT , que utiliza o pré-treinamento do BERTimbau, e uma versão BumbaBERT-*small* SC, treinada exclusivamente com dados jurídicos sem um modelo pré-treinado como ponto de partida e com uma arquitetura mais compacta, para reduzir o custo computacional e ainda assim manter alto desempenho em tarefas de PLN, como classificação de petições e identificação de precedentes jurídicos (CARMO, 2024).

Mais recentemente, Garcia et al. (2024) introduziram o RoBERTaLexPT, um modelo baseado na arquitetura RoBERTa pré-treinado em um *corpus* combinado de textos jurídicos (LegalPT, 125 GiB) e textos genéricos (CrawlPT, 163 GiB). O *corpus* LegalPT foi construído agregando múltiplas fontes públicas brasileiras (MultiLegalPile-PT, Ulysses-Tesemõ, ParlamentoPT, Iudicium Textum, Acórdãos TCU, DataSTF), resultando em 11,9 milhões de documentos únicos, com taxa de duplicação de 50,63%, que indica o percentual de textos repetidos no *corpus* antes do processo de deduplicação (GARCIA et al., 2024). O modelo foi pré-treinado com a arquitetura RoBERTabase (12 camadas, 768 unidades ocultas, 12 cabeças de atenção, 110 milhões de parâmetros) por 62.500 passos, com *batch size* de 2.048 sequências de 512 *tokens*, expondo o modelo a aproximadamente 65 bilhões de *tokens* durante o treinamento. Experimentos demonstraram que a combinação de *corpus* jurídico especializado com *corpus* genérico produziu melhor desempenho do que o pré-treinamento exclusivo em qualquer um dos *corpus* individualmente, com RoBERTaLexPT obtendo um F1-Score médio de 85,41% no *benchmark* PortuLex (composto por tarefas de NER e classificação textual).

A comparação no que diz respeito à arquitetura, ao treinamento, aos parâmetros e ao limite de *tokens* de entrada dos modelos supracitados está sintetizada na Tabela 4.

Essa trajetória de aprimoramento de modelos baseados em codificadores estabeleceu

Tabela 4 – Síntese comparativa dos modelos de linguagem do domínio jurídico brasileiro

Modelo	**Arquitetura**	**Corpus**	**Params.**	**Observações**
BERTikal	BERTimbau-Base (12L, 768H, 12A)	*Further pretraining* com ≈6 milhões de documentos jurídicos (2019–2020): acórdãos, decisões, peças processuais e legislação.	110M	*Checkpoint* inicial: BERTimbau-Base; 1 época MLM (*mask* 0,15); *batch size*=4; treino ≈ 1 semana em GPU Tesla T4
JurisBERT	BERT-Base (12L, 768H, 12A)	*From scratch* com ≈1,5 milhão de sentenças jurídicas (leis, decretos, súmulas, decisões, acórdãos e tratados).	110M	Vocabulário especializado (30k subpalavras WordPiece, *uncased*); 20 épocas de MLM; batch size = 128; ≈7 dias de treino em 2 x NVIDIA GeForce RTX 3080 (12 GB).
LegalBERT-PT (SC)	BERT-Base (12L, 768H, 12A)	*From scratch* com ≈1,5 milhão de documentos jurídicos (Codex CNJ): petições, decisões e sentenças de 10 tribunais.	110M	Vocabulário especializado de 36.345 subpalavras (5.977 identificadores de legislação); perplexidade = 3.822.
LegalBERT-PT (FP)	BERTimbau-Base (12L, 768H, 12A)	*Further pretraining* com mesmo conjunto da versão SC; 2,4 milhões de passos adicionais.	110M	Checkpoint inicial: BERTimbau-Base; vocabulário de 30k subpalavras; perplexidade = 3.700 (melhor que SC).
BumbaBERT-*small* (SC)	BERT (6L, 512H, 8A)	*From scratch* com ≈5,5 milhões de documentos jurídicos (legislação, jurisprudência, petições).	**42M**	Sem *checkpoint* inicial; taxa de aprendizagem = 1e-4, dropout = 0,1; *batch size = 8*; etapas: 1M
BumbaBERT-*base* (FT)	BERTimbau-Base (12L, 768H, 12A)	*Further pretraining* com os mesmos documentos da versão SC.	110M	taxa de aprendizagem = 1e-4, dropout = 0,1; *batch size = 8*; etapas: 400k
RoBERTaLexPT	RoBERTa-Base (12L, 768H, 12A)	Corpus combinado: LegalPT (jurídico) + CrawlPT (geral) = 11,9 milhões de documentos.	110M	62.500 passos; *batch size* = 2.048 x 512 *tokens* (~1 M tokens/etapa); ambiente: cluster DGX-A100 com 2 x Nvidia A100 (GB); F1 médio (PortuLex) = 85,41%.

Fonte: Elaborada pela autora com base em Polo et al. (2021), Viegas, Costa e Ishii (2023), Silveira et al. (2023), Carmo (2024), Garcia et al. (2024). Legenda: L = camadas, H = *hidden size*, A = cabeças de atenção, M = milhões, SC = *from scratch*, FT = *further pretraining*, FP = *further pretraining*.

a base para o avanço seguinte da área, representado pelos modelos generativos apresentados na Subseção a seguir.

2.5.3 Modelos generativos de linguagem

Modelos generativos de linguagem representam uma classe de modelos capazes de gerar texto de forma autorregressiva, produzindo sequências coerentes palavra por palavra. Diferentemente de modelos como BERT, que são primariamente *encoders* bidirecionais treinados para tarefas de compreensão, modelos generativos são tipicamente *decoders* unidirecionais ou *encoder-decoders* completos treinados para geração de texto (RADFORD et al., 2018; BROWN et al., 2020).

A família GPT, desenvolvida pela OpenAI, exemplifica essa abordagem. O GPT-2 e o GPT-3 utilizam arquiteturas *decoder-only* com bilhões de parâmetros, treinadas em vastos *corpora* textuais (RADFORD et al., 2019; BROWN et al., 2020). Esses modelos demonstraram capacidades emergentes de *few-shot learning*, em que conseguem realizar tarefas com poucos exemplos no *prompt*, sem a necessidade de *fine-tuning* explícito (BROWN et al., 2020). No contexto de código aberto, modelos como *Large Language*

Model Meta AI (LLaMA) (TOUVRON et al., 2023) e suas variantes (LLaMA-2, LLaMA-3) representam alternativas acessíveis aos modelos proprietários. O LLaMA foi treinado exclusivamente com dados públicos, com versões de 7B a 65B de parâmetros, demonstrando desempenho competitivo em relação a modelos proprietários maiores (TOUVRON et al., 2023).

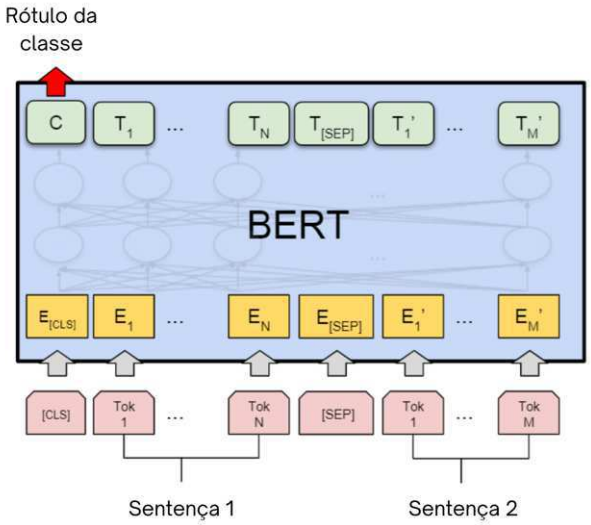
Considerando LLMs nacionais, modelos como Sabiá (PIRES et al., 2023) e Maritalk representam adaptações de LLMs para o português, embora sua aplicação específica em documentos jurídicos longos ainda demande investigação adicional. Uma limitação significativa desses modelos generativos é a questão da transparência dos dados de treinamento e da privacidade, frequentemente referidos como "caixas-pretas" (BENDER et al., 2021; BOMMASANI et al., 2021), além de custos computacionais elevados e de latência (CARLINI et al., 2021). Diante dessas limitações estruturais e operacionais, uma das estratégias eficazes para tornar modelos pré-treinados e preservar suas representações linguísticas em diferentes domínios é o processo de *fine-tuning*, descrito na Subseção a seguir.

2.5.4 *Fine-tuning*

No contexto de modelos de linguagem como o BERT, o *fine-tuning* envolve o treinamento adicional do modelo em um conjunto de dados específico da tarefa, ajustando seus pesos para otimizar o desempenho nessa tarefa em particular (HOWARD; RUDER, 2018). De forma mais técnica, esse processo é simplificado pela flexibilidade do mecanismo de autoatenção dos *Transformers*, que permite modelar diversas tarefas *downstream* ajustando apenas as entradas e as saídas apropriadas (DEVLIN, 2018). O BERT unifica a codificação de pares de texto e a atenção cruzada bidirecional em um único estágio, diferentemente de abordagens anteriores (PARIKH et al., 2016; SEO et al., 2017).

Durante o *fine-tuning*, as entradas e as saídas específicas da tarefa são inseridas no BERT, e todos os parâmetros são ajustados de ponta a ponta. As sentenças A e B do pré-treinamento são adaptadas para diversos tipos de tarefas, como paráfrase, inferência textual, sistemas de pergunta e resposta e classificação de texto (DEVLIN, 2018). Para fins de exemplo, essa configuração é apresentada na Figura 7.

Figura 7 – Ilustração do *fine-tuning* na tarefa de classificação de texto



Fonte: Adaptada de Devlin (2018).

Uma característica importante do processo de *fine-tuning* é a adição de uma camada de classificação específica para a tarefa ao topo do modelo BERT pré-treinado. Esta camada é tipicamente inicializada aleatoriamente e treinada junto com o ajuste fino dos parâmetros do BERT (DEVLIN, 2018). Para tarefas de classificação, a representação *[CLS]* da última camada é alimentada nesta camada, enquanto, para tarefas de nível de *token*, as representações de todos os *tokens* são utilizadas. O *fine-tuning* é computacionalmente eficiente, podendo ser realizado em cerca de uma hora em uma *Cloud Tensor Processing Unit* (TPU) ou em algumas horas em uma *Graphics Processing Unit* (GPU), a partir do modelo pré-treinado (DEVLIN, 2018). Essa eficiência torna os modelos de linguagem particularmente adequados para adaptação rápida a novas tarefas ou domínios específicos, permitindo uma ampla gama de aplicações com ajustes mínimos (QIU et al., 2020). Ainda assim, essa técnica apresenta limitações relevantes, como a suscetibilidade ao sobreajuste (*overfitting*) em domínios restritos, a dependência de anotações de qualidade e a necessidade de calibração precisa dos hiperparâmetros. A isso somam-se os custos computacionais associados ao treinamento repetido de grandes modelos, que podem restringir sua aplicação em contextos institucionais com infraestrutura limitada (DING et al., 2023).

Contudo, a eficácia do *fine-tuning* está diretamente relacionada à configuração adequada de hiperparâmetros, que controlam o comportamento do treinamento e a capacidade de generalização do modelo. Na Tabela 5 são sintetizados os principais hiperparâmetros utilizados no *fine-tuning* de modelos baseados em BERT para tarefas de classificação, acompanhados de suas funções.

Tabela 5 – Principais hiperparâmetros do *fine-tuning* de modelos BERT

Hiperparâmetro	**Descrição**
Learning rate	Taxa de aprendizado que controla a magnitude dos ajustes de pesos durante o treinamento. Valores altos aceleram o treinamento mas podem causar instabilidade; valores baixos aumentam a estabilidade mas tornam o treinamento mais lento (DEVLIN, 2018).
Batch size	Número de exemplos processados simultaneamente em cada passo de treinamento. Valores maiores estabilizam gradientes mas requerem mais memória GPU; valores menores aumentam a variabilidade do treinamento (SMITH, 2018).
Epochs	Número de passagens completas pelo conjunto de treinamento. Poucas épocas podem resultar em subajuste (*underfitting*); muitas épocas podem levar a sobreajuste (*overfitting*) (DEVLIN, 2018).
Dropout	Taxa de desativação aleatória de neurônios durante o treinamento para prevenir sobreajuste, forçando o modelo a aprender representações mais robustas (SRIVASTAVA et al., 2014).
Optimizer	Algoritmo de otimização utilizado para ajustar os pesos do modelo durante o treinamento. Adam combina as vantagens de AdaGrad e RMSProp, adaptando taxas de aprendizado individualmente para cada parâmetro (KINGMA; BA, 2014).
Warmup steps	Número inicial de passos com taxa de aprendizado crescente gradualmente antes de aplicar *decay*, estabilizando o início do treinamento (DEVLIN, 2018).
Scheduler	Estratégia de ajuste da taxa de aprendizado ao longo do treinamento. O decaimento linear reduz progressivamente o *learning rate* após o *warmup*, refinando o ajuste (DEVLIN, 2018).
Max sequence length	Comprimento máximo das sequências de entrada em *tokens*. Sequências maiores que esse limite são truncadas; sequências menores são preenchidas com *padding* (DEVLIN, 2018).
Random seed	Semente para inicialização de geradores de números pseudoaleatórios, garantindo reprodutibilidade dos experimentos ao fixar os estados aleatórios de divisões de dados, inicialização de pesos e *dropout* (DEVLIN, 2018).

Fonte: Elaborada pela autora com base em (DEVLIN, 2018; SMITH, 2018; SRIVASTAVA et al., 2014; LOSHCHILOV; HUTTER, 2019).

A definição dos valores desses hiperparâmetros pode ocorrer por meio de diferentes estratégias. Em contextos acadêmicos ou experimentais, é comum o uso de *grid search*, *random search* ou otimizadores bayesianos para explorar sistematicamente combinações possíveis (BERGSTRA; BENGIO, 2012). Em ambientes com menos recursos ou quando se utilizam modelos amplamente documentados, como o BERT base, empregam-se valores recomendados na literatura, como *learning rate* entre $2e^{-5}$ e $5e^{-5}$, *batch size* entre 16 e 32 e até 3-10 épocas de treinamento, como ponto de partida (DEVLIN, 2018). Esses valores servem como configurações padrão iniciais, que podem ser refinadas empiricamente conforme o tamanho do *corpus*, a complexidade da tarefa e as limitações computacionais disponíveis.

2.6 Considerações sobre o Capítulo

No presente Capítulo foram estabelecidas bases teóricas que sustentam o desenvolvimento da presente dissertação, articulando fundamentos jurídicos e computacionais sob uma perspectiva integrada. No âmbito jurídico, discutiu-se a estrutura e o fluxo processual das petições iniciais, bem como o papel dos precedentes judiciais, com ênfase nas IRDRs e na sua operacionalização no TJMA. Essa contextualização foi importante para compreender a natureza e a complexidade dos documentos analisados, bem como os desafios relacionados à sua padronização, categorização e automação no domínio.

No campo computacional, foram perpassados por conceitos referentes a PLN, a evolução das arquiteturas de aprendizado de máquina e o advento dos LLMs baseados em *Transformers*, culminando na discussão de modelos aplicados ao domínio jurídico brasileiro, desde o modelo pioneiro BERTimbau até os modelos especializados como BumbaBERT e RoBERTaLexPT, demonstrando dessa forma, os avanços na representação semântica e na adaptação linguística ao domínio. Do mesmo modo, foi apresentado o modelo metodológico DST, que orienta as etapas deste estudo, ao mesmo tempo em que descreve o papel do *fine-tuning* como estratégia para a adaptação desses modelos em tarefas de classificação de documentos.

Assim, o conjunto de conceitos aqui discutidos forneceu base necessária para a compreensão das decisões metodológicas e experimentais, uma vez visto que os modelos ainda mantêm a restrição estrutural herdada da arquitetura BERT, somada às características de extensão e a complexidade dos documentos jurídicos identificadas nos fundamentos das petições iniciais, o que motivou a investigação de estratégias alternativas capazes de processar os documento completos sem perda informacional, as quais serão apresentadas no próximo Capítulo.

3 Estratégias para classificação de documentos longos

Conforme abordado no Capítulo 2, modelos de linguagem baseados em *Transformers* apresentam uma restrição fundamental quanto ao comprimento de entrada, o que representa um desafio significativo para o processamento de documentos longos. Este Capítulo aprofunda as estratégias desenvolvidas pela comunidade científica para superar essa limitação, explorando tanto técnicas de pré-processamento quanto arquiteturas especializadas. Visando fornecer uma visão geral da área, o Capítulo foi organizado em três eixos: a Seção 3.1 apresenta três técnicas de sumarização automática como abordagens de redução de conteúdo: a extrativa, a abstrativa e a híbrida. A Seção 3.2 discute arquiteturas especializadas que processam diretamente documentos longos, sem necessidade de sumarização prévia. Finalizando com a Seção 3.3, que descreve as métricas de avaliação e os testes estatísticos que fundamentam a comparação empírica entre essas abordagens.

3.1 Sumarização de documentos

A sumarização de texto também é uma tarefa no PLN, voltada à condensação de um documento ou de um conjunto de documentos em uma versão mais curta, mantendo suas informações mais relevantes e coerentes (LIU; LAPATA, 2019; TSIRMPAS; GKIONIS; MADEMLIS, 2023). Diferentes métodos têm sido implementados, incluindo abordagens extrativas, que se baseiam na seleção de sentenças relevantes; métodos abstrativos, que visam reescrever a informação em nova forma textual; e combinações híbridas, que buscam equilibrar precisão e coesão de ambos os métodos (EL-KASSAS et al., 2021; LIU; LAPATA, 2019; KIRMANI et al., 2019). 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* (PRINCIPE; CHIARINI; VIVIANI, 2025).

3.1.1 Sumarização extrativa

A sumarização extrativa consiste em selecionar diretamente frases ou sentenças do texto original e concatená-las para formar o resumo final, com o objetivo de identificar as partes mais importantes do documento e apresentá-las em sua forma original, sem modificações linguísticas significativas (NALLAPATI et al., 2017). Métodos extrativos geralmente operam atribuindo uma pontuação de importância a cada sentença ou termo, com base em critérios como a frequência de palavras, a posição no texto ou as relações com

outros elementos do documento (MIHALCEA; TARAU, 2004). Esse tipo de sumarização é mais simples de implementar e pode produzir resumos diretamente rastreáveis ao texto-fonte, o que é uma vantagem em contextos em que a fidelidade ao original deve ser levada em conta (PARK; VYAS; SHAH, 2022).

Seus algoritmos podem variar a depender do tipo de técnica, incluindo: (i) os estatísticos, que utilizam métricas estatísticas como frequência de palavras ou frases; (ii) os baseados em gráficos, que modelam o documento em um gráfico de frases e, em seguida, utilizam conceitos da teoria dos gráficos como a centralidade e as medidas de detecção de comunidades; e (iii) os baseados em semântica, que modelam frases e seus termos em uma matriz de coocorrência, que é então analisada usando semântica distributiva (GIARELIS; MASTROKOSTAS; KARACAPILIDIS, 2023). Dentre essas abordagens há algumas mais proeminentes, a saber: *Term Frequency-Inverse Document Frequency* (TF-IDF), TextRank, LexRank e o Sentence-BERT (SBERT).

Algoritmos estatísticos, como o TF-IDF, usam uma técnica baseada no princípio de que termos frequentes em um documento, mas raros no *corpus*, carregam maior carga informativa; ou seja, o peso de um termo que ocorre em um documento é diretamente proporcional à sua frequência (RAMOS et al., 2003). As sentenças são ranqueadas pela soma dos pesos de seus termos (Eq. 3.1).

$$score_{TFIDF}(s) = \sum_{t \in s} \text{TFIDF}(t, d, D) \tag{3.1}$$

onde s representa uma sentença pertencente ao documento d, t é um termo contido em s, e D é o conjunto de documentos do corpus. O valor $\text{TFIDF}(t, d, D)$ corresponde ao peso do termo t, proporcional à sua frequência no documento e inversamente proporcional à sua frequência no corpus.

Métodos como TextRank (MIHALCEA; TARAU, 2004) e LexRank (ERKAN; RADEV, 2004) operam como algoritmos baseados em grafos, em que as sentenças são representadas como nós e as conexões entre elas como arestas, com base na similaridade. A importância de uma sentença é determinada por sua centralidade no grafo. O TextRank, por exemplo, identifica sentenças-chave ao calcular o "ranking" de cada sentença no texto, modelando o documento como um grafo em que as sentenças são representadas por nós e as similaridades léxicas por arestas. Aplicando *PageRank*, as sentenças mais centrais são selecionadas, conforme descrito na Eq. 3.2.

$$\text{Score}_{\text{T}}\text{extRank}(V_i) = (1-d) + d \sum_{V_j \in In(V_i)} \frac{\text{Sim}(V_i, V_j)}{\sum_{V_k \in Out(V_j)} \text{Sim}(V_j, V_k)} \text{Score}(V_j) \tag{3.2}$$

onde V_i representa o vértice (ou sentença) para o qual se calcula o valor de importância. $In(V_i)$ e $Out(V_j)$ correspondem, respectivamente, aos conjuntos de vértices

que apontam para V_i e para os quais V_j aponta. O d é o fator de amortecimento (*damping*), tipicamente fixado em 0,85, e o $\text{Sim}(V_i, V_j)$ é o peso da aresta entre V_i e V_j.

Em relação aos modelos baseados em semântica, há o SBERT, que "*fine-tuna*" o BERT para gerar *embeddings* de sentenças semanticamente significativas por meio de uma arquitetura siamesa (REIMERS; GUREVYCH, 2019). Diferentemente do BERT padrão, que requer processamento de pares de sentenças, cuja complexidade é $O(n^2)$ para comparar n sentenças, o SBERT gera representações independentes comparáveis por meio de similaridade de cosseno ($O(n)$).

3.1.2 Sumarização abstrativa

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 (RUSH; CHOPRA; PARIKH, 2015). Modelos abstrativos utilizam técnicas mais avançadas de PLN, frequentemente baseadas em arquiteturas *Encoder-Decoder* com mecanismos de atenção, como os *Transformers* (VASWANI, 2017; LEWIS et al., 2020). Embora a sumarização abstrativa possa gerar resumos de alta qualidade e mais naturais, ela é computacionalmente mais intensiva e apresenta desafios maiores relacionados à geração de informações incorretas, fenômeno chamado de alucinação de fatos, ou à perda de fidelidade ao texto original, um risco considerável em aplicações críticas, como diagnóstico médico ou direito processual (FABBRI et al., 2019). Atualmente, a pesquisa indica que, no contexto de lidar com as limitações dos *Transformers* em documentos longos, as técnicas abstrativas ainda não foram plenamente exploradas para esse propósito (PRINCIPE; CHIARINI; VIVIANI, 2025).

Modelos como Pegasus e T5 são exemplos de arquiteturas baseadas em *Transformer* projetadas para sumarização abstrativa. Eles são capazes de gerar resumos originais, mas sua aplicação para superar a limitação de comprimento de entrada em tarefas de classificação de documentos longos ainda é uma área em evolução na pesquisa (PRINCIPE; CHIARINI; VIVIANI, 2025). A integração desses modelos na classificação de documentos longos geralmente envolveria a geração de um resumo abstrativo como uma etapa inicial, e este resumo seria então alimentado em um modelo de classificação.

3.1.3 Sumarização híbrida

A sumarização híbrida busca combinar as vantagens das abordagens extrativa e abstrativa, selecionando primeiro as sentenças mais relevantes e, em seguida, reescrevendo-as de forma mais natural e coerente (EL-KASSAS et al., 2021; KIRMANI et al., 2019). Essa estratégia reduz o risco de alucinações típicas dos métodos puramente abstrativos, mantendo a cobertura de conteúdo. Recentemente, LLMs têm sido explorados para esse tipo

de abordagem, integrando capacidades de compressão semântica e de reescrita contextual (LUCENA et al., 2025; JIANG; YANG; RAO, 2024).

Em síntese, essas abordagens de sumarização, sejam extrativas, abstrativas ou híbridas, contribuem para a redução da complexidade textual e para a preservação das principais informações de documentos longos. No que diz respeito à integração híbrida, pode constituir parte da adaptação de modelos, oferecendo mecanismos computacionalmente eficientes para redução de dimensionalidade textual sem necessidade de modificação arquitetural dos modelos base. Na próxima Seção, essa discussão é ampliada ao apresentar uma taxonomia de métodos voltados ao processamento de documentos longos, na qual a sumarização se insere como uma das principais estratégias.

3.2 Processamento de documentos longos

Conforme estabelecido no Capítulo 1, a limitação de 512 *tokens* dos modelos baseados em BERT inviabiliza o processamento direto de documentos jurídicos extensos, que frequentemente ultrapassam milhares de *tokens*. Além dessa restrição arquitetural, o processamento de textos longos enfrenta desafios linguísticos e estruturais específicos que amplificam a complexidade da tarefa.

No contexto jurídico, um desafio adicional decorre da própria digitalização do acervo processual. A conversão de documentos físicos para formato digital pelos tribunais resulta na maioria dos casos em arquivos *Portable Document Format* (PDF) compostos por imagens digitalizadas obtidas por meio de *scanners*, sem camada textual nativa. A extração do conteúdo textual dessas imagens requer a aplicação de técnicas de Reconhecimento Óptico de Caracteres - do inglês *Optical Character Recognition*(OCR), processo que frequentemente introduz ruídos textuais que comprometem a qualidade dos dados (BAVISKAR et al., 2021; SUBRAMANI et al., 2020). Erros típicos de OCR incluem substituição de caracteres visualmente semelhantes, perda ou distorção de acentuação, segmentação inadequada de palavras, inserção de quebras de linha espúrias e corrupção de termos técnicos especializados (NGUYEN et al., 2021).

Esses problemas são particularmente críticos quando a digitalização envolve documentos físicos degradados, páginas com baixo contraste, variações na qualidade de impressão original ou equipamentos de digitalização com resolução inadequada, podendo resultar em sentenças fragmentadas ou semanticamente corrompidas (PHILIPS; TABRIZI, 2020; IGOREVNA et al., 2022). Uma vez que todas as estratégias subsequentes, sejam técnicas de sumarização, geração de *embeddings*, aplicação de janelas deslizantes ou segmentação hierárquica que operam sobre representações textuais, a presença de ruído na etapa de extração propaga-se e amplifica-se ao longo do *pipeline* de processamento, impactando diretamente a capacidade dos modelos de capturar dependências semânticas e estrutu-

rais (NGUYEN et al., 2021; BAVISKAR et al., 2021). Dessa forma, o pré-processamento cuidadoso e a normalização textual constituem etapas importantes e indispensáveis que antecedem qualquer estratégia de processamento de documentos longos no domínio jurídico.

Além desse desafio, a polissemia e a homonímia exigem que modelos diferenciem apropriadamente o significado de palavras, dependendo do contexto local (sentença, parágrafo) e global (documento). A linguagem figurada pode manifestar-se por meio de alegorias sustentadas ao longo do texto, como sarcasmos, ironia e metáforas (KARAMOUZAS; MADEMLIS; PITAS, 2022). A natureza não estruturada de textos longos implica que informações relevantes podem estar distribuídas por numerosos Capítulos ou seções, com contexto e convenções textuais em constante evolução (TSIRMPAS et al., 2024).

Embora não exista consenso formal na literatura sobre o que constitui um "documento longo", Tsirmpas et al. (2024) propõem uma definição operacional adotada em que um documento longo possui média de pelo menos 2.000 palavras, podendo alcançar 170.000 palavras ou mais, e não é semanticamente segmentado. Isso significa que o texto não é composto por partes independentes, como no caso de e-mails e postagens de fóruns de discussão, e sim por um fluxo contínuo de ideias. Assim, documentos jurídicos como petições iniciais, sentenças e acórdãos enquadram-se tipicamente nessa categoria, apresentando desafios específicos para processamento automatizado (KALAMKAR et al., 2022; FIELDS; CHOVANEC; MADIRAJU, 2024).

Para contornar os desafios identificados, arquiteturas especializadas em documentos longos frequentemente combinam estratégias de Janela Deslizante (*Slidding Window*) utilizando o princípio de capturar contexto local por meio do processamento de subconjuntos contíguos de *tokens*, baseando-se na premissa de que o significado semântico de um *token* deriva primariamente de seus vizinhos imediatos; Truncamento, dado pela minimização do tamanho do texto ou *embedding* de entrada para conformidade com limites dos modelos (DEVLIN, 2018); Atenção Esparsa (*Sparse Attention*), pela modificação da arquitetura *Transformer* subjacente para permitir entradas maiores por meio de padrões de atenção seletivos; e a criação iterativa de *embeddings* para produzir representação de documento de tamanho fixo, comum em arquiteturas hierárquicas (TSIRMPAS et al., 2024; FIELDS; CHOVANEC; MADIRAJU, 2024).

Principe, Chiarini e Viviani (2025) propõem taxonomia estruturada das estratégias para processamento de documentos longos, fundamentada em três paradigmas arquiteturais distintos que respondem de formas complementares aos desafios identificados. Esta taxonomia, corroborada e expandida por Tsirmpas et al. (2024) e Fields, Chovanec e Madiraju (2024), organiza as soluções segundo seus princípios operacionais fundamentais.

3.2.1 *Efficient Transformers*

Efficient Transformers mantém a arquitetura *Transformer* essencialmente inalterada, concentrando as modificações no mecanismo de autoatenção para reduzir sua complexidade computacional de $O(n^2)$ para aproximadamente $O(n \log n)$ ou $O(n)$ (ELOUARGUI et al., 2023). O pressuposto teórico é que a atenção completa entre todos os pares de *tokens* contém redundância significativa e que padrões de atenção mais esparsos podem aproximar suficientemente bem a atenção completa para a maioria das tarefas de NLP. Esta subdivide-se em três subcategorias, a saber: i) Atenção esparsa baseada em localização; ii) Atenção esparsa baseada em conteúdo; e iii) Aproximação de Low-Rank.

A atenção esparsa baseada em localização parte do princípio de que a semântica de um *token* deriva primariamente de seu contexto local imediato, combinado com acesso seletivo à informação global por meio de *tokens* especiais. Modelos como Longformer (BELTAGY; PETERS; COHAN, 2020) e BigBird (ZAHEER et al., 2020) implementam este conceito por meio da composição de três padrões de atenção complementares, a *atenção local*, a *global* e a *atenção dilatada/aleatória*. A *atenção local* utiliza janelas deslizantes para capturar o contexto imediato. Enquanto a *atenção global* é baseada em *tokens* especiais que interagem com toda a sequência, e por fim, a *atenção dilatada ou aleatória* introduz conexões esparsas para ampliar o campo receptivo sem aumento proporcional do custo computacional. Essa estratégia permite processar sequências de até 4.096 *tokens* mantendo tratabilidade computacional (PRINCIPE; CHIARINI; VIVIANI, 2025).

Já a atenção esparsa baseada em conteúdo diverge da abordagem anterior ao determinar padrões de atenção dinamicamente durante a inferência, com base na similaridade semântica entre *tokens* em vez de proximidade posicional fixa. Por exemplo, o Reformer agrupa *tokens* semanticamente semelhantes em *buckets* por meio de *Locality-Sensitive Hashing* (LSH) (KITAEV; KAISER; LEVSKAYA, 2020), permitindo que cada *token* atenda preferencialmente a outros *tokens* do mesmo agrupamento. Essa estratégia sacrifica determinismo posicional em favor de flexibilidade semântica, sendo particularmente adequada quando dependências importantes não seguem padrões posicionais regulares (PRINCIPE; CHIARINI; VIVIANI, 2025).

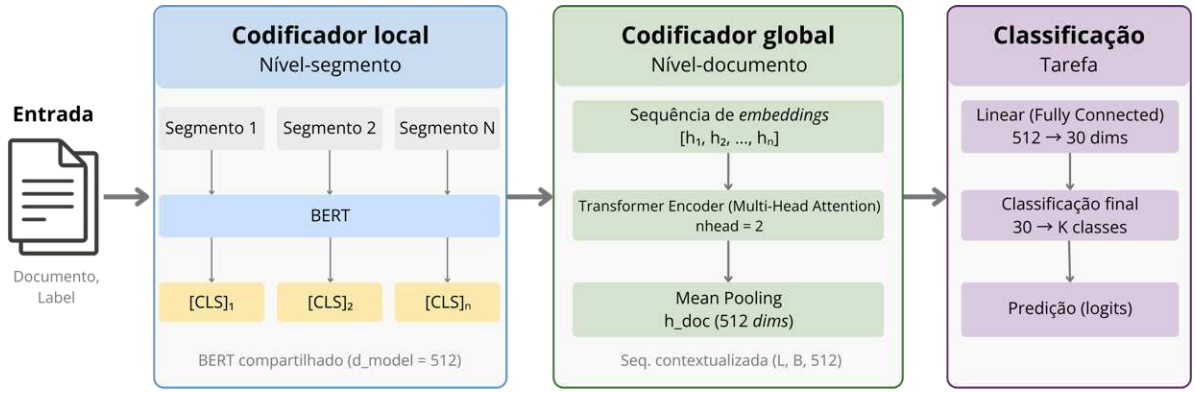
Por fim, a aproximação de *Low-Rank* baseia-se no princípio de decomposição matricial, em que a matriz de atenção completa é aproximada por um produto de matrizes de menor dimensionalidade (PRINCIPE; CHIARINI; VIVIANI, 2025; FIELDS; CHOVANEC; MADIRAJU, 2024). O Performer implementa este conceito por meio do algoritmo *Fast Attention por via positive Orthogonal Random features* (FAVOR+), que estima a atenção *softmax* completa com complexidade linear por meio de mapeamento em um espaço de características aleatórias ortogonais (CHOROMANSKI et al., 2021).

3.2.2 Modelos de decomposição-recomposição

Esses modelos preservam a autoatenção completa, mas dividem o documento em segmentos tratáveis, processando-os separadamente e agregando suas representações, em vez de modificar o mecanismo de atenção (TSIRMPAS et al., 2024). O desafio dessa abordagem é mitigar a fragmentação de contexto, em que cada segmento é processado isoladamente, o que pode resultar na perda de dependências entre segmentos distantes (PRINCIPE; CHIARINI; VIVIANI, 2025). Essa categoria é representada por métodos de *Transformers* hierárquicos e os recorrentes.

Os *Transformers* hierárquicos, fundamentam-se no reconhecimento de que documentos longos possuem uma estrutura hierárquica natural (palavras → sentenças → parágrafos → seções → documento), e que processar essa hierarquia explicitamente pode capturar melhor as relações semânticas em múltiplas escalas. Modelos como Hi-Transformer (WU et al., 2021), HATN (HU et al., 2021) e *Transformer over BERT* ToBERT (PAPPAGARI et al., 2019) implementam este conceito por meio de uma arquitetura em dois níveis de codificação. A Figura 8 ilustra a arquitetura geral de um *transformer* hierárquico.

Figura 8 – Ilustração do BERT hierárquico para classificação de documentos longos.



Fonte: Adaptado de Prasad (2024) e Pappagari et al. (2019). Arquitetura em três fases: (i) Codificação local por meio de BERT compartilhado produzindo *embeddings* $[CLS]_i$; (ii) Codificação global via *Transformer Encoder* ($nhead = 2$) e *mean pooling* gerando h_{doc}; (iii) Classificação por meio de camadas FC (512→30→K).

Conforme ilustrado na Figura 8, no codificador local (nível-segmento), o documento é dividido em segmentos de tamanho fixo, que são processados independentemente por um BERT compartilhado, gerando *embeddings* locais $\{[CLS]_1, [CLS]_2, \ldots, [CLS]_n\}$ que capturam contexto intra-segmento. No nível de documento, um *Transformer Encoder* processa a sequência de *embeddings* locais por meio de atenção *multi-head*, permitindo que informações de segmentos distantes interajam. A operação de *mean pooling* agrega a sequência contextualizada (L, B, d_{model}) em um *embedding* global único h_{doc} que representa o documento completo, utilizado pelas camadas finais de classificação.

As variantes dessa arquitetura se diferem principalmente na granularidade da seg-

mentação adotada. Como no caso do Hi-Transformer (WU et al., 2021) que utiliza sentenças como unidade básica, uma vez que cada sentença s_i é processada independentemente por um *sentence encoder* gerando representações $\{h_s^1, h_s^2, \ldots, h_s^m\}$ cientes do contexto local, enquanto um *document encoder* subsequente integra essas representações produzindo h_{doc} que incorpora contexto global. Por outro lado, o ToBERT opera com *chunks* de tamanho fixo (tipicamente 200 *tokens*), oferecendo maior flexibilidade para documentos não estruturados, em que os limites de sentenças podem ser ambíguos (PAPPAGARI et al., 2019). O HATN adota uma abordagem intermediária, com os parágrafos utilizados como granularidade (HU et al., 2021). Além dessas diferenças de segmentação, outras variantes incluem mecanismos de retroalimentação, onde representações de documento enriquecem iterativamente as representações de segmentos por meio de atenção bidirecional, e a fusão de múltiplas granularidades por meio de *gates* de atenção que combinam informações de palavras, sentenças e parágrafos simultaneamente (HU et al., 2022; PRINCIPE; CHIARINI; VIVIANI, 2025).

Já os *Transformers* recorrentes, em vez de processar texto como sequência linear ou em hierarquia estrita, constroem um grafo em que nós representam entidades textuais (palavras, sentenças, parágrafos) e as arestas codificam relacionamentos semânticos, estruturais ou posicionais (LI et al., 2023; PRINCIPE; CHIARINI; VIVIANI, 2025). Os *Graph Neural Networks* (GNNs) são então empregados para propagar informação por meio desse grafo, permitindo que o contexto de longo alcance seja capturado por meio de caminhos no grafo independentemente da distância sequencial, isso é particularmente adequado para documentos em que dependências importantes não seguem a ordem sequencial natural (PRINCIPE; CHIARINI; VIVIANI, 2025).

3.2.3 Modelos de síntese de conteúdo

Diferentemente dos demais, esse adota uma estratégia de pré-processamento, simplificando o documento original antes de aplicar modelos *Transformer* convencionais (TSIRMPAS et al., 2024). A premissa fundamental é que nem toda informação no documento é igualmente relevante para a tarefa em questão, e que identificar e preservar apenas o conteúdo relevante pode, simultaneamente, reduzir custos computacionais e melhorar o desempenho ao eliminar ruído (PRINCIPE; CHIARINI; VIVIANI, 2025; FIELDS; CHOVANEC; MADIRAJU, 2024). Este também é composto por duas subcategorias, como segue: (i) Sumarização baseada em seleção; (ii) Abstração de conceitos.

A sumarização baseada em seleção parte do pressuposto de que segmentos-chave do documento contêm informação suficiente para representar o todo. Abordagens simples utilizam métodos não supervisionados, como o TextRank (PARK; VYAS; SHAH, 2022), ou a seleção baseada em *scores* TF-IDF (RAMOS et al., 2003; DAI et al., 2022), para ranquear e selecionar os segmentos mais informativos, que são então concatenados e processados

por *Transformer* convencional (PRINCIPE; CHIARINI; VIVIANI, 2025);

Enquanto a abstração de conceitos, representa a estratégia mais radical de condensação, substituindo termos originais por conceitos latentes de alto nível derivados de modelos de linguagem pré-treinados (PRINCIPE; CHIARINI; VIVIANI, 2025). O LCF-IDF exemplifica essa subcategoria, em que *embeddings* contextualizados de palavras são clusterizados em um espaço latente de conceitos e cada palavra do *corpus* é substituída pelo identificador do *cluster* mais próximo (PRINCIPE; CHIARINI; VIVIANI, 2024). O método cria um *corpus* "traduzido" em que a diversidade lexical é reduzida enquanto relações semânticas são preservadas. Então, a vetorização TF-IDF é aplicada a esse *corpus* intermediário, resultando em representações que combinam a cobertura completa do documento com a consciência semântica derivada de *Pretrained Language Models* (PLMs) (PRINCIPE; CHIARINI; VIVIANI, 2025).

A escolha entre essas estratégias para o processamento de documentos longos depende não apenas das características dos documentos-alvo e dos recursos disponíveis, mas também da capacidade de mensurar e comparar objetivamente o desempenho dos modelos resultantes. Essa necessidade de avaliação sistêmica e comparável motiva a definição de métricas padronizadas e de procedimentos estatísticos, temas abordados na Seção a seguir.

3.2.4 Síntese comparativa das taxonomias

A escolha entre as três taxonomias apresentadas depende não apenas de suas características técnicas, mas também de considerações práticas relacionadas aos recursos computacionais disponíveis, aos riscos inerentes a cada abordagem e à adequação ao contexto de aplicação (PRINCIPE; CHIARINI; VIVIANI, 2024; PRINCIPE; CHIARINI; VIVIANI, 2025). Na Tabela 6 são sintetizadas essas dimensões para as principais categorias de estratégias para processamento de documentos longos.

Tabela 6 – Comparação multidimensional das taxonomias para processamento de documentos longos

Taxonomia	Eficácia Teórica	Custo Computacional	Limitações
Efficient Transformers (Longformer, BigBird, Reformer)	Alta para contextos longos. Modifica o mecanismo de autoatenção e reduzindo complexidade de $O(n^2)$ para $O(n \log n)$ ou $O(n)$, processando até 4.096-16.384 *tokens* (BELTAGY; PETERS; COHAN, 2020; ZAHEER et al., 2020).	Moderado-Alto. Requer GPUs e tempo de treinamento 3-5x maior que BERT base devido a sequências estendidas. Custo de inferência proporcional ao comprimento do documento (ELOUARGUI et al., 2023).	Dependência de modelos pré-treinados majoritariamente em inglês. Adaptação para português jurídico exige *continued pretraining* custoso. Complexidade arquitetural dificulta depuração e interpretabilidade (PRINCIPE; CHIARINI; VIVIANI, 2025).
Modelos de Decomposição-Recomposição (ToBERT, HiTransformer, HATN, HiPool)	Moderada-Alta. Preserva autoatenção completa mas divide documento em segmentos tratáveis. Escalável para documentos arbitrariamente longos (PAPPAGARI et al., 2019; WU et al., 2021).	Moderado. Custo proporcional ao número de segmentos ($n_{seg}\times$ custo BERT + custo agregação) e viável em GPUs; Paralelizável no nível de segmentos (PAPPAGARI et al., 2019; TSIRMPAS et al., 2024).	Fragmentação de contexto entre segmentos pode prejudicar captura de dependências de longo alcance. Qualidade depende da estratégia de segmentação (sentenças, *chunks* fixos, parágrafos). Risco de perda de coerência global (PRINCIPE; CHIARINI; VIVIANI, 2025).
Modelos de Síntese de Conteúdo (Sumarização Extrativa e Abstração de Conceitos)	Variável. Depende da qualidade da sumarização/seleção de conteúdo. **Extrativa:** preserva fidelidade mas pode perder contexto. **Abstrativa:** gera resumos coerentes mas com risco de alucinação.	Baixo-Moderado. **Extrativa:** pré-processamento $O(n^2)$ offline, *fine-tuning* similar ao BERT base. **Abstrativa:** Alto, requer GPUs e inferência mais lenta.	**Extrativa:** Perda de contexto se ranqueamento falhar e resumos podem carecer de coerência. **Abstrativa:** Risco de alucinação factual (geração de informações inexistentes). Perda de rastreabilidade (FABBRI et al., 2019).

Fonte: Elaborada pela autora com base em (BELTAGY; PETERS; COHAN, 2020; PAPPAGARI et al., 2019; LIU; LAPATA, 2019; PRINCIPE; CHIARINI; VIVIANI, 2025; TSIRMPAS et al., 2024).

Os *Efficient Transformers* maximizam o comprimento de contexto processável mediante modificações no mecanismo de atenção, mas dependem de modelos pré-treinados específicos e de infraestrutura computacional robusta. Os modelos de decomposição-recomposição oferecem escalabilidade arbitrária e adequação a recursos limitados, mas enfrentam o desafio da fragmentação de contexto entre segmentos, porém ainda são modelos que permitem o pré-treinamento a partir de modelos BERT. Já os modelos de síntese de conteúdo, por sua vez, apresentam a maior variabilidade, enquanto técnicas extrativas preservam fidelidade ao texto original, métodos abstrativos, embora teoricamente promissores, introduzem riscos de alucinação factual em aplicações que exigem rastreabilidade e conformidade estrita ao texto processual.

Para o contexto específico deste estudo, caracterizado por documentos jurídicos, recursos computacionais típicos de tribunais estaduais e necessidade de auditabilidade das decisões automatizadas, as taxonomias que permitem manter maior fidelidade possível ao texto e de ajuste a partir do modelo já pré-treinado pelo TJMA podem ser as mais adequada

por sua capacidade de processar documentos arbitrariamente longos mantendo tratabilidade computacional. Complementarmente, dentro da taxonomia de síntese de conteúdo, métodos extrativos podem ser mais explorados sobre abstrativos devido à preservação texto original, um requisito importante ao se tratar da classificação de documentos (JAIN; BORAH; BISWAS, 2021; ISLAM; MUHAMMAD; OUSSALAH, 2024).

3.3 Avaliação dos modelos

A etapa de avaliação dos modelos é relevante para mensurar o desempenho e assegurar a validade dos resultados obtidos. Essa Seção constitui, inicialmente, das principais métricas utilizadas na comparação entre modelos, seguidas dos procedimentos estatísticos aplicados para verificar a significância das diferenças observadas.

3.3.1 Métricas de avaliação

A avaliação do desempenho de modelos é importante para entender sua eficácia e comparar diferentes abordagens. Para modelos de linguagem pré-treinados, como BERT e suas variantes, a avaliação típica envolve algumas etapas, como a realização de tarefas de modelagem de linguagem (*e.g.*, MLM ou NSP) e a análise de métricas específicas (CASOLA; LAURIOLA; LAVELLI, 2022). As métricas comuns incluem a perplexidade, que mede quão bem o modelo prevê uma amostra de texto (WANG; LI; SMOLA, 2019); e a acurácia de MLM para modelos como BERT, que avalia a precisão na predição de *tokens* mascarados (DEVLIN, 2018). Enquanto um valor menor de perplexidade indica um melhor desempenho do modelo, valores maiores de acurácia indicam modelos com desempenho superior.

Outra abordagem envolve o *fine-tuning* do modelo para tarefas específicas e a avaliação de seu desempenho nessas tarefas (DEVLIN, 2018). As métricas de avaliação mais comuns para este propósito utilizam os seguintes conceitos: Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). Esses termos são mais utilizados na literatura por seus correspondentes em inglês, respectivamente: *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP) e *False Negatives* (FN) (POWERS, 2011). Com base nesses conceitos, as métricas incluem:

- Acurácia (*Acc*): É a proporção de previsões corretas em relação ao número total de casos examinados.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

- Precisão (P): Proporção de identificações positivas corretas.

$$P = \frac{TP}{TP + FP} \tag{3.4}$$

- Revocação (R): A proporção de positivos reais identificados corretamente.

$$R = \frac{TP}{TP + FN} \tag{3.5}$$

- F1-*score*: A média harmônica entre Precisão e Revocação, fornecendo uma medida balanceada do desempenho do modelo.

$$F1 - score = 2 \cdot \frac{P \cdot R}{P + R} \tag{3.6}$$

Outras métricas frequentemente empregadas na avaliação de problemas de classificação binária, multiclasse e multirrótulo incluem o *Weighted-F1* (F1-W) e o *Macro-F1* (F1-M). O F1-W calcula o F1-*score* de cada classe individualmente e obtém a média ponderada pelo número de amostras (*support*) de cada classe, refletindo o desempenho global em conjuntos de dados desbalanceados. Já o F1-M calcula o F1-*score* de cada classe e a média aritmética simples entre elas, atribuindo peso igual a todas as classes, independentemente de sua frequência (OPITZ; BURST, 2019).

Adicionalmente, é recomendado utilizar validação cruzada para estimar a variabilidade do desempenho (KOHAVI et al., 1995); reportar intervalos de confiança ou desvios padrão junto com as métricas (DROR et al., 2018); comparar com *baselines* apropriados e com o estado da arte (CHALKIDIS et al., 2020; SOUZA; NOGUEIRA; LOTUFO, 2020); e considerar a eficiência computacional, como o tempo de inferência e o uso de memória, especialmente para aplicações em tempo real (SCHWARTZ et al., 2020). Cabe ressaltar que a escolha das métricas deve ser guiada pela natureza específica da tarefa e do domínio. No contexto jurídico, em que a interpretação precisa é útil, métricas que penalizam erros graves, como classificar um documento em uma categoria completamente errada, podem ser particularmente relevantes (CHALKIDIS; ANDROUTSOPOULOS; MICHOS, 2019). Nesse sentido, a análise de métricas deve ser complementada por testes estatísticos que permitam verificar se as diferenças observadas entre os modelos são de fato significativas, conforme discutido na Subseção a seguir.

3.3.2 Testes de significância estatística

A avaliação de modelos de aprendizado de máquina não deve se limitar a métricas pontuais. É fundamental avaliar a significância estatística das diferenças de desempenho observadas entre modelos ou configurações experimentais (DROR et al., 2018; DEMŠAR, 2006). A questão não é apenas determinar qual modelo apresenta melhor desempenho

médio em uma métrica específica, mas sim estabelecer se as diferenças observadas são estatisticamente significativas ou podem ser atribuídas à variação aleatória inerente aos dados e ao processo de amostragem (DIETTERICH, 1998).

O processo envolve formular uma hipótese nula (H_0) que postula a ausência de diferença real entre os modelos, selecionar um nível de significância α (tipicamente 0,05) e calcular o valor-p, isto é, a probabilidade de obter um resultado ao menos tão extremo quanto o observado, sob a suposição de que H_0 é verdadeira. Quando $p < \alpha$, rejeita-se H_0 e conclui-se que há diferença estatisticamente significativa entre os modelos (DEMŠAR, 2006).

Sendo assim, dois tipos de erro podem ocorrer: erro tipo I, que consiste em rejeitar a hipótese nula (H_0) quando ela é verdadeira, e cuja probabilidade é α; e erro tipo II, que ocorre ao não rejeitar H_0 quando ela é falsa. Quando múltiplos modelos são comparados simultaneamente, é necessário controlar a taxa de erro *familywise* (do inglês, *Familywise error rate* - FWER) para evitar inflação da probabilidade de erro tipo I: para k modelos, há $k(k-1)/2$ comparações possíveis, e realizar testes pareados sem correção adequada eleva substancialmente a probabilidade de falsos positivos (SHAFFER, 1995; CARVALHO et al., 2022).

No contexto de aprendizado de máquina, os valores das métricas de avaliação são frequentemente pareados, pois os mesmos conjuntos de teste (*folds* em validação cruzada) são utilizados para todos os modelos (DEMŠAR, 2006). Adicionalmente, métricas limitadas a $[0, 1]$ frequentemente violam premissas de normalidade, especialmente com amostras pequenas ($n < 30$), e testes de normalidade como o *Shapiro-Wilk*, possuem baixo poder estatístico justamente nessas situações (RAINIO; TEUHO; KLÉN, 2024). Diante dessas características, Demšar (2006) recomenda testes não-paramétricos para a comparação de classificadores. O teste de *Wilcoxon signed-rank* para dois modelos e o teste de *Friedman* para três ou mais modelos. Contudo, quando as premissas se $p < \alpha$, rejeita-se H_0 e conclui-se que há evidência estatisticamente significativa de diferença entre os modelos (DEMŠAR, 2006).

Como mencionado, testes não paramétricos, como os de *Wilcoxon signed-rank* e *Friedman*, são recomendados quando as premissas de normalidade e de homocedasticidade não são atendidas. Contudo, em situações em que essas premissas são satisfeitas, é possível empregar métodos paramétricos como a Análise de Variância (ANOVA), descrita a seguir.

Análise de variância

A ANOVA é um método estatístico paramétrico utilizado para testar diferenças entre médias de três ou mais grupos (MONTGOMERY, 2017). No contexto de aprendizado de máquina, utiliza-se especificamente esse método com medidas repetidas (*repeated mea-*

sures ANOVA), também denominada "ANOVA para amostras relacionadas" ou "ANOVA intra-sujeitos", apropriada quando as mesmas unidades experimentais (neste caso, os mesmos *datasets* ou *folds*) são avaliadas sob diferentes condições (neste caso, diferentes modelos) (DEMŠAR, 2006; RAINIO; TEUHO; KLÉN, 2024).

A ANOVA *one-way* (unifatorial) decompõe a variabilidade total dos dados, tanto em variabilidade entre grupos, que reflete as diferenças atribuídas aos distintos modelos avaliados, quanto em variabilidade dentro dos grupos, associada à variabilidade aleatória inerente às observações (KIM, 2017). A estatística de teste F é definida pela razão entre essas variabilidades, conforme a Eq. 3.7.

$$F = \frac{\text{Intergroup variance}}{\text{Intragroup variance}} = \frac{\sum_{i=1}^{K} n_i(\bar{Y}_i - \bar{Y})^2/(K-1)}{\sum_{ij=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2/(N-K)} \tag{3.7}$$

onde $\bar{Y}_i$ é a média do grupo i, n_i é o número de observações do grupo i, $\bar{Y}$ é a média geral, K é o número de grupos (modelos), Y_{ij} é a j-ésima observação do grupo i, e N é o número total de observações ($N = \sum_{i=1}^{K} n_i$) (KIM, 2017).

A hipótese nula é rejeitada se $p < \alpha$ (tipicamente $\alpha = 0{,}05$), indicando que ao menos um par de grupos apresenta diferença estatisticamente significativa. Contudo, ANOVA é um teste *omnibus*, não identificando quais pares de grupos apresentam diferenças. Essa etapa é realizada por meio de testes *post-hoc*, que permitem identificar comparações específicas entre os grupos.

Dessa forma, a variabilidade da ANOVA depende de três pressupostos (MONTGOMERY, 2017; KIM, 2017).

- **Normalidade:** As observações de cada grupo devem seguir uma distribuição aproximadamente normal. Esse pressuposto pode ser verificado por meio do teste de *Shapiro-Wilk* (SHAPIRO; WILK, 1965) ou do teste de *Kolmogorov-Smirnov* (MONTGOMERY, 2017);
- **Independência:** As observações devem ser independentes entre si. Em validação cruzada *k-fold*, este pressuposto é atendido se cada *fold* for uma amostra independente (DEMŠAR, 2006);
- **Homocedasticidade:** As variâncias entre os grupos devem ser homogêneas. O teste de Levene pode verificar esse pressuposto (LEVENE, 1960). Quando violado, outras alternativas devem ser consideradas (KIM, 2017; MONTGOMERY, 2017).

Conforme mencionado anteriormente, quando esses pressupostos são violados, particularmente na normalidade em amostras pequenas, alternativas não paramétricas, como o teste de Friedman, são recomendadas (DEMŠAR, 2006).

Teste post-hoc

Uma limitação importante da ANOVA é que, ao rejeitar H_0, conclui-se apenas que ao menos um grupo difere dos demais, sem identificar especificamente quais pares de grupos apresentam diferenças significativas (KIM, 2017). Os testes *post-hoc* resolvem essa limitação por meio de comparações pareadas controlando a FWER (DEMŠAR, 2006; CARVALHO et al., 2022).

Os principais métodos para esse tipo de teste incluem o Teste de *Tukey's Honestly Significant Difference* (HSD), que controla o FWER para todas as $k(k-1)/2$ comparações pareadas, sendo apropriado quando todas as comparações são de interesse (CARVALHO et al., 2022; MONTGOMERY, 2017); a Correção de Bonferroni, que ajusta o nível de significância para $\alpha' = \alpha/m$, onde m é o número de comparações. Extremamente conservador, pode reduzir excessivamente o poder estatístico, mas é amplamente utilizado por sua simplicidade (MONTGOMERY, 2017; KIM, 2017; DEMŠAR, 2006). Outros, não tão menos relevantes, incluem o Teste de Dunnett, apropriado quando há um modelo de controle (*baseline*) e deseja-se comparar todos os demais apenas contra ele (DEMŠAR, 2006).

3.4 Considerações sobre o Capítulo

No presente Capítulo foram apresentadas as principais estratégias desenvolvidas para lidar com as limitações estruturais dos modelos baseados em *Transformers* no processamento de documentos extensos. Inicialmente, foram discutidas as abordagens de sumarização extrativas, abstrativas e híbridas, como mecanismos de redução de dimensionalidade textual, capazes de condensar informações mantendo a coerência e a relevância semântica. Em seguida, foram exploradas as arquiteturas especializadas, incluindo os *Efficient Transformers*, os modelos de decomposição-recomposição e os modelos de síntese de conteúdo, que representam diferentes caminhos para ampliar o contexto processável do escopo apresentado.

Essas estratégias, embora distintas em natureza e propósito, convergem no objetivo de permitir que os modelos de linguagem operem sobre textos longos de forma escalável e informativamente rica. Complementarmente, na Seção de avaliação foram introduzidos as métricas de desempenho e os testes estatísticos amplamente empregados para garantir a validade dos modelos. Dessa forma, esse Capítulo fornece a base metodológica para o alcance dos objetivos delineados nesta dissertação. No Capítulo seguinte, serão analisados os estudos que aplicaram essas estratégias em contextos reais, possibilitando identificar lacunas, limitações e oportunidades de aprimoramento para o domínio jurídico brasileiro.

4 Panorama de estudos relacionados

Nos Capítulos anteriores, foram apresentados conceitos sobre *Transformers* e a taxonomia de estratégias para o processamento de documentos longos. Este Capítulo complementa essa base, apresentando aplicações empíricas dessas estratégias no domínio jurídico, com ênfase em estudos que avaliaram a eficácia em documentos extensos, em comparação aos do presente trabalho. A organização mantém a taxonomia estabelecida, permitindo avaliar não apenas o desempenho reportado, mas também a viabilidade prática considerando recursos computacionais, disponibilidade de implementações e adaptabilidade ao contexto brasileiro. Para cada categoria, são analisados estudos representativos, destacando-se os *corpus* utilizados, as métricas alcançadas, os custos computacionais reportados e as limitações identificadas. Ao longo do Capítulo, são descritos estudos relacionados às arquiteturas *Efficient Transformers* na Seção 4.1. Na Seção 4.2 abordam-se os trabalhos que aplicaram os modelos de decomposição e recomposição. Em seguida, na Seção 4.3, discorre-se sobre os modelos baseados na síntese de conteúdo. Por fim, na Seção 4.4 sintetizam-se as lacunas na literatura que motivam as escolhas metodológicas deste estudo.

4.1 *Efficient Transformers*

A necessidade de processar documentos extensos levou ao desenvolvimento de arquiteturas capazes de lidar com janelas de contexto ampliadas sem comprometer a eficiência computacional (PRINCIPE; CHIARINI; VIVIANI, 2025). Nesse contexto, Beltagy, Peters e Cohan (2020) introduziram o Longformer combinando atenção local, estruturada por janelas deslizantes, com atenção global em *tokens* estratégicos, reduzindo a complexidade de $O(n^2)$ para $O(n)$. Essa formulação permitiu processar sequências de até 4.096 *tokens* e apresentou ganhos em relação ao *Robustly Optimized BERT Approach* (RoBERTa) em tarefas de raciocínio e de classificação de textos, com +2,6 pontos percentuais em WikiHop com 75,0% de acurácia *vs* 72,4% do RoBERTa e +7,4 pontos percentuais em Hyperpartisan com 94,8% de F1-*score vs* 87,4%. Apesar do avanço, o modelo requer pré-treinamento intensivo, como nesse caso foi rodado 65.000 passos em oito GPUs V100 que durou cerca de três dias, o que torna a adaptação a novos idiomas ou domínios computacionalmente custosa. Seguindo uma direção semelhante, Zaheer et al. (2020) propuseram o BigBird, que utiliza atenção esparsa baseada em grafos aleatórios, mantendo propriedades teóricas de atenção completa enquanto reduz a complexidade. O modelo alcançou desempenho superior em tarefas de *question answering* e de sumarização, mas exige *fine-tuning* dos hiperparâmetros de esparsidade conforme os domínios, além de demandar pré-treinamento

específico, o que limita sua aplicabilidade direta a novos contextos linguísticos.

No campo jurídico, esses princípios foram estendidos por Limsopatham (2021), que compararam, de forma sistemática, o BigBird e o Longformer com técnicas de truncamento e de *pooling* nas decisões da *European Court of Human Rights* (ECHR). O estudo mostrou que o BigBird obteve desempenho médio superior (μ-F1 de 73,08%), superando inclusive variantes de *pooling* aplicadas a modelos especializados, como o MaxPool-ECHR-Legal-BERT (72,13%). Por outro lado, técnicas simples, como o truncamento dos primeiros 512 *tokens*, obtiveram apenas 64,66% de μ-F1, evidenciando perdas de informação. Os autores identificaram que a posição da informação relevante influencia o desempenho. Por exemplo, documentos cuja argumentação se concentra no início beneficiam-se do corte simples, enquanto textos com conteúdo distribuído ao longo do corpo demandam modelos de atenção esparsa.

Avançando nessa linha, Mamakas et al. (2022) adaptaram o Longformer ao domínio jurídico anglófono, com capacidade de processar até 8.192 *tokens*. Partindo do LegalBERT (CHALKIDIS et al., 2020), ajustaram as janelas de atenção e inseriram *tokens* globais *[SEP]* estrategicamente ao final dos parágrafos para capturar a estrutura hierárquica de decisões judiciais. Avaliado nos conjuntos ECtHR e *Supreme Court of the United States* (SCOTUS), o modelo alcançou 76,5% μ-F1 (aproximadamente 7 p.p. acima do LegalBERT truncado). Os autores também exploraram uma abordagem alternativa, combinando LegalBERT com representações TF-IDF, removendo palavras duplicadas, reordenando-as por relevância TF-IDF e adicionando uma camada de *embeddings* TF-IDF. Essa variante, embora inferior ao LegalLongformer (72,3% μ-F1), supera modelos lineares tradicionais como o *Support Vector Machine* (SVM) com TF-IDF: 68,1% e com custo computacional menor.

No contexto brasileiro, Aguiar et al. (2021) exploraram a classificação de 16 mil petições iniciais e denúncias do Tribunal de Justiça do Ceará, em cinco classes do CNJ. Comparando modelos tradicionais, tais como: Random Forest, SVM e XGBoost e redes neurais como o BERT e LSTM-CNN, os autores identificaram que o BERT alcançou F1-*macro* de 0,88, quando comparado ao LSTM-CNN de 0,84 e ao Word2Vec+Random Forest de 0,82. Os autores atribuíram a vantagem à capacidade do BERT de capturar contexto bidirecional e nuances sintáticas próprias da linguagem jurídica. Entretanto, o estudo restringiu-se a documentos que cabem na janela de 512 *tokens*, não abordando estratégias para textos mais longos, o que limita a aplicabilidade do modelo a peças processuais completas.

Mais recentemente, Tinarrage et al. (2025) investigaram a eficiência de súmulas vinculantes no STF por meio de classificação de casos semelhantes, comparando abordagens baseadas em TF-IDF, LSTM, Longformer e expressões regulares em um *corpus* de 634.068 decisões (1989–2018). Embora modelos TF-IDF tenham alcançado melhor F1-*score* em classificação supervisionada, as redes neurais se mostraram mais sensíveis à detecção de

eventos jurídicos relevantes não capturados por métodos lexicais. A análise revelou que, ao contrário do esperado, o número de citações às súmulas aumentou após sua publicação, levantando hipóteses sobre fatores estruturais do sistema judicial. O trabalho reforça, assim, o papel dos métodos de classificação automatizada na compreensão de fenômenos normativos e processuais.

De forma geral, os estudos analisados convergem na constatação de que modelos baseados em atenção esparsa (*e.g.,* Longformer, BigBird e suas variantes) oferecem ganhos mensuráveis em tarefas com textos longos. Entretanto, esses ganhos vêm acompanhados de um custo elevado de implementação, que envolve não apenas o poder computacional necessário, mas também o esforço de pré-treinamento e ajuste em *corpora* especializados. No contexto brasileiro, a indisponibilidade de pesos adaptados ao português jurídico, aliada às restrições de infraestrutura computacional disponível em instituições públicas, torna essa adaptação pouco viável no curto prazo. Diante dessas constatações, o presente estudo prioriza estratégias que aproveitam modelos já especializados ao domínio jurídico nacional, aliando eficiência e aplicabilidade prática. A investigação de *Efficient Transformers* permanece como direção futura, caso recursos computacionais para pré-treinamento se tornem disponíveis, voltada à análise dos ganhos marginais que tais modelos poderiam oferecer em relação aos custos de adaptação linguística e operacional.

4.2 Modelos de decomposição-recomposição

A limitação de janelas de atenção fixas nos modelos baseados em *Transformer* também motivou o desenvolvimento de arquiteturas hierárquicas capazes de processar textos longos por meio de estratégias de decomposição e recomposição. Nesse sentido, autores como Pappagari et al. (2019) propuseram dois modelos hierárquicos para a classificação de documentos longos: o *Recurrence over BERT* (RoBERT) e o ToBERT (descritos no Capítulo 3). Ambos segmentam o texto em partes menores processadas individualmente pelo BERT, diferindo no mecanismo de agregação, uma vez que o RoBERT utiliza camada LSTM bidirecional para integrar as representações segmentadas, enquanto o ToBERT emprega *Transformer* reduzido a duas camadas com autoatenção sobre os *embeddings [CLS]* de cada segmento. Avaliados em conjuntos como *Customer Support Analysis Tasks* (CSAT), 20NewsGroups e Fisher, os resultados demonstraram ampla superioridade do ToBERT com 85,52% de acurácia em 20NewsGroups *vs.* 71,89% do RoBERT e 95,48% em Fisher, superando o RoBERT em 13,63 p.p. neste último. Além de apresentar convergência mais rápida durante o treinamento. Quando comparado ao BERT truncado, o ToBERT apresentou ganhos de 20,48 p.p. com 65,04% $\rightarrow$ 85,52%, evidenciando as perdas de informação decorrentes de cortes simples. Apesar dos avanços, os autores apontaram limitações quanto à generalização para domínios não vistos, dado que o agregador hierárquico tende a aprender padrões estruturais específicos do *corpus* de

origem.

Inspirados por essa ideia de segmentação hierárquica, Lv et al. (2023) apresentaram o HBert, que combina atenção lexical e entre sentenças para processar textos longos de forma eficiente. O modelo codifica palavras em nível local e utiliza um *Transformer* para integrar as sentenças em uma representação global. Avaliado em conjuntos como IMDb, Hyperpartisan, Reuters-21578 e WikiHop, o HBert superou ou igualou a métodos de referência, dentre eles o HAN, RoBERTa e o Longformer, com destaque para ganhos de 0,9 p.p. no F1-*score* no Hyperpartisan e de 0,2 p.p. no WikiHop, além de reduzir em mais de 60% o uso de memória em relação ao BERT-base.

Avançando nessa linha, Li et al. (2023) propuseram o HiPool, que utiliza redes neurais gráficas para capturar relações entre sentenças e parágrafos. O modelo divide o documento em blocos codificados por BERT ou RoBERTa e organiza suas representações em um grafo heterogêneo, com atenção especializada entre os níveis. Avaliado em seis conjuntos do *benchmark* de classificação de documentos longos, do inglês - *Long document classification* (LDC), incluindo o jurídico, o HiPool superou *baselines* como BigBird e ToBERT, com ganhos de até 4,8 p.p. no *F1-score*. Apesar do desempenho, sua complexidade estrutural e dependência da propagação de mensagens dificultam a adaptação a novos idiomas e aplicações práticas.

No contexto jurídico brasileiro, a decomposição de documentos foi explorada de forma pragmática em aplicações reais. Costa et al. (2025) propuseram uma abordagem supervisionada em duas etapas para a classificação de petições do Tribunal de Justiça de Alagoas. O método utiliza uma rede neural sequencial composta por *Long Short-Term Memory Bidirecional* (BiLSTM) e BERT-CNN para localizar e rotular segmentos relevantes em nível de palavra, cujas saídas alimentam o classificador SVM em nível de documento. Essa estratégia de *human-in-the-loop*, em que especialistas anotam frases juridicamente significativas, reduziu a dimensionalidade do problema e alcançou 95,49% de acurácia, superando em 12 p.p. o *baseline* TF-IDF+SVM. O sistema foi efetivamente implementado na 15ª Vara Cível de Alagoas, automatizando tarefas administrativas repetitivas.

Ainda em escala nacional, Pires et al. (2024) combinaram *Transformers* e redes complexas para a classificação hierárquica de petições judiciais. A motivação é que as citações legislativas presentes nas petições são importantes para compreender o conteúdo delas e definir o assunto relevante. O método constrói um grafo bipartido ponderado, conectando temas jurídicos e dispositivos legais citados, e gera *embeddings* usando o node2vec+, que capturam relacionamentos entre citações e tópicos. Esses *embeddings* são combinados com representações textuais de BERTimbau por meio da arquitetura *Hierarchy-aware Prompt Tuning* (HPT), que incorpora o conhecimento da hierarquia de rótulos por meio de GNNs. Avaliado com um *dataset* de 300.000 petições de tribunais brasileiros de múltiplos domínios, dentre eles o trabalhista, criminal e civil. O modelo

demonstrou ganhos de desempenho em tópicos específicos em que citações legislativas são particularmente informativas, notadamente em Direito do Trabalho (F1=0,89 *vs.* 0,85 apenas com texto), Direito Civil (F1=0,82 *vs.* 0,78) e Direito Processual Penal (F1=0,80 *vs.* 0,76). O método foi implementado como microsserviço na plataforma SINAPSES do CNJ, sendo utilizado por tribunais estaduais para sugerir automaticamente a classificação de temas em petições no Sistema Nacional de Processos (PJe).

A literatura demonstra que modelos hierárquicos de decomposição-recomposição equilibram o custo computacional e a preservação do contexto. Arquiteturas mais complexas, como o HiPool, embora superiores em desempenho, impõem barreiras de implementação e adaptação que excedem o escopo das aplicações institucionais brasileiras. Assim, o presente estudo pondera sobre a aplicação do ToBERT, que equilibra desempenho e simplicidade arquitetural, utilizando apenas duas camadas *Transformers* no agregador hierárquico (PAPPAGARI et al., 2019). Esta estratégia garante a escalabilidade linear em relação ao tamanho do documento e a viabilidade prática de replicação mediante a disponibilidade de código reimplementado por Park, Vyas e Shah (2022), Jaiswal e Milios (2023).

4.3 Modelos de síntese de conteúdo

Baseados na seleção de sentenças por meio de algoritmos de sumarização, como o TextRank, e de aleatoriedade, os autores Park, Vyas e Shah (2022) propuseram dois métodos, a saber: BERT+TextRank e BERT+Random, que concatenam os primeiros 512 *tokens* com sentenças adicionais (até 512 *tokens*), diferenciando-se apenas pelo critério de seleção. Enquanto BERT+TextRank seleciona sentenças por meio do algoritmo TextRank, BERT+Random seleciona aleatoriamente, servindo como controle. Avaliados em múltiplos conjuntos de dados, a saber: EURLEX-57K, Hyperpartisan, 20NewsGroups. Os resultados mostraram que, em certos contextos, a diversidade de conteúdo pode superar a relevância estimada, com BERT+Random aproximando-se do Longformer em desempenho (73,22% μ-F1 *vs.* 73,89%), porém com custo computacional significativamente menor. Isso evidencia que abordagens simples, quando bem estruturadas, podem ser competitivas em relação a modelos complexos.

A partir de um aprimoramento das representações semânticas tradicionais, Reimers e Gurevych (2019) desenvolveram *Sentence-BERT* (SBERT), que gera *embeddings* de sentenças semanticamente significativas em complexidade linear, permitindo identificar sentenças mais relevantes ou semanticamente próximas às sentenças âncora. Diferentemente do BERT padrão, que requer processamento de pares de sentenças juntas, com complexidade de $O(n^2)$ para comparar n sentenças, o SBERT gera representações independentes comparáveis por meio da similaridade de cosseno. Embora amplamente usado para a recuperação de informações e o agrupamento de documentos (THAKUR et al., 2021), sua

aplicação na classificação de textos jurídicos longos ainda é pouco explorada.

Já Ding et al. (2020) desenvolveram *Cognize Long Texts* (CogLTX), inspirado na teoria cognitiva de memória de trabalho humana (limitada a 5-9 itens). O método utiliza um mecanismo de seleção iterativa chamado MemRecall, em que um modelo "juiz" seleciona blocos relevantes do texto, que são concatenados e processados por um segundo modelo "raciocinador". Avaliado em NewsQA, HotpotQA, 20NewsGroups e Alibaba, o CogLTX apresentou desempenho superior ao do BERT-base, mantendo consumo de memória constante, independentemente do comprimento do documento. No entanto, a necessidade de dois modelos BERT e de ajustes iterativos de seleção aumenta o tempo de inferência e a complexidade de implementação.

Complementando essas abordagens, Wang e Yoshinaga (2023) propuseram SUMMaug, um método de aumento de dados que emprega a sumarização como estratégia de aprendizado curricular para a classificação de documentos longos. O método gera pseudo-exemplos sumarizando documentos originais e utilizando o *Bidirectional & Autoregressive Transformer* (BART), opcionalmente mesclando rótulos finos em categorias mais amplas. O modelo é treinado primeiro nos resumos concisos e, posteriormente, nos documentos completos, mimetizando a progressão humana na compreensão gradual de textos longos. A inferência ocorre exclusivamente sobre documentos completos, eliminando o custo de sumarização em produção. De maneira similar, Jain, Borah e Biswas (2024) propuseram DCESumm, que combina LegalBERT com *deep clustering* para a sumarização extrativa de documentos jurídicos, ajustando os escores de relevância com base em agrupamentos temáticos, demonstrando ganhos consistentes em BillSum e no *Forum for Information Retrieval Evaluation* (FIRE) em relação a métodos clássicos e recentes.

No contexto jurídico brasileiro, Medina et al. (2022) investigaram a sumarização extrativa como estratégia de redução dimensional para a classificação de documentos processuais. O modelo opera por meio de um *pipeline* que inicia com pré-processamento básico do texto, seguido da construção de um grafo de similaridade entre sentenças, cuja similaridade é medida pela distância Jaro-Winkler. O algoritmo PageRank identifica as 10 sentenças mais centrais nesse grafo, que são então representadas por *Bag-of-Words* e classificadas por *Support Vector Classifier* (SVC). Avaliado em 3.735 documentos distribuídos em seis classes, o estudo de ablação demonstrou progressão de desempenho, em que o modelo sem qualquer pré-processamento alcançou 94% de F1-*score*, a adição de pré-processamento elevou para 95%, e o *pipeline* completo com sumarização atingiu 96%. A melhoria de 2 pontos percentuais evidencia a contribuição da sumarização extrativa, embora o uso de SVM tradicional, ao invés de modelos pré-treinados, limite comparações diretas com abordagens baseadas em BERT.

Quanto aos métodos abstrativos, LLMs como LLaMA (TOUVRON et al., 2023) demonstram capacidade de executar tarefas por meio de *prompt* sem necessidade de

fine-tuning. Estudos recentes, como Zhang et al. (2024) e Lucena et al. (2025), avaliaram LLMs em documentos longos e legislativos brasileiros, constatando que modelos como LLaMA-3 e LLaMA2-13b produzem resumos coerentes e semanticamente fiéis, destacando o potencial de LLMs para sintetizar textos legais complexos, embora métricas tradicionais precisem ser complementadas por avaliações semânticas. Por fim, a literatura sugere que abordagens híbridas, que combinam métodos extrativos e abstrativos, oferecem equilíbrio entre cobertura e coesão do resumo (KIRMANI et al., 2019; EL-KASSAS et al., 2021). Liu e Lapata (2019) e Chen e Bansal (2018) demonstraram que *frameworks* híbridos, especialmente com seleção hierárquica seguida de geração abstrativa, superam métodos puros em *benchmarks* padrão, oferecendo caminhos promissores para sumarização de decisões jurídicas.

4.4 Lacunas identificadas e posicionamento do estudo

Embora existam modelos de linguagem especializados para o domínio jurídico, como o BERTikal (POLO et al., 2021) e o BumbaBERT (CARMO, 2024), nenhum estudo anterior realizou uma comparação sistemática de diferentes estratégias de processamento de documentos longos no contexto jurídico brasileiro. A literatura à fora concentra-se predominantemente em *corpora* europeus e norte-americanos, cujas características estruturais, linguísticas e de extensão diferem dos documentos jurídicos brasileiros, os quais 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).

Além da extensão dos documentos, há diferenças linguísticas importantes. O português jurídico brasileiro inclui terminologia técnica própria, estruturas sintáticas complexas, uso de tempos verbais específicos e arcaísmos preservados na linguagem forense. Estudos de transferência entre idiomas, como Chalkidis et al. (2020), indicam que adaptações superficiais de modelos treinados em outros idiomas podem degradar significativamente o desempenho, o que justifica a necessidade de avaliação específica neste domínio. Adicionalmente, métodos de seleção baseados em similaridade léxica, como o TextRank, podem comportar-se de forma diferente em textos com alta densidade de citações literais em comparação a paráfrases argumentativas (GIARELIS; MASTROKOSTAS; KARACAPILIDIS, 2023).

Tal complexidade organizacional torna o processamento automático de documentos jurídicos distinto de outros domínios textuais, demandando soluções que preservem tanto a precisão técnica quanto a coesão argumentativa (FAMA et al., 2024). De um lado, estudos focam na maximização absoluta do desempenho por meio de arquiteturas complexas, como *Efficient Transformers*, que modificam fundamentalmente o mecanismo de atenção. Beltagy,

Peters e Cohan (2020) e Zaheer et al. (2020) demonstraram ganhos consistentes, mas com custos computacionais elevados. De outro lado, o truncamento simples oferece eficiência, mas resulta em perdas de informação (LIMSOPATHAM, 2021). Park, Vyas e Shah (2022) demonstraram que métodos intermediários oferecem equilíbrio favorável em certos contextos, mas essa hipótese não foi testada sistematicamente em documentos jurídicos brasileiros longos, particularmente diante de restrições orçamentárias e de infraestrutura tecnológica dos tribunais estaduais.

O presente trabalho busca preencher essas lacunas mediante uma abordagem metodológica que combina rigor experimental, adaptação linguística e análise da viabilidade prática. Considerando seis modelos, um deles por meio da estratégia de decomposição-recomposição baseada no ToBERT e os demais baseados em síntese de conteúdo, nominalmente: BumbaBERT+TextRank, BumbaBERT+Random, BumbaBERT+SBERT, BumbaBERT+LexRank e BumbaBERT+LLaMA em *corpus* reais de petições iniciais brasileiras com controle experimental. Por meio da avaliação multidimensional, utilizando validação cruzada estratificada para preservação entre classes e combinando medidas preditivas com indicadores de eficiência computacional, como o tempo de treinamento, o tempo de inferência e o uso de memória da GPU. Os resultados obtidos foram submetidos a testes de significância estatística e a comparações *post-hoc*. Por último, utiliza o BumbaBERT como modelo base, aproveitando sua especialização em português jurídico brasileiro, treinado especificamente em documentos de tribunais brasileiros, incluindo o TJMA, no qual este estudo está inserido, garantindo alinhamento entre o vocabulário, a terminologia técnica e as estruturas sintáticas do modelo e as características dos dados de teste. Adicionalmente, a disponibilidade de pesos pré-treinados do BumbaBERT elimina a necessidade de pré-treinamento custoso, reduzindo as barreiras à adoção. Essa escolha metodológica facilita a transferência de conhecimento para outros tribunais brasileiros que desejem implementar soluções similares, uma vez que podem partir do mesmo modelo base.

Ao comparar métodos com diferentes níveis de complexidade arquitetural e custos computacionais, utilizando um conjunto real de documentos jurídicos brasileiros longos, e avaliando tanto desempenho quanto eficiência, este estudo fornece evidências empíricas para orientar decisões de implementação de sistemas de processamento de documentos longos em tribunais brasileiros. A sumarização dos trabalhos relacionados é apresentada nas Tabelas 7 e 8, destacando os modelos-base utilizados nos estudos anteriores e detalhando os métodos específicos propostos nos estudos para lidar com textos longos.

Tabela 7 – Síntese comparativa dos métodos para processamento de documentos longos (Parte 1)

Categoria	Método	Estudo	Dataset	Tokens	Justificativa de decisão
Efficient Transf.	Longformer	Beltagy, Peters e Cohan (2020)	WikiHop/ Hyperpartisan	4,096	Pré-treino 65k passos (8xV100, 3 dias); Inexiste para PT-BR jurídico.
	BigBird	Zaheer et al. (2020)	NaturalQ/ HotpotQA	4,096	Hiperparâmetros de esparsidade sensíveis ao domínio.
	BigBird (Legal)	Limsopatham (2021)	ECHR	4,096	Supera MaxPool-Legal-BERT (72,1%); Truncamento obtém 64,7%. ECHR: 73,1% μ-F1
	LegalLongformer-8192	Mamakas et al. (2022)	ECtHR/ SCOTUS	8,192	*corpus* anglófono; Inferência 2x mais lenta.
	BERTimbau	Aguiar et al. (2021)	TJCE (BR)	512	Petições brasileiras; Limitado a 512 *tokens*; Não aborda docs longos.
	TF-IDF+LSTM	Tinarrage et al. (2025)	STF (BR)	4,096	634k docs; TF-IDF supera redes neurais em classificação súmulas vinculantes
Decomp-Recomp	RoBERT	Pappagari et al. (2019)	Fisher	Ilimitado	LSTM bidirecional; 13,6 p.p. inferiores ao ToBERT em 20News.
	ToBERT†	Pappagari et al. (2019)	20News	Ilimitado	Autoatenção hierárquica (2 camadas); Complexidade $O(n^2/k^2)$; Implementado.
	HBert	Lv et al. (2023)	Hyperpartisan	Ilimitado	Ganho +0,9 p.p. *vs* Longformer; 60% menos memória; Implementação não pública.
	HiPool	Li et al. (2023)	ILDC (Índia)	Ilimitado	GNN hierárquico; +4,8 p.p. *vs* BigBird; Complexidade alta.
	BERT-CNN+SVM	Costa et al. (2025)	TJAL (BR)	512	*Human-in-the-loop*; Anotação de nível de palavra custosa. TJAL: 95,5% Acc
	BERTimbau+HPT	Pires et al. (2024)	CNJ (BR)	512	Grafo de citações legislativas (node2vec+); Específico CNJ. Média: 65,8% F1

Fonte: Elaborada pela autora.
†Métodos implementados neste estudo.

Tabela 8 – Síntese comparativa dos métodos para processamento de documentos longos (Parte 2)

Categoria	Método	Estudo	Dataset	Tokens	Justificativa de decisão
Síntese Cont.	BERT+TextRank†	Park, Vyas e Shah (2022)	EURLEX	1,024	Seleção por relevância; Tempo de inferência 2x BERT; Implementado. EURLEX: 72,9% μ-F1
	BERT+Random†	Park, Vyas e Shah (2022)	EURLEX	1,024	Controle: diversidade vs relevância; Supera TextRank (+0,3 p.p.). EURLEX: 73,2% μ-F1
	SBERT†	Reimers e Gurevych (2019)	SentEval	512	*Embeddings* semânticos; Complexidade $O(n)$; Pouco explorado no contexto jurídico. SentEval: 87,7% (média)
	CogLTX	Ding et al. (2020)	Alibaba	Ilimitado	2 BERT (juiz+raciocinador); Tempo de inferência 12,5x BERT.
	SUMMaug	Wang e Yoshinaga (2023)	IMDb-2 / IMDb-10	Variável	Aumento de dados por meio da sumarização com BART; inferência apenas em docs completos. IMDb-2: 95,5% Acc; IMDb-10: 57,6% Acc
	DCESumm	Jain, Borah e Biswas (2024)	BillSum / FIRE	Variável	Deep clustering LegalBERT; focado em sumarização extrativa.
	PageRank+SVC	Medina et al. (2022)	TJSP (BR)	Variável	Jaro-Winkler; 3,735 docs; SVM tradicional; 6 classes. TJSP: 96% F1
Atual	BumbaBERT + ToBERT, TextRank, LexRank, Random, SBERT, LLaMA	Presente estudo	TJMA IRDR (BR)	Ilimitado	Comparação sistemática PT-BR jurídico; validação cruzada 5-fold estratificada; ANOVA + *post-hoc* e métricas computacionais.

Fonte: Elaborada pela autora.
†Métodos implementados neste estudo.

5 Materiais e Métodos

Este Capítulo descreve detalhadamente os materiais, procedimentos e decisões metodológicas que orientaram a condução deste estudo, que segue a abordagem DST, conforme fundamentada no Capítulo 2. A adoção do DST justifica-se pela natureza aplicada da pesquisa, voltada à criação de artefato computacional, nesse caso, *pipeline* de processamento de textos jurídicos extensos.

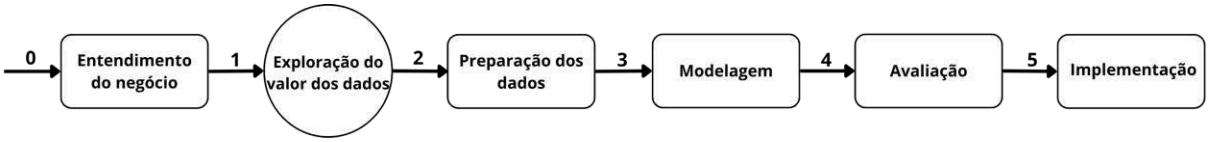
Dessa forma, o Capítulo é composto por seis seções principais, delineadas a partir da Seção 5.1, com a escolha da trajetória metodológica adotada à luz do DST, descrevendo as atividades e as justificativas que estruturam o percurso da dissertação. Na sequência, a Seção 2.1 apresenta o entendimento do domínio jurídico e a contextualização do problema investigado, enquanto a Seção 5.4 detalha os procedimentos de exploração, preparação, estratificação e tokenização do conjunto de dados. Na Seção 5.5 é descrito o processo de modelagem e de configuração experimental, incluindo os modelos de decomposição-recomposição e de síntese de conteúdo aplicados ao problema. Em seguida, na Seção 5.6 são apresentados os critérios de avaliação e a validação estatística empregados para mensurar o desempenho e a eficiência computacional dos métodos. Por fim, na Seção 5.7 discute-se a implementação prática e a integração dos resultados obtidos no contexto do TJMA, estabelecendo o elo entre o artefato desenvolvido e sua aplicação institucional.

5.1 Escolha da melhor trajetória do DST

No contexto do DST, a trajetória representa um grafo direcionado acíclico de atividades, predominantemente sequenciais, que podem ocasionalmente se ramificar para representar tarefas realizadas em paralelo (MARTÍNEZ-PLUMED et al., 2019).A adoção da trajetória DST neste estudo insere-se em um contexto metodológico mais amplo, estabelecido pelo grupo de pesquisa em cooperação técnica com o TJMA. Essa abordagem já vem sendo aplicada de forma consistente em outros projetos vinculados ao acordo de cooperação, incluindo o desenvolvimento do *framework* e demais iniciativas voltadas ao processamento de documentos jurídicos. Com isso, assegurando transparência, coerência e interoperabilidade entre os diferentes componentes tecnológico desenvolvidos no âmbito institucional, facilitando integração dos artefatos produzidos e compartilhamento de práticas entre os pesquisadores envolvidos.

Sendo assim, a definição da trajetória mais adequada foi orientada pelas necessidades específicas deste estudo, que demandam a integração entre o entendimento do domínio jurídico, a exploração do valor dos dados e a modelagem de soluções automatizadas, culminando na sua implementação prática.

Figura 9 – Trajetória metodológica do estudo



Fonte: Elaborado pela autora

Esse fluxo versa desde a compreensão do contexto prático das limitações dos modelos atuais até a implementação de soluções que contribuam para a celeridade e eficiência do sistema judiciário brasileiro. A sequência das etapas garante uma coerência metodológica que parte da identificação do problema, passa pela exploração e preparação adequadas dos dados jurídicos, avança para o desenvolvimento e a avaliação de métodos especializados e culmina na análise crítica dos resultados para sua posterior aplicação prática no sistema judiciário. Cada uma dessas etapas é descrita nas subseções seguintes, nas quais são apresentados os procedimentos específicos, os critérios adotados e as justificativas metodológicas que fundamentam a condução deste estudo.

5.2 Entendimento do domínio jurídico

Conforme ilustrado na Figura 9, a trajetória metodológica deste estudo tem início na etapa de entendimento do negócio, na qual foram definidos os objetivos e as necessidades específicas do domínio jurídico. Essa etapa inicial é estratégica para converter o conhecimento especializado do domínio em requisitos técnicos e conceituais que orientam as etapas subsequentes do estudo, notadamente a exploração do valor dos dados, a preparação, a modelagem e a avaliação (MARTÍNEZ-PLUMED et al., 2019). Como ponto de partida, foi conduzida em cooperação com o TJMA, no âmbito de um acordo técnico firmado com a UEMA. O trabalho conjunto permitiu identificar, de forma prática e contextualizada, os principais desafios enfrentados pelo tribunal na automatização do tratamento de grandes volumes de documentos judiciais, principalmente petições iniciais, que servem de insumo para a identificação de precedentes e a formação de teses no âmbito do IRDR.

A compreensão desse contexto institucional, já discutida no Capítulo 1, sustenta os objetivos deste estudo, que contribuem para a transformação digital da Justiça por meio de soluções que ampliem a eficiência e a celeridade processual, princípios basilares da iniciativa Justiça 4.0. Essa motivação foi traduzida em um problema técnico específico: a dificuldade de processar e classificar automaticamente documentos jurídicos longos, cuja superação é condição necessária para o avanço de sistemas de apoio à decisão em larga escala.

Como detalhado na Seção 2.1, as petições vinculadas a processos de IRDR apresentam particularidades linguísticas e estruturais que as diferenciam de outros tipos de

textos jurídicos. Elas contêm seções narrativas extensas, argumentos jurídicos complexos e citações jurisprudenciais que frequentemente ultrapassam o limite de entrada dos modelos baseados em *Transformer*, fixado em 512 *tokens*. Essa característica inviabiliza o uso direto de modelos convencionais, sem perda de informação contextual, o que impacta a precisão e a interpretabilidade dos resultados.

Observa-se que, no cenário nacional, embora existam modelos de linguagem especializados no português jurídico, como *BERTikal* (POLO et al., 2021), *JurisBERT* (VIEGAS; COSTA; ISHII, 2023), *LegalBERT-PT* (SILVEIRA et al., 2023), *RoBERTaLexPT* (GARCIA et al., 2024) e *BumbaBERT* (CARMO, 2024), nenhum deles foi originalmente projetado para lidar de forma eficiente com textos longos. Todos mantêm o mesmo limite estrutural de contexto curto, herança direta da arquitetura BERT, o que restringe sua aplicabilidade em petições e decisões completas. Além disso, o retreinamento integral de modelos de grande porte em novos contextos jurídicos implica custos computacionais e financeiros consideráveis, tornando inviável a atualização frequente em ambientes institucionais de curto prazo.

No contexto do acordo de cooperação técnica entre a UEMA e TJMA, o modelo *BumbaBERT* constitui o núcleo linguístico das tarefas de PLN relacionadas ao acordo. Embora eficaz na identificação de padrões e jurisprudências em textos de menor extensão, o modelo se limita ao processamento de documentos, comprometendo a extração de informações relevantes e a precisão na classificação de precedentes, questões a serem consideradas para a celeridade processual.

Esse diagnóstico das necessidades práticas do sistema judiciário direcionou a definição dos objetivos da investigação, com foco na adaptação e na otimização de métodos capazes de ampliar a capacidade de processamento de documentos jurídicos longos sem a necessidade de treinar modelos de larga escala a partir do zero. Com base nisso, a investigação concentrou-se em estratégias alternativas que exploram a decomposição hierárquica de textos e a síntese de conteúdo, buscando equilibrar a qualidade preditiva, a eficiência computacional e a viabilidade de implementação no ambiente real do TJMA. Essa compreensão inicial do negócio fundamentou as decisões metodológicas subsequentes e orientou o desenho experimental apresentado nas próximas seções.

5.3 Exploração do valor dos dados

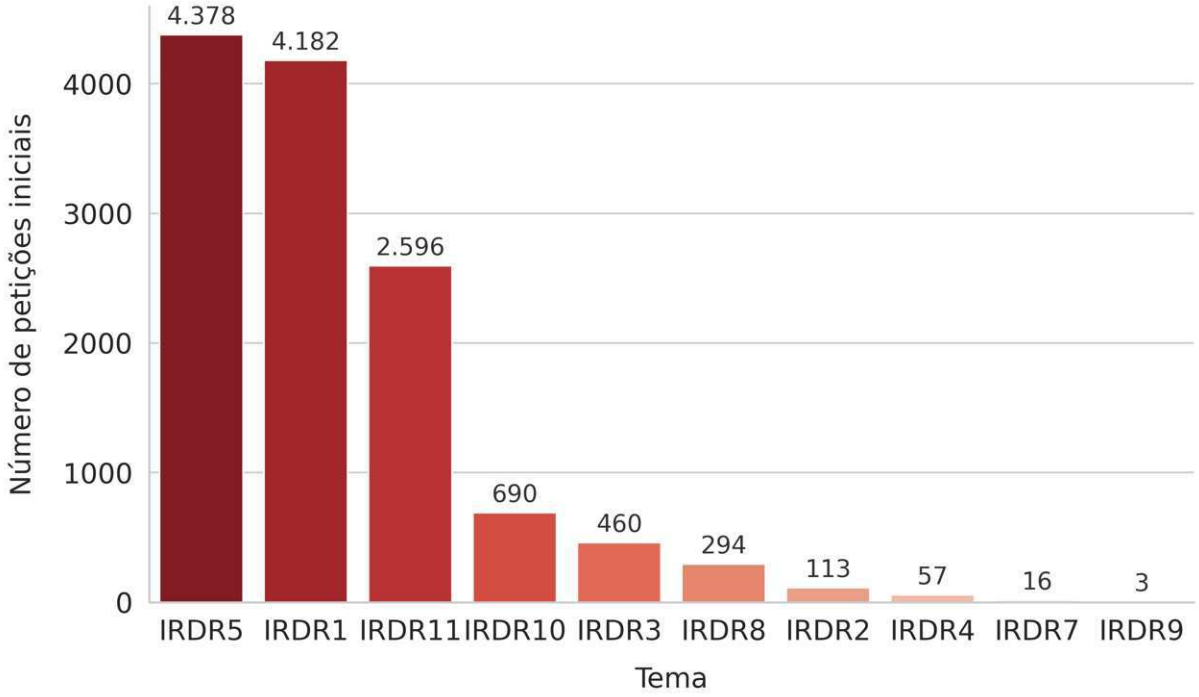
Em continuidade ao método de pesquisa definido, a exploração do valor dos dados corresponde à fase em que os dados são analisados de forma sistemática, visando extrair conhecimento aplicável, ou seja, a ênfase está na solução de problemas específicos do domínio de forma orientada a dados usados para criar modelos, projetar artefatos e, em geral, aumentar a compreensão do assunto (MARTÍNEZ-PLUMED et al., 2019). Assim

sendo, essa etapa incluiu a identificação de conjuntos de dados jurídicos relevantes, análise da estrutura e do conteúdo desses documentos, avaliação da qualidade e completude dos dados, e identificação de padrões ou características específicas dos textos jurídicos que possam impactar o processamento e a classificação de documentos.

Para fins práticos e de solução, esse estudo baseia-se em um *corpus* composto por petições iniciais disponibilizadas pelo TJMA, associadas a dez temas de IRDR, conforme delineado no Capítulo 2. Nessa etapa, procedeu-se à análise quantitativa e estrutural dos documentos, com o objetivo de caracterizar suas propriedades linguísticas e avaliar os fatores que comprometem o processamento integral pelos LLMs.

O conjunto de dados é composto por 12.789 petições iniciais, incluindo matérias de direito administrativo, processual civil, militar e do consumidor. A Figura 10 apresenta a distribuição das petições iniciais por tema de IRDR.

Figura 10 – Distribuição de petições iniciais por tema de IRDR



Fonte: Elaborado pela autora com base nos dados do TJMA.

Nota-se a predominância dos temas 5 e 1, que, juntos, concentram aproximadamente dois terços do conjunto, refletindo o caráter massivo das demandas envolvendo servidores públicos e contratos de consumo. Essa desproporção evidencia um forte desbalanceamento de classes, aspecto que será abordado nas estratégias de amostragem e validação descritas posteriormente.

Cada documento foi submetido à tokenização utilizando o vocabulário e o *tokenizador* nativo do modelo BumbaBERT (CARMO, 2024), o que permitiu estimar a extensão

efetiva de cada texto em *tokens* e mensurar a dispersão intra e interclasse entre os diferentes temas de IRDR. Essas estatísticas descritivas do número de *tokens* por tema do IRDR, antes do pré-processamento, estão resumidas na Tabela 9.

Tabela 9 – Estatísticas descritivas do número de *tokens* por tema IRDR antes do pré-processamento

IRDR	Amostras	Média ± DP	Mediana ($\times 10^3$)	Q1 ($\times 10^3$)	Q3 ($\times 10^3$)
1	4.182	11.314 ± 14.169	7,88	5,87	14,97
2	113	12.025 ± 8.157	9,31	5,50	19,85
3	460	10.001 ± 10.089	8,35	3,99	13,92
4	57	7.347 ± 6.164	6,72	4,12	9,05
5	4.378	9.688 ± 10.696	7,51	2,86	12,87
7	16	9.804 ± 5.568	7,88	7,44	11,46
8	294	19.059 ± 8.345	19,79	13,54	25,07
9	3	7.814 ± 3.098	6,30	6,03	8,84
10	690	14.178 ± 8.786	14,86	5,99	21,93
11	2.596	7.398 ± 6.298	6,30	1,81	8,85

Fonte: Elaborado pela autora. DP = Desvio padrão; Q1 = Primeiro quartil (25^o percentil); Q3 = Terceiro quartil (75^o percentil).

A distribuição apresentada demonstra que as petições iniciais constituem documentos intrinsecamente longos (99,91%), com médias que variam de aproximadamente 7.400 a mais de 19.000 *tokens*, a depender do tipo de IRDR, o que corresponde a um aumento de aproximadamente 14 a 37 vezes em relação ao limite dos modelos BERT. O IRDR tipo 8 apresenta os documentos mais extensos com média de 19.059 *tokens*, enquanto os tipos 4 e 11 apresentam documentos relativamente menores. Para além disso, observou-se um desbalanceamento acentuado em IRDRs dos tipos 1 e 5, que concentram 66,9% do *corpus* (8.560 documentos), enquanto o IRDR-9 possui apenas 3 amostras.

Conforme discutido na Subseção 2.1.2 do Capítulo 2, essa variabilidade correlaciona-se à natureza da controvérsia jurídica. O IRDR-8, que trata de questões de prescrição em matérias de direito administrativo e militar, demanda fundamentação doutrinária e jurisprudencial extensa sobre os marcos temporais, enquanto o IRDR-11, relacionado ao termo prescricional de sentença coletiva, envolve uma questão processual de caráter mais objetivo. A amplitude dos desvios-padrão (6.164 a 14.169 *tokens*) evidencia uma heterogeneidade interna significativa mesmo dentro de cada tema, o que sugere que os advogados adotam estratégias argumentativas diversas, variando entre a reprodução literal de trechos jurisprudenciais e a síntese de precedentes.

Essa constatação evidencia de forma contundente a magnitude do desafio no processamento de textos jurídicos, reforçando a importância de desenvolver métodos capazes de processar efetivamente documentos de diferentes extensões, considerando o tipo de tarefa a ser tratada (LIMSOPATHAM, 2021; KALAMKAR et al., 2022).

5.4 Preparação dos dados

A preparação dos dados sucede à exploração, envolvendo pré-processamento, estratificação e particionamento, tokenização e codificação dos conjuntos de dados jurídicos, de maneira que modelos como o BERT possam compreender e processar efetivamente a informação contida no texto, capturando relações semânticas e contextualizando as palavras, conforme as práticas recomendadas na literatura PLN e aderentes ao tipo de dados tratado (SIINO et al., 2025).

5.4.1 Pré-processamento textual

As petições foram obtidas em formato digital a partir de processo de digitalização utilizando OCR, o que introduziu ruídos textuais característicos de processos automáticos de reconhecimento de caracteres, especialmente em documentos digitalizados a partir de imagens com baixa resolução, marcas de carimbo, assinaturas ou formatações jurídicas, incluindo alinhamento, tabelas, caixas de texto. Esses documentos frequentemente apresentavam padrões como sequências de caracteres invisíveis (`\xa0`, `\n`, `\t`), quebras excessivas de linha, fragmentação de palavras (e.g., "muni- cipalidade"), cabeçalhos e rodapés replicados e artefatos de formatação irrelevantes para análise textual. Na Figura 11 é apresentado um exemplo real com os dados sensíveis anonimizados da saída bruta do OCR, evidenciando a presença desses ruídos, antes da aplicação de qualquer técnica de pré-processamento.

Figura 11 – Exemplo de saída bruta gerada pelo OCR antes do pré-processamento.

```
Poder Judiciário do Estado do Maranhão\n \n \n 1ª Vara da [Comarca Y]\n \n \n
\n \n\n TERMO\n \n \n DE\n \n VIRTUALIZAÇÃO DE AUTOS FÍSICOS\n \n \n \n\n\n \n
\n \n \n \n \n
\n\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\n \n
\n\xa0\xa0\xa0\xa0 [LOCALIDADE], 16 de março de 2020\n \n\xa0\xa0\xa0\xa0 \n
1ª Vara da [Comarca Y], em conformidade com os termos da PORTARIA-CONJUNTA nº
[XX]/2019 e PORTARIA-CONJUNTA nº [YY]/2019, foi concluída a digitalização das
peças encartadas nos autos físicos selecionados para migração, realizado o
cadastro dos metadados do processo judicial e feita a juntada dos arquivos
armazenados em formato eletrônico para fins de virtualização, formando os
respectivos autos digitais no Sistema Processo Judicial Eletrônico (PJe) do 1º
Grau, cujo resumo do protocolo contém as seguintes informações:\n\n \n \n
**********DADOS DE AUTUAÇÃO DO PROCESSO JUDICIAL:**********\n\n \n\t
\tÓrgão Julgador :\t \t1ª Vara da [Comarca Y]\n\t \tProcesso número :[NÚMERO
DO PROCESSO]\n\t \tClasse Judicial :\t \tPROCEDIMENTO COMUM CÍVEL [X]\n\t
\tAssunto Principal :\t \t[Assunto Oculto]\n\t \tData da Distribuição :\t
\t22/11/2016 00:00:00\n\t \tAutor(a)(es) :\t \t[NOME AUTOR]\n\t \tAdv.(a/s)
:\t \t\n\t \tProc.(a/s)(es) :\t \t\n\t \tAssist. Judiciária :\t \t\n\t
\tRéu(e)(es) :\t \t[INSTITUIÇÃO FINANCEIRA B]\n\t \tAdv.(a/s) :\t \t\n\t
\tProc.(a/s)(es) :\t \t\n\t \tAssist. Judiciária :\t \t\n\n Assim, para
constar, firmo o presente termo.\n\n \n \n [LOCALIDADE], 16 de março de 2020\n
\n \n [RESPONSÁVEL] C\n AUXILIAR JUDICIÁRIO
```

Fonte: NUGEP/TJMA.

Essas inconsistências não apenas aumentam artificialmente o comprimento dos documentos, como também afetam diretamente o processo de tokenização, acarretando estouro de sequência no BERT, perda de contexto semântico e elevação do vocabulário fora do domínio legal.

Por essa razão, o pré-processamento seguiu um protocolo padronizado para textos jurídicos (CHALKIDIS et al., 2020; SIINO et al., 2025), executado por meio de um *pipeline* de etapas sequenciais aplicado ao *corpus* de 12.789 petições iniciais. As operações visaram normalizar estruturas textuais heterogêneas resultantes de digitalizações, transcrições e formatações inconsistentes características de documentos jurídicos. As etapas que compõem esse processo são descritas a seguir, desde a limpeza estrutural até a tokenização dos dados.

- **Limpeza estrutural:** remoção de artefatos de formatação e digitalização mediante expressões regulares, incluindo: (i) caracteres invisíveis e de controle, como `\xa0`, `\t`, `\r` e sequências repetidas de `\n`; (ii) múltiplos sublinhados e linhas de carimbo provenientes do OCR (padrões como "_____" ou "—-"); (iii) hifenização indevida gerada por quebra de linha (*e.g.*, *"inde- pendente"* → *"independente"*); (iv) padronização de múltiplos espaços em um único espaço;. Por fim, o texto foi linearizado em blocos de parágrafos contínuos, preservando a estrutura semântica para compatibilidade com o *tokenizador subword* do BERT (DEVLIN, 2018);

- **Padronização e redução da dispersão vocabular:** todo o texto foi convertido para minúsculas e submetido a uma normalização Unicode do tipo NFKD, com o objetivo de remover sinais gráficos e símbolos não ASCII, preservando o conteúdo semântico. Essa etapa diminuiu a fragmentação do vocabulário sem comprometer a expressividade textual, facilitando o aprendizado de representações sublexicais nos modelos baseados em *subwords*. Optou-se por manter as *stopwords*, já que modelos baseados em atenção apreendem automaticamente sua relevância contextual, sobretudo em textos jurídicos;

- **Linearização e compatibilização com tokenização:** os documentos foram transformados em sequências textuais contínuas, livres de marcações estruturais não linguísticas. Essa linearização é essencial para a tokenização BPE e WordPiece, evitando a criação de *tokens* artificiais derivados de espaçamentos, linhas quebradas ou caracteres residuais comuns em documentos digitalizados;

- **Filtragem de qualidade:** exclusão de petições com menos de 50 palavras compostas predominantemente por menções como "em anexo" sem conteúdo relevante, valores nulos ou duplicados (`id_peticao`), resultando em remoção de 0,3% das amostras;

A representação dessas etapas anteriormente descritas pode ser visualizada melhor na Tabela 10, a partir de um exemplo real.

Tabela 10 – Exemplo das etapas de pré-processamento e limpeza de ruídos

Etapa	**Exemplo**
Texto bruto (OCR)	`Poder Judiciário do Estado do Maranhão\n \n \n 1ª Vara da [Comarca Y]\n \n \n \n \n\n TERMO\n \n \n \n DE\n \n \n VIRTUALIZAÇÃO DE AUTOS FÍSICOS\n \n \n \n \n\n\n \n \n \n \n \n \n \n \n\xa0\xa0\xa0\xa0... [LOCALIDADE], 16 de março de 2020...`
Limpeza estrutural	`Poder Judiciário do Estado do Maranhão 1ª Vara da Comarca Y TERMO DE VIRTUALIZAÇÃO DE AUTOS FÍSICOS [LOCALIDADE], 16 de março de 2020 1ª Vara da Comarca Y,`
Normalização (*Lower*)	`poder judiciário do estado do maranhão 1ª vara da comarca y termo de virtualização de autos físicos [localidade], 16 de março de 2020 1ª vara da comarca y,`
Normalização (NFKD)	`poder judiciário do estado do maranhão 1a vara da comarca y termo de virtualização de autos físicos [localidade], 16 de marco de 2020 1a vara da comarca y,`

Fonte: Elaborada pela autora.

Após a filtragem, procedeu-se à normalização dos rótulos e à organização final do *corpus*, na qual foi criada a variável `tema_completo` a partir da concatenação do tipo (`str_tipo_tema`) e do número de IRDR (`str_numero_tema`), que passou a ser utilizada como rótulo de classificação multiclasse. Colunas auxiliares, como identificadores técnicos e metadados não textuais, foram descartadas por não contribuírem para a modelagem.

Subsequentemente, conduziu-se uma tokenização diagnóstica para mensurar se houve impacto das transformações textuais antes e após o pré-processamento, permitindo assim estimar a proporção de documentos que excedem o limite de 512 *tokens* dos modelos BERT.

Considerando o desbalanceamento ainda observado na Tabela 9 entre as categorias, as classes com menos de 300 amostras (IRDR-2, IRDR-4, IRDR-7, IRDR-8 e IRDR-9) foram agrupadas sob o rótulo `Outros`, totalizando 483 amostras, e mantendo as demais: IRDR-5 (4379), IRDR-1 (4182), IRDR-11 (2596), IRDR-10 (690), IRDR-3 (460). Essa consolidação reduziu a assimetria entre categorias e preveniu viés de predição em classes sub-representadas, além de estabilizar o processo de treinamento. A decisão segue as recomendações de He e Garcia (2009) e Japkowicz e Stephen (2002), segundo as quais classes que representam menos de 2% do total de amostras formam *subclusters* minoritários que impedem aprendizado de padrões discriminativos, particularmente quando classes possuem menos de 50-100 exemplares, resultando em viés de predição favorável à classe majoritária e degradação de métricas específicas da classe minoritária, embora a acurácia global permaneça artificialmente alta.

A condução destas etapas resultou em um *corpus* limpo, consistente e representativo do domínio jurídico, preservando a integridade semântica dos textos e reduzindo o ruído estrutural. Além da análise quantitativa confirmando que embora o pré-processamento reduza de forma expressiva o comprimento textual, a maioria das petições ainda permanece extensa. Com a consolidação dos rótulos e a padronização textual concluída, o conjunto de dados foi preparado para as etapas subsequentes de estratificação, tokenização e codificação descritas a seguir.

5.4.2 Estratificação e validação cruzada

Considerando o desbalanceamento de classes identificado anteriormente, adotou-se o procedimento de validação cruzada estratificada com cinco partições (*5-fold stratified cross-validation*) (STONE, 1974), amplamente utilizada em cenários de classificação e implementadas em outros estudos (LIMSOPATHAM, 2021; REIMERS; GUREVYCH, 2019; AGUIAR et al., 2021; PIRES et al., 2024), para que assim a proporção de documentos por categoria de IRDR seja preservada em cada divisão, evitando viés amostral e garantindo comparabilidade estatística entre os modelos (DEMŠAR, 2006). A Figura 12 ilustra o esquema dessa estratificação (*5-fold*), com subdivisão interna em treinamento, validação e

teste.

Figura 12 – Visualização da validação cruzada e da divisão dos dados. *Teste*, *Val* e *Treino* representam os conjuntos de testes, de validação e de treinamento, respectivamente. Um retângulo representa 20% de todos os dados.

Fonte: Elaborado pela autora.

Dentro de cada partição, o subconjunto de treino (90% das amostras) foi subdividido, de forma estratificada e determinística, em 80% para treinamento e 10% para validação, conforme a semente fixa (`base_seed + fold_idx`). Essa divisão interna permaneceu constante em todas as épocas do treinamento, assegurando a reprodutibilidade e o controle estatístico sobre o ajuste dos hiperparâmetros, enquanto o conjunto de teste (10% das amostras) permaneceu completamente isolado até a avaliação final de cada *fold*.

Em cada rodada de validação cruzada, a semente aleatória foi ajustada de forma incremental, assegurando a independência entre os *folds* e a reprodutibilidade integral do processo experimental. Assim, cada amostra foi utilizada exatamente uma vez como instância de teste e quatro vezes em treinamento, maximizando o aproveitamento dos dados e reduzindo a variância das estimativas de desempenho (KOHAVI et al., 1995).

5.4.3 Tokenização e codificação

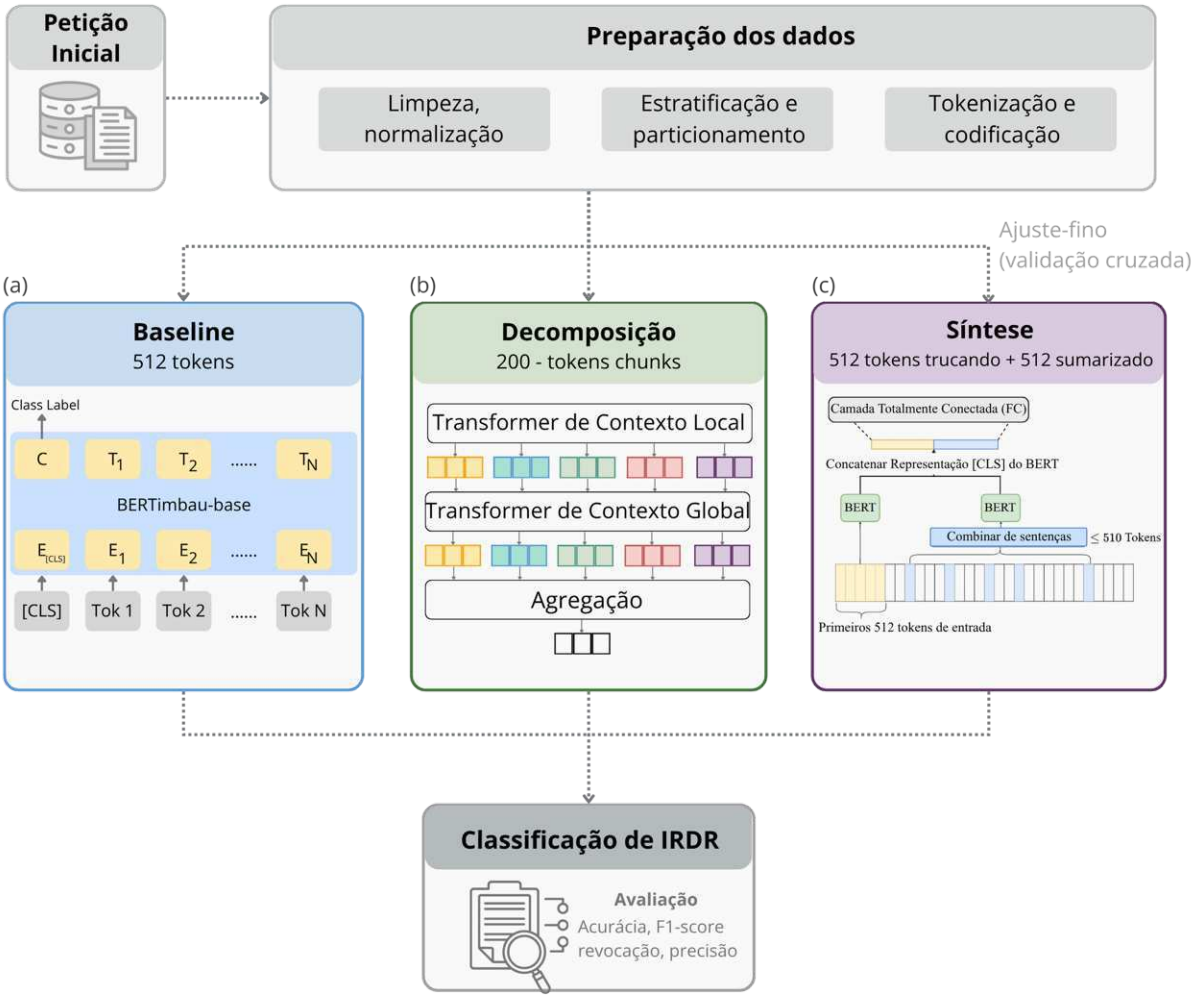
Os textos foram submetidos à tokenização e codificação utilizando o vocabulário nativo do modelo BumbaBERT (CARMO, 2024). A escolha desse *tokenizador* assegura compatibilidade entre o vocabulário e o domínio do *corpus*, além de coerência com o modelo-base empregado nas etapas experimentais. A tokenização foi realizada com base no método BPE, que segmenta as palavras em subunidades lexicais (*subwords*) de acordo com a frequência de coocorrência no *corpus* de pré-treinamento. Esse procedimento permite representar termos técnicos longos ou raros, comuns em textos jurídicos, sem aumentar excessivamente o tamanho do vocabulário. Por exemplo, o termo “constitucionalidade” pode ser decomposto em `[constitucion]` `[al]` `[idade]`, preservando regularidades morfológicas úteis ao modelo.

Durante a codificação, cada *token* foi convertido em seu respectivo índice numérico no vocabulário do modelo e foram inseridos os *tokens* especiais utilizados pelo BERT e suas variantes, `[CLS]` no início de cada sequência (representando a agregação semântica global do texto), `[SEP]` para demarcar o final e `[PAD]` para preenchimento de sequências menores que o tamanho máximo definido (`max_length`). Em paralelo, foram geradas as máscaras de atenção (*attention masks*), vetores binários que indicam quais posições contêm *tokens* válidos (1) e quais representam *padding* (0), assegurando que o mecanismo de atenção não atribua peso indevido a posições vazias (DEVLIN, 2018). A configuração de tokenização variou conforme o método de processamento de documentos longos, detalhado na Seção a seguir.

5.5 Modelagem e configuração experimental

A etapa de modelagem constitui o cerne desta dissertação, envolvendo a implementação e adaptação dos métodos selecionados para o processamento de textos longos no contexto jurídico. Considerando que a maioria das petições (86,03%) ainda excede o limite de *tokens* após o pré-processamento, esta fase concentra-se em estratégias para contornar a limitação da arquitetura *Transformer* baseada em BERT, conforme demonstrado no fluxo da Figura 13.

Figura 13 – Fluxo experimental da comparação dos métodos propostos no presente estudo



Fonte: Elaborado pela autora. O diagrama ilustra o fluxo do ajuste fino do modelo BumbaBERT aplicado à classificação de petições iniciais vinculadas a temas de IRDR, considerando três estratégias experimentais: (a) truncamento simples (*baseline*); (b) decomposição e recomposição hierárquica (*ToBERT-like*) adaptado com base em; e (c) síntese de conteúdo, combinando truncamento e sumarização, adaptado com base em Jaiswal e Milios (2023).

Isso inclui o *fine-tuning* de modelos de linguagem do domínio jurídico e a aplicação dos métodos mais relevantes explorados no Capítulo 4, envolvendo dois grupos principais: (i) modelos de decomposição-recomposição, que preservam o conteúdo integral do documento por meio da agregação hierárquica e (ii) modelos de síntese de conteúdo, que selecionam as seções mais relevantes do texto antes da classificação. A seguir, são apresentados os modelos e as configurações experimentais adotados, uma vez que as descrições conceituais completas já estão detalhadas nos Capítulos 2, 3 e 4.

5.5.1 Modelos base

A performance dos métodos voltados a documentos longos foi comparada a modelos de linguagem pré-treinados em português e especializados no domínio jurídico. Estes serviram como *baselines*, representando o limite inferior de desempenho sob truncamento de texto, conforme a Figura 13a.

- **BumbaBERT**: modelo de referência do domínio jurídico maranhense, baseado na arquitetura BERT-Base e na tokenização BPE, sendo utilizado como principal *baseline* e ponto de partida para adaptações. O modelo possui três variantes já descritas no Capítulo 2, treinadas com pesos do BERTimbau (BumbaBERT (*base*FT)) e treinadas a partir do zero com dados do domínio (BumbaBERT (*base e small*SC)). Sendo selecionada sua versão *small* para a aplicação neste estudo devido às suas vantagens de inferência. Nos testes de referência, o modelo recebeu apenas os primeiros 512 *tokens* de cada petição.
- **LegalBERT-PT**: modelo também voltado para o português jurídico, utilizado para comparação cruzada e validação dos resultados em um modelo externo ao contexto do TJMA.

5.5.2 Modelo hierárquico

Esta abordagem lida com textos longos dividindo-os em segmentos menores (*chunks*) que se encaixam no limite de entrada do BERT, e em seguida, combinando as representações contextuais desses segmentos por meio de uma camada de agregação hierárquica. Como representante dessa classe, foi selecionado o ToBERT (PAPPAGARI et al., 2019), devido à compatibilidade arquitetural com o BumbaBERT, ambos baseados em BERT-Base, à capacidade de processar sequências arbitrariamente longas sem perda de informação e à disponibilidade de implementação de referência validada empiricamente em contextos legais (PAPPAGARI et al., 2019; PARK; VYAS; SHAH, 2022).

A arquitetura do ToBERT, detalhada no Capítulo 3 opera segmentando o documento em *chunks* de 200 *tokens*, com sobreposição de 50 *tokens*, processando, de forma independente, cada *chunk* pelo BumbaBERT e gerando uma representação `[CLS]` de 512 dimensões. Assim, faz-se a agregação temporal por meio de uma camada *Transformer Encoder* com 2 cabeças de atenção e dimensão de modelo *d_model*=512, seguida de *mean pooling* e classificação por meio de *Multilayer Perceptron* (MLP) com arquitetura $512 \rightarrow 30 \rightarrow 6$ classes.

A escolha de *chunks* de 200 *tokens* justifica-se pela necessidade de incluir os *tokens* especiais, como `[CLS]`, `[SEP]`, *padding* e a sobreposição, garantindo que nenhum segmento exceda o limite máximo. A sobreposição de 50 *tokens* ($\approx$25%) preserva a coesão

contextual entre *chunks* adjacentes, mitigando a perda de informação nas fronteiras de segmentação (PARK; VYAS; SHAH, 2022; PAPPAGARI et al., 2019). Esses parâmetros também foram utilizados neste estudo, devido ao número excessivo das petições e a necessidade de manter a coerência textual jurídica, com isso resultando em uma média de 30 *chunks* por petição inicial.

Diferentemente dos métodos de síntese, o ToBERT não requer geração prévia de resumos, processando o texto integral de forma hierárquica. Essa característica o torna particularmente adequado para documentos com argumentação distribuída ao longo de múltiplas seções, cenário frequente em petições do IRDR-8, que apresentam uma média de 11.349 *tokens* (Tabela 9).

5.5.3 Modelos de sumarização

Esta classe de métodos busca condensar o texto original, retendo apenas as sentenças mais representativas, aplicando o modelo base (BumbaBERT) à concatenação entre o início do documento com os primeiros 512 *tokens* e as sentenças selecionadas de até 512 *tokens* adicionais, totalizando 1.024 *tokens* processados. A arquitetura segue o *pipeline* BERTPlus proposto por Park, Vyas e Shah (2022), adaptado ao contexto jurídico brasileiro e ao vocabulário do BumbaBERT, conforme ilustrado na Figura 13c.

Os resumos utilizados nos experimentos foram gerados a partir das petições iniciais completas, previamente normalizadas, com o objetivo de condensar informações mantendo a estrutura argumentativa e os elementos jurídicos. O processo foi implementado no ambiente Colab, com integração ao Google Drive e segmentação em *batches* de 50 documentos por vez, assegurando controle de memória e rastreabilidade dos resultados. Cada resumo foi vinculado ao identificador original da petição, garantindo a correspondência direta entre o texto integral e a versão condensada.

- **BumbaBERT+TextRank:** algoritmo extrativo baseado em grafo que seleciona sentenças centrais mediante análise de coocorrência lexical e ranqueamento por PageRank (MIHALCEA; TARAU, 2004), sendo escolhido por sua simplicidade, eficiência ($O(n^2)$) e ampla adoção em contextos jurídicos;
- **BumbaBERT+LexRank:** variante do TextRank que emprega similaridade de cosseno entre representações TF-IDF de sentenças para construção do grafo, incluindo mecanismo de *threshold* para eliminação de conexões fracas, reduzindo ruído na seleção (ERKAN; RADEV, 2004);
- **BumbaBERT+SBERT:** abordagem semântica utilizando o modelo *stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0*[1], derivado do BERTimbau-Large e adap-

[1] <https://huggingface.co/stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0>

tado ao domínio jurídico (MELO; SANTOS; DIAS, 2023), selecionando sentenças mais próximas da representação vetorial média do documento, preservando coesão semântica e argumentativa;

- **BumbaBERT+LLaMA:** modelo generativo LLaMA (3.1–8B) foi executado localmente para a geração de resumos abstratos controlados. O *prompt* foi estruturado com base no Art. 319 do CPC (Brasil, 2015), exigindo identificação explícita do objeto da ação; fatos relevantes; fundamentos jurídicos e pedidos. Essa parametrização visa preservar a estrutura argumentativa das petições enquanto compacta o conteúdo para 510-512 *tokens*;
- **BumbaBERT+Random:** *Baseline* de seleção aleatória de sentenças, sem critério semântico ou estatístico. Serve como controle experimental para mensurar o ganho informacional dos métodos estruturados (PARK; VYAS; SHAH, 2022).

Todos os métodos foram ajustados para gerar resumos de aproximadamente 184 a 375 *tokens*, compatíveis com a entrada do BumbaBERT e seguindo a abordagem proposta por Park, Vyas e Shah (2022), que concatena os primeiros 512 *tokens* do documento original com até 512 *tokens* selecionados, totalizando uma entrada máxima de 1.024 *tokens*.

5.5.4 Configuração de *fine-tuning*

O processo de *fine-tuning* e a definição dos hiperparâmetros seguiram protocolos consolidados na literatura para tarefas de classificação de textos longos (DEVLIN, 2018; PARK; VYAS; SHAH, 2022), com adaptações específicas ao domínio jurídico e à infraestrutura computacional disponível. Dada a elevada complexidade computacional dos modelos hierárquicos e de sumarização, optou-se por adotar os hiperparâmetros de otimização recomendados na literatura, validados por meio de etapas de experimentos iniciais (*exploratory runs*). A execução dos experimentos foi conduzida sob o rigor do protocolo de validação cruzada estratificada em cinco partes, garantindo a reprodutibilidade e o equilíbrio entre desempenho e a viabilidade operacional.

O tamanho do lote (*batch size*) foi o parâmetro ajustado empiricamente com base nas limitações de *hardware*, variando entre 8, 16 e 32, conforme o comprimento das sequências e a disponibilidade de memória GPU, e foi reduzido em modelos com entradas mais extensas, como o ToBERT, para evitar estouros de memória. A taxa de aprendizado foi explorada no intervalo de 3e-5 a 5e-5, com fase de aquecimento linear correspondente a 10% do total de passos de treinamento, seguindo as recomendações originais do BERT e de Park, Vyas e Shah (2022).

Foram definidas 10 épocas de treinamento, com monitoramento da perda de validação, preservando o melhor *checkpoint* com base em critérios de máxima acurácia

na validação. Os modelos hierárquicos demandaram mais iterações até a convergência devido à maior complexidade da camada de agregação (PAPPAGARI et al., 2019). Para regularização, foi mantida uma taxa de *dropout* de 0,1 nas camadas densas e de atenção, enquanto a otimização utilizou o algoritmo Adam com parâmetros $\beta_1 = 0,9$, $\beta_2 = 0,999$, $\epsilon =$ 1e-8 e *weight decay* de 0,01, mantendo os valores originais do BERT para garantir convergência estável.

A formatação de entrada dos textos seguiu a especificação do modelo BumbaBERT, empregando a tokenização baseada em subpalavras (*e.g.,* BPE) e a inserção dos marcadores especiais `[CLS]` no início e `[SEP]` ao final das sequências. As entradas foram convertidas em índices numéricos, truncadas ou preenchidas (*padding*) conforme o limite máximo definido para cada método e acompanhadas por máscaras de atenção binárias. Nos métodos de síntese (BERTPlus), a entrada foi composta pela concatenação do início do documento com as sentenças selecionadas, totalizando até 1.024 *tokens* na forma `[CLS] primeiros_512_tokens [SEP] sentenças_selecionadas [SEP]`.

Todos os experimentos foram executados no ambiente Python 3,8+ com PyTorch 1,12+ e Transformers 4,25+. Os modelos de maior demanda computacional, como o ToBERT e as variantes de síntese com múltiplas iterações, foram processados no Centro Tecnológico de Computação Científica Aplicada da Universidade Federal do Oeste do Pará (UFOPA), utilizando servidores com processadores AMD EPYC 7532 com 128 *Central Processing Unit* (CPUs) e 243 GiB de RAM. Experimentos de menor escala foram conduzidos em GPUs NVIDIA A100 (24 GB VRAM) no Google Colab Pro.

Com os modelos treinados e ajustados conforme as configurações descritas, procedeu-se à fase de avaliação sistemática (Seção 5.6), na qual o desempenho preditivo de cada método foi mensurado por meio de métricas de classificação multiclasse e de análise de eficiência computacional, permitindo identificar *trade-offs* entre precisão, custo e viabilidade de implementação no contexto do TJMA.

5.6 Avaliação e validação

Nessa etapa de avaliação, os métodos implementados são submetidos a testes com conjuntos de dados isolados ao longo de todo o processo de treinamento, permitindo estimar sua capacidade de generalização para documentos nunca antes vistos. A avaliação preditiva dos modelos baseou-se em métricas consolidadas para tarefas de classificação multiclasse (PAPPAGARI et al., 2019; PARK; VYAS; SHAH, 2022; CHALKIDIS; ANDROUTSOPOULOS; MICHOS, 2019), cuja fundamentação teórica foi apresentada no Capítulo 3. Compreendendo desde métricas de desempenho até métricas de eficiência computacional, bem como a comparação estatística dos modelos.

A escolha das métricas seguiu práticas consolidadas em pesquisas de classificação

de textos jurídicos (PAPPAGARI et al., 2019; PARK; VYAS; SHAH, 2022; CHALKIDIS; ANDROUTSOPOULOS; MICHOS, 2019). Por se tratar de um conjunto de dados desbalanceado, as principais medidas foram a precisão, a revocação e o F1-*score*, calculadas separadamente por classe e, depois, combinadas por médias *macro* e *micro* e ponderadas, e a acurácia, que indica a proporção de classificações corretas. Trazendo, assim, uma visão mais justa tanto do desempenho global quanto do comportamento nas categorias menos representadas.

Para compreender onde ocorrem os erros, utilizou-se ainda a matriz de confusão, que mostra quais classes tendem a ser confundidas entre si. Essa análise foi feita em três formatos: contagens absolutas, normalização por linha (revocação) e por coluna (precisão), permitindo a visualização do comportamento dos modelos. Todas as métricas foram calculadas com base nos conjuntos de teste de cada divisão da validação cruzada em cinco partes, conforme recomendado por Kohavi et al. (1995) e Demšar (2006).

Além da qualidade das previsões, a aplicabilidade prática dos modelos no ambiente do TJMA depende de sua eficiência computacional. Assim, foram analisados três aspectos: o tempo de treinamento, o tempo de inferência e o consumo de memória. O tempo de treinamento mede o tempo que cada modelo leva para ajustar seus parâmetros até convergir. Métodos mais complexos, como o ToBERT, exigem mais ciclos e recursos (PAPPAGARI et al., 2019). O tempo de inferência indica o tempo que o sistema leva para classificar uma petição individual, fator a ser considerado em aplicações que demandam resposta rápida. Já o consumo de memória (CPU e GPU) indica a viabilidade de execução em diferentes infraestruturas.

O monitoramento desses indicadores foi realizado de forma automatizada por meio da integração com a plataforma *Weights & Biases*[2] (WandB), ferramenta de rastreamento de experimentos que registra métricas de desempenho e de consumo de recursos em tempo real, facilitando análises comparativas posteriores e garantindo a reprodutibilidade dos resultados (BIEWALD et al., 2020).

Para garantir que as diferenças observadas entre os métodos fossem estatisticamente significativas e não decorressem de variações aleatórias inerentes aos dados ou ao processo de amostragem, aplicaram-se testes estatísticos apropriados para dados pareados. As métricas de cada modelo foram comparadas nos mesmos cinco *folds*, o que permitiu análises mais precisas e controlou a variabilidade entre as partições (DEMŠAR, 2006; DROR et al., 2018).

A análise estatística foi realizada com o auxílio da biblioteca *Autorank*[3], que automatiza a escolha e a execução dos testes mais adequados às propriedades dos dados, seguindo as diretrizes de Demšar (2006) para a comparação de classificadores. O *Autorank*

[2] <https://wandb.ai/>
[3] <https://sherbold.github.io/autorank/>

aplica, de forma integrada, testes de normalidade (Shapiro–Wilk) para verificar se as distribuições de desempenho por modelo seguem a normalidade dos dados. Também realiza o teste de homogeneidade de variâncias entre os modelos e o teste global correspondente, conforme as suposições. E quando há diferenças significativas, os testes *post-hoc* são aplicados para identificar quais pares de modelos diferem entre si.

Além do valor de p, calcularam-se os tamanhos de efeito para avaliar a relevância prática das diferenças. Esse conjunto de procedimentos garante que as conclusões sobre o desempenho relativo dos modelos estejam baseadas em evidências estatísticas e não apenas em variações pontuais (DEMŠAR, 2006).

5.7 Implementação e integração dos resultados

A etapa de implementação representa a fase final da trajetória metodológica do DST (Figura 9), na qual os resultados obtidos nas fases anteriores são consolidados, interpretados e preparados para aplicação prática no contexto do TJMA. Essa fase inclui a análise crítica dos resultados empíricos, destacando pontos fortes e limitações de cada método, bem como a discussão das implicações teóricas dos achados à luz do estado da arte em processamento de textos jurídicos longos, conforme a revisão sistemática apresentada no Capítulo 4; e a formulação de recomendações para integração dos métodos mais promissores ao *framework* Robô Maria Firmina, considerando desempenho, viabilidade operacional e custos de implementação.

Além de consolidar os resultados técnicos, a implementação também serve como elo entre o plano teórico e a aplicação prática. Ao analisar o comportamento dos modelos sob diferentes configurações e conjuntos de dados, foi possível identificar limitações, avanços e potenciais caminhos de aprimoramento. Essa reflexão final conecta os experimentos realizados às lacunas apontadas na revisão sistemática, contribuindo para o avanço do estado da arte e para a incorporação de abordagens de inteligência artificial mais eficazes e contextualizadas ao sistema de justiça brasileiro.

Os resultados dessa avaliação multidimensional, bem como suas implicações teóricas e práticas para o processamento de textos jurídicos longos, serão detalhados no Capítulo 6, no qual serão apresentados dados empíricos, análises estatísticas e uma discussão crítica dos achados em relação aos objetivos do estudo.

6 Análise e interpretação dos resultados

Neste Capítulo são apresentados e interpretados os resultados obtidos nas etapas experimentais e analíticas deste estudo, representando o produto de 40 experimentos individuais correspondendo a 8 modelos em 5 *folds* cada. As evidências aqui reunidas buscam atingir os objetivos gerais e específicos desta dissertação, validando empiricamente as escolhas metodológicas, oferecendo subsídios interpretativos sobre o comportamento dos modelos no contexto jurídico brasileiro e comparando estrategicamente o desempenho e a eficiência dos oito métodos implementados.

De modo geral, os resultados demonstram que abordagens hierárquicas (ToBERT) superam métodos baseados em síntese de conteúdo, alcançando melhor equilíbrio entre precisão e estabilidade, ainda que com maior custo computacional. Ademais, a análise estatística confirma que tais diferenças são significativas, reforçando a importância de preservar a estrutura argumentativa integral dos textos jurídicos.

A estrutura do Capítulo segue o encadeamento natural da investigação. Inicia-se pela Seção 6.1, que caracteriza o *corpus* experimental, quantificando o impacto das etapas de pré-processamento e validando a necessidade das estratégias de processamento de documentos longos implementadas. Em seguida, nas Seções 6.2 e 6.3 apresenta-se a análise comparativa de desempenho e eficiência computacional dos oito modelos avaliados. Na Seção 6.4 é discutida a validação estatística das diferenças observadas, por meio de testes de significância e de análise de tamanhos de efeito. Por fim, na Seção 6.5, aprofunda-se a discussão interpretativa, conectando os achados empíricos à literatura revisada e aos objetivos específicos desta dissertação.

6.1 Caracterização do *corpus* experimental

Nessa Seção, apresenta-se a caracterização empírica do *corpus* processado, quantificando o impacto das etapas de pré-processamento e validando a necessidade das estratégias de processamento de documentos longos implementadas. A análise estrutura-se em duas dimensões complementares: primeiro, na Subseção 6.1.1, examina-se o impacto do pré-processamento na distribuição de *tokens*, que evidencia a persistência da limitação estrutural dos modelos; e, por segundo, na Subseção 6.1.2, caracterizam-se os resumos gerados pelos métodos de síntese de conteúdo, avaliando suas taxas de compressão e a diversidade metodológica adotada para mitigar as restrições de entrada.

6.1.1 Impacto do pré-processamento na distribuição de *tokens*

O pré-processamento textual, descrito na Seção 5.4, resultou em alterações expressivas na distribuição de *tokens* dos documentos. Essas diferenças podem ser observadas na Tabela 11, que contém as estatísticas descritivas por tema de IRDR.

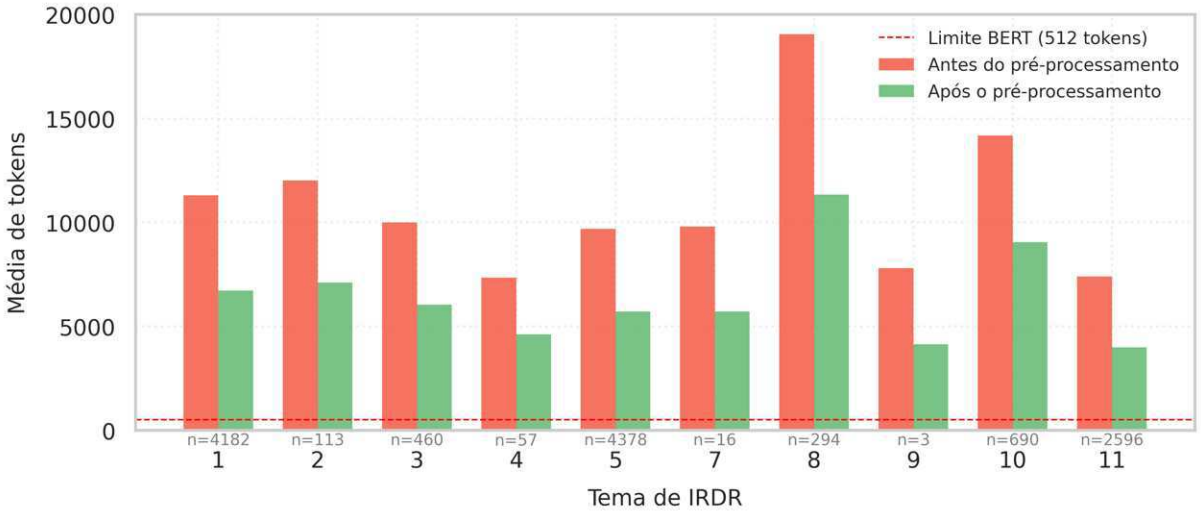
Tabela 11 – Estatísticas descritivas do número de *tokens* por tema IRDR após pré-processamento

IRDR	**Amostras**	**Média ± DP**	**Mediana** ($\times 10^3$)	**Q1** ($\times 10^3$)	**Q3** ($\times 10^3$)
1	4.182	6.720 ± 8.761	4,65	3,58	9,11
2	113	7.113 ± 4.813	6,24	3,28	11,45
3	460	6.043 ± 6.329	4,90	2,17	8,23
4	57	4.616 ± 3.966	4,09	2,33	5,73
5	4.378	5.722 ± 6.252	4,67	1,89	7,57
7	16	5.718 ± 3.316	4,64	4,27	6,73
8	294	11.349 ± 5.157	11,13	7,27	15,56
9	3	4.152 ± 1.359	3,73	3,39	4,70
10	690	9.060 ± 5.861	8,25	3,65	14,32
11	2.596	4.002 ± 3.638	3,09	0,72	5,02

Fonte: Elaborada pela autora. DP = Desvio padrão; Q1 = Primeiro quartil (25^o percentil); Q3 = Terceiro quartil (75^o percentil).

Essa distribuição pode ser visualizada de forma comparativa na Figura 14, que ilustra o impacto do pré-processamento antes e após, levando em conta o número de *tokens*.

Figura 14 – Comparação do tamanho médio das petições por tema de IRDR antes e após o pré-processamento



Fonte: Elaborado pela autora com base nos dados do TJMA.

A partir dessa análise comparativa, observa-se uma redução expressiva no comprimento médio global dos documentos, passando de 10.234 *tokens* como visto na Tabela 9

para 6.027 *tokens*, o que representa uma redução média de 41,1%. Essa diferença reflete a eliminação de ruídos estruturais e de redundâncias decorrentes de formatações incorretas e de caracteres residuais provenientes de digitalizações e de cópias processuais. Antes do pré-processamento, apenas 0,09% das petições encontravam-se abaixo do limite de 512 *tokens*, enquanto 99,91% ultrapassavam esse limite.

Além da análise estatística, tornou-se relevante observar como o pré-processamento afeta a estrutura textual real dos documentos. Nas Figuras 15 e 16 são apresentados exemplos extraído de uma petição inicial, comparando a versão original com a versão normalizada após a aplicação das etapas descritas na Seção 5.4.

Figura 15 – Comparação entre texto original e pré-processado de documento jurídico

Texto original (antes do pré-processamento):

EXCELENTÍSSIMO SENHOR DOUTOR JUIZ DE DIREITO DA __ VARA DA

FAZENDA PÚBLICA DA COMARCA DE SÃO LUÍS DO ESTADO DO

MARANHÃO.

Assistência Judiciária.

Prioridade especial na tramitação por ser idoso (a) - Lei 13.466/2017

Ação Ordinária nº. [NÚMERO DO PROCESSO PRINCIPAL].

[NOME DO EXEQUENTE], brasileiro, casado, funcionário público municipal,
CPF [NÚMERO DE CPF], RG [NÚMERO DE RG] SSP/MA, com endereço na
[ENDEREÇO COMPLETO], município de São Luís, CEP [NÚMERO DE CEP];

conforme Ação Coletiva Ordinária de nº. [NÚMERO DO PROCESSO PRINCIPAL]
que teve como Autor o SINDICATO DOS FUNCIONÁRIOS E SERVIDORES
PÚBLICOS MUNICIPAIS DE SÃO LUÍS - SINFUNSP-SL em face do
MUNICÍPIO DE SÃO LUÍS, pessoa jurídica de direito público interno, que
tramitou na 1ª VARA DA FAZENDA PÚBLICA, desta comarca, vem,

respeitosamente, à presença de Vossa Excelência, através do seu advogado
que assina eletronicamente, conforme procuração anexa documento nº. 01,
com amparo no Art. 534, da Lei Processual Civil brasileira, requerer o:

CUMPRIMENTO DE SENTENÇA

Conforme cópias de nº. 02/08, em anexo, a saber: [...]

Fonte: Elaborada pela autora. Exemplo de petição inicial antes do pré-processamento com 1385 *tokens*. Os dados sensíveis foram anonimizados.

Figura 16 – Comparação entre texto original e pré-processado de documento jurídico

Texto pré-processado (após normalização):

```
excelentíssimo senhor doutor juiz de direito da vara da fazenda pública da
comarca de são luís do estado do maranhão. assistência judiciária. prioridade
especial na tramitação por ser idoso (a) - lei 13.466/2017 ação ordinária nº.
[NÚMERO DO PROCESSO PRINCIPAL]. [NOME DO EXEQUENTE], brasileiro, casado,
funcionário público municipal, cpf [NÚMERO DO CPF], rg [NÚMERO DO RG] ssp/ma,
com endereço na [ENDEREÇO COMPLETO], município de são luís, cep [NÚMERO DE
CEP]; conforme ação coletiva ordinária de nº. [NÚMERO DO PROCESSO PRINCIPAL]
que teve como autor o sindicato dos funcionários e servidores públicos
municipais de são luís - sinfunsp-sl em face do município de são luís, pessoa
jurídica de direito público interno, que tramitou na 1ª vara da fazenda
pública, desta comarca, vem, respeitosamente, à presença de vossa excelência,
através do seu advogado que assina eletronicamente, conforme procuração anexa
documento nº. 01, com encosto no art. 534, da lei processual civil brasileira,
requerer o: cumprimento de sentença conforme cópias de nº. 02/08, em anexo, a
saber: [...]
```

Fonte: Elaborada pela autora. Exemplo de petição inicial após pré-processamento reduzido para 696 *tokens*.

Após as etapas de limpeza e normalização, o percentual de textos dentro do limite aumentou para 13,97%, embora a maioria, ainda com 86,03%, ultrapasse a capacidade de entrada de 512 *tokens* dos modelos. Essa persistência evidencia que a limitação arquitetural dos modelos *Transformer* não pode ser contornada apenas por técnicas convencionais de limpeza textual, o que justifica a investigação e comparação das estratégias de processamento de documentos longos apresentadas nas seções seguintes, em especial as abordagens hierárquicas e de síntese de conteúdo.

6.1.2 Caracterização dos resumos gerados

Os resumos gerados para os cinco métodos baseados em síntese de conteúdo, quais sejam: BumbaBERT+LexRank, +TextRank, +SBERT, +LLaMA e +Random, variaram quanto ao comprimento dos textos produzidos. As estatísticas descritivas correspondentes encontram-se na Tabela 12. O método Random não foi incluído na tabela por ter gerado seleções sem padrão estatístico consistente e por estar diretamente incluído no *pipeline*.

Tabela 12 – Estatísticas descritivas dos resumos gerados pelos métodos de síntese de conteúdo

Método	**Média ± DP**	**Taxa Compressão (%)**
LLaMA	374,58 ± 121,62	6,07
SBERT	323,44 ± 80,39	5,24
LexRank	319,10 ± 73,74	5,17
TextRank	183,50 ± 134,23	2,97

Fonte: Elaborada pela autora. Taxa de compressão calculada em relação ao comprimento médio original do documento.

A diversidade metodológica adotada, combinando abordagens extrativas baseadas em grafos, representações semânticas contextualizadas e modelos generativos, teve como objetivo avaliar o impacto do tipo de compressão textual sobre o desempenho de classificação, permitindo investigar se modelos de síntese mais elaborados oferecem ganhos significativos em relação a métodos tradicionais de truncamento ou de seleção aleatória.

Notavelmente, o método LLaMA produziu resumos consistentemente mais extensos, com média de 374,58 *tokens*, aproximadamente 16% maiores do que o SBERT que teve a média de 323,44 *tokens*, e 104% maiores do que o TextRank com 183,50 *tokens*, refletindo assim sua estratégia abstrativa que reformula o conteúdo em vez de apenas selecionar sentenças existentes. O *prompt* resultante completo para a formulação desses resumos abstrativos é apresentado na Figura 17.

Figura 17 – *Few-shot prompt* para sumarização estruturada de petições iniciais (LLaMA 3.1-8B)

Instrução para sumarização estruturada de petições iniciais

Você é um assistente jurídico especializado em analisar petições iniciais conforme o Código de Processo Civil (Art. 319). Com base no texto abaixo, gere um resumo técnico, claro e conciso:
[**PETIÇÃO**] {text}
[**TAREFA**] Produza um resumo em parágrafo contínuo com até {max_tokens} tokens identificando:

- OBJETO: tipo e natureza da ação
- FATOS: acontecimentos principais
- FUNDAMENTOS: dispositivos legais, jurisprudência e teses
- PEDIDOS: requerimentos e tutelas

[**FORMATO**] Parágrafo contínuo, sem explicações adicionais. Extraia apenas as informações presentes na petição.
[**EXEMPLO**] a autora, aposentada, propõe ação anulatória de empréstimo bancário cumulada com repetição de indébito [...] fundamenta-se nos arts. 5º, x da cf/88, 6º, viii do cdc [...] requer tramitação prioritária, justiça gratuita, declaração de nulidade do contrato [...] dá à causa o valor de R$ 7.480,40.

Fonte: Elaborado pela autora.

Para a parametrização do *prompt* preservou-se a estrutura argumentativa das petições, enquanto compactou-se o conteúdo para um limite compatível com BumbaBERT, uma estratégia alinhada às diretrizes de *few-shot prompting* para domínios especializados (BROWN et al., 2020). A inclusão de um exemplo concreto de resposta esperada, identificado pelo campo *[EXEMPLO]*, reduziu a ambiguidade interpretativa do modelo e aumentou a consistência na formatação dos resumos gerados, uma técnica amplamente validada em estudos de engenharia de *prompts* para LLMs (BROWN et al., 2020). A avaliação do impacto na geração de resumos é apresentada na Seção a seguir, com base nos resultados da tarefa de classificação das petições iniciais.

6.2 Análise comparativa de desempenho

Na Tabela 13 são apresentados as médias e os desvios-padrão das métricas de avaliação obtidas pelos oito modelos ao longo das cinco execuções da validação cruzada estratificada. Os valores reportados correspondem às médias aritméticas, acompanhadas de seus respectivos desvios-padrão, oferecendo não apenas uma estimativa pontual do desempenho esperado, mas também uma indicação da estabilidade de cada modelo diante de variações na composição dos conjuntos de treinamento e teste.

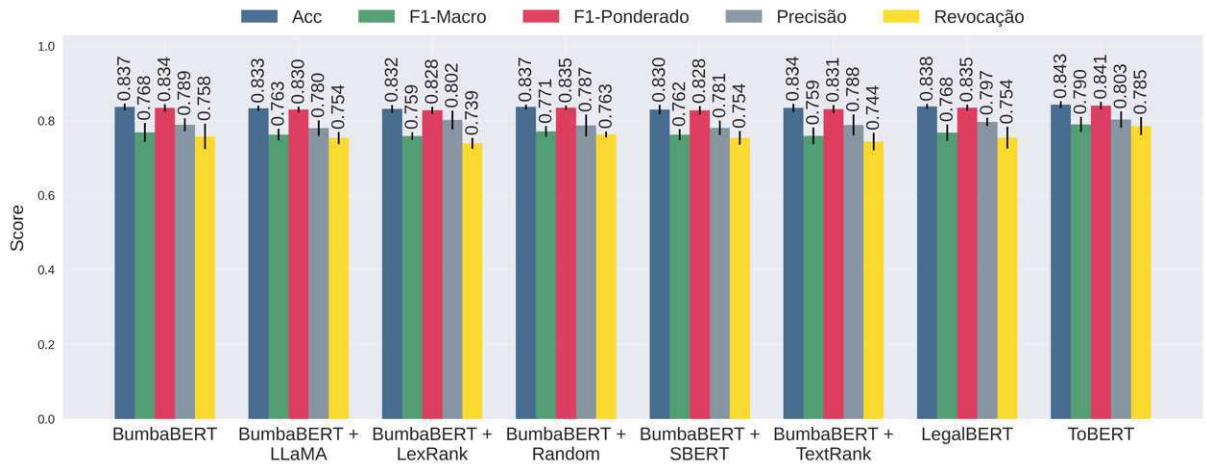
Tabela 13 – Desempenho médio dos modelos nas métricas de classificação de petições iniciais

Model	Acc	F1-M	F1-W	P	R
Baselines					
BumbaBERT	0,837±0,009	0,768±0,025	0,834±0,010	0,789±0,018	0,758±0,034
LegalBERT-PT	0,838±0,007	0,768±0,022	0,835±0,008	0,797±0,011	0,755±0,029
Modelos de decomposição-recomposição					
ToBERT	**0,843±0,009**	**0,790±0,020**	**0,841±0,010**	**0,803±0,022**	**0,785±0,024**
Modelos de síntese de conteúdo					
BumbaBERT+LexRank	0,832±0,010	0,759±0,010	0,828±0,010	0,802±0,025	0,739±0,015
BumbaBERT+TextRank	0,834±0,011	0,760±0,022	0,831±0,011	0,788±0,028	0,744±0,024
BumbaBERT+Random	0,837±0,006	0,771±0,014	0,835±0,006	0,787±0,029	0,763±0,008
BumbaBERT+SBERT	0,830±0,013	0,763±0,015	0,828±0,011	0,781±0,019	0,754±0,018
BumbaBERT+LLaMA	0,833±0,007	0,763±0,016	0,830±0,008	0,780±0,021	0,754±0,016

Acc = Acurácia, F1-M = F1-*score* Macro, F1-W = F1-*score* Ponderado, P= Precisão, R = Revocação. Fonte: Elaborado pela autora.

Os resultados indicam uma diferença consistente entre as famílias de modelos. O ToBERT, como modelo representativo da decomposição-recomposição, obteve desempenho superior em todas as métricas avaliadas, com valor acentuado de 0,790 (±0,020) no F1-*macro*. Essa diferença é de 1,9 pontos percentuais em relação ao modelo de síntese BumbaBERT+Random, com 0,771, e de 2,2 pontos percentuais em relação ao melhor método baseado em seleção algorítmica de sentenças, BumbaBERT+SBERT, com 0,763. Embora essa diferença possa parecer modesta em termos absolutos, no contexto prático do judiciário, ela representa um aumento na acurácia do serviço, correspondendo a um número menor de petições que precisam de triagem manual.

Figura 18 – Desempenho médio dos modelos na classificação de petições iniciais



Fonte: Elaborado pela autora

Esse ganho observado alinha-se aos achados de Pappagari et al. (2019) em documentos científicos longos e de Dai et al. (2019) em textos narrativos extensos, validando

a hipótese de que arquiteturas especificamente projetadas para capturar dependências de longo alcance mediante agregação hierárquica de representações superam adaptações *ad hoc* baseadas em truncamento. O desvio-padrão relativamente baixo do ToBERT de 0,020 no F1-*macro* sugere estabilidade do método frente a variações na composição dos conjuntos de treinamento.

No comparativo entre os *baselines*, o BumbaBERT e LegalBERT-PT apresentaram resultados praticamente idênticos, com 0,768 no F1-*macro*, provavelmente devido à arquitetura reduzida do BumbaBERT com 6 camadas e 8 cabeças de atenção e aos 42 milhões de parâmetros (CARMO, 2024), enquanto o LegalBERT-PT baseia-se em uma arquitetura *base* de 12 camadas com 110 milhões de parâmetros (POLO et al., 2021). Essa equivalência sugere que o BumbaBERT mesmo com menor capacidade paramétrica, é suficientemente adaptado ao domínio para compensar a diferença de escala estrutural.

Em termos técnicos, o truncamento pode ter eliminado justamente as seções nas quais o vocabulário especializado se manifesta com maior proeminência. Como visto, as petições jurídicas tipicamente seguem uma estrutura retórica padronizada e, ao truncar em 512 *tokens*, ambos os modelos processam predominantemente cabeçalhos e trechos iniciais do documento, regiões textuais em que o vocabulário tende a ser mais genérico e processual (PRINCIPE; CHIARINI; VIVIANI, 2025; TSIRMPAS et al., 2024). Essa hipótese é corroborada por Park, Vyas e Shah (2022), que demonstrou que métodos de truncamento simples perdem informação desproporcionalmente concentrada nas seções centrais e finais de documentos com estruturas complexas ou múltiplos blocos informativos.

Os cinco métodos de síntese de conteúdo apresentaram desempenho inferior ou, no máximo, equivalente aos *baselines* de truncamento simples, o que contradiz a premissa intuitiva de que "mais informação relevante" deveria produzir classificações mais precisas. O BumbaBERT+Random, que seleciona sentenças aleatoriamente para complementar os primeiros 512 *tokens*, foi o único método dessa categoria a superar ligeiramente os *baselines* com F1-*macro* de 0,771 *vs* 0,768, respectivamente.

Embora as mais de 40 matrizes de confusão geradas ao longo dos experimentos não tenham sido incluídas integralmente no documento, devido ao volume e à redundância visual, todas foram registradas e estão disponíveis para consulta interativa no relatório da plataforma WandB[1], bem como reproduzidas integralmente no Apêndice B. Assim, apresentamos aqui apenas a análise qualitativa dos padrões mais recorrentes e relevantes para a interpretação dos resultados.

De forma mais detalhada, a análise das matrizes de confusão geradas revelou três padrões sistemáticos principais. A confusão mais frequente observada ocorreu entre IRDR-1 (Classe 0) e IRDR-5 (Classe 4). O IRDR-1 trata de direitos remuneratórios de

[1] <https://api.wandb.ai/links/gabiaraujo-state-university-of-maranh-o/xhoeq5x2>

servidores públicos estaduais, especialmente reajustes salariais, percentuais de 21,7% e 6,1%, e diferenças de vencimentos, enquanto o o IRDR-5 discute a validade de cláusulas contratuais em empréstimos consignados, envolvendo instituições financeiras, descontos em folha, contratos bancários e tarifas. Embora pertençam a áreas jurídicas distintas (Direito Administrativo vs. Direito do Consumidor/Bancário), os modelos confundiram essas classes devido à presença recorrente de vocábulos como "servidor público", "estado", "desconto", "consignado" e "remuneração", além de ambos envolverem a Administração Pública como parte.

Essa sobreposição semântica manifesta-se bidirecionalmente, onde documentos verdadeiramente pertencentes ao IRDR-1 foram frequentemente preditos como IRDR-5, e vice-versa. Notavelmente, o modelo BumbaBERT+Random apresentou melhor desempenho na discriminação dessas classes, alcançando maior precisão na Classe 4 (IRDR-5) em comparação aos métodos de síntese algorítmica. Esse resultado sugere que a diversidade estrutural introduzida pela seleção aleatória de sentenças pode ter capturado seções discriminativas que algoritmos baseados em relevância lexical tendem a omitir.

A Classe 3 (IRDR-3), correspondente ao tema de nomeação de candidatos excedentes em concurso público, foi a com pior desempenho, refletindo sua condição de classe minoritária. A inspeção das predições errôneas revelou confusão predominante com a Classe 4 (IRDR-5), a classe majoritária do conjunto de dados, possivelmente devido ao compartilhamento de termos administrativos genéricos. Esse padrão é característico de desbalanceamento, em que o modelo, diante de incerteza, tende a favorecer a classe com maior representação nos dados de treinamento (HE; GARCIA, 2009). O modelo ToBERT conseguiu capturar melhor as representações semânticas nessa classe, sugerindo maior sensibilidade contextual.

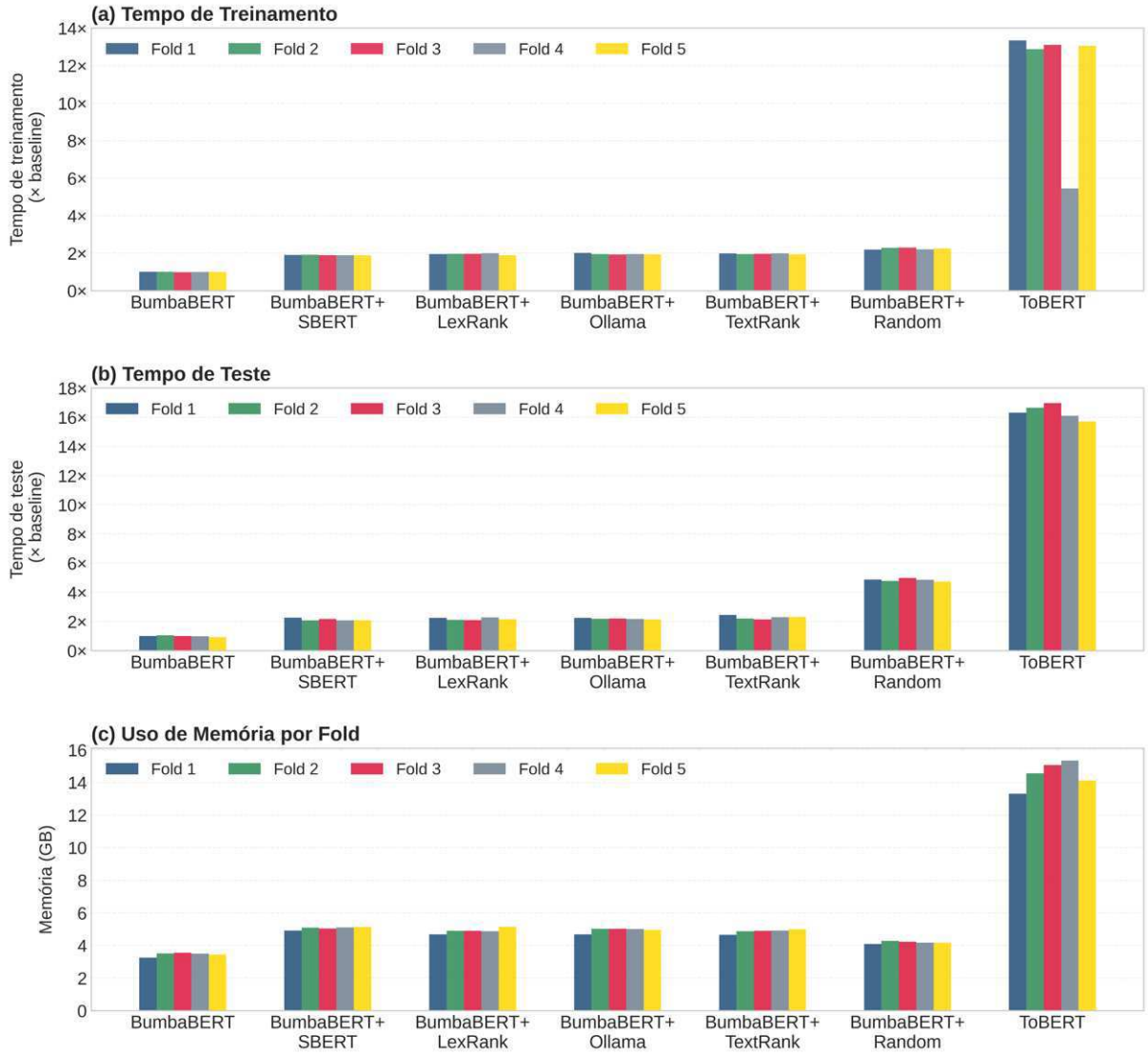
A Classe 5 ("Outros"), construída pela agregação de cinco temas minoritários (IRDR-2, IRDR-4, IRDR-7, IRDR-8, IRDR-9), apresentou valores intermediários mas com alta variabilidade entre os *folds*. Essa classe contém decisões de natureza mista (Administrativo, Bancário, Processual e Militar), o que dificulta a formação de um padrão semântico consistente para treinamento. Como consequência, documentos genuinamente pertencentes à classe "Outros" foram frequentemente classificados como IRDR-1 e IRDR-5, que representam as classes mais estruturadas e semanticamente densas do conjunto.

6.3 Análise da eficiência computacional

Além do desempenho preditivo, a análise de eficiência computacional permitiu avaliar o custo associado à obtenção desses resultados. Na Figura 19 podem ser observados o tempo de treinamento, o tempo de teste e o uso de memória média por *fold* para cada modelo. Os valores foram normalizados em relação ao desempenho do modelo *baseline*

(BumbaBERT), o que possibilita observar de forma comparativa a magnitude relativa do custo computacional entre as abordagens.

Figura 19 – Comparação da eficiência computacional dos modelos ao longo de todo o processo de treinamento e de inferência.



Fonte: Elaborado pela autora. (i) tempo de treinamento normalizado, (ii) tempo de teste normalizado e (iii) uso médio de memória (em GB).

Como se observa na Figura 19, os métodos hierárquicos, apesar do ganho expressivo em F1-*macro*, demandam recursos significativamente superiores aos dos modelos de síntese e truncamento. Essa tendência é visível na amplitude vertical dos gráficos, enquanto as variações entre os métodos de síntese permanecem próximas ao fator 2× do *baseline*. O ToBERT apresenta crescimento acentuado, ultrapassando uma ordem de grandeza em tempo de treinamento e de inferência. O uso de memória segue o mesmo padrão, evidenciando o impacto cumulativo da agregação hierárquica de representações no consumo de GPU.

Para traduzir esses valores relativos em medidas absolutas, na Tabela 14 são detalhados os tempos médios de execução e uso de memória convertidos para unidades compreensíveis.

Tabela 14 – Comparativo dos tempos de treinamento, de inferência e de uso de memória dos modelos avaliados. Os valores foram convertidos para minutos e horas, quando aplicável.

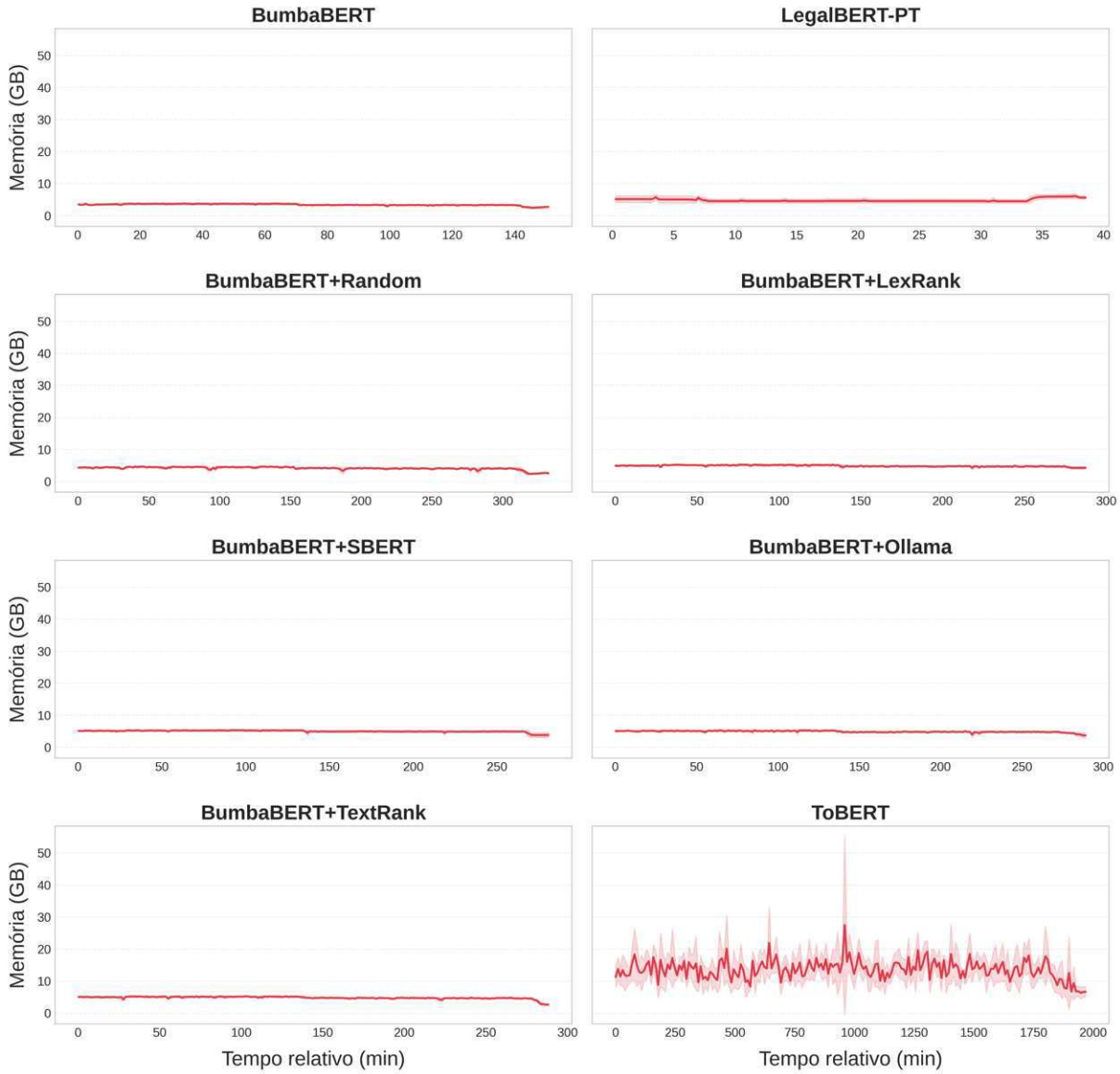
Modelo	**Treino (época)**	**Inferência (época)**	**Memória (GB)**
BumbaBERT	14,13 min	0,92 min	3,5
BumbaBERT+SBERT	26,86 min	1,97 min	5,1
BumbaBERT+LexRank	27,59 min	2,01 min	4,9
BumbaBERT+LLaMA	27,70 min	2,02 min	5,0
BumbaBERT+TextRank	27,81 min	2,10 min	4,9
BumbaBERT+Random	31,74 min	4,46 min	4,2
ToBERT	3h12min	15,05 min	14,9 (pico $\approx$ 118)

Fonte: Elaborado pela autora.

Os resultados confirmam que o BumbaBERT obteve o menor tempo por época de treinamento, totalizando 14 min 13 s, e 55 s de inferência, confirmando sua viabilidade operacional para cenários de alto volume, como triagens automáticas em larga escala. Em contraste, o ToBERT, que combina codificação hierárquica de *chunks* com um *Transformer* de agregação global, apresentou o maior custo computacional, com cerca de 3h 12 min de treinamento por época de treinamento e 15,05 min de inferência, além de um pico de consumo de memória de aproximadamente 118 GB, dez vezes superior ao do modelo de referência.

Os métodos de síntese treinaram entre 26 e 28 min e inferiram em 2 min, mantendo consumo de memória abaixo de 5 GB. Já o BumbaBERT+Random, embora marginalmente mais preciso que o *baseline*, exigiu aproximadamente 32 min de treinamento e 4,5 min de inferência, sendo, portanto, 2,25 vezes mais lento que o BumbaBERT na etapa de treino. A variação no consumo de memória ao longo do tempo reflete a complexidade e a estratégia de processamento de cada arquitetura, como ilustrado na Figura 20. É importante ressaltar que os tempos reportados consideram o ciclo completo de treinamento, compreendendo as etapas de *fine-tuning* do modelo base e o treinamento da camada de classificação.

Figura 20 – Evolução temporal de memória por modelo

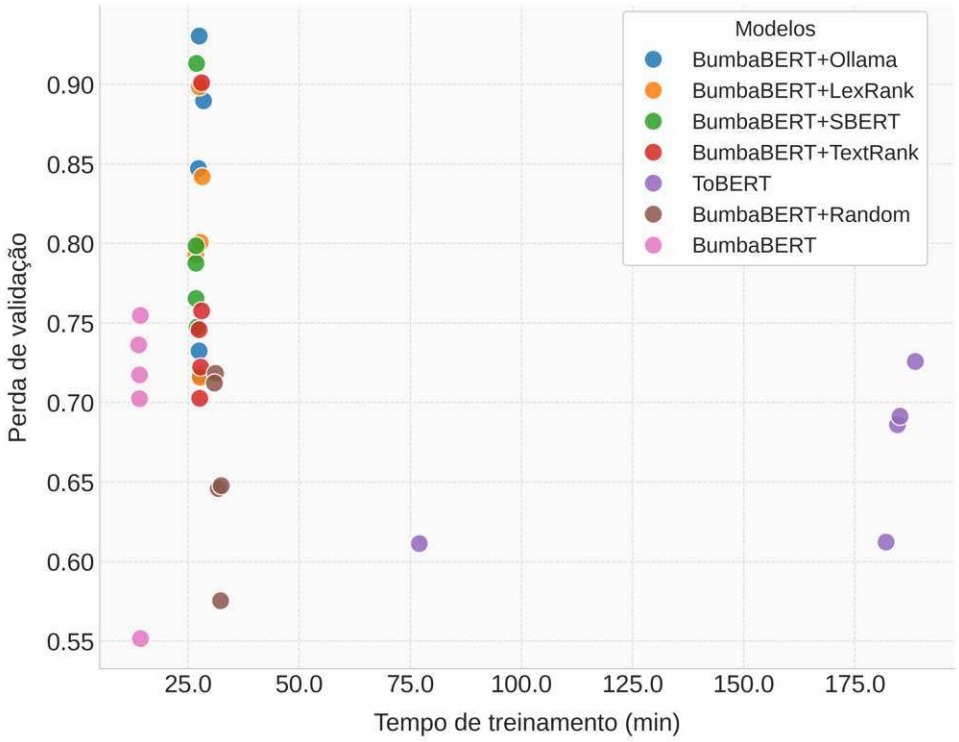


Fonte: Elaborado pela autora

Modelos de síntese e truncamento simples, como o BumbaBERT, mantêm um uso de memória estável e baixo. Em contrapartida, modelos que introduzem a codificação de documentos longos apresentam picos ao longo do tempo. O que reforça a natureza experimental e de alto custo dos métodos hierárquicos, como o ToBERT, que, apesar de sua superioridade em precisão, impõe restrições sérias de hardware para a implantação.

A análise de eficiência computacional deve considerar não apenas o consumo de memória, mas também a velocidade de convergência durante o treinamento. Nesse sentido, na Figura 21 é apresentada a evolução da perda de validação em função do tempo de treinamento para cada modelo, permitindo avaliar o equilíbrio entre custo temporal e qualidade do ajuste.

Figura 21 – Relação entre tempo de treinamento e perda de validação por modelo



Fonte: Elaborado pela autora

A partir da análise da Figura, é visto que os modelos de síntese de conteúdo e o BumbaBERT apresentam convergência rápida, atingindo perdas de validação entre 0,70 e 0,90 em tempos inferiores a 30 min. Em contraste, o ToBERT demonstra padrão de convergência significativamente mais lento, com perda de validação ainda acima de 0,60 após 75 min de treinamento e apenas aproximando-se de 0,70 após 192 min, o que representa aumento de ordem de grandeza no tempo necessário para a convergência comparável.

Essa diferença ilustra o limiar entre o desempenho e a eficiência, evidenciado pela correlação direta entre menor perda de validação e maior custo computacional. Tal análise sugere que, embora o ToBERT eventualmente alcance desempenho competitivo, seu custo temporal de convergência inviabiliza sua aplicação em cenários que demandam ciclos iterativos rápidos de experimentação ou retreinamento frequentes. Assim, uma vez identificados os padrões de desempenho e eficiência, a próxima Seção buscou verificar se as diferenças observadas entre os modelos são estatisticamente significativas.

6.4 Validação estatística das diferenças

Para assegurar que as diferenças de desempenho observadas entre os modelos não foram produto de flutuações amostrais, conduziu-se uma análise estatística inferencial,

conforme fundamentado na Subseção 3.3.2. A análise baseou-se nos resultados de acurácia, precisão, revocação e F1-macro obtidos em cada um dos 5 *folds* da validação cruzada para os oito modelos, totalizando 40 observações pareadas.

Para cada uma das cinco métricas avaliadas, foram verificadas a normalidade das distribuições por meio do teste de Shapiro-Wilk, com $\alpha_{ajustado} = 0,05/8 = 0,00625$) e a homogeneidade de variâncias por meio do teste de Bartlett, com $\alpha = 0,05$. Em todos os casos, as premissas foram satisfeitas. Nenhuma das 40 populações testadas rejeitou a hipótese nula de normalidade, e todos os cinco testes de homogeneidade não rejeitaram a hipótese nula, indicando que as variâncias dos modelos podem ser consideradas estatisticamente compatíveis.

Esse resultado não é trivial, pois métricas de avaliação limitadas ao intervalo [0,1], como as utilizadas neste estudo, frequentemente violam premissas de normalidade quando calculadas sobre amostras relativamente pequenas, especialmente se os valores se concentram próximos aos extremos do intervalo (DIETTERICH, 1998; STONE, 1974). A normalidade observada sugere que os desempenhos dos modelos, embora variáveis entre *folds*, seguem padrão de distribuição gaussiana centrada em suas respectivas médias populacionais, comportamento esperado quando a fonte de variação é predominantemente estocástica por composição aleatória dos *folds* e não sistemática com instabilidade do modelo (DEMŠAR, 2006).

A escolha pela ANOVA com medidas repetidas foi precedida da verificação de seus pressupostos. O teste de Shapiro-Wilk não indicou violação grave da normalidade nas distribuições de desempenho por *fold* (p > 0,05 para a maioria dos modelos), e o teste de Levene confirmou a homogeneidade das variâncias entre os grupos (p = 0,094). Ademais, a natureza pareada dos dados – em que os mesmos 5 *folds* foram aplicados a todos os modelos – atende ao pressuposto de dependência das medidas, tornando a ANOVA um método robusto e adequado para este contexto (DEMŠAR, 2006; KOEHN, 2009). Os resultados da ANOVA, sumarizados na Tabela 15, indicam um efeito principal estatisticamente significativo do modelo (fator intra-sujeitos) para todas as métricas (p < 0,001), permitindo rejeitar a hipótese nula de igualdade de desempenho entre os oito métodos avaliados.

Tabela 15 – Resultados dos testes ANOVA com medidas repetidas por métrica de desempenho

Métrica	**Teste**	***p*-valor**	**Resultado**
F1-*macro*	ANOVA + Tukey HSD	0,008	Diferenças significativas
Acurácia	ANOVA	0,076	*Sem diferenças*
F1-ponderado	ANOVA	0,078	*Sem diferenças*
Precisão	ANOVA	0,199	*Sem diferenças*
Revocação	ANOVA + Tukey HSD	0,018	Diferenças significativas

Fonte: Elaborada pela autora. Nível de significância $\alpha = 0,05$. Teste *post-hoc* aplicado apenas quando ANOVA rejeita H_0.

Foram detectadas diferenças estatisticamente significativas com $p < 0,05$ em duas das cinco métricas: F1-*score* macro com $p = 0,008$ e a Revocação com $p = 0,018$. Para Acurácia o valor foi $p = 0,076$), F1-*score* ponderado com $p = 0,078$ e a Precisão com $p = 0,199$, sendo assim, não foi possível rejeitar a hipótese nula. Diante disso, esses resultados indicam que, embora os modelos apresentem diferenças em sua capacidade de equilibrar precisão e revocação, evidenciadas no F1-*macro*, os testes não detectaram diferenças significativas no desempenho geral de acertos (Acurácia), no balanceamento ponderado por classes (F1-ponderado), ou no controle de falsos positivos (Precisão).

Dado que o F1-*macro* equilibra precisão e revocação sem viés em favor de classes majoritárias, a análise concentrou-se nessa métrica, sendo, portanto, a métrica mais informativa para conjuntos desbalanceados. Na Tabela 16 pode-se ver as estatísticas descritivas para a métrica F1-macro, que serve de referência central para as comparações seguintes, dada a sua natureza balanceada para problemas multiclasse.

Tabela 16 – Estatísticas descritivas para F1-*macro*

Modelo	**Média**	**DP**	**IC 95%**	**Cohen's d**	**Magnitude**
ToBERT	0,790	0,020	[0,771, 0,809]	—	(referência)
BumbaBERT+Random	0,771	0,014	[0,752, 0,790]	0,63	médio
BumbaBERT	0,768	0,025	[0,749, 0,788]	0,38	pequeno
LegalBERT	0,768	0,022	[0,749, 0,787]	0,37	pequeno
BumbaBERT+LLaMA	0,763	0,016	[0,743, 0,782]	0,17	negligível
BumbaBERT+SBERT	0,763	0,015	[0,743, 0,782]	0,16	negligível
BumbaBERT+LexRank	0,759	0,010	[0,740, 0,778]	0,03	negligível
BumbaBERT+TextRank	0,759	0,022	[0,740, 0,779]	0,00	negligível

Fonte: Elaborada pela autora. DP = desvio-padrão; IC = intervalo de confiança (95%); Cohen's d calculado como $|M_{ToBERT} - M_{modelo}|/SD_{pooled}$. Magnitude interpretada segundo convenção de Cohen (1988): negligível ($d < 0,2$), pequeno ($0,2 \leq d < 0,5$), médio ($0,5 \leq d < 0,8$), grande ($d \geq 0,8$).

Para identificar especificamente quais pares de modelos diferiam entre si, realizou-se

o teste post-hoc de Tukey HSD. Os resultados, sumarizados na Tabela 17, demonstram a formação de três grupos de desempenho estatisticamente distintos ($\alpha = 0{,}05$).

Tabela 17 – Agrupamento estatístico dos modelos pelo teste de Tukey HSD (métrica: F1-Macro)

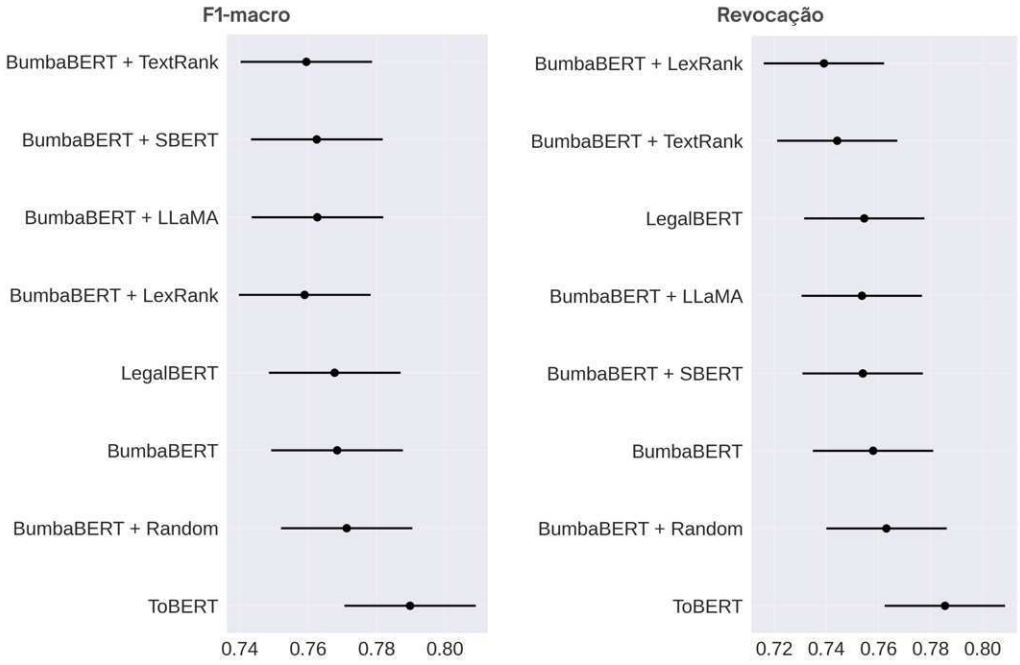
Modelo	**Média F1-Macro**	**Agrupamento**
ToBERT	0,790	A
BumbaBERT+Random	0,771	A B
BumbaBERT (baseline)	0,768	B
LegalBERT	0,768	B
BumbaBERT+LLaMA	0,763	C
BumbaBERT+SBERT	0,763	C
BumbaBERT+LexRank	0,759	C
BumbaBERT+TextRank	0,760	C

Fonte: Elaborada pela autora. Nota: Modelos que compartilham a mesma letra no agrupamento não apresentam diferenças estatisticamente significativas entre si ($p > 0{,}05$).

O ToBERT ocupa posição isolada no topo, com média de 0,790 e intervalo de confiança de [0,772, 0,808], sem sobreposição com os demais modelos. O segundo grupo, formado por BumbaBERT+Random com 0,771, BumbaBERT *baseline* com 0,769 e o LegalBERT-PT com 0,768, apresentou sobreposição parcial dos intervalos de confiança, indicando que não foram detectadas diferenças estatisticamente significativas entre eles, embora todos diferenciem-se do ToBERT. O terceiro grupo, composto pelos quatro métodos de síntese algorítmica com médias entre 0,759 e 0,763, também mostrou equivalência interna, mas difere significativamente do ToBERT.

Na Figura 22 pode-se visualizar esses agrupamentos por meio dos gráficos de intervalos de confiança ordenados. Nessa representação, modelos cujos intervalos de confiança se sobrepõem não apresentam diferenças estatisticamente significativas.

Figura 22 – Intervalos de confiança (95%) para F1-*score* e Revocação por modelo



Fonte: Elaborada pela autora. Os pontos representam as médias e as barras horizontais os intervalos de confiança de 95%.

Pode-se verificar na Figura 22, que apresenta os intervalos de confiança para F1-*score* e Revocação, que a análise *post-hoc* permite as seguintes conclusões:

- O modelo hierárquico (ToBERT) formou um grupo de desempenho estatisticamente superior (Grupo A), sendo significativamente melhor ($p < 0{,}05$) que todos os outros métodos, com exceção do BumbaBERT+Random, com o qual dividiu o agrupamento 'A B';
- Os *baselines* BumbaBERT e LegalBERT-PT, juntamente com os métodos BumbaBERT+LLaMA e BumbaBERT+SBERT, formaram o Grupo B, apresentando equivalência estatística entre si, sem diferenças significativas detectadas pelo teste de Tukey HSD ($p > 0{,}05$);
- Os métodos de síntese algorítmica BumbaBERT+TextRank e BumbaBERT+LexRank formaram o grupo de desempenho inferior (Grupo C), sendo estatisticamente indistinguíveis entre si, mas significativamente inferiores ao ToBERT ($p < 0{,}05$).

Desta forma, a validação estatística corrobora inequivocamente as conclusões da análise comparativa. A superioridade do modelo hierárquico (ToBERT) não é apenas perceptível nas médias de desempenho, mas é estatisticamente significativa, solidificando a premissa central deste estudo de que a preservação da estrutura argumentativa integral é fundamental para uma classificação confiável de documentos jurídicos longos. As diferenças

práticas entre os grupos de modelos, portanto, são robustas e não atribuíveis ao acaso. Esses achados fundamentam a discussão integrada apresentada na próxima Seção, que articula desempenho, eficiência e aplicabilidade prática no contexto institucional do acordo UEMA–TJMA.

6.5 Discussão dos resultados e implicações práticas

Os resultados reportados nas seções anteriores, quando interpretados à luz da literatura revisada no Capítulo 4 e dos objetivos específicos estabelecidos no Capítulo 1 levaram a três achados principais que merecem discussão maior, a saber: a superioridade do processamento hierárquico completo sobre métodos de síntese de conteúdo; a ineficácia relativa de algoritmos de seleção de sentenças neste domínio específico; a necessidade de equilibrar desempenho preditivo e viabilidade operacional para adoção prática em sistemas judiciais automatizados.

6.5.1 Preservação da coerência argumentativa

A superioridade estatística e prática do ToBERT sobre todos os métodos de síntese de conteúdo avaliados com ganho de 1,9 a 3,1 pontos percentuais em F1-*macro* demonstra que, no domínio jurídico brasileiro, a preservação integral da estrutura textual é determinante para o desempenho do modelo. Tal ganho se deve à capacidade da arquitetura de decomposição-recomposição de construir uma representação de documento que integra o contexto local de sentenças e parágrafos ao contexto global, sendo o documento inteiro, conforme proposto por Pappagari et al. (2019).

No processo judicial, a tese de um IRDR não é definida por uma única sentença, mas pela interação entre o relato fático no início, a fundamentação jurídica no meio e os pedidos específicos no final (Brasil, 2015). Essa interdependência entre seções explica por que abordagens que processam o documento integralmente, ainda que de forma segmentada, preservam melhor o raciocínio jurídico distribuído.

Ao dividir a petição em *chunks* e usar um segundo *Transformer*, o codificador de contexto global, para forçar a atenção entre eles, o ToBERT preserva essa coerência argumentativa distribuída (PAPPAGARI et al., 2019; MARINO et al., 2023; CHALKIDIS et al., 2022). A performance inferior de todos os métodos de síntese sugere que a perda de informação contextual nas seções descartadas, ou seja, a grande maioria do texto, não pode ser compensada, mesmo quando o resumo é gerado por algoritmos avançados ou LLMs como o LLaMA (BELTAGY; PETERS; COHAN, 2020; AKTER et al., 2025).

Essa baixa performance dos métodos de síntese reforça a ideia de que redução textual implica perda de inferência jurídica (JAIN; BORAH; BISWAS, 2021). Mesmo modelos generativos de larga escala, capazes de capturar relações semânticas amplas,

demonstraram limitações para preservar a coerência argumentativa e a estrutura lógica das decisões jurídicas, mesmo quando expostos ao texto integral (PEYKANI et al., 2025; AKTER et al., 2025). Essa constatação tem implicações práticas em aplicações judiciais, sintetizar demais pode ser arriscado, pois compromete a rastreabilidade e a justificabilidade da decisão e os princípios centrais de transparência e ética previstos na Resolução n^o 332/2020 do CNJ (CNJ, 2020).

6.5.2 Limitações dos métodos de relevância extrativa

O desempenho superior do BumbaBERT+Random em relação aos algoritmos de seleção TextRank, LexRank e SBERT é um achado contraintuitivo, mas revelador. O TextRank e o LexRank buscam a centralidade da informação , e o SBERT busca a proximidade semântica com a média vetorial. Contudo, no contexto jurídico, a redundância é intencional, uma vez que, advogados frequentemente reiteram argumentos, reproduzem longos trechos de leis, e transcrevem jurisprudência para reforçar sua argumentação (JR; BRAGA; OLIVEIRA, 2010).

Consequentemente, algoritmos baseados em frequência ou similaridade tendem a confundir recorrência com relevância, privilegiando trechos redundantes em detrimento de trechos mais distintos (GIARELIS; MASTROKOSTAS; KARACAPILIDIS, 2023). O método aleatório, ao introduzir diversidade estrutural, provavelmente consegue preservar exemplos de diferentes seções, como fatos, fundamentos e pedidos, o que favorece a generalização do classificador. Essa descoberta apoia os achados de Park, Vyas e Shah (2022), que mostraram que a diversidade lexical pode ser mais útil do que a relevância estimada em contextos de repetição temática.

Em termos teóricos, esse resultado reforça a hipótese de que o critério de relevância precisa ser redesenhado para textos jurídicos, incorporando métricas de argumentatividade e de hierarquia retórica, oferecendo direcionamentos para possíveis abordagens híbridas, combinando seleção supervisionada com pesos baseados em estrutura lógica, podendo, assim, oferecer ganhos sem comprometer a eficiência (KUŞ; ACI, 2024; LI et al., 2024).

6.5.3 Viabilidade operacional

Embora o ToBERT tenha demonstrado desempenho em todas as métricas, sua implementação em um ambiente real, como o TJMA, não está isenta de desafios. O modelo requer o processamento sequencial de múltiplos *chunks* de 200 *tokens* pelo BumbaBERT, seguido de uma camada de agregação *Transformer*, o que resulta em um tempo de inferência maior do que o obtido com o truncamento simples.

Para um sistema como o Robô Maria Firmina que prioriza a celeridade processual, o dilema entre o ganho de 2,2 p.p. em F1-*score* macro e o custo computacional precisa

ser ponderado, ainda assim esse custo não invalida o modelo, mas define seu nicho de aplicação. Os métodos de síntese são mais rápidos e mais leves, mas o risco de classificar uma petição de forma incorreta (Falso Negativo) é alto, o que no direito é um erro custoso. A adoção do ToBERT, portanto, implica em um investimento em infraestrutura com melhores GPUs ou TPUs, ou em um maior tempo de latência por petição, um ônus que se justifica apenas se a precisão for a prioridade máxima, como é o caso na identificação de precedentes vinculantes como as IRDRs.

Essa reflexão remete ao quarto objetivo específico deste estudo, que é avaliar o equilíbrio entre desempenho e escalabilidade. Esses achados sugerem que tribunais com restrições de hardware podem optar por uma arquitetura híbrida de uso seletivo, como o ToBERT, aplicado apenas em casos de alta incerteza probabilística, conforme o limiar de confiança do classificador. Essa proposta está em linha com as diretrizes da Justiça 4.0 e com práticas de uso responsável de IA no setor público, em que o desempenho deve ser ponderado com o custo, a transparência e a auditabilidade (CNJ, 2024).

Do ponto de vista de infraestrutura, a implementação prática do ToBERT demanda GPUs com memória superior a 16 GB ou execução distribuída. Embora técnicas de otimização, como *gradient checkpointing*, *mixed precision training* e quantização, possam ser alternativas para reduzir o consumo de memória, a validação empírica dessas adaptações no contexto de documentos jurídicos longos permanece como um trabalho futuro necessário (ELOUARGUI et al., 2023; CHUN et al., 2025). A adoção do ToBERT em produção demandaria, portanto, investimento em infraestrutura especializada (GPUs enterprise com 40-80 GB de VRAM, como as NVIDIA A100 ou H100) ou em redesenho arquitetural para processamento em lotes menores, com maior latência.

A análise integrada revela que arquiteturas hierárquicas superam métodos de síntese por meio da preservação de uma coerência argumentativa distribuída, validando a hipótese principal de que o desempenho depende da forma, não da quantidade, de processamento contextual. Contudo, o custo computacional associado demanda arquiteturas híbridas de uso seletivo para viabilizar a implantação prática em tribunais com restrições de infraestrutura, equilibrando precisão, eficiência e conformidade com as diretrizes de governança da Justiça 4.0 (CNJ, 2024).

6.6 Consolidação dos achados e implicações gerais

Em síntese, os resultados demonstraram que modelos hierárquicos, como o ToBERT, preservam a coerência argumentativa imprescindível ao raciocínio jurídico, superando métodos de síntese em desempenho e confiabilidade, ainda que exijam recursos computacionais de alto custo. A natureza redundante dos textos jurídicos desafia abordagens baseadas na relevância lexical, evidenciando a necessidade de novos critérios de seleção

contextual capazes de capturar a estrutura lógica e argumentativa das petições.

Dessa forma, esses achados indicam que o desenvolvimento de sistemas de IA para o judiciário deve priorizar modelos capazes de representar integralmente a estrutura argumentativa dos documentos. Além disso, sinalizam oportunidades de pesquisa voltadas à otimização de arquiteturas hierárquicas e à definição de métricas que conciliem precisão e eficiência, permitindo maior equilíbrio entre desempenho e viabilidade operacional. Por fim, a adoção prática dessas soluções deve considerar a compatibilidade entre desempenho e infraestrutura disponível, assegurando que o uso de IA no judiciário brasileiro seja tecnicamente eficaz, socialmente responsável e alinhado aos princípios da Justiça 4.0

7 Considerações finais

Um dos desafios mais prementes do sistema judiciário brasileiro reside no processamento eficiente de documentos jurídicos longos, cuja complexidade estrutural e extensão textual dificultam a automação de tarefas analíticas e decisórias. No contexto do acordo de cooperação técnica entre a UEMA e o TJMA, essa necessidade manifesta-se de forma concreta no desenvolvimento do *framework* Robô Maria Firmina, sistema de apoio à decisão judicial que visa ampliar a eficiência e a transparência na gestão de processos repetitivos. A presente dissertação insere-se como componente específico desse ecossistema tecnológico mais amplo, focando na etapa crítica de classificação automatizada de petições iniciais enquadráveis em temas de IRDR. Situada na interseção entre ciência da computação e direito, esta pesquisa propôs, implementou e avaliou métodos baseados em modelos de linguagem pré-treinados no domínio jurídico brasileiro, explorando adaptações arquiteturais que permitem superar as limitações de contexto dos modelos baseados em *Transformer*. Assim, o objetivo geral consistiu em investigar como as arquiteturas de PLN podem ser adaptadas para capturar a estrutura argumentativa de textos extensos, equilibrando desempenho, custo computacional e aplicabilidade institucional, contribuindo não apenas para o avanço científico, mas também para a concretização de uma ferramenta com impacto social direto no acesso à justiça.

A condução do projeto seguiu os princípios da DSR, combinando experimentação empírica e fundamentação teórica para o desenvolvimento de artefatos computacionais reprodutíveis. Foram realizados 40 experimentos envolvendo 8 modelos e 5 *folds* de validação cruzada, analisando o desempenho e a eficiência computacional sob diferentes estratégias de processamento. Essa abordagem multidimensional permitiu compreender não apenas "qual modelo é melhor", mas "por que e em quais condições" determinadas arquiteturas se mostram mais adequadas ao contexto jurídico.

7.1 Síntese dos principais achados

Os resultados empíricos confirmam a hipótese inicial de que a preservação do contexto global é determinante para o desempenho em tarefas jurídicas complexas. Os principais achados estão listados a seguir.

- **Superioridade de representação hierárquica:** A arquitetura hierárquica ToBERT superou todas as abordagens comparadas, alcançando F1-*macro* de 0,790 e demonstrando superioridade estatisticamente significativa em F1-*macro* e na revocação. Essa vantagem decorre da capacidade de integrar informações locais e globais

ao longo de todo o documento, preservando a coerência argumentativa essencial ao raciocínio jurídico. Em contraste, modelos baseados em síntese de conteúdo apresentaram desempenho inferior, evidenciando que a redução textual, ainda que orientada por algoritmos de relevância semântica ou por grandes modelos de linguagem, tende a eliminar trechos cruciais da fundamentação jurídica;

- **Limitações dos métodos de síntese de conteúdo:** Os cinco métodos baseados em compressão textual, nomeadamente: LLaMA, SBERT, LexRank, TextRank e Random, não apresentaram diferenças estatisticamente significativas entre si, com F1-*macro* variando apenas entre 0,759 e 0,771, sem diferenças significativas, mesmo entre a seleção aleatória e a sumarização supervisionada por LLM. Esse resultado reforça que, no âmbito jurídico, a redundância textual é intencional e estrutural, não constituindo ruído passível de eliminação. A análise sugere a necessidade de desenvolver abordagens que representem a estrutura lógica e retórica dos textos, e não apenas seu conteúdo lexical superficial;

- **Equilíbrio entre desempenho e viabilidade operacional:** O ToBERT, embora superior em precisão, apresentou custo computacional 13,6 vezes maior do que o BumbaBERT, com 192 min *vs.* 14 min por época, e consumo de memória 10 vezes maior, 118 GB *vs.* 11,8 GB, respectivamente. Diante desse *trade-off*, propôs-se uma arquitetura híbrida de uso seletivo, na qual o modelo hierárquico é aplicado apenas em casos de alta incerteza probabilística, conforme o limiar de confiança do classificador. Essa estratégia harmoniza precisão, custo e auditabilidade, alinhando-se às diretrizes de governança tecnológica da Justiça 4.0 (CNJ, 2024) e à viabilidade de implantação no contexto do acordo UEMA–TJMA.

Esses achados apresentados atendem ao cumprimento do objetivo geral desta dissertação, qual seja, "comparar e avaliar métodos baseados em modelos de linguagem para o processamento eficiente de documentos jurídicos longos, buscando superar as limitações inerentes aos modelos de arquitetura *Transformer*". Assim, os resultados validam empiricamente a hipótese de que o desempenho dos modelos de PLN em textos jurídicos depende não apenas da quantidade de informação processada, mas também da forma como o contexto é preservado e hierarquicamente representado.

7.2 Contribuições tecnológicas, científicas e institucionais

Do ponto de vista tecnológico, esta dissertação integra-se ao desenvolvimento do Robô Maria Firmina, *framework* de apoio à decisão judicial em construção pelo TJMA em parceria com a UEMA. Especificamente, o presente trabalho contribui para o módulo de triagem e classificação de petições iniciais, funcionalidade essencial para identificar

automaticamente temas de IRDR e suspender processos correlatos até o julgamento do precedente aplicável. O desenvolvimento, treinamento e avaliação dos modelos resultaram em um artefato computacional reprodutível, cuja integração ao ecossistema do Maria Firmina está prevista para a próxima fase do projeto, condicionada à validação em ambiente de produção e aos ajustes de infraestrutura identificados neste estudo e o registro do programa de computador correspondente ao *pipeline* desenvolvido junto ao INPI, de forma a assegurar a proteção intelectual e a reprodutibilidade do artefato criado, consolidando o produto técnico do projeto como referência para futuras iniciativas de automação e inovação tecnológica no âmbito do Poder Judiciário brasileiro.

Além dos resultados técnicos, as investigações se articularam com um conjunto de produções científicas que complementam e fortalecem os achados desta dissertação, demonstrando a maturidade e a integração do percurso de pesquisa. Durante o período de execução da dissertação, foram produzidos artigos diretamente e indiretamente relacionados à dissertação, em diferentes estágios de publicação. Essas produções evidenciam a consolidação do percurso científico e a integração entre a pesquisa experimental e o debate acadêmico sobre a aplicação responsável de IA nos setores público e jurídico.

Institucionalmente, o estudo consolida um modelo de parceria entre a universidade e o poder público. A cooperação UEMA–TJMA exemplifica como a pesquisa aplicada pode orientar políticas de transformação digital no setor judiciário, ao mesmo tempo em que gera benefícios sociais concretos, maior celeridade processual, redução de retrabalho e democratização do acesso à justiça.

Produções diretas derivadas da dissertação

1. ***The Artificial Intelligence Integration in the Brazilian Legal Sector: A Systematic Review*** - publicado nos anais do SBSI 2025[1]. Esse artigo apresentou a revisão sistemática inicialmente dissertação, mapeando 90 projetos de IA no sistema judicial brasileiro entre 2020 e 2024. No estudo, fundamentou-se o problema de pesquisa ao evidenciar a ausência de documentação técnica e a necessidade de metodologias transparentes para a avaliação de impacto. Com base na *Diffusion of Innovations Theory*, identificou-se que a maioria das iniciativas de IA concentra-se em classificação textual e automação processual, mas carece de padronização e de governança de dados. Essa revisão foi importante para contextualizar a contribuição dos modelos propostos como resposta à lacuna de soluções viáveis e reprodutíveis no contexto jurídico nacional;

2. ***Hybrid Summarization for Brazilian Judicial Decisions*** — submetido ao SBSI 2026. O estudo apresenta uma extensão experimental do presente trabalho,

[1] <https://doi.org/10.5753/sbsi.2025.246589>

aplicando os princípios de decomposição e síntese híbrida à tarefa de sumarização das decisões do STF. Propondo um *pipeline* de sumarização híbrido, integrando métodos extrativos e abstrativos fundamentado na *Task-Technology Fit* (TTF) e na DSR. Os resultados demonstraram que abordagens híbridas ampliam a aplicabilidade de modelos especializados, equilibrando a precisão técnica e a coesão textual. O manuscrito encontra-se submetido, e sua versão integral está incluída nos Apêndice A e Anexo A dissertação, acompanhada do comprovante de submissão;

3. ***Fine-Tuning BumbaBERT for Long Legal Document Classification in the Brazilian Judiciary*** - a ser submetido a um periódico internacional, o qual consolida os principais resultados desta dissertação, ampliando a discussão sobre a avaliação de modelos hierárquicos e suas implicações para o uso ético e eficiente da IA no Judiciário brasileiro, discutindo o papel da transparência e da auditabilidade em ambientes públicos sensíveis.

Produções indiretas e colaborações internacionais

Durante o período de execução desta dissertação, a autora também participou do Programa Abdias Nascimento (Edital CAPES n.º 16/2023), realizando estágio acadêmico de pesquisa na *Cooperative State University of Saxony*, conhecida em alemão como *Duale Hochschule Sachsen* (DHSN), localizada em Riesa, Alemanha, entre novembro de 2024 e agosto de 2025. O estágio foi supervisionado pelo Prof. Dr. Olaf Reinhold e integrou o projeto "Estudo Transcultural de Inclusão e Acessibilidade por meio de uma Ferramenta Georreferenciada", desenvolvido em parceria entre a UEMA, a Universidade Federal do Pará (UFPA), a UFOPA, a DHSN e a *Technische Universität Dortmund*. Embora esteja independente do escopo principal desta dissertação, o estágio representou um importante complemento à formação acadêmica e científica da autora, ampliando sua experiência internacional e consolidando parcerias que resultaram em publicações e projetos correlatos nas áreas de acessibilidade, tradução automática e governança de dados. Alguns desses destacados a seguir.

1. O artigo *"Hybrid Approaches for Pneumonia Detection in X-rays: Combining CNNs and ML Classifiers"*[2], publicado na *Conference on Computer Science and Intelligence Systems* (FedCSIS 2025), resultou da disciplina de "Introdução ao *Deep Learning*" do Mestrado Profissional, ministrada pelo Prof. Dr. Omar Andrés C. Cortes na UEMA. O estudo explorou metodologias híbridas de classificação de imagens médicas, cujos aprendizados metodológicos, principalmente quanto ao uso da validação cruzada estratificada, dos testes de significância estatística e da comparação entre arquiteturas de redes neurais, foram posteriormente aplicados ao desenho experimental

[2] <http://dx.doi.org/10.15439/2025F4310>

desta dissertação, subsidiando a avaliação comparativa entre modelos hierárquicos e métodos de síntese de conteúdo para o processamento de documentos jurídicos longos;

2. Em cooperação com a DHSN, foi desenvolvido o artigo *"Managing Customer Data in Data-driven Service Innovation: A Framework of Data Principles"*, atualmente em submissão a um periódico internacional. Esse trabalho propõe um arcabouço conceitual de princípios de governança de dados para inovação orientada por dados (*data-driven service innovation*), contribuindo para a discussão ética e regulatória que também permeia a aplicação de IA no direito;

3. Foi também produzido e submetido o artigo *"Evaluation of Proprietary and Open-Source Machine Translation for Domain-Specific Content on Accessibility"*, voltado à análise comparativa de sistemas de tradução automática. No estudo, foram avaliados APIs comerciais, como Google, DeepL e Microsoft, e modelos abertos, como o LLaMA, o Qwen e o NLLB, segundo métricas de qualidade, custo e privacidade. Os resultados mostraram que soluções abertas podem garantir soberania de dados, embora exijam maior custo computacional, o que configura um dilema análogo ao enfrentado nesta dissertação no contexto do Judiciário. O projeto foi conduzido em cooperação com os professores doutores Olaf Reinhold, Antônio Jacob Junior (UFOPA), Ronaldo Zampolo (UFPA), Simone Silva (UFPA) coordenadora do projeto Abdias, e sob a supervisão do orientador da autora; o desenho experimental foi proposto pela autora, por seu orientador e pelo Prof. Dr. Olaf Reinhold, enquanto os professores Ronaldo Zampolo e Antônio Jacob Junior contribuíram com a revisão de código e do manuscrito.

7.3 Impactos e implicações sociais

A presente dissertação produz impactos tangíveis em múltiplas dimensões da sociedade, transcendendo a contribuição acadêmica para promover transformações práticas no acesso à justiça, na eficiência do serviço público e na democratização do conhecimento tecnológico aplicado ao direito. No eixo da justiça e cidadania, a automação da classificação de petições por tema de IRDR contribui diretamente para reduzir a morosidade processual, um dos maiores entraves ao acesso à justiça no Brasil. A identificação automática de demandas repetitivas permite suspender processos correlatos até o julgamento do precedente aplicável, reduzindo o tempo médio de tramitação e assegurando tratamento isonômico a cidadãos em situações equivalentes. Assim, a tecnologia proposta reforça os princípios de celeridade e segurança jurídica previstos no CPC de 2015 e na Resolução nº 332/2020 do CNJ (Brasil, 2015; CNJ, 2020).

No âmbito do serviço público e da eficiência estatal, a dissertação oferece subsídios técnicos para a modernização do Judiciário no âmbito do programa Justiça 4.0, ao propor um método de avaliação multidimensional que considera desempenho, eficiência computacional e viabilidade de implantação. Essa abordagem orienta a escolha de soluções compatíveis com a infraestrutura de cada tribunal, evitando gastos desnecessários e otimizando o uso de recursos públicos. O modelo metodológico também se mostra replicável, permitindo sua adoção por outros tribunais e fortalecendo a sinergia entre a pesquisa acadêmica e a gestão pública.

E, em ciência, tecnologia e inovação, o trabalho preenche uma lacuna relevante na literatura sobre PLN para o português jurídico, apoiando futuras pesquisas e contribuindo para a formação de recursos humanos especializados em IA aplicada ao direito.

7.4 Limitações e perspectivas futuras

Insta, de antemão, a salientar as limitações e os desafios potenciais identificados ao longo do projeto. A complexidade e a variabilidade estrutural dos documentos jurídicos, somadas às diferenças regionais de redação e terminologia, representam obstáculos à generalização dos modelos propostos. Além disso, as questões éticas associadas ao uso de IA no direito, especialmente no que se refere à transparência, à responsabilização e à privacidade de dados, demandam atenção contínua. Os desafios de integração dos modelos em sistemas judiciais consolidados exigem adaptações graduais, conciliando inovação tecnológica com os fluxos processuais já estabelecidos. Ademais, o intervalo entre a conclusão dos experimentos e a finalização desta dissertação impossibilitou a entrega do artefato computacional ao TJMA para validação em ambiente de produção, o que constitui etapa importante para avaliar a viabilidade operacional e a aceitação institucional da solução proposta, permanecendo como desdobramento necessário do acordo de cooperação UEMA–TJMA.

Embora este trabalho tenha cumprido com seus objetivos, reconhece-se que os resultados constituem contribuição inicial, passível de aprimoramentos em estudos futuros. Diante disso perspectivas futuras, vislumbra-se aprimorar o modelo hierárquico por meio de técnicas de otimização, como *gradient checkpointing*, quantização e *knowledge distillation*, a fim de reduzir o custo computacional e ampliar sua viabilidade de implantação em ambientes de infraestrutura limitada. A expansão do *corpus* para incluir decisões e petições de outros tribunais brasileiros é igualmente necessária para testar a robustez e a capacidade de generalização do modelo diante de diferentes estilos redacionais e realidades jurídicas. Além de tudo, a incorporação de mecanismos de explicabilidade, como *Explainable AI* (XAI), devem permitir auditoria transparente das decisões automatizadas. A extensão para outras tarefas jurídicas como sumarização de decisões, extração de entidades, e

reconhecimento de entidades nomeadas permitiria avaliar a transferibilidade e consolidar ecossistema integrado de automação judicial.

A estratégia de versionamento e adequação contínua do modelo constitui aspecto fundamental para sua sustentabilidade de longo prazo. Propõe-se a implementação de um protocolo de retreinamento periódico que incorpore novas petições classificadas, permitindo que o modelo acompanhe a evolução da jurisprudência, a criação de novos temas de IRDR e as mudanças no vocabulário jurídico ao longo do tempo. Esse processo deve ser documentado por meio de controle de versão rigoroso, registrando as alterações no *dataset* de treinamento, os hiperparâmetros ajustados e as métricas de desempenho de cada versão, garantindo rastreabilidade completa e possibilitando auditoria externa. A adoção de práticas de MLOps, incluindo testes automatizados de regressão de desempenho e monitoramento contínuo de *drift* de dados, assegurará que atualizações do modelo não degradem sua precisão em classes já estabilizadas. Adicionalmente, recomenda-se a criação de um comitê consultivo composto por magistrados, servidores do TJMA e membros do acordo de cooperação, responsável por revisar periodicamente o desempenho do sistema, validar as atualizações propostas e garantir que a evolução tecnológica permaneça alinhada às necessidades institucionais e aos princípios éticos de uso responsável de IA no Judiciário.

Para além das melhorias técnicas, é fundamental considerar a adequação do modelo a diferentes contextos jurisdicionais. Embora treinado com dados do TJMA, a arquitetura proposta é suficientemente genérica para ser adaptada a outros tribunais estaduais, desde que retreinada com *corpus* localmente representativo. Essa capacidade de transferência (*transfer learning*) pode ser testada em estudos colaborativos envolvendo múltiplos tribunais, avaliando o quanto o modelo pré-treinado em um estado consegue generalizar para demandas de outras regiões sem necessidade de retreinamento completo. Essa agenda de pesquisa contribuiria para a criação de um modelo nacional de classificação de IRDRs, reduzindo a fragmentação de esforços e promovendo economia de escala no desenvolvimento de soluções de IA para o Judiciário brasileiro.

Por fim, as perspectivas futuras delineadas nesta seção transcendem o aprimoramento técnico incremental, apontando para uma agenda transformadora que posiciona a ciência de dados como instrumento estratégico de democratização do acesso à justiça. A integração dos avanços aqui propostos ao ecossistema do Robô Maria Firmina representa oportunidade concreta de consolidar o TJMA como referência nacional em inovação judicial baseada em IA responsável, ética e auditável. Mais do que automatizar processos, trata-se de criar infraestrutura tecnológica que amplie a capacidade do Estado de garantir direitos fundamentais com celeridade, isonomia e transparência. O legado desta dissertação, portanto, não se limita aos modelos treinados ou às métricas alcançadas, mas materializa-se na construção de um caminho metodológico replicável, cientificamente fundamentado e socialmente comprometido, que pode inspirar e orientar tribunais de todo

o país na jornada de transformação digital do Judiciário brasileiro. Que este trabalho, desenvolvido em parceria entre academia e poder público, sirva como evidência de que a pesquisa aplicada, quando conduzida com rigor metodológico, responsabilidade ética e compromisso com a equidade, pode ser vetor potente de mudança social, fortalecendo a confiança da população nas instituições democráticas e reafirmando o papel da ciência como instrumento de construção de uma sociedade mais justa.

Referências

AGGARWAL, C. C.; ZHAI, C. A survey of text classification algorithms. *Mining text data*, Springer, p. 163–222, 2012. Citado na página 37.

AGUIAR, A.; SILVEIRA, R.; PINHEIRO, V.; FURTADO, V.; NETO, J. A. Text classification in legal documents extracted from lawsuits in brazilian courts. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2021. p. 586–600. Citado 3 vezes nas páginas 65, 72 e 82.

AKTER, M.; ÇANO, E.; WEBER, E.; DOBLER, D.; HABERNAL, I. A comprehensive survey on legal summarization: Challenges and future directions. *arXiv preprint arXiv:2501.17830*, 2025. Citado 2 vezes nas páginas 109 e 110.

ALMEIDA, N. D. de; PINTO, P. A. L. de A. O uso da inteligência artificial como ferramenta de eficiência e acesso à justiça em revisão sistemática da literatura. *Research, Society and Development*, v. 11, n. 11, p. e349111133674–e349111133674, 2022. Citado na página 18.

ANDERLUCCI, L.; GUASTADISEGNI, L.; VIROLI, C. Classifying textual data: shallow, deep and ensemble methods. *arXiv preprint arXiv:1902.07068*, 2019. Citado na página 37.

ARAÚJO, G. S.; CORTES, O. A. C.; REINHOLD, O.; LOBATO, F. M. F. Hybrid approaches for pneumonia detection in x-rays: Combining cnns and ml classifiers. In: BOLANOWSKI, M.; GANZHA, M.; MACIASZEK, L.; PAPRZYCKI, M.; ŚLęZAK, D. (Ed.). *Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2025. (Annals of Computer Science and Information Systems, v. 43), p. 461–470. Disponível em: <http://dx.doi.org/10.15439/2025F4310>. Citado na página 26.

ARAÚJO, G. S.; JUNIOR, A. F. L. J.; SANTANA, E. E. C.; LOBATO, F. M. F. 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: [s.n.], 2025. Citado 4 vezes nas páginas 19, 25, 26 e 32.

ARIOZO, C. R.; DOMINGOS, M. dos S. Um olhar sobre a esfera jurídica: Modelo teórico do gênero textual petição inicial. *REVISTA DIÁLOGO E INTERAÇÃO*, v. 19, n. 1, p. 277–296, 2025. Citado na página 32.

BAVISKAR, D.; AHIRRAO, S.; POTDAR, V.; KOTECHA, K. Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. *Ieee Access*, IEEE, v. 9, p. 72894–72936, 2021. Citado 2 vezes nas páginas 52 e 53.

BELTAGY, I.; PETERS, M. E.; COHAN, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. Citado 7 vezes nas páginas 22, 54, 58, 64, 71, 72 e 109.

BENDER, E. M.; GEBRU, T.; MCMILLAN-MAJOR, A.; SHMITCHELL, S. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the*

2021 ACM Conference on Fairness, Accountability, and Transparency, p. 610–623, 2021. Citado 2 vezes nas páginas 23 e 45.

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JAUVIN, C. A neural probabilistic language model. *Journal of machine learning research*, v. 3, n. Feb, p. 1137–1155, 2003. Citado na página 22.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *The journal of machine learning research*, JMLR. org, v. 13, n. 1, p. 281–305, 2012. Citado na página 47.

BEZERRA, E. V.; ROSÁRIO, P. G. T. T. do; PEREIRA, L. M. et al. Inteligência artificial aplicada ao judiciário: A gestão de precedentes como ferramenta de otimização no tjma por meio da ia maria firmina: Precedent management as an optimization tool in the maranhão state court through the maria firmina ai. *Interfaces Científicas-Humanas e Sociais*, v. 12, n. 3, p. 253–265, 2025. Citado na página 20.

BIEWALD, L. et al. Experiment tracking with weights and biases. 2020. Disponível em: <https://wandb.ai/>. Citado na página 90.

BOMMASANI, R. et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. Citado 2 vezes nas páginas 23 e 45.

BRASIL. *Constituição da República Federativa do Brasil de 1988*. 1988. <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 10 out. 2025. Citado 2 vezes nas páginas 29 e 30.

Brasil. *Lei nº 13.105, de 16 de março de 2015*: Código de processo civil. Brasília, DF: [s.n.], 2015. Diário Oficial da União. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm>. Acesso em: 10 jan. 2025. Citado 7 vezes nas páginas 28, 29, 30, 32, 88, 109 e 117.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A. et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1901, 2020. Citado 2 vezes nas páginas 44 e 97.

CARLINI, N.; TRAMER, F.; WALLACE, E.; JAGIELSKI, M.; HERBERT-VOSS, A.; LEE, K.; ROBERTS, A.; BROWN, T.; SONG, D.; ERLINGSSON, et al. Extracting training data from large language models. In: *30th USENIX Security Symposium (USENIX Security 21)*. [S.l.: s.n.], 2021. p. 2633–2650. Citado 2 vezes nas páginas 23 e 45.

CARMO, F. A. do. *Representações Embeddings Orientadas à Linguagem Jurídica Brasileira*. 84 p. Dissertação (Mestrado em Engenharia da Computação e Sistemas) — Universidade Estadual do Maranhão, São Luís - MA, 2024. Disponível em: <https://repositorio.uema.br/jspui/handle/123456789/3399>. Citado 12 vezes nas páginas 19, 20, 21, 23, 42, 43, 44, 70, 76, 77, 83 e 99.

CARVALHO, A. M. X. d.; SOUZA, M. R. d.; MARQUES, T. B.; SOUZA, D. L. d.; SOUZA, E. F. M. d. Familywise type i error of anova and anova on ranks in factorial experiments. *Ciência Rural*, SciELO Brasil, v. 53, p. e20220146, 2022. Citado 2 vezes nas páginas 61 e 63.

CASOLA, S.; LAURIOLA, I.; LAVELLI, A. Pre-trained transformers: an empirical comparison. *Machine Learning with Applications*, Elsevier, v. 9, p. 100334, 2022. Citado na página 59.

CAVUS, N.; GOKSU, M.; OKTEKIN, B. Real-time fake news detection in online social networks: Fandc cloud-based system. *Scientific Reports*, Nature Publishing Group UK London, v. 14, n. 1, p. 25954, 2024. Citado na página 33.

CHALKIDIS, I.; ANDROUTSOPOULOS, I.; MICHOS, A. Large-scale multi-label text classification on eu legislation. *arXiv preprint arXiv:1906.02192*, 2019. Citado 3 vezes nas páginas 60, 89 e 90.

CHALKIDIS, I.; DAI, X.; FERGADIOTIS, M.; MALAKASIOTIS, P.; ELLIOTT, D. An exploration of hierarchical attention transformers for efficient long document classification. *arXiv preprint arXiv:2210.05529*, 2022. Citado na página 109.

CHALKIDIS, I.; FERGADIOTIS, M.; MALAKASIOTIS, P.; ALETRAS, N.; ANDROUTSOPOULOS, I. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020. Citado 5 vezes nas páginas 23, 60, 65, 70 e 80.

CHEN, Y.-C.; BANSAL, M. Fast abstractive summarization with reinforce-selected sentence rewriting. In: *Proceedings of ACL 2018*. [S.l.: s.n.], 2018. p. 675–686. Citado na página 70.

CHI, W. W.; TANG, T. Y.; SALLEH, N. M.; MUKRED, M.; ALSALMAN, H.; ZOHAIB, M. Data augmentation with semantic enrichment for deep learning invoice text classification. *IEEE Access*, IEEE, v. 12, p. 57326–57344, 2024. Citado na página 33.

CHOROMANSKI, K. M.; LIKHOSHERSTOV, V.; DOHAN, D.; SONG, X.; GANE, A.; SARLOS, T.; HAWKINS, P.; DAVIS, J. Q.; MOHIUDDIN, A.; KAISER, L. et al. Rethinking attention with performers. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2021. Citado na página 54.

CHUN, H.; YAZHU, R.; NUAN, Q.; BAO, Y. Towards efficient transformers for large-scale applications. 2025. Citado na página 111.

CNJ. *Resolução n. 332, de 21 de agosto de 2020*. 2020. *Diário da Justiça do Conselho Nacional de Justiça*, Brasília, DF. Acesso em: 2 set. 2024. Disponível em: <https://atos.cnj.jus.br/files/original191707202008255f4563b35f8e8.pdf>. Citado 3 vezes nas páginas 18, 110 e 117.

CNJ. 2024. Painel Estatísticas do Poder Judiciário. Disponível em: <https://justica-em-numeros.cnj.jus.br/painel-estatisticas/>. Acesso em: 26 ago. 2024. Citado na página 18.

CNJ. *Justiça 4.0*. Brasília: CNJ, 2024. <https://www.cnj.jus.br/tecnologia-da-informacao-e-comunicacao/justica-4-0/>. Acesso em: 27 ago. 2024. Citado na página 19.

CNJ. *Pesquisa uso de inteligência artificial (IA) no Poder Judiciário: 2023*. Brasília: Conselho Nacional de Justiça, 2024. 120 p. Disponível em: <https://bibliotecadigital.cnj.jus.br/jspui/handle/123456789/858>. Acesso em: 27 ago. 2024. ISBN 978-65-5972-141-2. Citado 6 vezes nas páginas 19, 20, 32, 37, 111 e 114.

CNJ; PNUD. *Cartilha Justiça 4.0*. Brasília: CNJ, 2021. Acesso em: 2 set. 2024. Disponível em: <https://www.cnj.jus.br/tecnologia-da-informacao-ecomunicacao/justica-4-0/cartilhas/>. Citado 2 vezes nas páginas 19 e 32.

COSTA, Y. D.; OLIVEIRA, H.; JR, V. N.; MASSA, L.; YANG, X.; BARBOSA, A.; OLIVEIRA, K.; VIEIRA, T. Automating petition classification in brazil's legal system: A two-step deep learning approach. *Artificial Intelligence and Law*, Springer, v. 33, n. 1, p. 227–251, 2025. Citado 2 vezes nas páginas 67 e 72.

DAI, X.; CHALKIDIS, I.; DARKNER, S.; ELLIOTT, D. Revisiting transformer-based models for long document classification. *arXiv preprint arXiv:2204.06683*, 2022. Citado na página 56.

DAI, Z.; YANG, Z.; YANG, Y.; CARBONELL, J.; LE, Q. V.; SALAKHUTDINOV, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. Citado na página 98.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, v. 7, n. Jan, p. 1–30, 2006. Citado 8 vezes nas páginas 60, 61, 62, 63, 82, 90, 91 e 105.

DEVLIN, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Citado 13 vezes nas páginas 20, 21, 39, 40, 41, 45, 46, 47, 53, 59, 80, 84 e 88.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. [S.l.: s.n.], 2019. p. 4171–4186. Citado na página 40.

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 10, n. 7, p. 1895–1923, 1998. Citado 2 vezes nas páginas 61 e 105.

DING, M.; ZHOU, C.; YANG, H.; TANG, J. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, v. 33, p. 12792–12804, 2020. Citado 3 vezes nas páginas 23, 69 e 73.

DING, N.; QIN, Y.; YANG, G.; WEI, F.; YANG, Z.; SU, Y.; HU, S.; CHEN, Y.; CHAN, C.-M.; CHEN, W.; YI, J.; ZHAO, W.; WANG, X.; LIU, Z.; ZHENG, H.-T.; CHEN, J.; LIU, Y.; TANG, J.; LI, J.; SUN, M. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, v. 5, n. 3, p. 220–235, 2023. ISSN 2522-5839. Disponível em: <https://doi.org/10.1038/s42256-023-00626-4>. Citado na página 46.

DONG, X.; YU, Z.; CAO, W.; SHI, Y.; MA, Q. A survey on ensemble learning. *Frontiers of Computer Science*, Springer, v. 14, n. 2, p. 241–258, 2020. Citado na página 37.

DROR, R.; BAUMER, G.; SHLOMOV, S.; REICHART, R. The hitchhiker's guide to testing statistical significance in natural language processing. *arXiv preprint arXiv:1812.06216*, 2018. Citado 2 vezes nas páginas 60 e 90.

EL-KASSAS, W. S.; SALAMA, C. R.; RAFEA, A. A.; MOHAMED, H. K. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, Elsevier, v. 165, p. 113679, 2021. Citado 3 vezes nas páginas 49, 51 e 70.

ELOUARGUI, Y.; ZYATE, M.; SASSIOUI, A.; CHERGUI, M.; KAMILI, M. E.; OUZZIF, M. A comprehensive survey on efficient transformers. In: IEEE. *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*. [S.l.], 2023. p. 1–6. Citado 3 vezes nas páginas 54, 58 e 111.

ERKAN, G.; RADEV, D. R. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, v. 22, p. 457–479, 2004. Citado 2 vezes nas páginas 50 e 87.

FABBRI, A.; LI, I.; RADEV, D.; LI, X.; LU, R.; LI, X.; LI, M.; FU, M.; ZHOU, W.; JHA, S. et al. Multi-news: A large-scale multi-document summarization dataset and baselines. *arXiv preprint arXiv:1906.01749*, 2019. Citado 2 vezes nas páginas 51 e 58.

FAMA, I.; BUENO, B.; ALCOFORADO, A.; FERRAZ, T.; MOYA, A.; COSTA, A. H. 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*. Porto Alegre, RS, Brasil: SBC, 2024. p. 129–138. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/31124>. Citado na página 70.

FIELDS, J.; CHOVANEC, K.; MADIRAJU, P. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, IEEE, v. 12, p. 6518–6531, 2024. Citado 3 vezes nas páginas 53, 54 e 56.

GARCIA, E. A.; SILVA, N. F.; SIQUEIRA, F.; ALBUQUERQUE, H. O.; GOMES, J. R.; SOUZA, E.; LIMA, E. A. Robertalexpt: a legal roberta model pretrained with deduplication for portuguese. In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*. [S.l.: s.n.], 2024. p. 374–383. Citado 4 vezes nas páginas 42, 43, 44 e 76.

GIARELIS, N.; MASTROKOSTAS, C.; KARACAPILIDIS, N. Abstractive vs. extractive summarization: An experimental review. *Applied Sciences*, MDPI, v. 13, n. 13, p. 7620, 2023. Citado 3 vezes nas páginas 50, 70 e 110.

GLENN, H. P. *Legal Traditions of the World: Sustainable Diversity in Law*. 5th. ed. Oxford, UK: Oxford University Press, 2014. Citado na página 70.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, Ieee, v. 21, n. 9, p. 1263–1284, 2009. Citado 2 vezes nas páginas 82 e 100.

HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. *Science*, American Association for the Advancement of Science, v. 349, n. 6245, p. 261–266, 2015. Citado 2 vezes nas páginas 36 e 37.

HOWARD, J.; RUDER, S. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018. Citado na página 45.

HU, Y.; CHEN, P.; LIU, T.; GAO, J.; SUN, Y.; YIN, B. Hierarchical attention transformer networks for long document classification. In: IEEE. *2021 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2021. p. 1–7. Citado 2 vezes nas páginas 55 e 56.

HU, Y.; DING, W.; LIU, T.; GAO, J.; SUN, Y.; YIN, B. Hierarchical multiple granularity attention network for long document classification. In: IEEE. *2022 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2022. p. 1–7. Citado na página 56.

HUA, W.; ZHANG, Y.; CHEN, Z.; LI, J.; WEBER, M. Legalrelectra: Mixed-domain language modeling for long-range legal text comprehension. *arXiv preprint arXiv:2212.08204*, 2022. Citado na página 23.

IGOREVNA, A. E.; BULATOVICH, B. K.; PETROVICH, N. D.; OLEGOVNA, P. O.; IGOREVICH, S. B.; ANATOLEVICH, S. O. Document image analysis and recognition: a survey. , . . . , v. 46, n. 4, p. 567–589, 2022. Citado na página 52.

ISLAM, M. M.; MUHAMMAD, U.; OUSSALAH, M. Evaluating text summarization techniques and factual consistency with language models. In: IEEE. *2024 IEEE International Conference on Big Data (BigData)*. [S.l.], 2024. p. 116–122. Citado na página 59.

JAIN, D.; BORAH, M. D.; BISWAS, A. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, Elsevier, v. 40, p. 100388, 2021. Citado 2 vezes nas páginas 59 e 109.

JAIN, D.; BORAH, M. D.; BISWAS, A. A sentence is known by the company it keeps: improving legal document summarization using deep clustering. *Artificial Intelligence and Law*, Springer, v. 32, n. 1, p. 165–200, 2024. Citado 2 vezes nas páginas 69 e 73.

JAISWAL, A.; MILIOS, E. Breaking the token barrier: Chunking and convolution for efficient long text classification with bert. *arXiv preprint arXiv:2310.20558*, 2023. Citado 2 vezes nas páginas 68 e 85.

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study. *Intelligent data analysis*, SAGE Publications Sage UK: London, England, v. 6, n. 5, p. 429–449, 2002. Citado na página 82.

JIANG, Z.; YANG, J.; RAO, D. An empirical study of leveraging plms and llms for long-text summarization. In: SPRINGER. *Pacific Rim International Conference on Artificial Intelligence*. [S.l.], 2024. p. 424–435. Citado na página 52.

JR, D. J. D.; BRAGA, P. S.; OLIVEIRA, R. A. de. *Curso de direito processual civil*. [S.l.]: Juspodivm Salvador, 2010. v. 5. 385 p. Citado 3 vezes nas páginas 29, 30 e 110.

JURAFSKY, D.; MARTIN, J. H. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. [S.l.]: Prentice Hall, 2000. Citado na página 40.

Jusbrasil. *Incidente de Resolução de Demandas Repetitivas (IRDR): o que é, como funciona e quais são seus benefícios*. Jusbrasil, 2025. Acesso em: 01 out. 2025. Disponível em: <https://www.jusbrasil.com.br/artigos/incidente-de-resolucao-de-demandas-repetitivas-irdr-o-que-e-como-funciona-e-quais-sao-seus-beneficio 1267723623>. Citado na página 32.

KALAMKAR, P.; TIWARI, A.; AGARWAL, A.; KARN, S.; GUPTA, S.; RAGHAVAN, V.; MODI, A. Corpus for automatic structuring of legal documents. *arXiv preprint arXiv:2201.13125*, 2022. Citado 6 vezes nas páginas 18, 20, 23, 32, 53 e 78.

KARAMOUZAS, D.; MADEMLIS, I.; PITAS, I. Neural knowledge transfer for sentiment analysis in texts with figurative language. In: IEEE. *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*. [S.l.], 2022. p. 1–6. Citado na página 53.

KEARNS, M.; ROTH, A. Ethical algorithm design should guide technology regulation. *Brookings Institute*, 2019. Citado na página 23.

KIM, T. K. Understanding one-way anova using conceptual figures. *Korean journal of anesthesiology*, v. 70, n. 1, p. 22, 2017. Citado 2 vezes nas páginas 62 e 63.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Citado na página 47.

KIRMANI, M.; HAKAK, N. M.; MOHD, M.; MOHD, M. Hybrid text summarization: A survey. Springer, p. 63–73, 2019. Citado 3 vezes nas páginas 49, 51 e 70.

KITAEV, N.; KAISER, L.; LEVSKAYA, A. Reformer: The efficient transformer. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2020. Citado na página 54.

KOEHN, P. *Statistical machine translation*. [S.l.]: Cambridge University Press, 2009. Citado na página 105.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, v. 14, n. 2, p. 1137–1145, 1995. Citado 3 vezes nas páginas 60, 83 e 90.

KOWSARI, K.; MEIMANDI, K. J.; HEIDARYSAFA, M.; MENDU, S.; BARNES, L.; BROWN, D. Text classification algorithms: A survey. *Information*, MDPI, v. 10, n. 4, p. 150, 2019. Citado na página 37.

KUŞ, A.; ACI, Ç. İ. A hybrid approach to automatic text summarization of turkish texts: Integrating extractive methods with llms. In: IEEE. *2024 Innovations in Intelligent Systems and Applications Conference (ASYU)*. [S.l.], 2024. p. 1–6. Citado na página 110.

LEVENE, H. Robust tests for equality of variances. *Contributions to probability and statistics*, Stanford University Press, p. 278–292, 1960. Citado na página 62.

LEWIS, M.; LIU, Y.; GOYAL, N.; GHAZVININEJAD, M.; MOHAMED, A.; LEVY, O.; STOYANOV, V.; ZETTLEMOYER, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2020. p. 7871–7880. Citado na página 51.

LI, I.; FENG, A.; RADEV, D.; YING, R. Hipool: Modeling long documents using graph neural networks. In: *The 61st Annual Meeting Of The Association For Computational Linguistics*. [S.l.: s.n.], 2023. Citado 4 vezes nas páginas 22, 56, 67 e 72.

LI, Z.; LI, C.; ZHANG, M.; MEI, Q.; BENDERSKY, M. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. [S.l.: s.n.], 2024. p. 881–893. Citado na página 110.

LIMSOPATHAM, N. Effectively leveraging bert for legal document classification. In: *Proceedings of the Natural Legal Language Processing Workshop 2021*. [S.l.: s.n.], 2021. p. 210–216. Citado 5 vezes nas páginas 65, 71, 72, 78 e 82.

LIU, Y.; LAPATA, M. Text summarization with pretrained encoders. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. [S.l.: s.n.], 2019. p. 3730–3740. Citado 4 vezes nas páginas 37, 49, 58 e 70.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. Citado na página 39.

LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2019. Citado na página 47.

LUCENA, D. C. de; SOUZA, E.; ALBUQUERQUE, H. O.; FÉLIX, N.; OLIVEIRA, A. L.; CARVALHO, A. C. de. Performance analysis of llms for abstractive summarization of brazilian legislative documents. In: *Conference on Digital Government Research*. [S.l.: s.n.], 2025. v. 1. Citado 2 vezes nas páginas 52 e 70.

LV, X.; LIU, Z.; ZHAO, Y.; XU, G.; YOU, X. Hbert: A long text processing method based on bert and hierarchical attention mechanisms. *International Journal on Semantic Web and Information Systems (IJSWIS)*, IGI Global, v. 19, n. 1, p. 1–14, 2023. Citado 2 vezes nas páginas 67 e 72.

MAMAKAS, D.; TSOTSI, P.; ANDROUTSOPOULOS, I.; CHALKIDIS, I. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. *arXiv preprint arXiv:2211.00974*, 2022. Citado 2 vezes nas páginas 65 e 72.

MARINO, G.; LICARI, D.; BUSHIPAKA, P.; COMANDÉ, G.; CUCINOTTA, T. et al. Automatic rhetorical roles classification for legal documents using legal-transformeroverbert. In: CEUR-WS. *CEUR WORKSHOP PROCEEDINGS*. [S.l.], 2023. v. 3441, p. 28–36. Citado na página 109.

MARTÍNEZ-PLUMED, F.; CONTRERAS-OCHANDO, L.; FERRI, C.; HERNÁNDEZ-ORALLO, J.; KULL, M.; LACHICHE, N.; RAMÍREZ-QUINTANA, M. J.; FLACH, P. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering*, IEEE, v. 33, n. 8, p. 3048–3061, 2019. Citado 7 vezes nas páginas 33, 34, 35, 36, 74, 75 e 76.

MEDINA, M. C. C.; OLIVEIRA, L. M. D. S.; FERREIRA, J. F. C.; SILVA, L. H. D. S.; RODRIGUES, C. M. O.; OLIVEIRA, J. F. L. D.; SOBRAL, P. C.; SOUZA, B.; FEITOSA, D.; FERNANDES, B. J. Classification of legal documents in portuguese language based on summarization. In: IEEE. *2022 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*. [S.l.], 2022. p. 1–6. Citado 2 vezes nas páginas 69 e 73.

MELO, R.; SANTOS, P. A.; DIAS, J. A semantic search system for the supremo tribunal de justiça. In: MONIZ, N.; VALE, Z.; CASCALHO, J.; SILVA, C.; SEBASTIÃO, R. (Ed.). *Progress in Artificial Intelligence*. Cham: Springer Nature Switzerland, 2023. p. 142–154. ISBN 978-3-031-49011-8. Citado na página 88.

MENDES, A. G. de C.; TEMER, S. O incidente de resolução de demandas repetitivas do novo código de processo civil. *Revista de Processo/ vol*, v. 243, n. 2015, p. 283–331, 2015. Citado na página 30.

MENDES, J. B. Precedente judicial como fonte do direito. IDP/EDB, 2015. Citado na página 29.

MIHALCEA, R.; TARAU, P. Textrank: Bringing order into text. In: *Proceedings of the 2004 conference on empirical methods in natural language processing*. [S.l.: s.n.], 2004. p. 404–411. Citado 2 vezes nas páginas 50 e 87.

MINAEE, S.; KALCHBRENNER, N.; CAMBRIA, E.; NIKZAD, N.; CHENAGHLU, M.; GAO, J. Deep learning–based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 54, n. 3, p. 1–40, 2021. Citado na página 37.

MIROŃCZUK, M. M.; PROTASIEWICZ, J. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, Elsevier, v. 106, p. 36–54, 2018. Citado na página 37.

MONTGOMERY, D. C. *Design and analysis of experiments*. [S.l.]: John wiley & sons, 2017. Citado 3 vezes nas páginas 61, 62 e 63.

NALLAPATI, R.; ZHOU, B.; GULCEHRE, C.; XIANG, B. Summarization with long short-term memory recurrent neural networks. *arXiv preprint arXiv:1704.01342*, 2017. Citado na página 49.

NEVES, D. A. A. Manual de direito processual civil. *Editora JusPODIVM*, 2015. Citado 2 vezes nas páginas 28 e 32.

NGUYEN, T. T. H.; JATOWT, A.; COUSTATY, M.; DOUCET, A. Survey of post-ocr processing approaches. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 54, n. 6, p. 1–37, 2021. Citado 2 vezes nas páginas 52 e 53.

OPITZ, J.; BURST, S. Macro f1 and macro f1. In: *arXiv preprint arXiv:1911.03347*. [S.l.: s.n.], 2019. Citado na página 60.

PAGLIARDINI, M.; PALIOTTA, D.; JAGGI, M.; FLEURET, F. Faster causal attention over large sequences through sparse flash attention. *arXiv preprint arXiv:2306.01160*, 2023. Citado na página 21.

PAPPAGARI, R.; ZELASKO, P.; VILLALBA, J.; CARMIEL, Y.; DEHAK, N. Hierarchical transformers for long document classification. In: IEEE. *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. [S.l.], 2019. p. 838–844. Citado 14 vezes nas páginas 20, 22, 55, 56, 58, 66, 68, 72, 86, 87, 89, 90, 98 e 109.

PARIKH, A. P.; TÄCKSTRÖM, O.; DAS, D.; USZKOREIT, J. A decomposable attention model for natural language inference. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2016. p. 2249–2255. Citado na página 45.

PARK, H.; VYAS, Y.; SHAH, K. Efficient classification of long documents using transformers. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022. p. 702–709. Disponível em: <https://aclanthology.org/2022.acl-short.79>. Citado 13 vezes nas páginas 23, 50, 56, 68, 71, 73, 86, 87, 88, 89, 90, 99 e 110.

PEYKANI, P.; RAMEZANLOU, F.; TANASESCU, C.; GHANIDEL, S. Large language models: A structured taxonomy and review of challenges, limitations, solutions, and future directions. *Applied Sciences*, MDPI, v. 15, n. 14, p. 8103, 2025. Citado na página 110.

PHILIPS, J. P.; TABRIZI, N. Historical document processing: historical document processing: a survey of techniques, tools, and trends. *arXiv preprint arXiv:2002.06300*, 2020. Citado na página 52.

PIRES, R.; ABONIZIO, H.; ALMEIDA, T. S.; NOGUEIRA, R. Sabiá: Portuguese large language models. In: SPRINGER. *Brazilian conference on intelligent systems*. [S.l.], 2023. p. 226–240. Citado na página 45.

PIRES, R. S.; SILVEIRA, R.; FERNANDES, C. G.; NETO, J. A. M.; FURTADO, V. Using complex networks to improve legal text hierarchical classification. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2024. p. 476–490. Citado 3 vezes nas páginas 67, 72 e 82.

POLO, F. M.; MENDONÇA, G. C. F.; PARREIRA, K. C. J.; GIANVECHIO, L.; CORDEIRO, P.; FERREIRA, J. B.; LIMA, L. M. P. de; MAIA, A. C. do A.; VICENTE, R. Legalnlp-natural language processing methods for the brazilian legal language. In: SBC. *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2021. p. 763–774. Citado 6 vezes nas páginas 19, 42, 44, 70, 76 e 99.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2011. Citado na página 59.

PRASAD, N. *Large Language Models and their Hierarchical adaptation on Long Documents for Classification and their Explanation: a case of Legal NLP*. Tese (Doutorado) — Université de Toulouse, 2024. Citado na página 55.

PRINCIPE, R. A.; CHIARINI, N.; VIVIANI, M. Long document classification in the transformer era: A survey on challenges, advances, and open issues. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 15, n. 2, p. e70019, 2025. Citado 12 vezes nas páginas 22, 23, 49, 51, 53, 54, 55, 56, 57, 58, 64 e 99.

PRINCIPE, R. A. A.; CHIARINI, N.; VIVIANI, M. An lcf-idf document representation model applied to long document classification. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. [S.l.: s.n.], 2024. p. 1129–1135. Citado na página 57.

QIU, X.; SUN, T.; XU, Y.; SHAO, Y.; DAI, N.; HUANG, X. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, Springer, v. 63, p. 1872–1897, 2020. Citado 2 vezes nas páginas 39 e 46.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. et al. Improving language understanding by generative pre-training. San Francisco, CA, USA, 2018. Citado 2 vezes nas páginas 39 e 44.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. et al. Language models are unsupervised multitask learners. *OpenAI blog*, v. 1, n. 8, p. 9, 2019. Citado na página 44.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M.; ZHOU, Y.; LI, W.; LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, v. 21, p. 1–67, 2020. Citado na página 39.

RAINIO, O.; TEUHO, J.; KLÉN, R. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, Nature Publishing Group UK London, v. 14, n. 1, p. 6086, 2024. Citado 2 vezes nas páginas 61 e 62.

RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: NEW JERSEY, USA. *Proceedings of the first instructional conference on machine learning*. [S.l.], 2003. v. 242, n. 1, p. 29–48. Citado 2 vezes nas páginas 50 e 56.

REIMERS, N.; GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. [S.l.], 2019. p. 3982. Citado 4 vezes nas páginas 51, 68, 73 e 82.

RIBEIRO, M. T.; WU, T.; GUESTRIN, C.; SINGH, S. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020. Citado na página 23.

RUSH, A. M.; CHOPRA, S.; PARIKH, A. A neural attention model for abstractive summarization. *arXiv preprint arXiv:1509.00600*, 2015. Citado na página 51.

RUSSO, A. R. *Uma moderna gestão de pessoas no Poder Judiciário*. Tese (Doutorado) — Estado do Rio Grande do Sul, Poder Judiciário, Tribunal de Justiça, 2009. Citado na página 28.

SCHWARTZ, R.; DODGE, J.; SMITH, N. A.; ETZIONI, O. Green ai. In: ACM NEW YORK, NY, USA. *Communications of the ACM*. [S.l.], 2020. v. 63, n. 12, p. 54–63. Citado na página 60.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 34, n. 1, p. 1–47, 2002. Citado na página 37.

SEO, M.; KEMBHAVI, A.; FARHADI, A.; HAJISHIRZI, H. Bidirectional attention flow for machine comprehension. In: *5th International Conference on Learning Representations, ICLR 2017*. [S.l.: s.n.], 2017. Citado na página 45.

SHAFFER, J. P. Multiple hypothesis testing. *Annual review of psychology*, v. 46, n. 1, p. 561–584, 1995. Citado na página 61.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, Oxford University Press, v. 52, n. 3-4, p. 591–611, 1965. Citado na página 62.

SIINO, M.; FALCO, M.; CROCE, D.; ROSSO, P. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, IEEE, 2025. Citado 2 vezes nas páginas 79 e 80.

SILVEIRA, R.; PONTE, C.; ALMEIDA, V.; PINHEIRO, V.; FURTADO, V. Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2023. p. 268–282. Citado 4 vezes nas páginas 42, 43, 44 e 76.

SISKA, F.; OKTAVIA, T. Predictive analysis of hypertensive heart disease using a machine learning approach. In: IEEE. *2024 International Conference on ICT for Smart Society (ICISS)*. [S.l.], 2024. p. 1–9. Citado na página 33.

SMITH, L. N. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018. Citado na página 47.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2020. p. 403–417. Citado 2 vezes nas páginas 41 e 60.

SOUZA, F. C. de. *BERTimbau: pretrained BERT models for Brazilian Portuguese BERTimbau: modelos BERT pré-treinados para Português Brasileiro*. Tese (Doutorado) — Universidade Estadual de Campinas, 2020. Citado na página 41.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, JMLR. org, v. 15, n. 1, p. 1929–1958, 2014. Citado na página 47.

STONE, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, Wiley Online Library, v. 36, n. 2, p. 111–133, 1974. Citado 2 vezes nas páginas 82 e 105.

SUBRAMANI, N.; MATTON, A.; GREAVES, M.; LAM, A. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534*, 2020. Citado na página 52.

TAUK, C. S.; SALOMÃO, L. F. Inteligência artificial no judiciário brasileiro. *Diké-Revista Jurídica*, v. 22, n. 23, p. 2–32, 2023. Citado na página 19.

TELLA, M. J. F. y. *Lições de teoria geral do direito*. São Paulo: Revista dos Tribunais, 2011. Citado na página 29.

TEMER, S. O. et al. Incidente de resolução de demandas repetitivas. Universidade do Estado do Rio de Janeiro, 2019. Citado na página 30.

THAKUR, N.; REIMERS, N.; RÜCKLÉ, A.; SRIVASTAVA, A.; GUREVYCH, I. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021. Citado na página 68.

TINARRAGE, R.; ENNES, H.; RESCK, L.; GOMES, L. T.; PONCIANO, J. R.; POCO, J. Empirical analysis of binding precedent efficiency in brazilian supreme court via case classification: R. tinarrage et al. *Artificial Intelligence and Law*, Springer, p. 1–67, 2025. Citado 2 vezes nas páginas 65 e 72.

TJMA, T. d. J. d. E. d. M. *IRDR Admitido*. TJMA, 2025. Acesso em: 07 out. 2025. Disponível em: <https://www.tjma.jus.br/hotsite/nugepnac/item/1992/0/irdr-admitido>. Citado na página 31.

TOUVRON, H.; LAVRIL, T.; IZACARD, G.; MARTINET, X.; LACHAUX, M.-A.; LACROIX, T.; ROZIÈRE, B.; GOYAL, N.; HAMBRO, E.; AZHAR, F. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. Citado 2 vezes nas páginas 45 e 69.

TSIRMPAS, D.; GKIONIS, I.; MADEMLIS, I. Neural natural language processing for long texts: A survey of the state-of-the-art. *arXiv Preprint*, 2023. Citado na página 49.

TSIRMPAS, D.; GKIONIS, I.; PAPADOPOULOS, G. T.; MADEMLIS, I. Neural natural language processing for long texts: A survey on classification and summarization. *Eng. Appl. Artif. Intell.*, Pergamon Press, Inc., USA, v. 133, n. PC, jul. 2024. ISSN 0952-1976. Disponível em: <https://doi.org/10.1016/j.engappai.2024.108231>. Citado 5 vezes nas páginas 53, 55, 56, 58 e 99.

VASWANI, A. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. Citado 5 vezes nas páginas 21, 37, 38, 39 e 51.

VIEGAS, C. F.; COSTA, B. C.; ISHII, R. P. Jurisbert: a new approach that converts a classification corpus into an sts one. In: SPRINGER. *International Conference on Computational Science and Its Applications*. [S.l.], 2023. p. 349–365. Citado 3 vezes nas páginas 42, 44 e 76.

WANG, C.; LI, M.; SMOLA, A. J. Language models with transformers. *arXiv preprint arXiv:1904.09408*, 2019. Citado na página 59.

WANG, Y.; CUI, L.; ZHANG, Y. Using dynamic embeddings to improve static embeddings. *CoRR arXiv*, 1911. Citado 2 vezes nas páginas 40 e 41.

WANG, Y.; YOSHINAGA, N. Summarization-based data augmentation for document classification. In: *Proceedings of EMNLP Workshop*. [S.l.: s.n.], 2023. p. 49. Citado 2 vezes nas páginas 69 e 73.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. In: MANCHESTER. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. [S.l.], 2000. v. 1, p. 29–39. Citado 2 vezes nas páginas 33 e 35.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M. et al. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:2003.00075*, 2020. Citado na página 38.

WU, C.; WU, F.; QI, T.; HUANG, Y. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. *arXiv preprint arXiv:2106.01040*, 2021. Citado 3 vezes nas páginas 55, 56 e 58.

YANG, Z.; DAI, Z.; YANG, Y.; CARBONELL, J.; SALAKHUTDINOV, R.; LE, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, v. 32, 2019. Citado na página 39.

ZAHEER, M.; GURUGANESH, G.; DUBEY, K. A.; AINSLIE, J.; ALBERTI, C.; ONTANON, S.; PHAM, P.; RAVULA, A.; WANG, Q.; YANG, L. et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, v. 33, p. 17283–17297, 2020. Citado 5 vezes nas páginas 54, 58, 64, 71 e 72.

ZHANG, L.; WANG, S.; LIU, B. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 8, n. 4, p. e1253, 2018. Citado na página 37.

ZHANG, T.; LADHAK, F.; DURMUS, E.; LIANG, P.; MCKEOWN, K.; HASHIMOTO, T. B. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA ..., v. 12, p. 39–57, 2024. Citado na página 70.

ZHONG, C.; WANG, J.; HUANG, M.; ZHANG, K.; XIAO, Y.; LIU, Z.; WANG, D.; YIN, J.; ZHOU, X.; LUO, Y. et al. Does it make sense? and why? a pilot study for sense making and explanation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2020. p. 4020–4026. Citado na página 23.

ZUFELATO, C.; OLIVEIRA, F. A. A teoria da tipologia das partes de galanter e a prática do irdr no brasil: O poder judiciário como um jogador? *Nucleus*, Fundação Educacional Ituverava, v. 1, n. 1, p. 9, 2024. Citado 2 vezes nas páginas 30 e 32.

Apêndices

APÊNDICE A – Artigo submetido ao SBSI 2026

Hybrid Summarization for Brazilian Judicial Decisions

[1]

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 Global South countries. This situation challenges not only legal professionals but also information systems (IS) that support the processing and organization of large document volumes.* ***Scientific and/or Practical Problem:*** *Many systems struggle with lengthy texts due to language model length limits 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, composed of STF decisions. Five summarization methods were compared: TF-IDF (baseline), <omitido para revisão>, 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

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 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 pesquisadores interessados em compreender o raciocínio por trás das deliberações. Em Sistemas de Informação (SI), esse cenário representa um desafio clássico de organização e 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, contando com mais de 140 projetos aplicados no judiciário brasileiro, com foco em classificação de documentos, triagem processual e automação de atividades repetitivas [CNJ 2024, Casimiro and Teixeira 2024, Moreira and de Souza Moura 2023]. Entre as atividades que mais demandam inovação está a produção e utilização de sumários jurídicos, como as ementas. Elas sintetizam os fundamentos e o dispositivo de decisões judiciais, funcionando como elemento central para indexação em sistemas de busca, classificação de casos semelhantes e 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 [Fama et al. 2024, Bhattacharya et al. 2019]. Diferentes métodos têm sido implementados, desde abordagens baseadas na seleção de sentenças relevantes (extrativas), métodos que reescrevem a informação em nova forma textual (abstrativas) e combinações híbridas que buscam equilibrar precisão e coesão de ambos os métodos [El-Kassas et al. 2021, Liu and Lapata 2019, Kirmani et al. 2018]. Apesar dos avanços, persistem desafios específicos na aplicação dessas técnicas ao domínio jurídico, incluindo textos longos que excedem a capacidade de processamento de modelos tradicionais, linguagem altamente especializada e a necessidade de preservar terminologia técnica e relações argumentativas complexas [Bhattacharya et al. 2019, Lai et al. 2024, El-Kassas et al. 2021].

Este estudo busca responder a esse desafio ao comparar cinco métodos de sumarização aplicados a decisões do Supremo Tribunal Federal (STF): (i) abordagem extrativa baseada em TF-IDF; (ii) abordagem extrativa com *<omitido para revisão>* com processamento baseado *chunks* [Alva Principe et al. 2025] ; (iii) 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 desenvolvido consiste em um *pipeline* de sumarização híbrida adaptado ao domínio jurídico brasileiro, capaz de processar documentos longos e gerar sumários consistentes, equilibrando qualidade técnica e viabilidade computacional.

O processamento híbrido reduz inicialmente o tamanho textual através de métodos extrativos antes de aplicar técnicas de IA generativa, otimizando o uso de recursos limitados enquanto preserva a qualidade dos sumários [Jiang et al. 2024, Kuş and Acı 2024]. Embora estudos anteriores tenham investigado combinações similares de modelos especializados com técnicas generativas, sua aplicação ao contexto jurídico brasileiro permanece pouco explorada [Zhang et al. 2025]. Este trabalho utiliza o *dataset* RulingBR, corpus bem estabelecido e amplamente referenciado na literatura de sumarização jurídica, que contém decisões do STF estruturadas em seções específicas (relatório, fundamentação, votos e ementa), permitindo análise controlada e comparação com tra-

balhos anteriores [Feijó and Moreira 2018].

As principais contribuições deste estudo está 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 Abertos, sobretudo aqueles relacionados à decisão escaláveis, transparentes e interoperáveis, 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 decomposição-recomposição hierárquica que permite ao <*omitido para revisão*> processar documentos longos sem truncamento, preservando representações contextualizadas através de agregação por *mean pooling* <*omitido para revisão*> e também considerando limitações de recursos financeiros e computacionais enfrentados por instituições ao lidar com a complexidade computacional desses sistemas.

O restante do artigo está organizado da seguinte forma: a Seção 2 discute trabalhos relacionados; a Seção 3 apresenta a metodologia; a Seção 4 detalha resultados experimentais e suas implicações práticas; e a Seção 5 apresenta considerações finais e direções futuras.

2. Trabalhos relacionados

A sumarização de textos jurídicos apresenta desafios únicos relacionados à complexidade linguística, extensão documental e necessidade de preservação de terminologia técnica. Esta seção apresenta a revisão de trabalhos relevantes em relação a 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 indicado ganhos em diversas tarefas de processamento de linguagem natural (PLN). Em [Chalkidis et al. 2020] desenvolveram Legal-BERT para textos jurídicos em inglês, demonstrando superioridade sobre BERT [Devlin et al. 2019] generalista em tarefas de classificação em 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].

No contexto brasileiro, os autores [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 para outros modelos, em especial no domínio jurídico, como o LegalBert-pt desenvolvido por [Silveira et al. 2023]. O LegalBert-pt foi pré-treinado em um corpus de 1.5 milhões de documentos legais de dez tribunais brasileiros, esses documentos consistiam em petições iniciais, petições, decisões e sentenças. O modelo apresentou vantagens em tarefas PLN direcionadas ao jurídico quando comparado ao BERTimbau-base.

Ainda neste cerne, os autores <*omitido para revisão*> apresentaram o <*omitido para revisão*>, modelo também pré-treinado para o domínio jurídico brasileiro, treinado em corpus extenso de legislações e jurisprudências nacionais. Diferente do LegalBert-pt, o <*omitido para revisão*> constitui de um corpus extenso com mais de 5 milhões de

documentos jurídicos composto por legislações e jurisprudências nacionais. Os resultados sugerem capacidades superiores em PLN jurídica comparado a modelos generalistas, confirmando as evidências percebidas em outros estudos.

Porém, devido a arquitetura *Transformer*, esses modelos baseados em BERT pecam ao lidar com o processamento de documentos longos, onde o mecanismo de autoatenção, que requer que cada *token* na sequência de entrada atenda a todos os outros *tokens*, resultando em complexidade computacional quadrática em relação ao 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 baseado em (*chunks*) têm sido exploradas, dividindo documentos longos em segmentos processáveis e agregando representações resultantes [Alva Principe et al. 2025], se tornando viável para documentos jurídicos extensos e permitindo a aplicação de modelos especializados mantendo suas representações semânticas ricas[Bhattacharya et al. 2019, de Castro and Ralha 2025].

Em [Silva Junior et al. 2025] foram conduzidas a avaliação de 16 métodos de representação textual para similaridade semântica em português brasileiro no domínio legal, incluindo representações esparsas (TF-IDF, LDA), *embeddings* estáticos (word2vec, fastText, doc2vec) e contextualizados (ELMo, BERT, Longformer, SBERT, SimCSE, DiffCSE). O objetivo foi identificar quais representações são mais efetivas para recuperação de documentos similares em cenários não supervisionados. A metodologia comparou documentos do STJ e TCU usando conjuntos rotulados heuristicamente e por especialistas. Os resultados revelam que representações simples como TF-IDF e BM25 ainda produzem resultados competitivos, enquanto SBERT apresentou maior correlação (Pearson: 0.50) com anotações de especialistas. Adicionalmente, o estudo evidenciou que *Domain Adaptive Pre-training* (DAPT) modifica o espaço vetorial de modelos *Transformer* mais significativamente que mudanças no mecanismo de atenção, observação relevante ao comparar BERT com Longformer.

Estes 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 também é uma tarefa em PLN, voltada para a condensação de um documento ou conjunto de documentos em uma versão mais curta, mantendo suas informações mais relevantes e coerentes [Liu and Lapata 2019]. 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 *Transformer* [Alva Principe et al. 2025]. As técnicas de sumarização podem ser integradas de diferentes formas.

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 decisões judiciais americanas utilizando *Term Frequency-Inverse Document Frequency* (TF-IDF) e análise estrutural. Seus resultados

demonstraram eficácia na captura de informações, mas limitações na geração de sumários coesos, desafio comum em abordagens puramente extrativas [Bhattacharya et al. 2019].

Os autores [Jain et al. 2024] propuseram DCESumm para sumarização extrativa de documentos jurídicos longos, combinando classificador supervisionado baseado em LegalBERT com ajuste de escores via *deep clustering* temático. A hipótese central é que sentenças relevantes tendem a se agrupar, permitindo reforço ou atenuação de relevância baseado em contexto global. Avaliado em BillSum (legislação norte-americana) e FIRE (Suprema Corte da Índia), superou métodos clássicos (TextRank, SummaRuNNer) e neurais (Legal Pegasus) com ganhos de 1-6 pontos ROUGE no BillSum e 6-12 pontos no FIRE, demonstrando efetividade em documentos extensos não estruturados.

Em [Feijó and Moreira 2018] criaram o RulingBR, primeiro corpus brasileiro para sumarização jurídica contendo 10.623 decisões do STF estruturadas em *ementa* (sumário, 7%), *acórdão* (2%), *relatório* (22%) e *voto* (69%). Estabeleceram *baselines* extrativos (TextRank, LexRank, Luhn) com ROUGE-1 entre 0.21-0.31, constatando que dispersão semântica em textos jurídicos invalida estratégias posicionais típicas de notícias. Análise revelou correlação fraca (r=0.39) entre comprimento de documento e ementa, evidenciando complexidade na determinação de tamanho ideal de sumário. Esse estudo estabeleceu fundação metodológica para pesquisas subsequentes, incluindo 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 demonstrado resultados superiores em sumarização abstrativa. BART utiliza uma arquitetura de *auto-encoder denoising* que aprende a reconstruir textos corrompidos, enquanto PEGASUS é pré-treinado especificamente para sumarização usando um 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 sumários abstrativos, mas projeta o resultado em sentenças do documento original, eliminando completamente possibilidades de alucinação. O objetivo foi combinar expressividade de modelos generativos com garantias factuais de métodos extrativos. Para isso, foi utilizado o GPT-3.5-turbo para gerar sumário abstrativo inicial, seguido de algoritmo de projeção que identifica no documento original sentenças que melhor correspondem a cada sentença gerada (via similaridade semântica). O método proposto retorna então sequência de sentenças originais que preserva 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 relacionados à alucinação de fatos (geração de informações incorretas) ou à perda de fidelidade ao texto original, um risco considerável em aplicações sensíveis ao dados [Fabbri et al. 2019]. Atualmente, no contexto de lidar com as limitações dos *Transformers* para documentos longos, as técnicas abstrativas ainda não foram plenamente exploradas para esse propósito, uma das motivações para esse estudo [Alva Principe et al. 2025].

Abordagens híbridas para sumarização. A literatura sugere que a combinação de métodos extrativos e abstrativos oferece equilíbrio entre cobertura de conteúdo e 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 *framework* híbrido combinando seleção extrativa hierárquica com geração abstrativa baseada em BERT, obtendo resultados superiores a métodos puros em *benchmarks* padrão. Com isso, sugerem que pré-treinamento extrativo inicial auxilia geração abstrativa subsequente. Em CNN/DailyMail, BERTSUMEXTABS alcançou ROUGE-1=42.13 (*vs* 41.72 do modelo puramente abstrativo). Análise revelou que modelo híbrido produz menos n-gramas novos que versão puramente abstrativa, indicando viés extrativo herdado do pré-treinamento inicial, adequado para *datasets* com sumários parcialmente extrativos.

[Bae et al. 2019] investigaram integração específica para documentos legais, evidenciando superioridade híbrida em métricas de coerência e preservação técnica. Os autores relatam que a combinação de extração e abstração é particularmente efetiva em documentos jurídicos devido à necessidade de preservar terminologia técnica enquanto se mantém coesão textual.

Em [Janakiraman and Ghoraani 2025], foram comparados 17 modelos através de *framework* multidimensional (consistência factual 35%, similaridade semântica 25%), identificando DeepSeek-v3 como superior em acurácia factual (*SummaC=0.68*), Gemini-1.5-Flash em custo-benefício ($0.00012/resumo), e tensão entre consistência factual (ótima em 50 *tokens*) *versus* 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), indicando melhor 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 ao contexto jurídico brasileiro permanece pouco explorada . Principalmente, a integração de modelos especializados como *<omitido para revisão>* com técnicas híbridas de sumarização para decisões judiciais brasileiras representa uma contribuição significativa, considerando as especificidades linguísticas, estruturais e jurídicas do sistema judiciário nacional.

Este trabalho se diferencia ao propor um *pipeline* híbrido que combina: (i) processamento baseado em *chunks* para superar limitações de tamanho de modelos especializados; (ii) utilização de *<omitido para revisão>* para representações semânticas ricas do domínio jurídico; (iii) refinamento através de modelos generativos; e (iv) avaliação usando múltiplas métricas de qualidade. Esta abordagem busca equilibrar precisão técnica, viabilidade computacional e qualidade dos sumários gerados, considerando limitações de recursos financeiros e computacionais frequentemente presentes em contextos de sistemas de informação no setor público brasileiro.

3. Metodologia

Este estudo foi guiado pela abordagem metodológica de *Design Science Research* (DSR), que contém etapas voltadas para o desenvolvimento e avaliação de artefatos

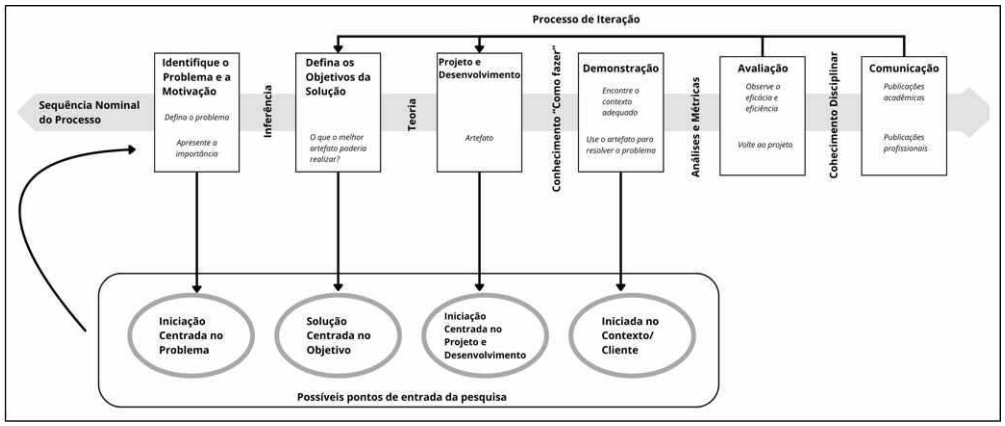


Figura 1. Etapas do DSR [Peffers et al. 2007]

tecnológicos para resolver problemas organizacionais identificados [Hevner et al. 2004, Peffers et al. 2007]. A DSR é adequada para pesquisas em SI que buscam criar 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.

Conforme a Figura 1, o DSR compreende seis etapas correlacionadas. As duas primeiras (identificação do problema e definição dos objetivos) foram apresentadas na Seção 1, contextualizando as especificidades de decisões judiciais: extensão significativa (frequentemente milhares de *tokens*), dispersão de informações relevantes ao longo de múltiplas seções, e complexidade estrutural que dificulta tanto análise manual por profissionais quanto processamento automatizado. Essas características motivam a exploração de estratégias de sumarização como solução viável considerando tecnologias existentes e limitações computacionais de modelos *Transformer* convencionais. As demais etapas que complementam esse processo serão descritas nas subseções a seguir.

3.1. Projeto e desenvolvimento

Esta etapa envolve o projeto e a construção do artefato tecnológico [Hevner et al. 2004]. 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 theory*).

Cada documento está estruturado em oitos campos distintos dos quais compreendem: *ementa* (sumário de referência), *relatório* (descrição do caso), *voto* (argumentação jurídica), e *acórdão* (decisão final). Os metadados classificatórios incluem *área* (ramo do direito), *classe* (instrumento processual), *relator* (ministro responsável), e *extrato* (informações complementares).

Na Tabela 1 são apresentadas as principais características estatísticas do corpus, calculadas utilizando o tokenizador do <*omitido para revisão*>.

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	99,89
Docs $>$ 1.500 *tokens* (%)	91,17	–
Docs $>$ 3.000 *tokens* (%)	64,81	–
Razão de compressão		
Média (%)	14,09	
Mediana (%)	11,43	

Devido a limitações computacionais do ambiente Google Colab, foram processados 5.792 documentos (54,8% do corpus) selecionados aleatoriamente (`random.seed(42)`). Esta amostra mantém distribuição representativa do corpus original e excede volumes utilizados em estudos similares de sumarização jurídica [Arfat et al. 2024, Preti et al. 2024], fornecendo poder estatístico adequado para comparação entre métodos.

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

Pré-processamento textual. O pré-processamento foi implementado priorizando a preservação de características linguísticas relevantes para o 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 pontuação especializada relevante para o domínio jurídico; normalização de espaços em branco e remoção de caracteres de controle; aplicação condicional de conversão para minúsculas baseada nos requisitos específicos de cada modelo; e filtragem de sentenças excessivamente curtas, estabelecendo limite mínimo de 20 caracteres. As *stopwords* foram obtidas da biblioteca spaCy para português, garantindo remoção adequada de termos não informativos específicos do idioma, preservando terminologia técnica relevante, nesse caso aplicadas apenas para o TF-IDF.

Implementação dos métodos de sumarização. Complementando essa aderência metodológica, escolha e design dos métodos de sumarização baseiam-se na teoria de *Task-Technology Fit* (TTF), que postula que o desempenho de SI é otimizado quando

existe alinhamento entre as características da tecnologia e os requisitos da tarefa [Goodhue and Thompson 1995]. No contexto deste estudo, a tarefa (produção de ementas jurídicas) demanda preservação de terminologia técnica, síntese de fundamentos legais dispersos, e manutenção de coesão argumentativa. Os métodos propostos (extrativos, abstrativos e híbridos) foram selecionados e configurados para maximizar esse ajuste tecnologia-tarefa, considerando *trade-offs* entre qualidade de saída, precisão jurídica, e viabilidade computacional. Isso justifica a avaliação multimétrica adotada e orienta decisões de design do *pipeline*, assegurando alinhamento entre capacidades tecnológicas e 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 métodos extrativos (e.g., TF-IDF e <*omitido para revisão*>) e 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 <*omitido para revisão*> foi implementado 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

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

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 que permite a seleção de sentenças com maior relevância semântica global, possibilitando o processamento de documentos extensos sem perda das representações contextuais ricas do modelo <*omitido para revisão*> [Polsley et al. 2016].

Para o método abstrativo, utilizou-se um modelo de linguagem de grande escala (`google/Gemini-2.5-flash-lite`) acessado via API. O *prompt* foi construído

seguindo a abordagem *few-shot*, incorporando: (i) 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 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 precisão na captura de conteúdo relevante com coesão textual na apresentação final. Duas variantes foram implementadas: a primeira utiliza seleção TF-IDF de oito sentenças seguida por refinamento via modelo generativo; a segunda emprega seleção <*omitido para revisão*> de oito sentenças mais similares ao documento completo, também seguida por refinamento abstrativo. A escolha de oito sentenças para a etapa extrativa baseia-se em análise empírica preliminar que identificou este volume como adequado para preservar informação suficiente sem exceder limites de contexto do modelo generativo.

3.2. Demonstração e avaliação do artefato

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

Métricas de avaliação. ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) nas variantes ROUGE-1, ROUGE-2 e ROUGE-L [Lin 2004] foi empregado para avaliar similaridade lexical através da sobreposição de unigramas, bigramas e subsequências comuns mais longas entre sumários gerados e ementas de referência. ROUGE-N mede sobreposição de n-gramas entre resumo gerado e referência, enquanto ROUGE-L considera a subsequência comum mais longa (LCS), capturando correspondências de ordem sequencial. O escore F1 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 similaridade baseada em *embeddings* contextualizados de modelos BERT. Esta métrica captura correspondências semânticas além da sobreposição lexical superficial, sendo consideráveis para o domínio jurídico onde diferentes formulações podem expressar conceitos equivalentes. O escore combina precisão e revocação através 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 precisão e revocação das correspondências de *embeddings*, respectivamente.

METEOR (*Metric for Evaluation of Translation with Explicit ORdering*) [Banerjee and Lavie 2005] foi incluído por considerar correspondências exatas, correspondências de radicais através de stemming, correspondências de sinônimos através de WordNet, e ordem das palavras. Combina precisão (P) e revocação (R) em F-média com penalização por 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 em ambiente Google Colab utilizando GPU NVIDIA A100 para processamento dos modelos. Para a implementação utilizou-se *Transformers* (<*omitido para revisão*>), PyTorch, spaCy (`pt_core_news_sm`), scikit-learn (TF-IDF), e 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 através 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 pelo repositório anônimo do presente estudo[1].

4. Resultados e discussão

Esta seção apresenta os resultados da avaliação comparativa dos cinco métodos de sumarização implementados: TF-IDF extrativo, <*omitido para revisão*> extrativo, abstrato puro (Gemini-2.5), híbrido TF-IDF e híbrido <*omitido para revisão*>. A discussão contextualiza os achados em relação aos resultados baseline reportados por Feijó e Moreira [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 através das métricas ROUGE (1, 2, L), BERTScore, e METEOR. Todas as métricas foram calculadas utilizando as ementas de referência produzidas por especialistas jurídicos como *gold standard*.

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 melhores métricas (<*omitido para revisão*>) e tenho um ganho de aproximadamente 45.9% sobre o melhor resultado *baseline* (LexRank com 4 sentenças: 0.31) evidenciado por [Feijó and Moreira 2018].

[1] `https://anonymous.4open.science/r/hybrid_summarization-A7D2`

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
<*omitido para revisão*>	0.3579	0.1677	0.2070	0.1621	0.2730
Gemini-2.5	0.4524	0.2315	0.2964	0.1418	0.2726
Híbrido TF-IDF	0.4080	0.1871	0.2609	0.1236	0.2329
Híbrido <*omitido para revisão*>	0.4053	0.1841	0.2584	0.1257	0.2267

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 para domínios especializados, consistente com achados de [Preti et al. 2024] em sumarização de decisões judiciais italianas.

Entre os métodos extrativos, <*omitido para revisão*> demonstrou superioridade em métricas semânticas (BERTScore: 0.1621, METEOR: 0.2730) comparado ao TF-IDF (BERTScore: 0.1377, METEOR: 0.2387), validando a hipótese de que representações contextualizadas especializadas para o domínio jurídico capturam relações semânticas mais ricas que abordagens estatísticas clássicas. O <*omitido para revisão*> 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 <*omitido para revisão*> seja mais sofisticado tecnicamente, TF-IDF manteve rendimento relativo, resultado consistente com achados [Silva Junior et al. 2025] que demonstraram efetividade de representações simples baseadas em frequência para similaridade textual em documentos jurídicos brasileiros, particularmente em cenários com dados não rotulados.

Por fim, os métodos híbridos apresentaram desempenho intermediário consistente, com resultados superiores aos métodos extrativos puros mas inferiores ao método abstrativo. O híbrido TF-IDF obteve ROUGE-1 de 0.4080 e ROUGE-2 de 0.1871, enquanto o híbrido <*omitido para revisão*> alcançou 0.4053 e 0.1841, respectivamente. A proximidade nos resultados entre ambas as variantes híbridas sugere que o refinamento pelo modelo generativo tende a homogeneizar as diferenças oriundas da etapa extrativa inicial. Isso aponta para uma possível redundância no *pipeline* híbrido, onde os modelos extrativos como mecanismo de redução prévia ao invés de seleção estratégica de conteúdo. Esta hipótese é explorada em maior profundidade na análise qualitativa apresentada a seguir.

4.2. Análise das limitações dos métodos híbridos

A Figura 2 apresenta 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.

Os métodos híbridos, embora não tenham superado o método abstrativo em

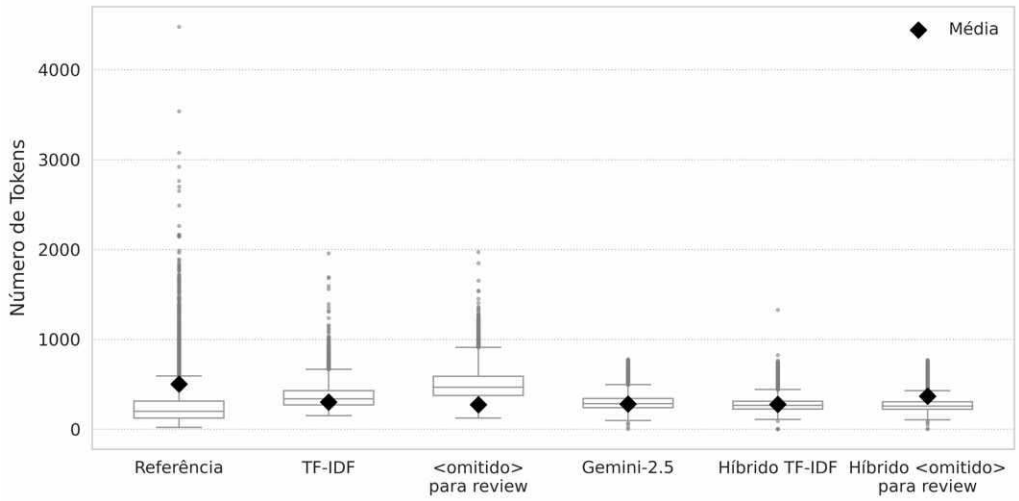


Figura 2. Comparação do números de *tokens* dos sumários gerados.

métricas ROUGE, apresentam atributos favoráveis para aplicações com restrições computacionais. Geraram resumos com médias de 300-350 *tokens*, representando redução de 35-40% em relação às ementas de referência (média: 540 *tokens*) e maior consistência comparado ao método abstrativo (média: 280 *tokens* com alta variabilidade). Apesar dessa maior compressão, o híbrido TF-IDF (ROUGE-1: 0,4080) ficou apenas 9,8% abaixo do Gemini-2.5, enquanto o híbrido <*omitido para revisão*> (ROUGE-1: 0,4053) apresentou *déficit* de 10,4%, comportamento consistente com observações de [Bae et al. 2019] sobre efetividade de estratégias híbridas para balancear fidelidade factual e coesão textual em domínios especializados.

A diferença qualitativa entre os métodos abstrativo e híbrido pode ser observada por meio de exemplo representativo de decisão sobre competência originária do STF em matéria previdenciária na Figura 3.

Isso evidencia que o Gemini-2.5 realiza reformulação do sumário, estruturando conteúdo em tópicos numerados e reorganizando logicamente fundamentos jurídicos, facilitando leitura hierárquica. Levando os métodos híbridos a preservarem maior fidelidade lexical, mantendo terminologia específica vantajosa para análises requerendo citação literal [Jain et al. 2024]. Essas atributos tornam híbridos atrativos para interfaces de busca jurisprudencial com visualização rápida; sistemas de classificação com janelas de contexto restritas; ou processamento em larga escala onde redução de *tokens* viabiliza análise mantendo qualidade adequada.

4.3. Implicações práticas e viabilidade computacional

Além da efetividade medida por métricas automáticas, considerações práticas de implementação devem ser ponderadas. O Gemini-2.5, embora superior em desempenho, apresenta custos significativamente maiores, onde o processamento de contexto extenso implica latência de vários segundos e custos por requisição. Esta tensão entre qualidade e viabilidade econômica foi sistematicamente documentada por [Janakiraman and Ghoraani 2025], que identificaram Gemini-1.5-Flash oferecendo me-

Referência: direito previdenciário. agravo regimental em recurso extraordinário com agravo. benefício previdenciário. natureza alimentar. recebimento de boa-fé em decorrência de decisão judicial. tutela antecipada revogada. devolução. 1. a jurisprudência do supremo tribunal federal já assentou que o benefício previdenciário recebido de boa-fé pelo segurado, em decorrência de decisão judicial, não está sujeito à repetição de indébito, em razão de seu caráter alimentar. precedentes. 2. decisão judicial que reconhece a impossibilidade de descontos dos valores indevidamente recebidos pelo segurado não implica declaração de inconstitucionalidade do art. 115 da lei nº 8.213/1991. precedentes. 3. agravo regimental a que se nega provimento.

Gemini-2.5: AGRAVO REGIMENTAL. RECURSO EXTRAORDINÁRIO. PREVIDENCIÁRIO. BENEFÍCIO RECEBIDO DE BOA-FÉ. REPETIÇÃO DE INDÉBITO. 1. O benefício previdenciário recebido de boa-fé pelo segurado, em decorrência de decisão judicial, não está sujeito à repetição de indébito, em virtude de seu caráter alimentar. 2. A decisão judicial que reconhece a impossibilidade de descontos dos valores indevidamente recebidos pelo segurado não implica declaração de inconstitucionalidade do art. 115 da Lei nº 8.213/1991. 3. Agravo regimental desprovido.

Híbrido TFIDF: AGRAVO REGIMENTAL. RECURSO EXTRAORDINÁRIO. PREVIDENCIÁRIO. 1. Benefício previdenciário recebido de boa-fé pelo segurado, em decorrência de decisão judicial, não está sujeito à repetição de indébito, em razão de seu caráter alimentar. 2. Decisão judicial que reconhece a impossibilidade de descontos dos valores indevidamente recebidos pelo segurado não implica declaração de inconstitucionalidade do art. 115 da Lei nº 8.213/1991. 3. Viola a cláusula de reserva de plenário (art. 97 da CF) a decisão de órgão fracionário de tribunal que, sem declarar expressamente a inconstitucionalidade de lei, afasta sua incidência. Agravo regimental desprovido.

Híbrido <omitido para revisão>: AGRAVO REGIMENTAL. RECURSO EXTRAORDINÁRIO. PREVIDENCIÁRIO. RESTITUIÇÃO DE VALORES. 1. Benefício previdenciário recebido de boa-fé pelo segurado, em decorrência de decisão judicial, não está sujeito à repetição de indébito, em razão de seu caráter alimentar. 2. Decisão judicial que reconhece a impossibilidade de descontos dos valores indevidamente recebidos pelo segurado não implica declaração de inconstitucionalidade do art. 115 da Lei nº 8.213/1991. 3. Viola a cláusula de reserva de plenário (art. 97 da CF) a decisão de órgão fracionário de tribunal que afasta a incidência de lei ou ato normativo, sem declará-la inconstitucional. Agravo regimental desprovido.

Figura 3. Exemplo comparativo de resumos gerados pelos diferentes métodos

lhor custo-benefício ($0,00012/resumo) entre 17 modelos avaliados, porém ainda assim sendo uma das opções custo-benefício.

Os métodos extrativos, particularmente TF-IDF, oferecem alternativa computacionalmente eficiente que pode ser executada em hardware convencional sem dependência de serviços externos. O *<omitido para revisão>*, embora requeira GPU para processamento eficiente, permite implantação local. Esta autonomia é particularmente relevante para instituições públicas brasileiras que podem enfrentar restrições orçamentárias ou políticas de privacidade que desencorajam 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 diferentes contextos de aplicação. Sistemas de produção com volumes elevados podem beneficiar-se de métodos extrativos rápidos, enquanto aplicações onde 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 avaliação humana estruturada considerando a precisão jurídica,

completude de fundamentos, conformidade com padrões estilísticos do CNJ, e adequação para diferentes perfis de uso (advogados, juízes, pesquisadores).

Para além, o método abstrativo dependeu de modelo proprietário específico (Gemini-2.5) cujos detalhes arquiteturais e de treinamento não são públicos. Investigação futura deve explorar modelos abertos como LLaMA, Mistral, Sabiá, permitindo análise mais profunda de características que contribuem para desempenho e possibilitando ajuste fino com dados jurídicos brasileiros.

Quanto ao escopo, todos os métodos foram avaliados exclusivamente em decisões do STF (5.792 documentos, 54,8% do corpus), sem considerar variações estilísticas entre relatores ou áreas do direito. Generalização para outros gêneros (petições, contratos, pareceres, sentenças de instâncias inferiores) ou tribunais requer validação adicional, considerando que aspectos estruturais variam substancialmente entre tipos documentais.

Por fim, o corpus RulingBR foi coletado em 2018, anterior à Recomendação CNJ n.º 154/2024 que estabelece novos padrões de estruturação de ementas. Trabalhos futuros devem investigar 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 *<omitido para revisão>*), abstrativa pura (Gemini-2.5) e híbridas utilizando ambos os métodos. Os resultados indicam que o método abstrativo alcançou desempenho superior em métricas ROUGE (ROUGE-1: 0.4524), representando ganho de 45.9% sobre *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* para *<omitido para revisão>*, viabilizando análise de documentos longos sem perda de representações semânticas especializadas. Metodologicamente, o estudo estabelece como equilibrar qualidade técnica com viabilidade computacional em contextos de recursos limitados, um desafio ainda inerente na infraestrutura do setor público. Os métodos híbridos mostraram-se adequados para aplicações como interfaces de busca jurisprudencial e sistemas de classificação, onde restrições de contexto e latência são determinantes.

Assim, o estudo alinha-se aos Grandes Desafios de SI no Mundo Aberto [Boscarioli et al. 2017], especificamente aos desafios de relacionados a governança de dados no setor público, propondo estratégias de processamento que consideram limitações sistêmicas; transparência institucional, ao viabilizar acesso mais eficiente a decisões judiciais através 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, redução de tempo em revisão de precedentes e pesquisa jurisprudencial; para instituições judiciais, apoio à produção de ementas e classificação documental em larga escala; para cidadãos, melhoria no acesso à informação jurídica através de sistemas mais eficientes. Ao fornecer evidências empíricas sobre estratégias

eficazes considerando limitações reais do contexto brasileiro e de países do Sul Global, o estudo contribui para desenvolvimento de sistemas de apoio à decisão escaláveis, transparentes e contextualmente adequados.

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

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.

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 grandsibr: Grand research challenges in information systems in brazil 2016-2026.

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.

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.

APÊNDICE B – Relatório WandB

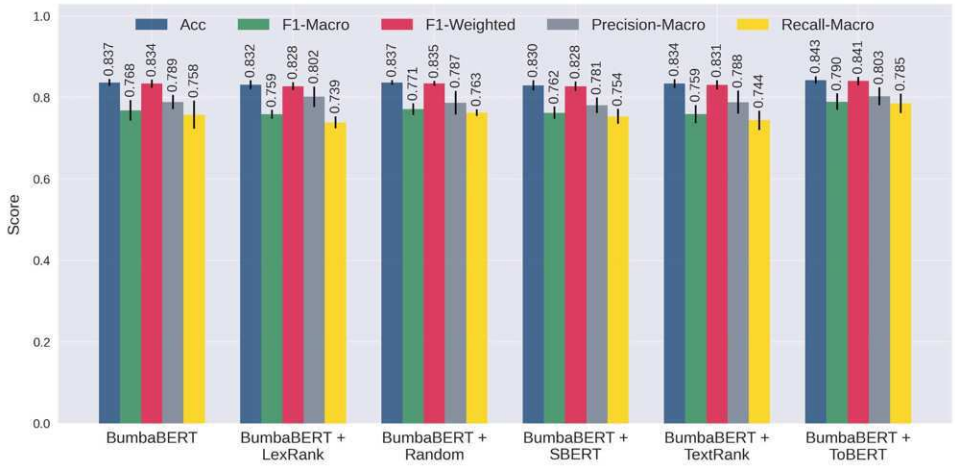
Classificação de Documentos Jurídicos Longos

Este relatório contém os registros experimentais da dissertação de Mestrado da discente Gabriele de Sousa Araújo apresentada ao Programa de Pós-Graduação em Engenharia da Computação e Sistemas (PECS) da Universidade Estadual do Maranhão (UEMA). No estudo foram avaliados métodos para classificação de petições iniciais longas em temas de IRDR. Metodologia: Validação Cruzada Estratificada (5-folds). Total de Experimentos: 40 execuções (8 modelos x 5).

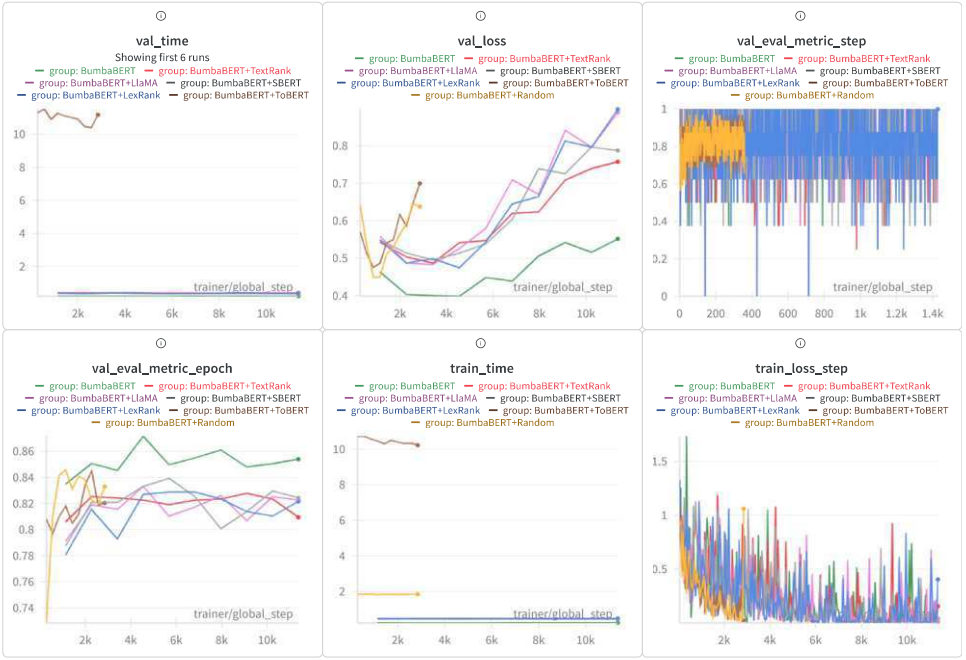
Gabriele Araújo

Created on September 18 | Last edited on November 27

Performance da classificação de petições iniciais (predições)



Dados de treinamento e validação do melhor *fold*

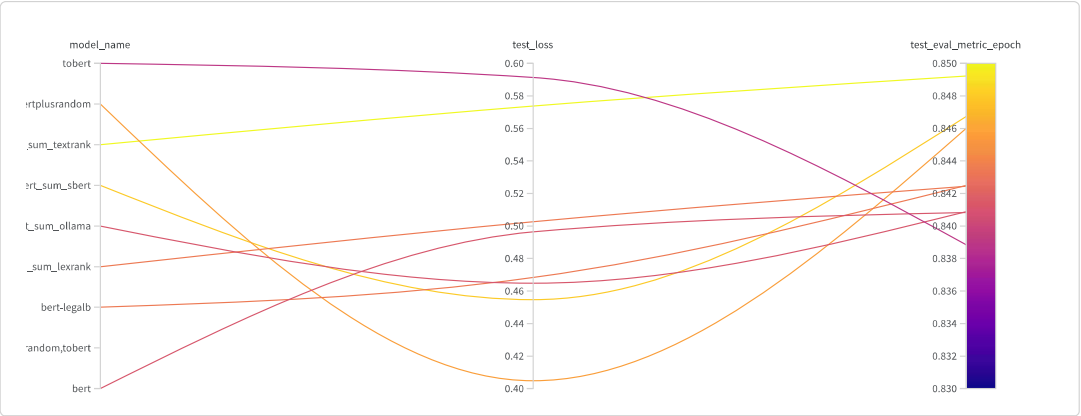


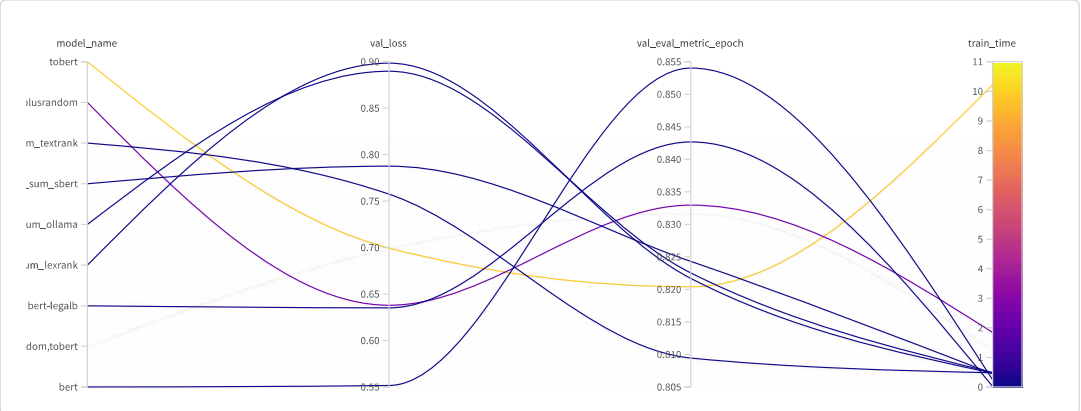
Matriz de confusão dos melhores *folds*

É possível controlar a exibição na tabela após esse painel em Filtro1

Resultado detalhado das matrizes

Comparativo dos modelos (*folds*)





Eficiência Computacional

Acordo de Cooperação Técnica N° 002/2021 celebrado entre o Tribunal de Justiça do Maranhão (TJMA) e a Universidade Estadual do Maranhão (UEMA)

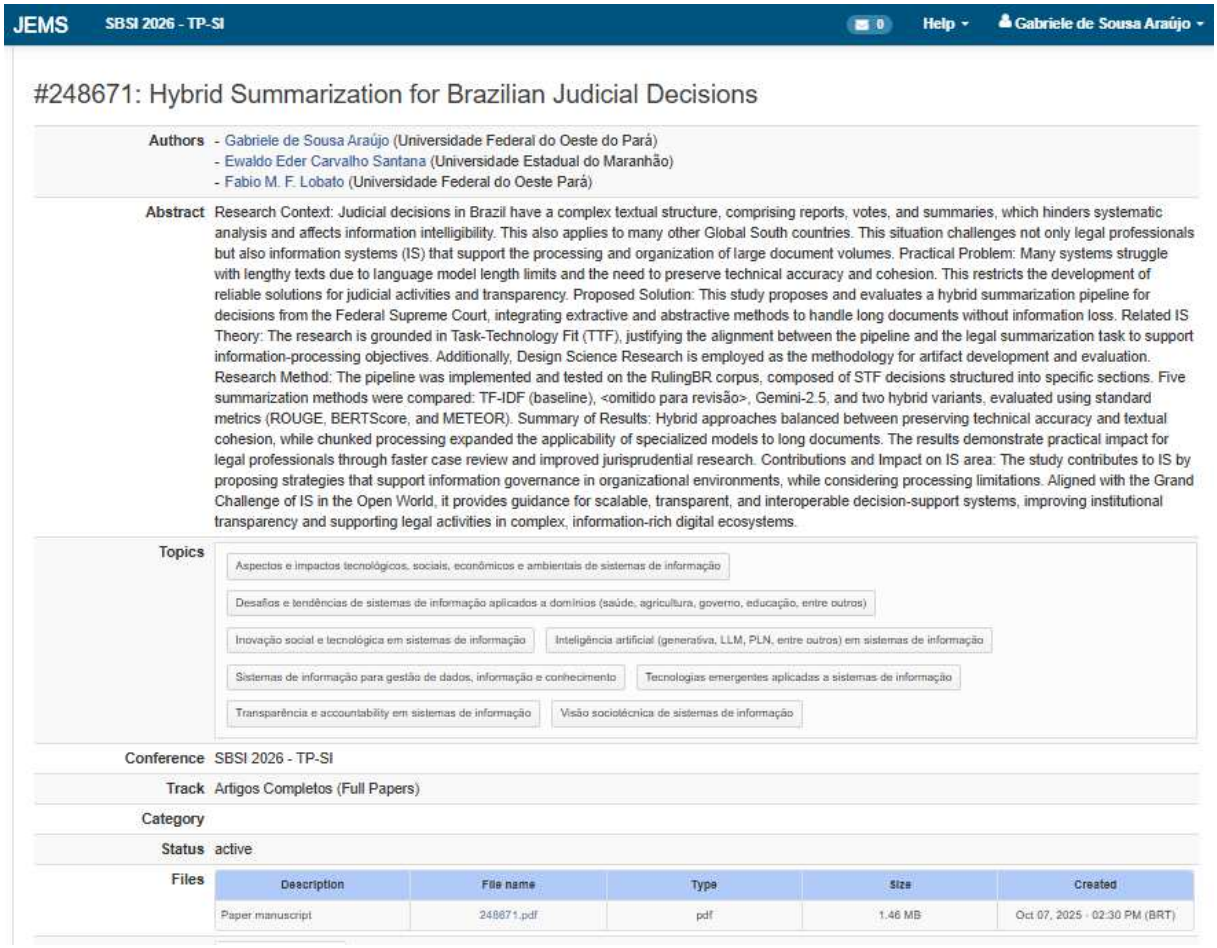
Created with ❤️ on Weights & Biases.

https://wandb.ai/gabiaraujo-state-university-of-maranh-o/train_long_docs_kfold_/reports/Classifica-o-de-Documentos-Jur-dicos-Longos--VmlldzoxNDQzOTI2Mw?accessToken=dxzm1kngcsf29srwukbc762p5ve24cr77p2xyd254nko7h7o23194h3d405qc07r

Anexos

ANEXO A - Comprovante do artigo submetido ao SBSI 2026

Comprovante de submissão do artigo *"Hybrid Summarization for Brazilian Judicial Decisions"* ao SBSI 2026.



JEMS SBSI 2026 - TP-SI 0 Help Gabriele de Sousa Araújo

#248671: Hybrid Summarization for Brazilian Judicial Decisions

Authors
- Gabriele de Sousa Araújo (Universidade Federal do Oeste do Pará)
- Ewaldo Eder Carvalho Santana (Universidade Estadual do Maranhão)
- Fabio M. F. Lobato (Universidade Federal do Oeste Pará)

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 Global South countries. This situation challenges not only legal professionals but also information systems (IS) that support the processing and organization of large document volumes. Practical Problem: Many systems struggle with lengthy texts due to language model length limits and the need to preserve technical accuracy and cohesion. This restricts the development of reliable solutions for judicial activities and transparency. Proposed Solution: 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, composed of STF decisions structured into specific sections. Five summarization methods were compared: TF-IDF (baseline), <omitido para revisão>, Gemini-2.5, and two hybrid variants, evaluated using standard metrics (ROUGE, BERTScore, and METEOR). Summary of Results: Hybrid approaches balanced between preserving technical accuracy and textual cohesion, while chunked processing expanded the applicability of specialized models to long documents. The results demonstrate practical impact for legal professionals 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, information-rich digital ecosystems.

Topics
- Aspectos e impactos tecnológicos, sociais, econômicos e ambientais de sistemas de informação
- Desafios e tendências de sistemas de informação aplicados a domínios (saúde, agricultura, governo, educação, entre outros)
- Inovação social e tecnológica em sistemas de informação
- Inteligência artificial (generativa, LLM, PLN, entre outros) em sistemas de informação
- Sistemas de informação para gestão de dados, informação e conhecimento
- Tecnologias emergentes aplicadas a sistemas de informação
- Transparência e accountability em sistemas de informação
- Visão sociotécnica de sistemas de informação

Conference SBSI 2026 - TP-SI

Track Artigos Completos (Full Papers)

Category

Status active

Files

Description	File name	Type	Size	Created
Paper manuscript	248671.pdf	pdf	1.46 MB	Oct 07, 2025 - 02:30 PM (BRT)