



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

Figura 1 – Arquitetura do <i>Transformer</i>	29
Figura 2 – Arquitetura do SBERT durante inferência	31
Figura 3 – Diagrama metodológico detalhado do <i>pipeline</i> de sumarização proposto	44
Figura 4 – Interface principal do Módulo de Explicabilidade Visual	59
Figura 5 – Realce de Texto	59
Figura 6 – Distribuição Posicional	60
Figura 7 – Nuvens de Palavras Comparativas	60
Figura 8 – Mapa de Calor do Foco Global	61
Figura 9 – Visualização da Matriz de Atenção	61
Figura 10 – Exploração Interativa do Grafo	62
Figura 11 – Mapeamento Textual de Comunidades	62
Figura 12 – Espaço Semântico PCA	63
Figura 13 – Matriz de Similaridade Cruzada	63
Figura 14 – Distribuição da quantidade de tokens do texto completo (Dataset Completo)	65
Figura 15 – Distribuição da proporção (em %) da ementa em relação ao texto completo	66
Figura 16 – Desempenho Comparativo entre Modelos Generalistas e Especialistas (F1-Score)	69
Figura 17 – Desempenho comparativo dos Algoritmos de Detecção de Comunidades	75
Figura 18 – Desempenho Comparativo das Estratégias de Busca e Seleção de <i>Clusters</i>	77
Figura 19 – Desempenho Comparativo das Métricas de Ranqueamento de Sentenças	78
Figura 20 – Desempenho Comparativo das Estratégias de Compressão de Texto	79
Figura 21 – Mapa de Calor da Atenção Global	87
Figura 22 – Detecção de Comunidades e Seleção de Sentenças	89
Figura 23 – Detecção de Comunidades e Seleção de Sentenças	90
Figura 24 – Mapa de Calor da Atenção Global	91

Lista de Tabelas

Tabela 1 – Datasets Jurídicos Brasileiros	26
Tabela 2 – Estatísticas descritivas de tokens e proporção de compressão da base RulingBR	65
Tabela 3 – Desempenho comparativo dos métodos de Agregação de Cabeças (<i>Heads</i>)	71
Tabela 4 – Desempenho comparativo dos métodos de Agregação de Camadas (<i>Layers</i>)	71
Tabela 5 – Desempenho comparativo dos métodos de Filtragem de Ruído (Token [CLS])	72
Tabela 6 – Desempenho comparativo da Direcionalidade do Grafo	73
Tabela 7 – Desempenho comparativo dos Limiares de Poda de Arestas	74
Tabela 8 – Desempenho comparativo das Heurísticas de Empacotamento (<i>Knapsack</i>)	80
Tabela 9 – Estudo de Ablação: Impacto da Limpeza de Texto no Desempenho do Modelo	81
Tabela 10 – Avaliação do Alinhamento Semântico com diferentes Textos-Alvo	82
Tabela 11 – Desempenho comparativo contra <i>baselines</i> , literatura, abordagens guia- das e Oráculo	85

Lista de abreviaturas e siglas

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

Sumário

1	INTRODUÇÃO	16
1.1	Contextualização	16
1.2	Definição do Problema	17
1.3	Justificativa	18
1.4	Questões de Pesquisa	19
1.5	Objetivos	19
1.5.1	Objetivo Geral	19
1.5.2	Objetivos Específicos	19
1.6	Contribuições	20
1.7	Organização do Trabalho	20
2	FUNDAMENTAÇÃO TEÓRICA	22
2.1	Processamento de Linguagem Natural e Sumarização	22
2.1.1	A Evolução do Processamento de Linguagem Natural	22
2.1.2	Fundamentos da Sumarização Automática	23
2.1.3	Desafios do Domínio Jurídico e Recursos no Brasil	24
2.2	Métodos Clássicos e Não Supervisionados	25
2.2.1	Sumarização Não Supervisionada no Contexto Jurídico	25
2.2.2	Algoritmos Clássicos e Modelos de Base (<i>Baselines</i>)	27
2.3	Arquitetura Transformer e Mecanismos de Atenção	29
2.3.1	O Modelo BERT e o Mecanismo de Autoatenção	29
2.3.2	Adaptações do BERT: SBERT, BERTimbau e Legal-BERTimbau	30
2.3.3	Extração de Atenção e Processamento de Documentos Longos	32
2.4	Teoria de Grafos e Detecção de Comunidades	33
2.4.1	Representação de Textos em Grafos	33
2.4.2	Modularidade e Detecção de Comunidades	33
2.4.3	Algoritmos de Particionamento e Otimização	34
2.4.4	Filtragem via Método de Otsu	35
2.5	Redução de Dimensionalidade e Análise de Componentes Principais (PCA)	36
2.6	Métricas de Avaliação de Sumarização	37
2.6.1	ROUGE	37
2.6.2	BERTScore	38
2.6.3	BLEU	38
2.7	Considerações Finais	39

3	TRABALHOS RELACIONADOS	40
3.1	Abordagens Lexicais e Supervisionadas Iniciais	40
3.2	Clusterização Semântica e Modelos Não Supervisionados	41
3.3	Atenção, Estrutura Semântica e Explicabilidade	42
3.4	Modelos para Documentos Longos e Eficiência Computacional	43
4	MATERIAIS E MÉTODO	44
4.1	Aquisição e Estruturação da Base de Dados	45
4.2	Fundamentos do Método Proposto	46
4.3	Pré-processamento, Tokenização e Segmentação	47
4.4	Processamento do Cubo de Atenção e Extração de Sentenças	48
4.4.1	Agregação do Cubo de Atenção	49
4.4.2	Filtragem de Ruído Global e Extração de Sentenças	51
4.5	Modelagem em Grafos e Detecção de Comunidades	52
4.5.1	Topologia do Grafo	52
4.5.2	Detecção de Comunidades	54
4.6	Alinhamento Semântico e Seleção de Sentenças	54
4.6.1	Estratégia de Busca e Alinhamento Semântico	55
4.6.2	Ranqueamento Interno e Critério de Compressão	56
4.7	Módulo de Explicabilidade Visual e API de Introspecção	57
4.7.1	Projeções Lexicais e Posicionais	59
4.7.2	Projeções Topológicas e Estruturais	60
4.7.3	Projeções Semânticas e Avaliação	62
5	RESULTADOS E DISCUSSÕES	64
5.1	Base de Dados e Estatísticas Descritivas	64
5.2	Configuração Experimental e Métricas de Avaliação	66
5.3	Otimização Sequencial de Hiperparâmetros	67
5.3.1	Seleção dos Modelos de Linguagem (<i>Reader</i> e <i>Encoder</i>)	68
5.3.2	Agregação de Atenção e Filtragem de Ruído	70
5.3.3	Topologia do Grafo e Detecção de Comunidades	73
5.3.4	Alinhamento Semântico e Seleção de Sentenças	75
5.3.5	Critérios de Compressão e Formatação Final	78
5.3.6	Estudo de Ablação do Pré-processamento (Limpeza de Texto)	80
5.3.7	Alinhamento com Padrão-Ouro e Sumarização Guiada por LLMs	81
5.4	Discussão Geral e Estudos de Caso	83
5.4.1	Comparação com Algoritmos Clássicos e Estado da Arte	83
5.4.2	Análise Diagnóstica dos Extremos: Limitações e Falhas (Piores Casos)	86
5.4.3	Análise Diagnóstica dos Extremos: Sucesso e Alinhamento Fático (Melhores Casos)	89

5.4.4	Validação das Questões de Pesquisa e Contribuições Metodológicas	92
6	CONCLUSÃO	94
6.1	Vantagens e Limitações	95
6.2	Contribuições	96
6.3	Trabalhos Futuros	96
6.4	Produção Científica	97
6.5	Declaração de Uso de Inteligência Artificial	97
	 REFERÊNCIAS BIBLIOGRÁFICAS	 99