UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Pedro Vinnícius Bernhard

Método Não Supervisionado de Sumarização Extrativa de Textos Jurídicos com Alinhamento de Grafos Semânticos Guiados por Atenção

São Luís
2026

Pedro Vinnícius Bernhard

Método Não Supervisionado de Sumarização Extrativa de Textos Jurídicos com Alinhamento de Grafos Semânticos Guiados por Atenção

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Programa de Pós-Graduação em Ciência da Computação

Universidade Federal do Maranhão

Orientador: Prof. Dr. João Dallyson Sousa de Almeida

Coorientador: Prof. Dr. Anselmo Cardoso de Paiva

São Luís - MA

2026

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Bernhard, Pedro Vinnícius.
Método Não Supervisionado de Sumarização Extrativa de Textos Jurídicos com Alinhamento de Grafos Semânticos Guiados por Atenção / Pedro Vinnícius Bernhard. - 2026.
105 f.

Coorientador(a) 1: Anselmo Cardoso de Paiva.
Orientador(a): João Dallyson Sousa de Almeida.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, Auditório do Nca Ufma, São Luís, Ma, 2026.

1. Sumarização Extrativa. 2. Processamento de Linguagem Natural Jurídico. 3. Grafos Semânticos. 4. Atenção. 5. Aprendizado Não Supervisionado. I. Almeida, João Dallyson Sousa de. II. Paiva, Anselmo Cardoso de. III. Título.

Pedro Vinnícius Bernhard

Método Não Supervisionado de Sumarização Extrativa de Textos Jurídicos com Alinhamento de Grafos Semânticos Guiados por Atenção

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Prof. Dr. João Dallyson Sousa de Almeida
Orientador
Universidade Federal do Maranhão

Prof. Dr. Anselmo Cardoso de Paiva
Coorientador
Universidade Federal do Maranhão

Prof. Dr. Darlan Bruno Pontes Quintanilha
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. Leandro Balby Marinho
Examinador Externo
Universidade Federal de Campina Grande

São Luís - MA
2026

Para Rosália, minha constante em meio às variáveis.

Agradecimentos

Agradeço à minha esposa, Rosália, pelo apoio, incentivo constante e por estar ao meu lado nesta jornada. Aos amigos e colegas, pela companhia e pela colaboração neste percurso. A todos os meus professores, cujos ensinamentos abriram meus horizontes e fortaleceram minha formação científica. Agradeço aos meus orientadores, João Dallyson Sousa de Almeida e Anselmo Cardoso de Paiva, pela paciência, pela confiança que depositaram em mim e pela orientação durante a graduação e mestrado. A todos os integrantes do NCA, pela colaboração e pelas valiosas discussões acadêmicas. À Universidade Federal do Maranhão (UFMA) e PPGCC, pela infraestrutura e oportunidade oferecidas. Agradeço à FAPEMA pelo apoio financeiro necessário à viabilização deste trabalho.

"O verdadeiro perigo não é que os computadores passem a pensar como os homens, mas sim que os homens passem a pensar como os computadores."

(Sydney Harris)

Resumo

O volume massivo e a complexidade técnica dos documentos jurídicos no Brasil impõem um grande desafio à celeridade do sistema judiciário. A sumarização automática surge como uma alternativa para mitigar essa sobrecarga e auxiliar o trabalho de magistrados e advogados. No entanto, a aplicação de modelos de aprendizado profundo no Direito enfrenta obstáculos críticos: a opacidade algorítmica ("caixa-preta"), o risco inaceitável de alucinações factuais em modelos generativos e a severa escassez de dados rotulados para treinamento. Dessa forma, o desenvolvimento de soluções que unam fidelidade fática e interpretabilidade é essencial. Neste contexto, este trabalho propõe um método não supervisionado de sumarização extrativa focado no domínio jurídico, estruturado na modelagem de grafos semânticos guiados por mecanismos de atenção. O método extrai os pesos de autoatenção de um modelo de linguagem especialista (Legal-BERTimbau) e filtra conexões ruidosas via binarização dinâmica pelo método de Otsu. O texto é convertido em um grafo direcionado, particionado tematicamente pelo algoritmo Infomap Hierárquico para isolar os eixos argumentativos. O alinhamento dos tópicos é realizado em um espaço vetorial denso (Sentence-BERT), e as sentenças são ranqueadas pela heurística de Atenção Máxima, respeitando um limite estrito de compressão de 10%. Na avaliação utilizando a base de dados RulingBR, o modelo proposto superou os algoritmos clássicos não supervisionados nas métricas ROUGE-1 (36,61%) e ROUGE-L (20,74%). Experimentos adicionais com um Oráculo Extrativo demarcaram o limite superior da tarefa em um ROUGE-1 de 65,21% e ROUGE-L de 47,37%, enquanto uma abordagem híbrida extrativa guiada por um LLM (GPT-5 mini) alcançou um ROUGE-L de 21,31%. Assim, o método desenvolvido demonstra-se promissor ao garantir a integridade do texto original, livre de alucinações processuais, oferecendo adicionalmente uma interface de explicabilidade visual que torna a seleção de sentenças totalmente auditável.

Palavras-chave: Sumarização Extrativa, Processamento de Linguagem Natural Jurídico, Grafos Semânticos, Atenção, Interpretabilidade de Modelos, Aprendizado Não Supervisionado.

Abstract

The massive volume and technical complexity of legal documents in Brazil impose a major challenge to the efficiency of the judicial system. Automatic summarization emerges as an alternative to mitigate this overload and assist the work of judges and lawyers. However, the application of deep learning models in Law faces critical obstacles: algorithmic opacity ("black-box"), the unacceptable risk of factual hallucinations in generative models, and the severe scarcity of labeled data for training. Thus, the development of solutions that unite factual fidelity and interpretability is essential. In this context, this work proposes an unsupervised extractive summarization method focused on the legal domain, structured on the modeling of semantic graphs guided by attention mechanisms. The method extracts self-attention weights from an expert language model (Legal-BERTimbau) and filters noisy connections via dynamic binarization using Otsu's method. The text is converted into a directed graph, thematically partitioned by the Hierarchical Infomap algorithm to isolate the argumentative axes. Topic alignment is performed in a dense vector space (Sentence-BERT), and sentences are ranked by the Maximum Attention heuristic, respecting a strict compression limit of 10%. In the evaluation using the RulingBR dataset, the proposed model outperformed classical unsupervised algorithms in the ROUGE-1 (36.61%) and ROUGE-L (20.74%) metrics. Additional experiments with an Extractive Oracle demarcated the upper bound of the task at a ROUGE-1 of 65.21% and ROUGE-L of 47.37%, while a hybrid extractive approach guided by an LLM (GPT-5 mini) achieved a ROUGE-L of 21.31%. Thus, the developed method proves promising by ensuring the integrity of the original text, free from procedural hallucinations, additionally offering a visual explainability interface that makes sentence selection fully auditable.

Keywords: Extractive Summarization, Legal Natural Language Processing, Attention Graphs, Model Interpretability, Unsupervised Learning.

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas e siglas

BERT	*Bidirectional Encoder Representations from Transformers* – Representações de Codificador Bidirecional de *Transformers*
BLEU	*Bilingual Evaluation Understudy – Assistente de Avaliação Bilíngue*
LCS	*Longest Common Subsequence – Maior Subsequência Comum*
LLMs	*Large Language Models – Grandes Modelos de Linguagem*
LSA	*Latent Semantic Analysis – Análise Semântica Latente*
MLM	*Masked Language Modeling – Modelagem de Linguagem Mascarada*
MVM	*Masked Language Modeling – Modelagem de Linguagem Mascarada*
NSP	*Next Sentence Prediction – Previsão da Próxima Sentença*
OFAT	*One-Factor-At-a-Time – Um Fator Por Vez*
PCA	*Principal Component Analysis – Análise de Componentes Principais*
PLN	*Processamento de Linguagem Natural*
RGPD	*Regulamento Geral sobre a Proteção de Dados*
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation – Avaliação de Resumo Orientado à Recuperação*
SBERT	*Sentence-BERT*
STS	*Semantic Textual Similarity – Similaridade Textual Semântica*
SVD	*Singular Value Decomposition – Decomposição em Valores Singulares*
TF-IDF	*Term Frequency-Inverse Document Frequency – Frequência do Termo-Inverso da Frequência nos Documentos*

Sumário