UNIVERSIDADE FEDERAL DO TOCANTINS
CÂMPUS DE PALMAS
PROGRAMA DE PÓS-GRADUAÇÃO MESTRADO PROFISSIONAL EM MODELAGEM COMPUTACIONAL

MÁRCIA HASIMOTO

ANÁLISE TEXTUAL DAS DECISÕES PROFERIDAS EM COLEGIADO DO TRIBUNAL DE JUSTIÇA DO TOCANTINS

Palmas / TO
2025

Márcia Hasimoto

ANÁLISE TEXTUAL DAS DECISÕES PROFERIDAS EM COLEGIADO DO TRIBUNAL DE JUSTIÇA DO TOCANTINS

Dissertação apresentada ao Programa de Pós-Graduação em Mestrado Profissional em Modelagem Computacional de Sistemas da Universidade Federal do Tocantins, como pré-requisito parcial para obtenção do título de Mestre em Modelagem Computacional de Sistemas.

Orientador: Prof. Dr. Rogério Nogueira de Sousa

Palmas / TO
2025

https://sistemas.uft.edu.br/ficha/

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

H348a Hasimoto, Márcia.

Análise Textual das Decisões Proferidas em Colegiado do Tribunal de Justiça do Tocantins. / Márcia Hasimoto. – Palmas, TO, 2025.

62 f.

Dissertação (Mestrado Profissional) - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Pós-Graduação (Mestrado Profissional) em Governança e Transformação Digital - PPGGTD, 2025.

Orientador: Rogério Nogueira de Sousa

1. Processamento de Linguagem Natural. 2. Agrupamento de Textos. 3. SBERT. 4. Inteligência Artificial. I. Título

CDD 004

Elaborado pelo sistema de geração automática de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

Márcia Hasimoto

ANÁLISE TEXTUAL DAS DECISÕES PROFERIDAS EM COLEGIADO DO TRIBUNAL DE JUSTIÇA DO TOCANTINS

Dissertação apresentada ao Programa de Pós-Graduação em Mestrado Profissional em Modelagem Computacional de Sistemas da Universidade Federal do Tocantins, como pré-requisito parcial para obtenção do título de Mestre em Modelagem Computacional de Sistemas.

Orientador: Prof. Dr. Rogério Nogueira de Sousa

Data de aprovação: 05 /09 /2025

Banca Examinadora

_____

Prof. Dr. Rogério Nogueira de Souza - UFT

_____

Prof. Dr. David Nadler Prata – PPGGTD/UFT

_____

Prof. Dr. Jefferson David Asevedo Ramos - TJTO

À minha mãe, in memoriam, por todo o amor, força e exemplo que me acompanham até hoje. Esta conquista também é sua.

AGRADECIMENTOS

Agradeço a minha família pela paciência e apoio.

Aos professores do Programa de Pós-graduação de Modelagem Computacional de Sistema da Universidade Federal do Tocantins, em especial ao Prof. Dr. Marcelo Lisboa, Prof. Dr. Gentil Veloso e Prof. Dr. Rogério Nogueira.

RESUMO

A crescente demanda por celeridade e eficiência no Poder Judiciário tem impulsionado a adoção de soluções baseadas em Inteligência Artificial. Neste contexto, esta dissertação propõe uma metodologia fundamentada em Processamento de Linguagem Natural (PLN) e aprendizado não supervisionado para o agrupamento de acórdãos judiciais, com foco na triagem de pedidos de recurso submetidos ao gabinete da presidência de tribunais. Parte-se da hipótese de que a organização automatizada de decisões judiciais por similaridade semântica pode contribuir para a identificação de temas repetitivos ou já pacificados, evitando o encaminhamento de recursos desnecessários às instâncias superiores. A metodologia abrangeu as etapas de pré-processamento textual, vetorização com modelos da família SBERT com destaque para o modelo jurídico *stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0*, aplicação do algoritmo K-Means para clusterização, redução de dimensionalidade com t-SNE e avaliação da qualidade dos agrupamentos por meio das métricas internas *Silhouette Score*, *Davies-Bouldin Index e Calinski-Harabasz Index*. A análise qualitativa foi conduzida por meio de técnicas de modelagem de tópicos com *Latent Dirichlet Allocation (LDA)*, visualizações interativas com a ferramenta *PyLDAvis* e geração de nuvens de palavras, permitindo uma interpretação aprofundada dos agrupamentos. Como resultado, observou-se a formação de grupos tematicamente consistentes, com destaque para categorias como Direito Fiscal, Sucessório, Administrativo e Responsabilidade Civil. Os achados demonstram que a abordagem proposta é eficaz na organização de acórdãos por similaridade semântica, apresentando potencial prático para apoiar a triagem e a organização de processos judiciais especialmente em contextos com grande volume documental e ausência de rotulagem prévia.

Palavras-chaves: Processamento de Linguagem Natural; Agrupamento de Textos; SBERT; Acórdãos Judiciais; Inteligência Artificial.

ABSTRACT

The growing demand for speed and efficiency in the Judiciary has driven the adoption of solutions based on Artificial Intelligence. In this context, this dissertation proposes a methodology grounded in Natural Language Processing (NLP) and unsupervised learning for clustering judicial decisions, with a focus on screening appeals submitted to the office of court presidencies. The underlying hypothesis is that automated organization of judicial decisions based on semantic similarity can contribute to identifying repetitive or settled themes, thereby avoiding the submission of unnecessary appeals to higher courts. The proposed methodology comprises several stages, including text preprocessing, vectorization using SBERT family models with emphasis on the legal-domain model *stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0* clustering with the K-Means algorithm, dimensionality reduction via t-SNE, and evaluation of clustering quality through internal metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. Qualitative analysis was conducted using topic modeling with Latent Dirichlet Allocation (LDA), interactive visualizations with PyLDAvis, and word cloud generation, allowing a deeper interpretation of the resulting clusters. The results showed the formation of thematically consistent groups, with emphasis on categories such as Tax Law, Probate Law, Administrative Law, and Civil Liability. These findings demonstrate that the proposed approach is effective in organizing judicial decisions by semantic similarity, offering practical potential to support the screening and organization of judicial processes especially in contexts involving large document volumes and lack of labeled data.

Keywords: Natural Language Processing; Text Clustering; SBERT; Judicial Decisions; Artificial Intelligence.

LISTA DE ILUSTRAÇÃO

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CNJ	Conselho Nacional de Justiça
BNP	Banco Nacional de Precedentes
NUGEP	Núcleos de Gerenciamento de Precedentes
CPC	Código de Processo Civil
PLN	Processamento de Linguagem Natural
BERT	Bidirectional Encoder Representations from Transformers
SBERT	Sentence-BERT
IA	Inteligência Artificial
LDA	Latent Dirichlet Allocation
t-SNE	Technique Distributed Stochastic Neighbor Embedding
NLU	Natural Language Understanding
NLG	Natural Language Generation
SVM	Máquinas de Vetor de Suporte
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
BoW	Bag-of-Words
TF-IDF	Term Frequency-Inverse Document Frequency
Word2Vec	Técnica que transforma palavras em vetores
K-Means	Algoritmo de Agrupamento Não Supervisionado
PCA	Análise de Componentes Principais
WCSS	Within-Cluster Sum of Squares
GloVe	Global Vectors for Word Representation
e-Proc	Processo Eletrônico
PDF	Portable Document Format (Formato Portátil de Documento)
EPUB	*Electronic Publication* (Publicação Eletrônica)
XPS	*XML Paper Specification* (Especificação de Documento XML)
NLTK	Natural Language Toolkit
AM	Aprendizado de Máquina
DBI	Davies-Bouldin Index
CHI	Calinski-Harabasz Index

LISTA DE SÍMBOLOS

$\cos(\theta)$	Cosseno do ângulo θ entre os vetores A e B
θ	*Ângulo entre os vetores A e B*
$A \cdot B$	Produto escalar entre os vetores A e B
K	Número de clusters utilizados no algoritmo de agrupamento

SUMÁRIO

1 INTRODUÇÃO

A padronização das informações jurídicas e a uniformização da jurisprudência configuram desafios centrais para o aprimoramento do sistema judiciário brasileiro. Diante disso, o Conselho Nacional de Justiça (CNJ) estabeleceu diretrizes e regulamentações voltadas à consolidação de precedentes, com o objetivo de promover maior coerência, previsibilidade e eficiência na tramitação dos processos. Entre essas iniciativas, destacam-se a Portaria nº 116/2022, que define os requisitos para a padronização das informações a serem fornecidas pelos tribunais e pela Turma Nacional de Uniformização dos Juizados Especiais Federais, visando à alimentação do Banco Nacional de Precedentes (BNP); a Resolução nº 235/2016, que institui os Núcleos de Gerenciamento de Precedentes (NUGEP); e a Resolução nº 444/2022, que formaliza o BNP como repositório digital acessível ao público e aos órgãos do Judiciário. Tais medidas buscam assegurar maior segurança jurídica e reduzir o tempo de tramitação dos processos, favorecendo decisões mais eficientes e alinhadas aos entendimentos já consolidados (BRASIL, 2016; 2022a; 2022b).

Nesse contexto, o Tribunal de Justiça do Tocantins enfrenta o desafio de realizar manualmente a triagem e análise de acórdãos nos processos que apresentam pedidos de recurso antes de seu encaminhamento às instâncias superiores. Conforme dispõe o artigo 204 do Código de Processo Civil (CPC), acórdão é a decisão colegiada proferida pelos tribunais, possuindo caráter vinculante em determinados casos. A análise desses documentos exige a verificação da existência de temas jurídicos ou súmulas previamente estabelecidas sobre a matéria em discussão. No entanto, devido ao elevado volume de processos e à complexidade inerente aos textos jurídicos, essa atividade demanda tempo significativo e está sujeita a inconsistências interpretativas.

Novas abordagens têm sido exploradas com o intuito de otimizar tarefas de categorização e recuperação da informação no âmbito jurídico, especialmente por meio do uso de técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina. Dentre essas técnicas, destacam-se os métodos baseados em modelos de linguagem profunda, como o BERT (*Bidirectional Encoder Representations from Transformers*) e sua variação SBERT (*Sentence-BERT*), que têm se mostrado eficazes na análise semântica de textos. Tais modelos possibilitam tanto o agrupamento de documentos por similaridade quanto a modelagem de tópicos, contribuindo significativamente para a organização e interpretação de grandes volumes de dados jurídicos (Chalkidis et al., 2020).

Diante desse cenário, este trabalho propõe o desenvolvimento de uma ferramenta para o agrupamento automático de acórdãos com base em sua similaridade semântica, além da modelagem de tópicos e geração de nuvem de palavras. A abordagem empregada utiliza técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina, visando aprimorar a análise preliminar dos pedidos de recurso. Para a construção do corpus de dados, foram considerados os acórdãos proferidos no Tribunal de Justiça do Tocantins, extraídos da base do e-Proc 2º grau, no período de 1º de abril de 2024 a 13 de novembro de 2024.

1.1 Problema de pesquisa

O gabinete da presidência dos tribunais é responsável por analisar os pedidos de recurso, verificando a existência de precedentes jurisprudenciais que justifiquem ou inviabilizem o seu encaminhamento às instâncias superiores. Atualmente, essa análise é realizada de forma manual, o que torna o processo moroso, sujeito a falhas humanas e ineficiente diante do crescente volume de processos. A ausência de mecanismos automatizados pode levar ao envio indevido de recursos que já possuem tema ou súmula definida, contribuindo para a sobrecarga do sistema judiciário.

Considerando o exposto, esta pesquisa busca responder à seguinte questão: o uso de técnicas de Inteligência Artificial pode contribuir para a eficiência da triagem de recursos judiciais, evitando o envio de casos cuja matéria já esteja pacificada por precedentes ou súmulas?

1.2 Justificativa

A crescente digitalização dos processos judiciais e o avanço das técnicas de Inteligência Artificial (IA) abrem novas possibilidades para a otimização do trabalho jurídico. A aplicação de técnicas de PLN para a classificação e agrupamento de acórdãos pode proporcionar um ganho significativo na eficiência do trabalho do gabinete da presidência, reduzindo o tempo de análise e melhorando a qualidade da triagem dos pedidos de recurso.

O crescimento exponencial do volume de processos judiciais e a complexidade dos textos jurídicos têm exigido soluções tecnológicas que contribuam para a celeridade e eficiência no trâmite processual. No contexto dos tribunais superiores, especialmente nos gabinetes da presidência, é comum a análise de acórdãos com pedidos de recurso, com a finalidade de identificar a existência de temas já pacificados ou súmulas previamente estabelecidas. Este processo, quando realizado de forma manual, é moroso e sujeito a inconsistências. O agrupamento

automatizado de acórdãos por similaridade apresenta-se como uma alternativa viável para aperfeiçoar esse fluxo de trabalho, possibilitando a organização e o direcionamento mais eficiente dos processos. Diferentemente da classificação supervisionada, que exige dados previamente rotulados, o agrupamento permite a identificação de padrões e temas recorrentes por meio de algoritmos não supervisionados, o que se alinha melhor à realidade da maioria dos tribunais, aonde o dado não vêm previamente categorizados.

Além disso, a adoção de modelos baseados em redes neurais profundas, como o BERT e o SBERT, tem possibilitado uma compreensão mais sofisticada do conteúdo semântico dos textos. Esses modelos foram projetados para capturar relações contextuais complexas e gerar representações vetoriais que preservam com precisão a similaridade semântica entre sentenças (DEVLIN et al., 2019; REIMERS; GUREVYCH, 2019), favorecendo a formação de agrupamentos mais coesos e semanticamente relevantes. A incorporação dessas tecnologias ao contexto da Justiça constitui um avanço significativo rumo à transformação digital do sistema judiciário, contribuindo para o aumento da eficiência e para a elevação da qualidade na prestação jurisdicional.

1.3 Objetivos

Diante do problema identificado e da relevância da aplicação de técnicas de Inteligência Artificial no apoio à análise de recursos judiciais, foram definidos os seguintes objetivos, os quais norteiam a condução deste trabalho. O objetivo geral define a finalidade central da pesquisa, enquanto os objetivos específicos detalham as etapas que viabilizam o alcance desse propósito.

1.3.1 Objetivo Geral

Desenvolver um modelo computacional utilizando técnicas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina com o propósito de agrupar acórdãos judiciais por similaridade textual, a fim de otimizar a triagem e análise dos pedidos de recurso no âmbito do gabinete da presidência de tribunais.

1.3.2 Objetivos Específicos

1. Realizar a extração e normalização das informações textuais contidas nos acórdãos, por meio de técnicas de Processamento de Linguagem Natural (PLN), a fim de aprimorar a representação textual e, consequentemente, o desempenho dos algoritmos de Aprendizado de Máquina.
2. Identificar e organizar decisões judiciais com conteúdo semanticamente semelhante por meio da aplicação de algoritmos de agrupamento (*clustering*), baseados em aprendizado não supervisionado.
3. Avaliar a qualidade dos agrupamentos por meio de métricas de validação
4. Aplicar métodos de modelagem de tópicos, com destaque para o modelo *Latent Dirichlet Allocation* (LDA), com o propósito de evidenciar os principais temas recorrentes nos grupos de acórdãos previamente formados.

1.4 Estrutura Da Dissertação

A presente dissertação está organizada em cinco capítulos, estruturados de forma a proporcionar uma exposição lógica e progressiva do desenvolvimento da pesquisa.

O Capítulo 1 - Introdução apresenta a contextualização do problema de pesquisa, os objetivos gerais e específicos, a justificativa e a relevância do estudo para o contexto jurídico, especialmente no âmbito do Tribunal de Justiça do Tocantins.

O Capítulo 2 - Fundamentação Teórica aborda os conceitos essenciais relacionados ao Processamento de Linguagem Natural (PLN), Aprendizado de Máquina, vetorização de textos, técnicas de clusterização, redução de dimensionalidade, modelagem de tópicos com LDA, bem como o uso de nuvens de palavras na análise textual de documentos jurídicos. Este capítulo estabelece o embasamento teórico necessário à compreensão dos métodos aplicados.

O Capítulo 3 - Metodologia descreve detalhadamente o percurso metodológico adotado, incluindo a coleta e pré-processamento dos dados, a aplicação do modelo SBERT para vetorização semântica, aplicação do algoritmo *K-Means* para clusterização, a redução de dimensionalidade com t-SNE, a geração das nuvens de palavras, e a modelagem de tópicos com LDA.

O Capítulo 4 - Resultados e Análise apresenta os resultados obtidos a partir das técnicas aplicadas. São descritos o comportamento do conjunto de dados, as métricas de avaliação da

clusterização, a comparação entre diferentes modelos de *embeddings*, a visualização gráfica dos agrupamentos, e a interpretação qualitativa dos tópicos extraídos.

Por fim, o Capítulo 5 - Conclusão discute as principais contribuições do trabalho, as limitações encontradas, as respostas à questão de pesquisa e as possibilidades de trabalhos futuros que podem aprofundar ou ampliar as abordagens aqui desenvolvidas.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos teóricos que sustentam o desenvolvimento da presente pesquisa, estruturando os conceitos essenciais relacionados às técnicas de Processamento de Linguagem Natural (PLN), aprendizado de máquina não supervisionado e suas aplicações no contexto jurídico. O objetivo é fornecer um embasamento sólido que justifique as escolhas metodológicas adotadas, bem como contextualizar os principais modelos, algoritmos e métricas utilizados ao longo do trabalho.

Inicialmente, são abordados os princípios do PLN, com destaque para as etapas de pré-processamento textual e as estratégias de representação vetorial de textos, fundamentais para a transformação de linguagem natural em dados computacionalmente manipuláveis. Em seguida, explora-se o funcionamento de modelos avançados baseados em redes neurais profundas, com ênfase no BERT e suas variantes, como o SBERT, cuja arquitetura siamesa tem se destacado em tarefas de similaridade semântica.

Além disso, são discutidos os principais métodos de redução de dimensionalidade, como o t-SNE, e técnicas de agrupamento de dados, com ênfase no algoritmo *K-Means*, amplamente utilizado para clusterização de textos. Por fim, são apresentados os critérios de avaliação de qualidade de agrupamentos *Silhouette Score,* Índice de *Davies-Bouldin* e Índice de *Calinski-Harabasz*, bem como a modelagem de tópicos com LDA e o uso de nuvens de palavras como recursos complementares de análise qualitativa.

A fundamentação teórica aqui desenvolvida visa sustentar, do ponto de vista técnico e científico, a proposta metodológica adotada, demonstrando sua pertinência diante do problema de pesquisa e dos objetivos estabelecidos neste estudo.

2.1 Acórdão

No contexto do processo civil brasileiro, o acórdão é uma das formas de pronunciamento judicial, ao lado da sentença e da decisão interlocutória. Apesar de ser proferido por órgão colegiado (normalmente por um conjunto de desembargadores ou ministros), a legislação o inclui, sistematicamente, entre os pronunciamentos atribuídos genericamente ao juiz (BRASIL, 2015, art. 203, § 1º).

O acórdão constitui a decisão proferida no âmbito dos tribunais, sendo o resultado do julgamento de recursos ou de processos de competência originária desses órgãos (BRASIL, 2015, art. 204 e art. 932). A sua elaboração compete ao relator do processo, a quem incumbe

dirigir e ordenar o andamento processual no tribunal, podendo inclusive homologar a autocomposição das partes, nos casos em que isso for cabível (BRASIL, 2015, art. 932, I).

Do ponto de vista de sua estrutura, o acórdão deve conter uma ementa, a qual será publicada no órgão oficial no prazo de dez dias após a lavratura, conforme previsto pelo Código de Processo Civil (BRASIL, 2015, art. 941, § 1º). Além disso, o conteúdo do acórdão deve abranger de forma expressa a análise dos fundamentos relevantes da tese jurídica discutida, garantindo a publicidade e a coerência da decisão judicial (BRASIL, 2015, art. 489, § 1º, e art. 926).

Em determinadas hipóteses, como nas decisões proferidas sob o regime de assunção de competência ou incidente de resolução de demandas repetitivas (IRDR), o acórdão poderá vincular todos os juízes e órgãos fracionários do tribunal, assegurando a uniformidade da jurisprudência, salvo nos casos em que houver posterior revisão de tese (BRASIL, 2015, arts. 947 e 985).

2.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) constitui um dos campos mais dinâmicos e interdisciplinares da Inteligência Artificial (IA), envolvendo conhecimentos de ciência da computação, linguística e estatística para possibilitar a comunicação entre humanos e máquinas de forma mais natural e eficiente.

Segundo Alcarde (2021), o PLN visa desenvolver modelos e algoritmos capazes de interpretar, compreender e gerar linguagem humana, aproximando as capacidades de processamento computacional das habilidades comunicativas humanas. Essa área tem sido responsável por avanços notáveis, como a tradução automática, análise de sentimentos, assistentes virtuais e sistemas de atendimento automatizado.

De acordo com os autores de Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português (CASELI; NUNES, 2024), o termo “natural” refere-se às línguas humanas, e o PLN pode ser dividido em duas grandes subáreas: *Natural Language Understanding* (NLU), voltada à interpretação da linguagem, e *Natural Language Generation* (NLG), voltada à geração textual.

Ainda que tenha possibilitado soluções inovadoras em diferentes setores, o PLN enfrenta desafios como a ambiguidade semântica e a interpretação contextual da linguagem. A constante evolução tecnológica e o aprimoramento dos modelos de linguagem têm contribuído para a superação de tais desafios, consolidando o PLN como ferramenta essencial na mediação

homem-máquina. Dando continuidade à contextualização teórica, é possível identificar um conjunto estruturado de etapas e técnicas fundamentais que viabilizam a conversão de dados textuais brutos em representações compreensíveis pelas máquinas. Conforme Alcarde (2021), essas etapas são interdependentes e visam aprimorar a qualidade e a profundidade da análise automatizada da linguagem.

2.2.1 Etapas Fundamentais do PLN

O processo de PLN pode ser dividido em etapas sequenciais, conforme descrito a seguir:

Coleta de Dados: Consiste na obtenção de dados textuais ou de fala que servirão como base para o treinamento e avaliação dos modelos de PLN. Esses dados podem ser provenientes de documentos jurídicos, redes sociais, transcrições ou bancos de dados institucionais.

Pré-processamento Textual: Etapa essencial para padronização e limpeza dos textos, envolvendo:

Conversão para letras minúsculas;

Remoção de pontuações, números e símbolos irrelevantes;

Eliminação de palavras irrelevantes (*stopwords*);

Lematização ou *stemming* para redução das palavras às suas formas base.

Segundo (CAMACHO, 2020), um bom pré-processamento é determinante para o desempenho dos algoritmos subsequentes.

Tokenização: Processo que divide o texto em unidades menores (*tokens*), como palavras, frases ou símbolos, que são analisadas individualmente pelos modelos de linguagem.

Análise Morfológica e Sintática: Visa identificar as classes gramaticais das palavras e suas funções dentro da estrutura frasal, possibilitando uma representação mais estruturada da linguagem.

Análise Semântica e Pragmática: Busca interpretar o significado das palavras no contexto e identificar a intenção comunicativa, elemento essencial em tarefas como resposta automática e compreensão de comandos.

Geração de Resposta: Fase em que a máquina, com base no conteúdo processado, gera uma resposta textual, realiza uma ação ou toma uma decisão, como ocorre em sistemas de recomendação ou assistentes virtuais.

2.3 Aprendizagem de Máquina (*Machine Learning*)

Com o crescente volume de dados, técnicas de Aprendizado de têm sido amplamente utilizadas para realizar tarefas simples e complexas, mas trabalhosas devido ao grande volume de dados.

Aprendizado de Máquina refere-se a um processo em que as máquinas aprendem de forma autônoma, sem programação explícita. Aprendem com a entrada de informações que oferecemos, com base nessa entrada, identificarão tendências e padrões e tomarão decisões com base em suas observações exclusivas sobre eventos dentro do conjunto informado, e podem aprender com seus erros e fazer previsões (HERRON,2020). Podemos definir um modelo com alguns parâmetros, treinar esse modelo com os dados, para que ele possa ser otimizado. Esse modelo pode ser preditivo para fazer previsões, ou descritivo para obter conhecimento dos dados, ou ambos. (ALPAYDIN, 2010).

Os algoritmos de machine learning disponíveis, normalmente podem ser divididos em quatro categorias principais:

a) aprendizagem supervisionada: dado um conjunto de categorias pré-definidas, compreender a relação entre as entradas e saídas fornecidas, para classificar, ou rotular, uma determinada instância; (THEOBALD,2020)
b) aprendizagem não supervisionada: análise das relações entre as variáveis de entrada e na descoberta de padrões ocultos que podem ser extraídos para criar novos rótulos sobre possíveis saídas; (THEOBALD,2020)
c) aprendizagem por reforço: constrói seu modelo de predição, obtendo feedback de tentativa e erro aleatório e aproveitando o insight de iterações anteriores. O objetivo do aprendizado por reforço é atingir um objetivo específico (resultado) testando aleatoriamente um vasto número de combinações possíveis de entrada e avaliando seu desempenho; (THEOBALD,2020)
d) aprendizagem semi-supervisionada: uma mistura de supervisionada com não supervisionada; (THEOBALD,2020)

2.3.1 Aprendizagem Supervisionada

O aprendizado supervisionado é uma das principais categorias de aprendizado de máquina e baseia-se na análise de dados rotulados para identificar padrões entre variáveis de

entrada e saída. Esse processo permite que um modelo aprenda a prever resultados a partir de novos dados, com base em exemplos anteriores.

Segundo THEOBALD (2020), esse tipo de aprendizado imita a capacidade humana de aprender por meio de observação e repetição de padrões conhecidos. A construção de um modelo supervisionado exige um conjunto de dados em que cada instância contenha variáveis independentes (entrada – X) e uma variável dependente (saída – y), permitindo que o algoritmo relacione os dois conjuntos de maneira estatística.

Exemplos comuns de variáveis de saída incluem a previsão do valor de mercado de um imóvel, a identificação de um objeto em uma imagem ou a classificação de um produto. Entre os algoritmos mais utilizados no aprendizado supervisionado estão regressão linear, regressão logística, árvores de decisão, k-vizinhos mais próximos, redes neurais e máquinas de vetor de suporte (SVM).

Para que o modelo seja construído de forma eficaz, é necessário um conjunto de dados completamente rotulado. Dados não rotulados não podem ser utilizados diretamente no treinamento supervisionado, pois o modelo preciso da correspondência explícita entre entrada e saída para aprender a realizar predições.

2.3.2 Aprendizagem Não Supervisionada

O aprendizado não supervisionado é uma vertente do aprendizado de máquina que opera sem a necessidade de dados rotulados, buscando identificar padrões, estruturas e agrupamentos ocultos em grandes volumes de dados (MITCHELL, 1997). Diferentemente do aprendizado supervisionado, não há uma variável de saída definida, e o objetivo é descobrir relações entre os dados de entrada com base em medidas de similaridade ou distância.

Uma das principais aplicações do aprendizado não supervisionado é o agrupamento (clustering), técnica que visa particionar um conjunto de dados em grupos que compartilham características semelhantes entre si e são distintos entre grupos diferentes. Esse método é eficaz em contextos sem categorias pré-definidas, como na análise de textos jurídicos ou dados não estruturados.

Entre os algoritmos mais utilizados para clusterização está o *K-Means*, que distribui os dados em um número pré-determinado de clusters com base na minimização da distância entre os pontos e os centroides de cada grupo (MACQUEEN, 1967). O algoritmo opera de forma iterativa, ajustando os agrupamentos até atingir um critério de convergência.

A eficácia do *K-Means* depende de fatores como a escolha do número de clusters e a forma de inicialização dos centróides. Apesar de sua simplicidade e eficiência computacional, ele pode apresentar limitações em conjuntos de dados com clusters de formas irregulares ou com ruídos. Nesses casos, outras técnicas como HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) ou DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)* podem ser mais apropriadas.

No contexto jurídico, a aplicação de algoritmos de agrupamento em acórdãos judiciais permite organizar decisões com base em similaridades semânticas, contribuindo para a identificação de temas recorrentes e a uniformização da jurisprudência. Essa abordagem torna-se ainda mais eficaz quando combinada com representações vetoriais geradas por modelos de linguagem como BERT ou SBERT, capazes de capturar nuances semânticas complexas dos textos jurídicos. Diante desse cenário, neste trabalho propõe-se a utilização de técnicas de aprendizado não supervisionado como estratégia para a análise e agrupamento automatizado dos acórdãos.

2.4 Representação Vetorial de Texto

A representação vetorial de textos é uma etapa essencial nas tarefas de Processamento de Linguagem Natural (PLN), pois permite que textos em linguagem natural sejam convertidos em formatos numéricos que podem ser manipulados por algoritmos de aprendizado de máquina. Desde os primeiros modelos até os mais recentes baseados em redes neurais profundas, diversas abordagens foram desenvolvidas para capturar o significado e a estrutura dos textos. A seguir, apresentam-se algumas das principais abordagens utilizadas:

- *Bag-of-Words* (BoW): representa um texto por meio da frequência de ocorrência de palavras em um vocabulário fixo, ignorando a ordem ou o contexto em que as palavras aparecem. Apesar de sua simplicidade, é útil em tarefas como classificação e análise de sentimentos, embora sofra com alta dimensionalidade e perda de informações semânticas.
- TF-IDF (*Term Frequency-Inverse Document Frequency*): aperfeiçoa o modelo BoW ao ponderar a frequência de um termo considerando sua importância relativa em um conjunto de documentos. Palavras muito comuns recebem menor peso, enquanto termos mais representativos ganham destaque (MANNING; RAGHAVAN; SCHÜTZE, 2008). No entanto, o TF-IDF ainda não considera a semântica ou o contexto das palavras.

- Word2Vec: proposto por Mikolov et al. (2013), este modelo representa palavras como vetores densos e contínuos, treinados para preservar relações semânticas com base nos contextos em que aparecem. Essa técnica trouxe avanços importantes, permitindo, por exemplo, cálculos de similaridade entre palavras, mas ainda apresenta limitações quanto à compreensão de significados contextuais distintos (MIKOLOV et al., 2013).
- BERT (*Bidirectional Encoder Representations from Transformers*): desenvolvido por Devlin et al. (2018), o BERT introduz uma representação contextualizada de palavras, considerando os dois lados do contexto por meio de atenção bidirecional. Revolucionou o PLN ao possibilitar interpretações mais precisas do significado das palavras conforme o contexto da frase. No entanto, não foi originalmente projetado para calcular diretamente a similaridade entre sentenças, pois não gera embeddings fixos para elas (DEVLIN et al., 2018).
- SBERT (*Sentence-BERT*): proposto por REIMERS; GUREVYCH (2019), o SBERT adapta o BERT para tarefas de similaridade textual ao utilizar uma arquitetura *siamese* ou *triplet* que permite gerar *embeddings* vetoriais fixos para sentenças. Essa abordagem é especialmente eficaz em tarefas como: similaridade textual, agrupamento temático de documentos, classificação por tema.

O modelo SBERT apresenta avanços significativos em termos de desempenho e eficiência na comparação semântica entre sentenças, superando as limitações de abordagens anteriores que operavam apenas no nível de palavra (REIMERS; GUREVYCH, 2019). Sua eficiência decorre da capacidade de gerar representações vetoriais fixas e contextualizadas para sentenças inteiras, o que permite o uso direto de métricas como a similaridade do cosseno para mensurar o grau de similaridade semântica entre textos. Diferentemente do BERT original, que exige o pareamento de sentenças a cada comparação (o que aumenta o tempo computacional), o SBERT permite pré-computar os *embeddings* de forma independente, o que reduz significativamente o custo computacional em tarefas que envolvem grandes volumes de dados ou múltiplas comparações simultâneas. Em razão dessas características, este será o modelo adotado no presente trabalho para a geração de representações vetoriais de sentenças.

2.5 *K-Means*: Algoritmo de Agrupamento Não Supervisionado

O *K-Means* é um dos algoritmos mais utilizados para a realização de clusterização, uma técnica de aprendizado não supervisionado cujo objetivo é particionar um conjunto de dados

em k grupos ou clusters com base em características de similaridade (ALPAYDIN, 2010). O método baseia-se em um processo iterativo de partição dos dados em k grupos. Inicialmente, são escolhidos *k* centros (ou centroides) de forma aleatória ou por meio de técnicas auxiliares, como o uso de médias dos dados ou resultados de uma análise de componentes principais. Em seguida, cada instância do conjunto de dados é atribuída ao centro mais próximo, com base em uma medida de distância — geralmente a distância Euclidiana. Após essa etapa, os centroides de cada grupo são recalculados como a média dos pontos atribuídos a eles. Esse processo de atribuição e atualização é repetido até que os grupos deixem de mudar significativamente, indicando a convergência do algoritmo. Ressalta-se que, por se tratar de uma técnica de busca local, o resultado final pode ser sensível à escolha inicial dos centroides (ALPAYDIN, 2010).

Uma limitação significativa do algoritmo *K-Means* refere-se à necessidade de definição prévia do número de clusters (*k*), o que pode impactar substancialmente a qualidade da clusterização. A determinação do valor ideal de *k* é um desafio comum em tarefas de agrupamento e, por isso, diversas métricas têm sido propostas, entre as quais se destacam o Método do Cotovelo (*Elbow Method*), o Coeficiente de Silhueta (*Silhouette Score*) e o Índice de *Davies-Bouldin*. No domínio da análise textual, o *K-Means* tem sido amplamente utilizado em tarefas de agrupamento de documentos, especialmente quando aplicado sobre representações vetoriais derivadas de modelos de linguagem, tais como TF-IDF, Word2Vec ou *Sentence*-BERT. Essa abordagem permite a identificação de agrupamentos semânticos latentes, contribuindo significativamente para a classificação temática e estruturação de extensos conjuntos textuais (corpora), viabilizando a descoberta automatizada de tópicos e análises exploratórias em grandes volumes de dados jurídicos.

2.6 Redução de Dimensionalidade: PCA, t-SNE

A redução de dimensionalidade constitui uma etapa essencial no Processamento de Linguagem Natural (PLN) e no Aprendizado de Máquina, sendo especialmente relevante em contextos nos quais os dados possuem elevada complexidade e redundância. Essa necessidade decorre da chamada "maldição da dimensionalidade", expressão que denota os problemas causados pela presença de um número excessivo de variáveis (ou atributos), dificultando a generalização dos modelos, elevando o custo computacional e comprometendo a acurácia dos resultados (BENGIO; COURVILLE; VINCENT, 2013; VERZELLA et al., 2021).

No caso de dados textuais, a dimensionalidade elevada é particularmente comum, visto que cada termo de um vocabulário pode ser considerado como uma dimensão distinta. Essa

estrutura esparsa e de alta dimensionalidade não apenas prejudica a performance de algoritmos, como também inviabiliza a visualização e a interpretação direta dos dados (JINDAL; LIU, 2006).

A redução de dimensionalidade busca projetar os dados originais em um espaço de menor dimensão, com o objetivo de preservar ao máximo suas propriedades estruturais e semânticas. Tal processo pode ser realizado por métodos de seleção de atributos, que escolhem subconjuntos representativos das variáveis originais, ou por métodos de projeção, que transformam os dados em novas variáveis latentes (JOLLIFFE; CADIMA, 2016).

Duas técnicas amplamente adotadas nesse contexto são a Análise de Componentes Principais (PCA) e o *t-Distributed Stochastic Neighbor Embedding* (t-SNE), as quais serão descritas a seguir:

- Análise de Componentes Principais (PCA): A Análise de Componentes Principais é um método estatístico de redução de dimensionalidade linear e não supervisionado. Seu objetivo é encontrar uma nova base ortogonal que maximize a variância dos dados projetados, preservando, assim, o máximo de informação possível em menos dimensões (JOLLIFFE; CADIMA, 2016). O PCA opera por meio da decomposição da matriz de covariância dos dados, selecionando os autovetores associados aos maiores autovalores. A técnica minimiza o erro de reconstrução entre os dados originais e os dados projetados, sendo, portanto, indicada para situações em que os dados estão distribuídos em um subespaço linear (HAN; KAMBER; PEI, 2011). Além disso, o PCA pode ser utilizado para facilitar a visualização de dados em gráficos bidimensionais ou tridimensionais, e também como etapa preparatória para outros métodos que exigem dados não singulares, como é o caso da Análise Discriminante Linear (MARTÍNEZ; KAK, 2001). Existe ainda a versão Kernel PCA, que estende o método ao domínio não linear, permitindo trabalhar com estruturas mais complexas (SCHÖLKOPF; SMOLA; MÜLLER, 1998).
- *t-Distributed Stochastic Neighbor Embedding* (t-SNE): é uma técnica não linear de redução de dimensionalidade, particularmente eficaz para a visualização de dados de alta dimensão em espaços bidimensionais ou tridimensionais.

Ao contrário do PCA, o t-SNE busca preservar a estrutura local dos dados, mantendo relações de proximidade entre instâncias similares (VAN DER MAATEN; HINTON, 2008). O algoritmo constrói distribuições de probabilidade baseadas nas distâncias entre pontos no espaço original e no espaço projetado, minimizando a divergência de *Kullback-Leibler* entre essas

distribuições. Essa abordagem resulta em mapas de dispersão nos quais padrões e agrupamentos tornam-se mais evidentes, o que tem levado à sua ampla aplicação em tarefas exploratórias e análises visuais de agrupamentos textuais (WATTENBERG; VIEGAS; JOHNSON, 2016). Estudos empíricos apontam que o t-SNE frequentemente produz visualizações mais interpretáveis do que técnicas lineares, especialmente em conjunto com algoritmos de clusterização, como o *K-Means*, os quais podem ser utilizados para atribuir cores aos pontos com base nos grupos identificados (LAURIOLA; LAVELLI; AIOLLI, 2022).

Em síntese, enquanto o PCA se destaca pela simplicidade e eficiência em cenários lineares, o t-SNE oferece vantagens em tarefas exploratórias e de visualização de estruturas complexas, com forte preservação das vizinhanças locais. Ambas as técnicas são úteis para mitigar os desafios impostos pela alta dimensionalidade de dados textuais, permitindo a aplicação eficiente de algoritmos de clusterização e classificação em espaços reduzidos, sem perda significativa de informação relevante.

Em tarefas de Processamento de Linguagem Natural (PLN), técnicas de redução de dimensionalidade são frequentemente aplicadas para visualizar e interpretar representações vetoriais de palavras ou documentos, como embeddings gerados por modelos como Word2Vec, GloVe (*Global Vectors for Word Representation*) ou BERT. O PCA pode ser utilizado para identificar direções principais de variação semântica, enquanto o t-SNE é eficaz para visualizar agrupamentos semânticos e relações complexas entre palavras ou documentos. Neste trabalho, utilizaremos o t-SNE para projetar os embeddings em duas dimensões e apoiar a visualização e a interpretação dos agrupamentos (com cores por cluster obtidos via K-Means).

2.7 Extração de Tópicos com LDA

A técnica de *Latent Dirichlet Allocation* (LDA) é um dos métodos mais amplamente utilizados para modelagem de tópicos em grandes corpora textuais, sendo particularmente eficaz para identificar padrões latentes e descobrir temas ocultos em documentos não rotulados. Essa abordagem é útil em contextos em que há grande volume de dados textuais, como em acórdãos judiciais, jornais, artigos acadêmicos ou redes sociais, facilitando a organização e a compreensão semântica do conteúdo textual (BLEI; NG; JORDAN, 2003).

O LDA é um modelo generativo probabilístico que assume que cada documento é uma mistura de diversos tópicos, e que cada tópico é representado por uma distribuição de probabilidade sobre um conjunto de palavras. Assim, a geração de documentos ocorre por meio de um processo em que tópicos são escolhidos com base em distribuições de *Dirichlet* e,

posteriormente, palavras são selecionadas de acordo com essas distribuições (STEYVERS; GRIFFITHS, 2007).

De maneira mais formal, o LDA considera que: cada documento possui uma distribuição sobre tópicos, extraída de uma distribuição *Dirichlet*; cada tópico possui uma distribuição sobre palavras, também extraída de uma *Dirichlet*; e, para cada palavra em um documento, seleciona-se um tópico com base na distribuição de tópicos do documento, e uma palavra com base na distribuição do tópico selecionado.

O objetivo do modelo é inferir as distribuições e associações latentes entre palavras e documentos. Para isso, utilizam-se algoritmos de inferência como *Gibbs Sampling* ou *Variational Bayes*, que estimam as variáveis ocultas a partir de observações (BLEI et al., 2003).

A principal utilidade do LDA está na redução da complexidade semântica dos dados, permitindo identificar os principais tópicos tratados em um conjunto de documentos; classificar documentos com base nos temas predominantes; realizar visualizações temáticas para fins exploratórios; e apoiar tarefas de recuperação de informação, resumo automático, classificação textual, entre outras (ALGHAMDI; ALFALQI, 2015).

No contexto jurídico, como o analisado neste trabalho, o LDA tem se mostrado promissor para mapear temas jurídicos recorrentes, auxiliar na classificação temática de acórdãos, e fomentar estudos sobre a jurisprudência ao longo do tempo (CARDELLINO et al., 2017).

Contudo, apesar de sua popularidade, o LDA apresenta limitações. Ele assume que as palavras em um documento são independentes e desconsidera a ordem das palavras, o que pode afetar a acurácia temática em textos mais complexos. Além disso, a escolha do número de tópicos é uma etapa sensível e depende de análise exploratória e validação com métricas como coerência de tópicos ou perplexidade (RÖDER; BOTH; HINNEBURG, 2015).

Para complementar a análise dos resultados obtidos com LDA, foi utilizada a ferramenta *PyLDAvis*. Trata-se de uma biblioteca interativa de visualização que permite explorar graficamente a distribuição e a relação entre os tópicos e as palavras mais relevantes de cada um. O *PyLDAvis* representa os tópicos em um plano bidimensional, onde sua posição e sobreposição indicam similaridades temáticas, e sua área representa a prevalência no corpus. Essa visualização contribui para a avaliação qualitativa da modelagem de tópicos, ampliando a compreensão das estruturas temáticas inferidas pelo modelo (SIEVERT; SHIRLEY, 2014).

2.8 Visualização de Conteúdo Textual

A nuvem de palavras (*word cloud*) é uma técnica exploratória de visualização textual utilizada para representar graficamente a frequência de palavras em um conjunto de documentos. Trata-se de um recurso bastante difundido na análise inicial de dados textuais, especialmente no campo do Processamento de Linguagem Natural (PLN), por permitir uma compreensão rápida dos termos mais recorrentes em um corpus (VIÉGAS; WATTENBERG; FEINBERG, 2009; HEIMERL et al., 2014).

Visualmente, as palavras mais frequentes aparecem em tamanhos maiores e mais destacados, enquanto palavras com menor ocorrência são representadas em proporções reduzidas. Essa estratégia facilita a identificação de padrões semânticos, termos-chave e possíveis tópicos predominantes nos textos, mesmo antes da aplicação de técnicas mais sofisticadas como modelagem de tópicos ou clusterização (HEIMERL et al., 2014).

Na análise de agrupamentos textuais, como no caso de acórdãos judiciais clusterizados, a nuvem de palavras pode ser aplicada por cluster ou tema, permitindo visualizar os termos mais representativos de cada grupo. Essa abordagem contribui para uma interpretação qualitativa dos clusters, ajudando o pesquisador a atribuir significados aos agrupamentos gerados automaticamente por algoritmos como o *K-Means* ou HDBSCAN.

Entre as principais aplicações das nuvens de palavras, destacam-se: exploração inicial de dados textuais; comparação de vocabulários entre diferentes grupos temáticos; identificação de termos discriminativos em análises comparativas; apoio à rotulagem de clusters ou categorias emergentes; comunicação de resultados de maneira acessível e visualmente atrativa.

Entretanto, a técnica possui limitações, pois não leva em conta o contexto das palavras, nem as relações semânticas entre os termos. Palavras de alta frequência, mas pouco informativas podem ocupar espaço visual desproporcional caso não sejam removidas na etapa de pré-processamento (ex.: *stopwords*, nomes próprios repetitivos etc.). Por isso, é recomendável que a construção da nuvem de palavras seja precedida por uma limpeza textual criteriosa, que inclua lematização, remoção de palavras irrelevantes e normalização de termos.

Ferramentas como *WordCloud (Python), Voyant Tools*, ou bibliotecas em R como *tm* e *wordcloud*, oferecem meios automatizados de gerar essas visualizações de forma customizada, sendo amplamente utilizadas tanto em pesquisas acadêmicas quanto em aplicações corporativas de análise textual.

2.9 Índices de Validação de Agrupamentos

Nos métodos de agrupamento não supervisionado, como a clusterização aplicada nesta pesquisa, torna-se essencial avaliar a qualidade dos agrupamentos obtidos. Como os algoritmos de clusterização não recebem previamente os rótulos dos dados, utiliza-se um conjunto de métricas de validação interna para mensurar a coesão e a separação dos clusters formados (HASSAN et al., 2024; GÖSGENS et al., 2019).

Dentre as métricas utilizadas neste trabalho, destacam-se:

- *SilhouetteScore*: Apresentado por Rousseeuw (1987), o *Silhouette Score* quantifica simultaneamente o quão próximo um elemento está de outros elementos do seu próprio cluster (coesão) e o quão distante está dos elementos de outros clusters (separação). O valor da métrica varia de -1 a 1, sendo que valores próximos a 1 indicam forte separação entre os agrupamentos e alta coesão interna.
- *Davies-Bouldin Index*: Proposto por *Davies* e *Bouldin* (1979), esse índice mede a média da razão entre a dispersão intra-cluster e a separação inter-cluster. Valores menores indicam melhor qualidade de agrupamento, pois representam maior compactação dentro dos clusters e maior distância entre eles.
- *Calinski-Harabasz Index*: Introduzido por *Caliński* e *Harabasz* (CALINSKI,1974), também conhecido como *Variance Ratio Criterion*, calcula a razão entre a dispersão inter-cluster e intra-cluster. Valores mais elevados indicam maior separação entre grupos e melhor definição dos agrupamentos (ALI et al., 2021).

A utilização combinada dessas métricas possibilita uma avaliação mais robusta dos agrupamentos formados, considerando diferentes aspectos geométricos e estatísticos dos dados (HASSAN et al., 2024; GÖSGENS et al., 2019).

3 METODOLOGIA

O presente estudo adota uma abordagem quantitativa, caracterizada pela análise de dados numéricos por meio de procedimentos estatísticos, com ênfase na objetividade durante as etapas de coleta e análise (UFRGS, 2009). Fundamenta-se em técnicas de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina, com o objetivo de agrupar e analisar semanticamente acórdãos judiciais. Para assegurar a qualidade e a robustez dos resultados, a metodologia foi estruturada em etapas sequenciais que incluem: coleta dos dados, limpeza e pré-processamento textual, vetorização semântica com o modelo SBERT, cálculo de similaridade entre textos, aplicação de algoritmos de agrupamento não supervisionados (*K-Means*), redução de dimensionalidade com t-SNE, visualização dos agrupamentos, geração de nuvens de palavras e modelagem de tópicos com LDA. Cada uma dessas fases será detalhada ao longo deste trabalho.

Trata-se de uma pesquisa de natureza aplicada, pois visa gerar conhecimento voltado à solução de um problema específico, relacionado à triagem e organização de decisões judiciais no âmbito do Tribunal de Justiça do Tocantins. Como destaca a UFRGS (2009), esse tipo de pesquisa envolve interesses e realidades locais, propondo soluções práticas e contextualizadas.

Quanto aos seus objetivos, esta pesquisa possui caráter exploratório e descritivo, conforme a classificação de Silva e Menezes (2005). A abordagem exploratória se justifica pela necessidade de aprofundar o entendimento sobre a aplicação de técnicas computacionais no tratamento de textos jurídicos, um campo ainda em desenvolvimento no cenário brasileiro. Por sua vez, o caráter descritivo busca observar, registrar, analisar e correlacionar os padrões temáticos presentes nos acórdãos, por meio de métodos como vetorização com SBERT, clusterização com *K-Means* e modelagem de tópicos com LDA. Com isso, pretende-se contribuir para a organização e compreensão do conteúdo jurídico, promovendo maior eficiência na análise documental (UFRGS, 2009).

No que se refere aos procedimentos técnicos adotados, este trabalho configura-se como um estudo de caso, por estar centrado na realidade do Tribunal de Justiça do Tocantins, onde se propõe aplicar e validar a metodologia desenvolvida. Enquadra-se também como pesquisa bibliográfica, por apoiar-se em obras já publicadas nas áreas de PLN, aprendizado de máquina e direito, utilizadas para embasar teoricamente as decisões metodológicas adotadas. Adicionalmente, trata-se de uma pesquisa documental, uma vez que utiliza como base analítica acórdãos judiciais emitidos e armazenados em bancos institucionais, considerados fontes primárias ainda não exploradas de forma analítica aprofundada.

As fases do desenvolvimento podem ser visualizadas na figura 1.

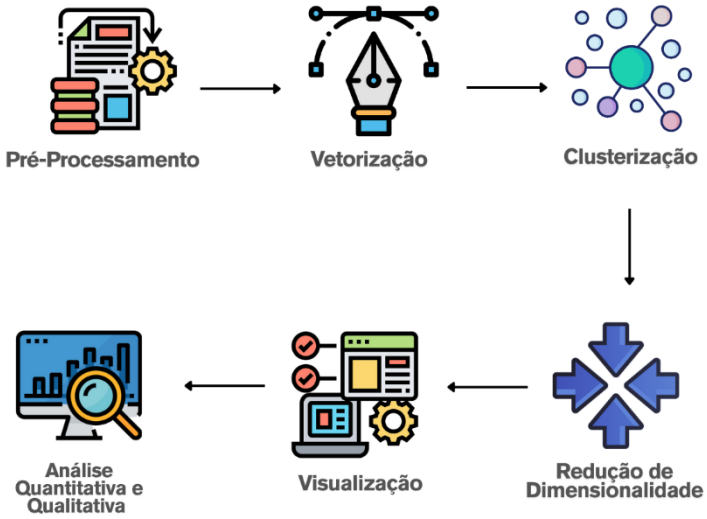

Figura 1 - Fluxo de desenvolvimento

Fonte: Elaborado pela autora (2025).

3.1 Coleta dos dados

O corpus desta pesquisa é composto por 863 acórdãos extraídos do sistema eletrônico de processos do Tribunal de Justiça do Tocantins (e-Proc/TJTO). Foram considerados processos de 2ª instância, na competência Cível, que registraram o evento "Tipo de Petição Judicial Recurso Especial (SREC)" entre 1º de abril e 13 de novembro de 2024. Para fins deste estudo, apenas a ementa do acórdão foi considerada, replicando o procedimento atualmente empregado na triagem manual.

Os documentos no e-Proc são em formato PDF de acordo com a legislação interna (TJTO, 2011). Para a extração e manipulação de arquivos PDF, utilizou-se a biblioteca *PyMuPDF* (*FITZ*), que permite o acesso programático ao conteúdo de documentos em formatos como PDF, EPUB e XPS, incluindo texto, imagens e metadados (PYMUPDF, 2024).

A figura 2 mostra o fluxo realizado na coleta dos dados.

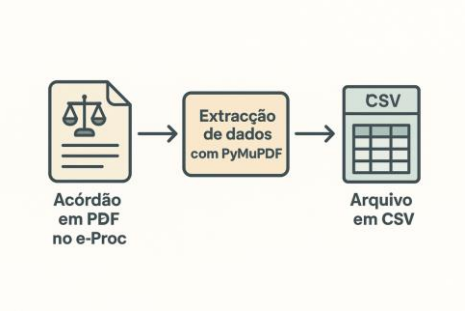

Figura 2 - Fluxo da coleta dos dados

Fonte: Elaborado pela autora (2025).

3.2 Pré-processamento textual

A análise de textos requer uma etapa de pré-processamento dos dados, a fim de garantir a qualidade dos resultados posteriores de vetorização e clusterização. Segundo ALCARDE (2021) nessa fase do pré-processamento os dados coletados passam por uma série de processos, como remoção de ruídos, pontuação, conversão para minúsculas, entre outros, para tornar os dados mais uniformes e facilitar as análises subsequentes.

Neste trabalho, foram aplicadas técnicas como normalização dos textos para minúsculas, remoção de acentos, pontuação, caracteres especiais, números, eliminação de palavras irrelevantes. utilizando o modelo *`pt_core_news_sm` do spaCy y* (https://spacy.io/) e a biblioteca NLTK (*Natural Language Toolkit*) (https://www.nltk.org/).

A etapa de limpeza e pré-processamento textual representa uma fase fundamental no pipeline de PLN, especialmente quando se trabalha com dados textuais, como acórdãos judiciais. O principal objetivo dessa etapa é remover ruídos, padronizar o conteúdo e garantir que apenas informações semanticamente relevantes sejam preservadas para análises posteriores de vetorização, similaridade e agrupamento (JURAFSKY; MARTIN, 2020).

Para essa finalidade, foi desenvolvida uma função em Python, que transforma os textos brutos em dados estruturados e consistentes. O processo inicia-se com a normalização básica do texto, que inclui: conversão de todos os caracteres para minúsculas; remoção de acentuação e caracteres especiais com base na normalização Unicode; exclusão de números, pontuação e espaços em branco excessivos. Essas operações visam reduzir a variabilidade léxica e melhorar a uniformidade do corpus textual, facilitando a aplicação dos algoritmos de aprendizado de máquina.

A etapa seguinte consiste na tokenização e lematização, utilizando o modelo pré-treinado pt_*core*_news_sm da biblioteca *spaCy*, especialmente desenvolvido para a língua

portuguesa. A tokenização é responsável por segmentar o texto em unidades léxicas, enquanto a lematização reduz as palavras à sua forma canônica (por exemplo, “correndo” para “correr”), permitindo uma representação mais concisa e generalizável do vocabulário (JURAFSKY; MARTIN, 2020).

Em seguida, é realizada uma filtragem criteriosa dos tokens, com a remoção de termos irrelevantes para a análise semântica. Para isso, empregou-se uma combinação entre a lista de stopwords da biblioteca NLTK e uma lista personalizada de palavras comuns em textos jurídicos, porém com baixo valor informativo, como “relator”, “tribunal”, “voto”, “art” e “nº”. Também foram descartados: tokens compostos por espaços, pontuações ou números; palavras com menos de três ou mais de quinze caracteres; termos que contenham dígitos.

A última fase do pré-processamento consiste na recombinação dos tokens filtrados em um novo texto limpo, pronto para ser submetido à vetorização. Além disso, nos casos em que o texto de entrada esteja vazio ou não produza conteúdo relevante após o processamento, a função retorna a string “sem conteúdo” como marcador de ausência de informação útil.

Essa metodologia está em consonância com as boas práticas recomendadas por (JURAFSKY; MARTIN, 2020), que enfatizam a importância da normalização e da lematização como etapas fundamentais para melhorar o desempenho dos algoritmos de PLN. A limpeza adequada do corpus é decisiva para a eficiência das fases seguintes, como a vetorização semântica com SBERT, a clusterização com algoritmos como *K-Means*, e a análise temática com modelos como o LDA.

3.3 Vetorização semântica com o modelo SBERT

A representação vetorial dos textos foi realizada por meio do modelo *Sentence*-BERT (SBERT), uma adaptação do BERT para gerar *embeddings* de sentenças com alta qualidade semântica (REIMERS; GUREVYCH, 2019). O SBERT transforma cada acórdão em um vetor denso de alta dimensão, permitindo a comparação direta entre documentos.

A vetorização de textos é uma etapa no fluxo de trabalho de Processamento de Linguagem Natural (PLN), visando converter dados textuais em representações numéricas densas e de alta dimensionalidade, conhecidas como *embeddings*. Essas representações são essenciais para que modelos de Aprendizado de Máquina (AM) possam processar e analisar o significado semântico dos textos, permitindo tarefas como classificação, clusterização, análise de similaridade e recuperação de informações (GOLDBERG, 2017).

Diferentemente dos vetores tradicionais baseados em frequência (como TF-IDF ou *Bag-of-Words*), os *embeddings* gerados por modelos como o SBERT são contextualizados, o que significa que a representação vetorial de uma palavra ou sentença depende de seu contexto sintático e semântico no enunciado (DEVLIN et al., 2018). Isso torna possível capturar nuances de sentido e similaridade textual com maior precisão, um aspecto crucial na análise de textos legais, frequentemente ambíguos e densos.

O processo técnico consistiu no carregamento do modelo pré-treinado por meio da biblioteca *SentenceTransformer*, seguido da vetorização dos textos armazenados. Ao término da operação, foi verificada a correspondência entre o número de embeddings gerados e a quantidade de textos processados, garantindo a integridade dos dados vetorizados para as etapas subsequentes de análise.

A literatura recente tem destacado os avanços proporcionados por modelos baseados em *transformers* para tarefas complexas de PLN. De acordo com Raffel et al. (2020), essas arquiteturas têm revolucionado a forma como os textos são processados, ao permitirem um aprendizado profundo baseado em atenção contextualizada. Além disso, o uso de representações densas e sem supervisionamento direto, como proposto por Gao, Yao e Chen (2021), permite ganhos substanciais em tarefas de recuperação semântica e agrupamento de documentos, tornando modelos como SBERT indispensáveis em pipelines modernos de PLN.

3.3.1 Comparação entre Modelos de Vetorização Semântica

Com o objetivo de avaliar o impacto da escolha do modelo de vetorização na qualidade da clusterização semântica dos acórdãos judiciais, este trabalho realizou uma análise comparativa entre diferentes modelos baseados em arquiteturas *transformer*. Cada modelo foi utilizado para gerar *embeddings* a partir dos textos previamente processados, sendo esses vetores posteriormente empregados nas etapas de cálculo de similaridade, redução de dimensionalidade, clusterização e modelagem de tópicos.

Na tabela 1, temos os modelos utilizados, tamanho e quantidade de dimensões de cada modelo.

Tabela 1 - Especificações de Modelos de Linguagem

Modelo	Tamanho (Parâmetros)	Dimensões (Tamanho do Embedding)
paraphrase-multilingual-mpnet-base-v2	110M	768
paraphrase-multilingual-MiniLM-L12-v2	33M	384
stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0	340M	1024
Alibaba-NLP/gte-multilingual-base	120M	768
rufimelo/Legal-BERTimbau-sts-large-ma-v3	340M	1024

Fonte: Elaborado pela autora (2025).

A escolha desses modelos fundamenta-se tanto em sua disponibilidade pública quanto em sua capacidade de gerar representações vetoriais densas e contextualizadas. Os modelos *paraphrase-multilingual* são amplamente utilizados por sua robustez em tarefas multilíngues, conforme destacado por Reimers e Gurevych (2019). Já os modelos *Legal-BERTimbau* e *STJIRIS*, treinados especificamente em corpora jurídicos em língua portuguesa, apresentam potencial para melhor desempenho em domínios especializados, conforme sugerido por Souza, Nogueira e Lotufo (2020).

A avaliação comparativa baseou-se em métricas quantitativas, como o *Silhouette Score*, o Índice de *Davies-Bouldin*, *Calinski-Harabasz Index* e a coerência de tópicos gerados via LDA. Adicionalmente, realizou-se uma análise qualitativa por meio da visualização dos agrupamentos com a técnica t-SNE, permitindo observar a separabilidade semântica dos clusters produzidos por cada modelo.

Essa análise busca identificar qual modelo de embeddings oferece maior fidelidade semântica na representação de textos jurídicos, contribuindo para uma clusterização mais coerente e interpretável.

3.4 Aplicação do algoritmo de agrupamento não supervisionados (*K-Means*)

Após a obtenção das representações vetoriais dos acórdãos, aplicou-se o algoritmo de agrupamento *K-Means*, uma técnica amplamente utilizada em aprendizado não supervisionado

com o objetivo de particionar os dados em k grupos de forma a minimizar a soma das distâncias intra-cluster (MACQUEEN, 1967). O K-Means baseia-se em um processo iterativo que envolve a atribuição de cada instância ao centroide mais próximo e a posterior atualização dos centroides com base na média dos pontos associados, repetindo essas etapas até a convergência.

Para determinar o número ideal de clusters (k), uma etapa crucial para a eficácia do agrupamento, foram empregadas duas métricas complementares: o Método do Cotovelo (*Elbow Method*) e o *Silhouette Score*. O Método do Cotovelo analisa a Soma das Distâncias Internas aos Clusters (WCSS - *Within-Cluster Sum of Squares*) em função de diferentes valores de k. O ponto de "cotovelo" na curva, onde a diminuição da WCSS começa a se estabilizar, sugere um valor apropriado para k. Concomitantemente, o *Silhouette Score* foi utilizado para avaliar a coesão e separação dos clusters. Essa métrica varia de -1 a 1, onde valores próximos de 1 indicam clusters bem definidos e separados, enquanto valores próximos de 0 ou negativos sugerem sobreposição ou má formação dos clusters.

O código implementado iterou sobre valores de k entre 2 e 14, calculando para cada um os valores de WCSS e do *Silhouette Score*. Os resultados obtidos foram visualizados em dois gráficos: um representando a curva do cotovelo e outro apresentando a pontuação de silhueta. A partir da análise desses gráficos, observou-se que o valor k = 5 proporcionou uma estrutura de agrupamento coerente, sendo, portanto, adotado para a clusterização final dos dados.

A escolha do algoritmo *K-Means* para a etapa de agrupamento baseou-se em sua eficiência computacional e na sua capacidade de particionar grandes volumes de dados de forma interpretável. Trata-se de um algoritmo amplamente consolidado na literatura para tarefas de agrupamento de textos, o que permite comparações diretas com estudos semelhantes e confere maior robustez metodológica a presente pesquisa (SUYAL; SHARMA, 2024). Além disso, sua aplicação sobre vetores semânticos como os gerados por modelos do tipo SBERT tem se mostrado eficaz na identificação de padrões latentes e estruturas semânticas em conjuntos textuais extensos.

3.5 Redução de dimensionalidade com t-SNE

A técnica t-SNE é um método de redução de dimensionalidade amplamente utilizado para visualização de dados de alta dimensão em espaços bidimensionais ou tridimensionais. Proposto por Van der Maaten e Hinton (2008), o t-SNE preserva principalmente as relações locais entre os pontos no espaço original, de modo que instâncias semelhantes permaneçam próximas no novo espaço vetorial. Essa característica o torna especialmente eficaz para

representar visualmente agrupamentos de dados complexos, como vetores semânticos oriundos de textos.

Neste trabalho, o algoritmo de agrupamento K-Means foi aplicado diretamente sobre os embeddings vetoriais, gerando os rótulos dos clusters sem prévia redução de dimensionalidade. Em seguida, utilizou-se o algoritmo t-SNE com o objetivo de reduzir a dimensionalidade dos dados para duas dimensões, permitindo a visualização dos agrupamentos de forma mais intuitiva. A técnica t-SNE é amplamente empregada em contextos de visualização por sua capacidade de preservar relações locais no espaço vetorial original (VAN DER MAATEN;HINTON,2008). A implementação foi realizada por meio da biblioteca Scikit-learn, com os parâmetros n_components=2, para obtenção de duas dimensões, init='random', e random_state=42, garantindo reprodutibilidade dos resultados. Essa etapa teve papel fundamental na visualização dos dados vetorizados, contribuindo para a interpretação da qualidade dos clusters formados, conforme orientam VAN DER MAATEN e HINTON (2008) e PEDREGOSA et al. (2011).

A escolha do t-SNE, em detrimento de outras técnicas de redução de dimensionalidade como a Análise de Componentes Principais (PCA), justifica-se pelas características específicas dos dados vetoriais utilizados e pelos objetivos da análise. Enquanto o PCA é eficiente na compressão linear dos dados e na explicitação da variância global, ele não preserva bem as relações locais entre os pontos quando aplicado a vetores de alta dimensionalidade e com estrutura semântica complexa, como é o caso dos embeddings textuais gerados por modelos como SBERT.

O t-SNE, por sua vez, é especialmente projetado para manter a proximidade local entre os dados, o que o torna mais apropriado para representar relações de similaridade semântica em espaços vetoriais derivados de linguagem natural. Em outras palavras, instâncias textuais que compartilham conteúdos semanticamente semelhantes tendem a permanecer agrupadas após a redução dimensional com t-SNE, o que favorece a identificação e a visualização dos padrões de agrupamento. Essa propriedade é fundamental no contexto deste trabalho, cujo objetivo inclui a identificação de grupos semânticos latentes em acórdãos judiciais.

Além disso, estudos recentes demonstram que o t-SNE apresenta desempenho superior ao PCA na visualização de agrupamentos em dados de linguagem natural, especialmente quando combinado com algoritmos de clusterização, como o *K-Means*. XIA et al. (2021) realizaram uma análise comparativa entre diferentes técnicas de redução de dimensionalidade e evidenciaram que métodos não lineares, como o t-SNE, proporcionam visualizações mais eficazes para a identificação de estruturas de agrupamento em espaços vetoriais de alta dimensionalidade. Assim, sua adoção neste estudo visa não apenas a redução de dimensionalidade, mas

também a facilitação da interpretação dos resultados por meio de representações visuais mais coerentes com a estrutura semântica dos dados.

3.6 Visualização dos agrupamentos

Após a aplicação do algoritmo *K-Means*, sobre os vetores semânticos dos acórdãos, procedeu-se à etapa de visualização gráfica dos agrupamentos. O t-SNE foi utilizado para projetar os dados de alta dimensionalidade em um espaço bidimensional, mantendo a estrutura local dos dados e preservando relações de proximidade entre os pontos. Essa técnica é especialmente eficaz para dados oriundos de representações vetoriais densas, como os embeddings gerados por modelos de linguagem.

A visualização foi realizada por meio de um gráfico de dispersão, no qual cada ponto representa um acórdão posicionado segundo suas coordenadas no espaço reduzido. Os pontos foram coloridos conforme o rótulo do cluster atribuído pelo *K-Means*, o que permitiu uma avaliação qualitativa da distribuição dos documentos e da coesão interna dos grupos. A visualização dos cinco clusters possibilitou identificar padrões semânticos latentes e avaliar a separabilidade entre os agrupamentos, contribuindo para a interpretação dos resultados e validação da metodologia.

A utilização combinada de *K-Means* e t-SNE para visualização é amplamente reconhecida na literatura como uma abordagem eficaz para análise exploratória em tarefas de agrupamento textual (VANDER MAATEN; HINTON, 2008; WATTENBERG et al., 2016), especialmente em domínios como o jurídico, onde a semântica desempenha papel central na organização dos dados.

3.7 Geração de nuvens de palavras (*wordclouds*)

A geração de nuvens de palavras constitui uma técnica visual amplamente utilizada na análise exploratória de dados textuais, permitindo a identificação rápida e intuitiva dos termos mais frequentes em um corpus. No contexto deste estudo, as nuvens de palavras foram geradas individualmente para cada um dos cinco clusters resultantes da aplicação do algoritmo K-Means, com o objetivo de facilitar a interpretação semântica dos agrupamentos.

Cada nuvem de palavras representa, de forma gráfica, os termos mais recorrentes dentro de um grupo, sendo que o tamanho de cada palavra está proporcionalmente relacionado à sua frequência no conjunto de textos pertencentes ao respectivo cluster. Essa visualização auxilia

na identificação de temas dominantes, padrões linguísticos e possíveis subtemas que caracterizam cada agrupamento, oferecendo uma síntese visual que complementa a análise quantitativa.

A utilização das nuvens de palavras é particularmente eficaz na fase de validação qualitativa dos agrupamentos, pois permite verificar a coesão temática dentro dos clusters e observar se os termos destacados são semanticamente coerentes com o domínio jurídico analisado. Essa abordagem também favorece a comunicação dos resultados para públicos não especializados, por se tratar de uma forma acessível e de fácil compreensão.

Conforme apontado por Silva e Ribeiro (2020), as nuvens de palavras constituem uma estratégia eficiente de exploração textual, ao passo que permitem evidenciar rapidamente os tópicos centrais em grandes volumes de dados. Assim, sua incorporação nesta pesquisa contribui não apenas para a interpretação dos clusters, mas também para a apresentação visual dos achados de maneira clara e objetiva.

3.8 Modelagem de tópicos com *Latent Dirichlet Allocation* (LDA)

A modelagem de tópicos por meio do algoritmo LDA constitui uma técnica amplamente empregada na extração de temas latentes a partir de grandes volumes de dados textuais. Proposto por Blei, Ng e Jordan (2003), o LDA é um modelo generativo que assume que cada documento é composto por uma distribuição de tópicos, e que cada tópico é representado por uma distribuição de palavras.

No contexto desta pesquisa, a aplicação do LDA foi realizada após a clusterização dos acórdãos judiciais por meio do algoritmo *K-Means*, com o objetivo de identificar os principais tópicos presentes em cada grupo de textos semanticamente semelhantes. Ao aplicar o LDA de forma separada em cada cluster, é possível extrair os temas predominantes em cada agrupamento, contribuindo tanto para a interpretação semântica quanto para a validação qualitativa dos resultados obtidos na etapa de agrupamento.

Essa abordagem permite não apenas confirmar a coerência temática dos clusters formados, mas também refinar a análise exploratória ao evidenciar subtemas ou padrões discursivos específicos que não são imediatamente perceptíveis apenas pela visualização das nuvens de palavras ou pela leitura dos documentos. Dessa forma, a modelagem de tópicos atua como um complemento analítico, enriquecendo a compreensão dos dados textuais e fortalecendo a consistência das conclusões obtidas.

Adicionalmente, a utilização do LDA mostra-se particularmente adequada ao domínio jurídico, cuja linguagem se caracteriza por alta densidade terminológica e complexidade

semântica. A extração automatizada de tópicos recorrentes favorece a categorização temática de decisões judiciais, podendo subsidiar práticas mais eficazes de gestão documental e recuperação da informação.

Após a aplicação do modelo LDA, utilizou-se a ferramenta *PyLDAvis* para a visualização interativa dos tópicos. O *PyLDAvis* permite representar os tópicos como círculos distribuídos em um espaço bidimensional, onde a distância entre eles reflete a similaridade temática e o tamanho indica sua prevalência no corpus. A visualização também apresenta, para cada tópico, a lista de palavras mais relevantes com base em métricas de frequência e exclusividade. Essa abordagem contribui para uma análise mais aprofundada das estruturas temáticas, reforçando a interpretação qualitativa dos agrupamentos. A implementação foi realizada com a versão *pyLDAvis*.*lda_model,* utilizando os vetores de contagem de palavras gerados por *CountVectorizer*.

3.9 Avaliação da Qualidade dos Agrupamentos

Após a aplicação dos algoritmos de clusterização, foi realizada a avaliação quantitativa da qualidade dos agrupamentos obtidos, com o objetivo de comparar o desempenho dos diferentes modelos de vetorização semântica testados. Considerando-se a natureza não supervisionada do problema, foram empregadas métricas de validação interna de agrupamentos, as quais permitem mensurar a compactação intra-cluster e a separação inter-cluster com base exclusivamente na distribuição dos dados (ALI et al., 2021; HASSAN et al., 2024; GÖSGENS et al., 2019).

As métricas aplicadas neste estudo foram:

- *Silhouette Score*, que avalia simultaneamente a coesão e separação dos clusters, variando de -1 a 1, onde valores próximos de 1 indicam melhor qualidade de agrupamento (ROUSSEEUW, 1987);
- *Davies-Bouldin Index*, o qual mede a razão entre a variabilidade intra-cluster e a separação inter-cluster, sendo que valores menores indicam melhor desempenho (DAVIES; BOULDIN, 1979);
- *Calinski-Harabasz Index*, que calcula a razão entre a dispersão inter-cluster e intra-cluster, onde valores maiores indicam melhor qualidade dos agrupamentos (CALIŃSKI; HARABASZ, 1974).

O cálculo dessas métricas foi realizado utilizando a biblioteca *scikit-learn* da linguagem Python, versão 1.2.2. As métricas foram aplicadas sobre os agrupamentos formados pelo algoritmo *K-Means*, considerando os embeddings gerados por cada modelo de representação

vetorial previamente testado. A utilização conjunta dessas métricas permitiu avaliar de forma mais robusta a qualidade dos agrupamentos formados, considerando múltiplos aspectos estatísticos e estruturais dos dados.

4 RESULTADOS E ANÁLISE

As análises apresentadas neste capítulo foram conduzidas com base nos procedimentos metodológicos descritos no Capítulo 3, abrangendo as etapas de coleta e pré-processamento textual, vetorização semântica por meio do modelo SBERT (com a seleção do modelo mais adequado ao conjunto de dados), aplicação do algoritmo K-Means para clusterização, redução de dimensionalidade com t-SNE, além da avaliação quantitativa e qualitativa dos agrupamentos formados.

A seguir, são apresentados os resultados obtidos com a aplicação das técnicas propostas ao agrupamento de acórdãos judiciais, considerando a similaridade semântica capturada por modelos de linguagem baseados em redes neurais. O capítulo descreve o conjunto de dados analisado, os procedimentos de vetorização, a definição do número ideal de clusters, a visualização dos agrupamentos gerados e as avaliações quantitativas, como o *Silhouette Score*, o Índice de *Davies-Bouldin* e o Índice de *Calinski-Harabasz*. Complementarmente, são apresentados os resultados qualitativos obtidos por meio de nuvens de palavras e modelagem de tópicos com LDA, que permitem a interpretação semântica dos agrupamentos formados. Por fim, discute-se a eficácia da metodologia aplicada, destacando suas contribuições para a organização e análise de acórdãos judiciais.

4.1 Características do Conjunto de Dados

O conjunto de dados analisado neste estudo é composto por 863 acórdãos judiciais, extraídos do sistema e-Proc da segunda instância, e processados conforme as etapas metodológicas descritas no Capítulo 3. Os documentos contemplam exclusivamente matérias de competência cível, assegurando a homogeneidade temática da base textual. Cada acórdão foi submetido a um processo de pré-processamento textual, que incluiu normalização, remoção de *stopwords* e lematização, com o objetivo de obter textos mais limpos e semanticamente representativos. Em seguida, os documentos foram transformados em representações vetoriais semânticas e, posteriormente, submetidos à etapa de clusterização por similaridade.

Embora a amostra possa ser considerada limitada em termos quantitativos, mostrou-se adequada e representativa para os objetivos da pesquisa, fornecendo uma base empírica consistente para a avaliação dos métodos de agrupamento e análise semântica aplicados.

4.2 Resultados da Vetorização Semântica

A vetorização semântica dos textos foi realizada com o modelo *Sentence*-BERT (SBERT), sendo aplicados cinco modelos distintos pertencentes a essa família. Ao final do processo, foram gerados 863 embeddings para cada modelo, correspondendo ao número total de acórdãos analisados. Esses vetores, armazenados em formato matricial, serviram como base para as etapas subsequentes da análise, tais como clusterização, avaliação quantitativa e visualização. A correspondência entre o número de vetores e o número de documentos originais foi devidamente verificada, assegurando a integridade e a completude da transformação dos dados textuais em representações numéricas.

A escolha por modelos da família SBERT justifica-se por sua arquitetura siamesa, que possibilita comparações diretas entre sentenças com maior eficiência computacional e sem a necessidade de grandes volumes de dados rotulados. Essa característica é particularmente relevante no contexto jurídico, marcado pela escassez de corpora anotados. Segundo estudos recentes, modelos baseados em SBERT demonstram desempenho superior em tarefas de recuperação de informação e agrupamento de textos, em comparação com abordagens tradicionais como TF-IDF ou Word2Vec (GAO; YAO; CHEN, 2021).

A análise comparativa dos resultados obtidos pelos modelos será apresentada em seção específica deste capítulo, com o objetivo de identificar a abordagem mais adequada ao agrupamento dos acórdãos judiciais.

4.3 Definição do Número de Clusters

A definição do número de clusters (k) foi realizada segundo as diretrizes descritas na seção 3.5, com apoio do Método do Cotovelo e do *Silhouette Score*, de modo a identificar o ponto de equilíbrio entre complexidade e desempenho dos agrupamentos.

No gráfico da esquerda (Figura 3), apresenta a Soma das Distâncias Internas (*Within-Cluster Sum of Squares* – WCSS) para diferentes valores de *k*, variando de 2 a 14. O método busca identificar o ponto em que o ganho na compactação dos clusters começa a diminuir significativamente, ou seja, onde há uma "quebra" na curva. Observa-se que a redução de WCSS é mais acentuada até o ponto k = 6, a partir do qual as quedas tornam-se menos expressivas. Essa inflexão indica um possível número ótimo de agrupamentos, sugerindo que seis clusters representam um bom equilíbrio entre granularidade e coesão interna.

Adicionalmente, o gráfico da direita apresenta os valores do *Silhouette Score*, métrica que avalia a coesão interna e a separação entre os clusters. Embora o valor máximo tenha sido registrado em k = 2, essa configuração tende a gerar agrupamentos muito genéricos, com perda de granularidade semântica. Ao considerar a combinação entre o valor de *Silhouette* e a análise do cotovelo, nota-se que o valor de k = 6 mantém um índice razoável de coesão (em torno de 0.075) e oferece maior discriminação temática nos grupos. Com base nas análises combinadas do Método do Cotovelo e do Índice de *Silhouette*, optou-se por definir k = 6 como o número de clusters a ser utilizado nas etapas seguintes da análise.

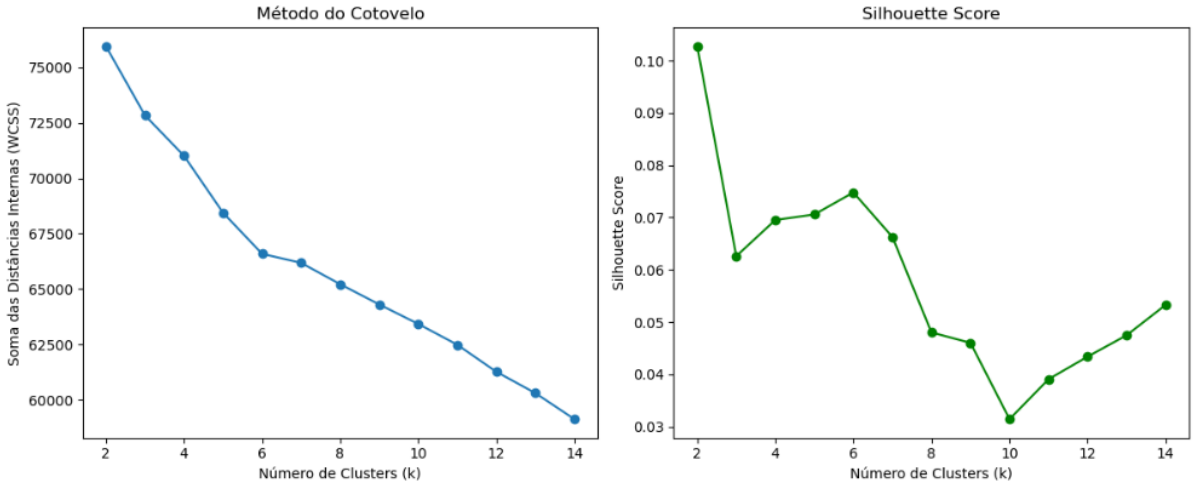

Figura 3 - Análise do número ideal de clusters por meio do Método do Cotovelo e *Silhouette Score*

Fonte: Elaborado pela autora (2025).

Essa decisão também está alinhada com a proposta exploratória deste estudo, que visa balancear granularidade interpretativa e coerência semântica entre os grupos de acórdãos.

4.4 Visualização dos Clusters

A visualização dos agrupamentos foi realizada com o auxílio do algoritmo t-SNE, após a aplicação do *K-Means*, sendo que o valor de k foi previamente definido com base nas análises do método do cotovelo e do coeficiente de silhueta. O objetivo dessa etapa é proporcionar uma representação bidimensional dos embeddings vetoriais, permitindo a inspeção visual da separação entre os grupos.

A Figura 4 apresenta a visualização bidimensional dos agrupamentos formados com o algoritmo *K-Means*, aplicada sobre os embeddings gerados. Para viabilizar a interpretação gráfica, foi utilizada a técnica de redução de dimensionalidade t-SNE, que projeta os dados

originalmente em alta dimensão para um espaço de duas dimensões, mantendo a estrutura de vizinhança local.

Cada ponto no gráfico representa um acórdão, posicionado com base na sua similaridade semântica relativa aos demais documentos. As cores indicam os rótulos atribuídos pelo *K-Means*, o que permite verificar a distribuição dos documentos nos agrupamentos.

Observa-se que há formações coesas em determinados grupos, indicando que o *K-Means* foi capaz de separar adequadamente documentos com características semânticas semelhantes. No entanto, também é possível notar regiões de sobreposição entre clusters, especialmente nas áreas centrais do gráfico, o que pode sugerir a existência de temas jurídicos com interseções conceituais, ou limitações do *K-Means* em capturar fronteiras complexas de decisão.

A técnica t-SNE não preserva distâncias absolutas, mas é eficaz na preservação da topologia local, sendo, portanto, apropriada para inspeção visual da qualidade dos agrupamentos. De modo geral, o gráfico indica que a combinação SBERT + *K-Means* + t-SNE é capaz de oferecer uma representação coerente dos agrupamentos semânticos de acórdãos judiciais.

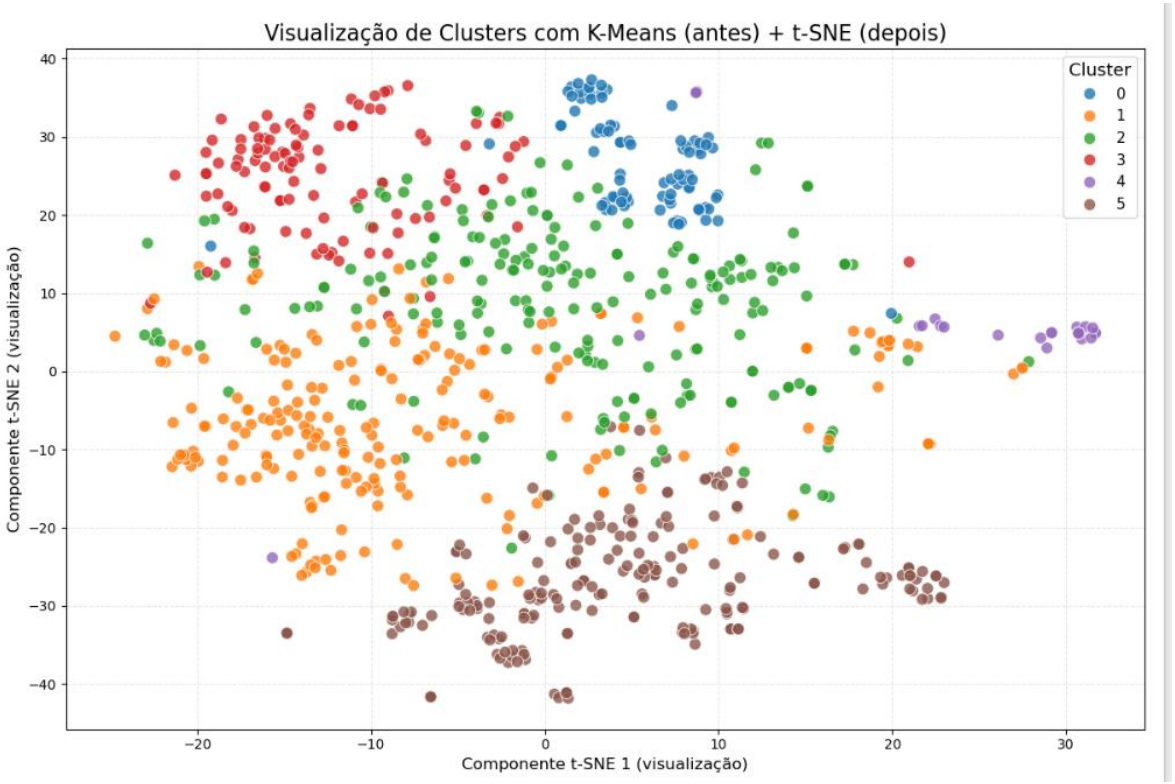

Figura 4 - Visualização dos Clusters com *K-Means* e t-SNE. Modelo utilizado: stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0

Fonte: Elaborado pela autora (2025).

Os agrupamentos formados constituem a base para as próximas etapas da análise, que envolvem a interpretação semântica dos clusters com o auxílio de nuvens de palavras e modelagem de tópicos, conforme descrito nas seções subsequentes

4.5 Comparação entre Modelos de Embeddings

Com o objetivo de identificar o modelo de vetorização semântica mais adequado para o agrupamento de acórdãos judiciais, foram avaliados cinco modelos baseados em *Sentence*-BERT (SBERT), como mostra a Tabela 2.

Para cada modelo, os embeddings foram gerados a partir dos mesmos textos pré-processados. A seguir, realizou-se a clusterização com o algoritmo *K-Means*, utilizando redução dimensional via t-SNE para visualização. A determinação do número de clusters (*k*) foi realizada por meio do Método do Cotovelo.

A Tabela 2 apresenta os resultados obtidos na avaliação quantitativa da qualidade dos agrupamentos gerados a partir de cinco modelos distintos de embeddings.

As métricas utilizadas nessa avaliação, previamente descritas na Seção 3.10, incluem o *Silhouette Score*, o *Davies-Bouldin Index* e o *Calinski-Harabasz Index*, as quais permitem mensurar, de forma combinada, a coesão interna e a separação entre os clusters gerados.

Tabela 2 - Desempenho dos Modelos de *Embeddings* segundo Métricas de Validação de Clusterização

Modelo	*Silhouette Score*	*Davies-Bouldin Index*	*Calinski-Harabasz Index*
Alibaba-NLP/gte-multilingual-base	0.2125	2.4547	310.37
paraphrase-multilingual-mpnet-base-v2	0.2027	1.4474	351.27
paraphrase-multilingual-MiniLM-L12-v2	0.1948	1.7755	357.74
stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0	**0.2319**	**1.1978**	**431.08**
rufimelo/Legal-BERTimbau-sts-large-ma-v3	0.1866	2.4381	252.22

Fonte: Elaborado pela autora (2025).

Com base nos resultados, observa-se que o modelo *stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0* obteve os melhores desempenhos nas três métricas analisadas. Esse modelo apresentou o maior *Silhouette Score* (0.2319), o menor *Davies-Bouldin Index* (1.1978) e o maior *Calinski-Harabasz Index* (431.08), o que indica, respectivamente, maior coesão interna dos clusters, menor sobreposição entre grupos e maior separação entre os agrupamentos formados.

Tais resultados são consistentes com a natureza do modelo, que foi treinado em corpora jurídicos em língua portuguesa, o que lhe confere maior capacidade de representar adequadamente a semântica dos acórdãos analisados. Dessa forma, o modelo *stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0* foi selecionado para as análises subsequentes no presente estudo.

Na figura 5, observa-se que o modelo *stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0* apresentou o maior valor de *Silhouette Score* (0.2319), o que indica que seus *clusters* apresentam maior coesão interna e estão mais bem separados uns dos outros, em comparação aos demais modelos. Os modelos *paraphrase-multilingual-mpnet-base-v2* e *Alibaba-NLP/gte-multilingual-base* também apresentaram desempenhos razoáveis, enquanto o modelo *rufimelo/Legal-BERTimbau* teve o menor valor, sugerindo agrupamentos menos definidos.

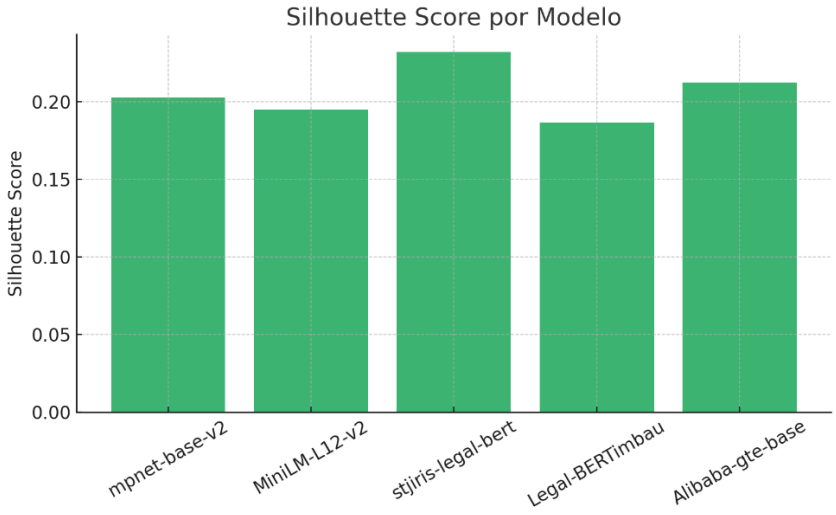

Figura 5 - Valores de *Silhouette Score* obtidos para cada modelo de *embeddings*.

Fonte: Elaborado pela autora (2025).

O *Davies-Bouldin Index* (DBI) mede a média da razão entre dispersão *intra-cluster* e separação *inter-cluster*. Ao contrário do *Silhouette*, valores menores de DBI indicam melhor qualidade nos agrupamentos.

Como pode ser observado na Figura 6, novamente o modelo *stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0* obteve o melhor desempenho, com o menor índice (1.1978). Esse resultado corrobora a análise anterior de que seus agrupamentos são mais bem definidos. Os piores desempenhos foram apresentados pelos modelos *Legal-BERTimbau* e *Alibaba-gte-base*, com índices superiores a 2.4, sugerindo sobreposição e menor separação entre clusters.

.

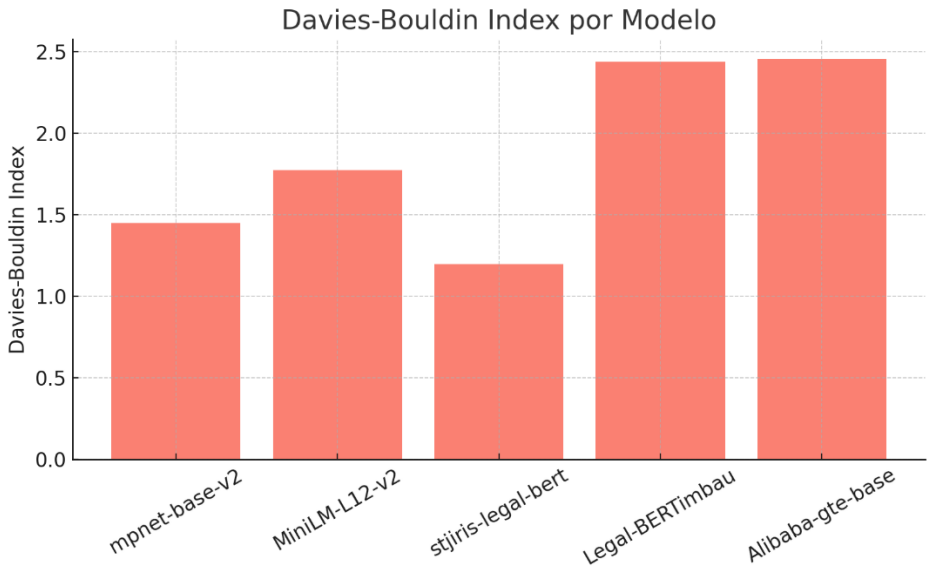

Figura 6 - Valores de *Davies-Bouldin Index* obtidos para cada modelo de *embeddings*.

Fonte: Elaborado pela autora (2025).

O *Calinski-Harabasz Index* (CHI) avalia a razão entre a dispersão entre os clusters e a dispersão dentro dos clusters. Quanto maior o valor, melhor a qualidade dos agrupamentos.

Na Figura 7, o modelo *stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0* novamente se destacou com o maior valor de CHI (431.08), indicando uma separação bem definida entre os grupos. Em contrapartida, o modelo *Legal-BERTimbau* teve o menor valor de CHI (252.22), o que reforça os indícios de baixa performance já apontados pelas outras métricas.

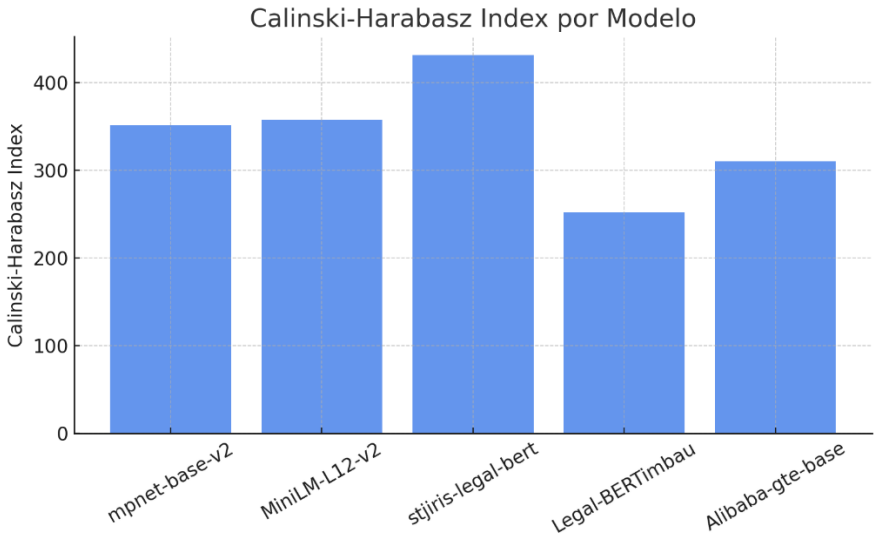

Figura 7 - Valores de *Calinski-Harabasz Index* obtidos para cada modelo de *embeddings*.
Fonte: Elaborado pela autora (2025).

A análise comparativa das três métricas confirmam que o modelo *stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0* foi o que apresentou os melhores resultados de clusterização, sendo, portanto, o mais adequado para representar semanticamente os acórdãos no contexto jurídico avaliado. Tal constatação será levada em conta nas análises posteriores desta dissertação, como a visualização dos agrupamentos e a modelagem de tópicos.

4.6 Análise dos Resultados com *Wordclouds* e Modelagem de Tópicos (LDA)

Como forma complementar de avaliação qualitativa dos agrupamentos obtidos, foram geradas nuvens de palavras e aplicada a modelagem de tópicos com o algoritmo LDA. Além disso, utilizou-se a ferramenta *PyLDAvis* para visualizar a distribuição e a relevância dos tópicos identificados, permitindo uma interpretação mais aprofundada das estruturas temáticas presentes em cada cluster.

Como etapa preliminar da análise temática, foram geradas nuvens de palavras para cada cluster. Essa abordagem visual destaca os termos mais frequentes em cada grupo, evidenciando padrões lexicais que caracterizam semanticamente os conjuntos de acórdãos. Por exemplo, a nuvem referente ao Cluster 0 na Figura 8, revelou forte predominância dos termos *execução*, *espólio*, *fiscal*, *extinção* e *inventariante*, indicando que esse cluster agrega decisões relacionadas a processos de execução fiscal e representação legal pós-falecimento.

As nuvens de palavras, apesar de não expressarem relações entre os termos, fornecem uma visão exploratória útil, especialmente para avaliar a coerência temática de cada agrupamento.

Figura 8 - Nuvem de palavras representativa do Cluster 0.

Fonte: Elaborado pela autora (2025).

Em seguida, foi aplicada a modelagem LDA com a extração de dois tópicos por cluster, utilizando o vetor de contagem de palavras (*CountVectorizer*) sobre os textos pré-processados. Os tópicos extraídos são compostos por lista de palavras com maior probabilidade condicional de ocorrência dentro daquele tema. A Figura 9 ilustra a distribuição desses tópicos identificados pelo modelo LDA para cada cluster analisado.

```
Principais tópicos por cluster (LDA com random_state=42):

Cluster 0:
  Tópico 1: espolio | fiscal | extincao | inventariante | representante | indicacao | ausencia | pressuposto | determinacao | representacao
  Tópico 2: espolio | fiscal | preclusao | representacao | representante | provir | indicacao | extincao | judicato | passivo

Cluster 1:
  Tópico 1: dano | saude | moral | ente | medico | autor | tratamento | energia | estado | eletrico
  Tópico 2: valor | dano | administrativo | prova | contrato | autor | provir | juro | consumidor | moral

Cluster 2:
  Tópico 1: execucao | divida | valor | credito | fiscal | banco | publicar | honorario | decisao | imovel
  Tópico 2: execucao | fiscal | provir | prescricao | credito | tributario | instrumento | prazo | valor | citacao

Cluster 3:
  Tópico 1: omissao | contradicao | terceiro | rediscussao | inexistencia | materia | impugnacao | decisao | honorario | embargar
  Tópico 2: materia | instrumento | omissao | rediscussao | decisao | execucao | vicio | contradicao | alegacao | obscuridade

Cluster 4:
  Tópico 1: tema | retratacao | conta | brasil | legitimidade | banco | vincular | saldo | programa | obrigacao
  Tópico 2: tema | favor | honorario | defensoria | obrigacao | advocaticio | retratacao | condenacao | possibilidade | estadual

Cluster 5:
  Tópico 1: municipal | feria | servidor | municipio | prescricao | constitucional | pagamento | terco | federal | ano
  Tópico 2: municipal | tempo | servico | adicional | servidor | municipio | complementar | prever | legal | despesa
```

Figura 9 - Tópicos extraídos pelo modelo LDA para os clusters analisados.

Fonte: Elaborado pela autora (2025).

Além das representações textuais, a ferramenta *PyLDAvis* foi utilizada para examinar a composição interna de cada tópico. A Figura 10, por exemplo, apresenta os dois tópicos extraídos do Cluster 0. Observa-se que os círculos vermelhos e azuis representam a distância semântica entre os tópicos, enquanto as barras horizontais à direita mostram os termos com maior relevância (em vermelho) e frequência geral (em azul claro). Palavras como *espólio*, *fiscal*, *extinção*, *inventariante* e *representante* aparecem com alta saliência e relevância no Tópico 1, reforçando a interpretação temática já indicada pela nuvem de palavras.

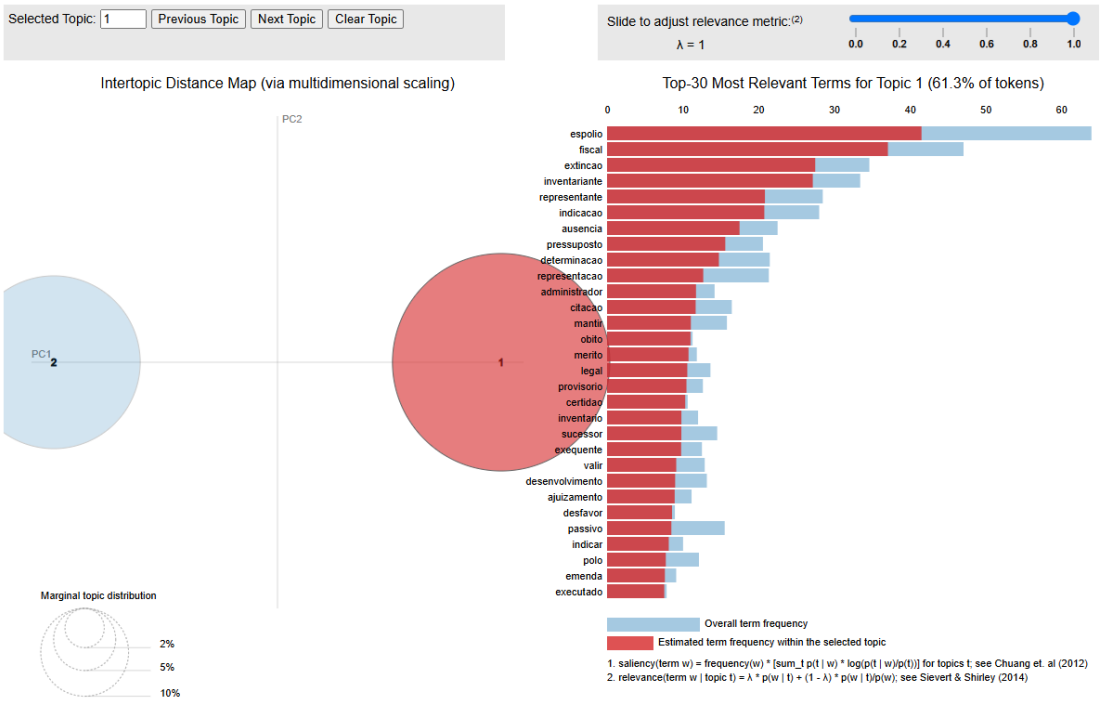

Figura 10 – Visualização interativa de tópicos com *PyLDAvis* (Cluster 0).
Fonte: Elaborado pela autora (2025).

A análise qualitativa dos agrupamentos, realizada com apoio das nuvens de palavras, da modelagem LDA e da visualização com *PyLDAvis*, permite inferir a composição temática predominante de cada cluster. De forma geral, o Cluster 0 apresenta forte presença de temas relacionados à execução fiscal contra espólio, como evidenciado pelos termos fiscal, execução e espólio. O Cluster 1 reúne majoritariamente acórdãos envolvendo responsabilidade civil e danos morais, com destaque para termos como dano, indenização e saúde. No Cluster 2, observa-se a concentração de processos de execução e dívida tributária. Já o Cluster 3 agrupa matérias associadas à impugnação e rediscussão de mérito, enquanto o Cluster 4 apresenta uma

combinação de temas cíveis e do trabalho, e o Cluster 5, por sua vez, trata majoritariamente de Direito Administrativo e questões ligadas a servidor público.

Essa interpretação integrada reforça que a metodologia aplicada conseguiu organizar semanticamente o corpus de acórdãos em categorias jurídicas com significado prático, preservando coerência temática e alinhamento com o contexto judicial analisado. Além disso, os resultados foram validados por um profissional da área do Direito, o que confirma a aplicabilidade e a pertinência da abordagem proposta no ambiente jurídico.

.

5 CONCLUSÃO

O presente trabalho teve como objetivo central o desenvolvimento e avaliação de uma metodologia baseada em Processamento de Linguagem Natural (PLN) e aprendizado não supervisionado para o agrupamento de acórdãos judiciais, a fim de contribuir para a triagem e análise de decisões no contexto jurídico.

A metodologia proposta demonstrou-se eficaz para atingir o objetivo geral: desenvolver um modelo computacional capaz de agrupar acórdãos por similaridade semântica, contribuindo para a otimização da análise de pedidos de recurso no gabinete da presidência do Tribunal de Justiça.

Em relação aos objetivos específicos, os resultados obtidos permitiram concluir:

Objetivo 1: Extrair e normalizar as informações textuais presentes nos acórdãos, por meio de técnicas de PLN. Essa etapa foi devidamente executada com o uso de bibliotecas consolidadas como *spaCy* e *NLTK*, contemplando procedimentos de normalização textual, remoção de ruídos e lematização, essenciais para garantir a qualidade da representação semântica dos textos.

Objetivo 2: Aplicar algoritmos de agrupamento (*clustering*) baseados em aprendizado não supervisionado para identificar e organizar decisões judiciais semanticamente semelhantes. A vetorização foi realizada com o modelo SBERT e suas variantes, seguida de clusterização com o algoritmo *K-Means* e redução de dimensionalidade com t-SNE para visualização. Os agrupamentos obtidos mostraram-se semanticamente coerentes, permitindo a formação de grupos temáticos bem definidos.

Objetivo 3: Avaliar a qualidade dos agrupamentos por meio de métricas de validação internas e visualização dos clusters. A avaliação quantitativa dos agrupamentos foi realizada com base nas métricas internas *Silhouette Score*, *Davies-Bouldin Index* e *Calinski-Harabasz Index*, associadas à visualização gráfica dos clusters após redução de dimensionalidade com t-SNE. Os resultados demonstraram que o modelo *stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0*, treinado especificamente para o domínio jurídico, obteve o melhor desempenho geral, destacando-se por apresentar o maior *Silhouette Score* (0,2319) e o menor *Davies-Bouldin Index* (1,1978), indicando maior coesão *intra-cluster* e melhor separação entre grupos. A comparação com outros modelos de embeddings, como os multilinguais generalistas *paraphrase-multilingual-mpnet-base-v2 e MiniLM-L12-v2*, bem como modelos alternativos como o Legal-BERTimbau e o GTE-base, revelou que modelos especializados tendem a capturar nuances semânticas mais relevantes em textos jurídicos. Entretanto, a performance variou

conforme a métrica utilizada, evidenciando a importância de uma abordagem comparativa e multimétrica na avaliação da qualidade dos agrupamentos. A análise visual reforçou essas conclusões ao evidenciar uma segmentação clara e semanticamente consistente entre os grupos formados.

Objetivo 4: Realizar uma análise qualitativa dos agrupamentos utilizando modelagem de tópicos e nuvens de palavras. As nuvens de palavras permitiram a visualização dos termos mais recorrentes por cluster, oferecendo uma visão exploratória inicial sobre os temas tratados. Em seguida, foi aplicada a modelagem de tópicos com o algoritmo LDA, com extração de dois tópicos por cluster. Essa abordagem permitiu uma interpretação mais precisa dos conteúdos temáticos predominantes, destacando-se, por exemplo, a associação de determinados clusters a temas como execução fiscal, responsabilidade civil, rediscussão de mérito e Direito Administrativo. Complementarmente, o uso da ferramenta *PyLDAvis* possibilitou uma visualização interativa da distribuição dos tópicos e da relevância dos termos, contribuindo para uma análise qualitativa mais robusta e aprofundada.

Em síntese, a metodologia desenvolvida demonstrou-se eficaz para organizar automaticamente acórdãos judiciais com base em sua similaridade semântica, atingindo de forma satisfatória todos os objetivos da pesquisa. A combinação de análises quantitativas e qualitativas revelou o potencial dos modelos baseados em SBERT para capturar relações semânticas complexas em textos jurídicos, mesmo na ausência de dados rotulados. Importa registrar que, para cada documento, utilizou-se exclusivamente a ementa do acórdão, a fim de manter compatibilidade com a prática de triagem manual.

Embora o estudo tenha sido realizado com um corpus restrito a 863 acórdãos cíveis, os resultados obtidos são promissores. Como direcionamentos para pesquisas futuras, recomenda-se a ampliação do conjunto de dados, a aplicação do modelo em diferentes ramos do Direito e a exploração de *embeddings* jurídicos mais avançados, treinados especificamente para a língua portuguesa. Tais ações poderão fortalecer ainda mais a robustez e a aplicabilidade dos resultados em ambientes reais de triagem e análise judicial.

Este trabalho não propõe automação da triagem. O sistema desenvolvido atua como suporte à análise preliminar, cabendo ao servidor/magistrado a decisão final.

Dessa forma, a abordagem proposta configura-se como uma contribuição prática e tecnicamente fundamentada para a modernização da análise de recursos judiciais, oferecendo uma alternativa eficiente para apoiar a tomada de decisão nos gabinetes da presidência e contribuir para a racionalização dos fluxos de trabalho no Poder Judiciário.

REFERÊNCIAS

ALCARDE, Charles. Processamento de Linguagem Natural: o poder da linguagem na era da inteligência artificial. Edição do Kindle, 2021.

ALGHAMDI, R.; ALFALQI, K. A Survey of Topic Modeling in Text Mining. International Journal of Advanced Computer Science and Applications (IJACSA), v. 6, n. 1, p. 147–153, 2015. DOI: 10.14569/IJACSA.2015.060121. Disponível em: https://thesai.org/Downloads/Volume6No1/Paper_21-A_Survey_of_Topic_Modeling_in_Text_Mining.pdf. Acesso em: 25 ago. 2025.

ALI, L. et al. A comprehensive review of recent advancements in clustering validation measures. IEEE Access, v. 9, p. 95421-95437, 2021. DOI: https://doi.org/10.1109/ACCESS.2021.3093175.

ALPAYDIN, Ethem. Introdução ao aprendizado de máquina. 2. ed. Porto Alegre: Bookman, 2010.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 35, n. 8, p. 1798–1828, 2013.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. Journal of Machine Learning Research, v. 3, p. 993–1022, 2003. Disponível em: https://www.jmlr.org/papers/v3/blei03a.html. Acesso em: 4 jun. 2025.

BRASIL. Lei nº 13.105, de 16 de março de 2015. Código de Processo Civil. Diário Oficial da União: seção 1, Brasília, DF, ano 152, n. 51, p. 1-34, 17 mar. 2015. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm. Acesso em: 12 maio 2025.

BRASIL. Conselho Nacional de Justiça. Resolução nº 235, de 13 de julho de 2016. Dispõe sobre a organização e as atribuições do Núcleo de Gerenciamento de Precedentes (NUGEP) nos tribunais e dá outras providências. Brasília, DF: CNJ, 2016. Disponível em: https://atos.cnj.jus.br/atos/detalhar/2312. Acesso em: 25 ago. 2025.

BRASIL. Conselho Nacional de Justiça. Portaria nº 116, de 6 de abril de 2022. Estabelece requisitos para a padronização das informações a serem apresentadas pelos tribunais e pela Turma Nacional de Uniformização para alimentação do Banco Nacional de Precedentes. Brasília, DF: CNJ, 2022a. Disponível em: https://atos.cnj.jus.br/atos/detalhar/4475. Acesso em: 25 ago. 2025.

BRASIL. Conselho Nacional de Justiça. Resolução nº 444, de 25 de fevereiro de 2022. Institui o Banco Nacional de Precedentes (BNP). Brasília, DF: CNJ, 2022b. Disponível em: https://atos.cnj.jus.br/atos/detalhar/4415. Acesso em: 25 ago. 2025.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. Communications in Statistics, v. 3, n. 1, p. 1-27, 1974.

CAMACHO-COLÓN, D. A.; DÍAZ, J. A. C. Natural Language Processing with Python and spaCy: A Practical Introduction. O'Reilly Media, 2020.

CARDELLINO, C. et al. Low-cost Resources and Evaluation for Named Entity Recognition in the Legal Domain. arXiv preprint, arXiv:1707.02270, 2017. Disponível em: https://arxiv.org/abs/1707.02270. Acesso em: 25 ago. 2025.

CASELI, H.M.; Nunes, M.G.V. (org.) Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. 2 ed. BPLN, 2024. Disponível em: https://brasileiras-pln.com/livro-pln/2a-edicao.

CHALKIDIS, Ilias; FERGADIOTIS, Manos; MALAKASIOTIS, Prodromos; ALETRAS, Nikolaos; ANDROUTSOPOULOS, Ion. LEGAL-BERT: os Muppets saídos diretamente da Faculdade de Direito. In: *FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: EMNLP 2020*, 2020, Online. Anais [...]. [S.l.]: Association for Computational Linguistics, 2020. p. 2898–2904. Disponível em: https://aclanthology.org/2020.findings-emnlp.261/. Acesso em: 19 maio 2025.

DAVIES, D. L.; BOULDIN, D. W. *A cluster separation measure.* IEEE Transactions on Pattern Analysis and Machine Intelligence, v. PAMI-1, n. 2, p. 224-227, 1979.

DEVLIN, Jacob et al. 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. Minneapolis: ACL, 2019. p. 4171–4186. Disponível em: https://arxiv.org/abs/1810.04805. Acesso em: 17 maio 2025.

GAO, L.; YAO, X.; CHEN, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. arXiv preprint arXiv:2104.08821, 2021. Disponível em: https://arxiv.org/abs/2104.08821. Acesso em: 25 maio 2025.

GOLDBERG, Yoav. *Neural Network Methods for Natural Language Processing*. San Rafael: Morgan & Claypool Publishers, 2017. (Synthesis Lectures on Human Language Technologies, v. 10, n. 1). DOI: https://doi.org/10.2200/S00762ED1V01Y201703HLT037

GÖSGENS, M.; TIKHONOV, A.; PROKHORENKOVA, L. *Systematic Analysis of Cluster Similarity Indices.* arXiv preprint arXiv:1911.04773, 2019.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3. ed. San Francisco: Morgan Kaufmann, 2011.

HASSAN, B. A.; ABBAS, A.; HADI, K. J.; OMAR, K. I. From A-to-Z review of clustering validation indices. arXiv preprint arXiv:2407.20246, 2024. Disponível em: https://arxiv.org/pdf/2407.20246

HEIMERL, Florian; LOHMANN, Steffen; LANGE, Simon; ERTL, Thomas. Word Cloud Explorer: Text Analytics Based on Word Clouds. In: 2014 47th Hawaii International Conference on System Sciences (HICSS). Washington, DC: IEEE, 2014. p. 1833–1842. DOI: 10.1109/HICSS.2014.231.

HERRON, James . Machine Learning: The Ultimate Guide for Beginners to Programming and Deep Learning With Python. . Edição do Kindle

JINDAL, N.; LIU, B. Identifying comparative sentences in text documents. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A, v. 379, n. 2191, p. 20200202, 2016.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd ed. Draft, 2023. Disponível em: https://web.stanford.edu/~jurafsky/slp3. Acesso em: 25 maio 2025.

LAURIOLA, I.; LAVELLI, A.; AIOLLI, F. An introduction to deep learning in natural language processing: Models, techniques, and tools. Neurocomputing, v. 470, p. 443–456, 2022.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

MARTÍNEZ, A. M.; KAK, A. C. PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 23, n. 2, p. 228–233, 2001.

MIKOLOV, T; SUTSKEVER, I; CHEN, K; CORRADO, G. S.; DEAN, J. Distributed Representations of Words and Phrases and their Compositionality. In: Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS 2013). Lake Tahoe, NV: Curran Associates, 2013. p. 3111–3119. Disponível em: https://proceedings.neurips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf. Acesso em: 25 ago. 2025.

MITCHELL, T. M. Machine Learning. New York: McGraw-Hill, 1997.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011.

PYMUPDF. PyMuPDF Documentation. Disponível em: https://pymupdf.readthedocs.io/. Acesso em: 19 maio 2025.

REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084, 2019. Disponível em: https://arxiv.org/abs/1908.10084. Acesso em: 17 maio 2025.

RÖDER, M.; BOTH, A.; HINNEBURG, A. Exploring the Space of Topic Coherence Measures. In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM 2015), Shanghai, China. New York: ACM, 2015. p. 399–408. DOI: 10.1145/2684822.2685324.

ROUSSEEUW, P. J. *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, v. 20, p. 53–65, 1987.

SCHÖLKOPF, B.; SMOLA, A.; MÜLLER, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, v. 10, n. 5, p. 1299–1319, 1998.

SIEVERT, Carson; SHIRLEY, Kenneth. LDAvis: A method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore: Association for Computational Linguistics, 2014. p. 63–70. Disponível em: https://aclanthology.org/W14-3110/. Acesso em: 30 jun. 2025.

SILVA, D. P.; RIBEIRO, F. G. Visualização de dados textuais com nuvem de palavras: uma proposta metodológica para análise exploratória. Revista Brasileira de Biblioteconomia e Documentação, São Paulo, v. 16, n. 1, p. 177–196, 2020. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/1364. Acesso em: 4 jun. 2025.

STEYVERS, M.; GRIFFITHS, T. Probabilistic Topic Models. In: LANDAUER, T. K.; MCNAMARA, D. S.; DENNIS, S.; KINTSCH, W. (org.). Handbook of Latent Semantic Analysis. New York: Psychology Press, 2007. DOI: 10.4324/9780203936399-29. Disponível em: https://doi.org/10.4324/9780203936399-29. Acesso em: 25 ago. 2025.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Brazilian Conference on Intelligent Systems (BRACIS), 2020. Disponível em: https://arxiv.org/abs/2009.11553. Acesso em: 04 jun. 2025.

SUYAL, M.; SHARMA, S. A review on analysis of K-Means clustering machine learning algorithm based on unsupervised learning. Journal of Artificial Intelligence and Systems, v. 6, p. 85–95, 2024. Disponível em: https://www.researchgate.net/publication/379878557_A_Review_on_Analysis_of_K-Means_Clustering_Machine_Learning_Algorithm_based_on_Unsupervised_Learning. Acesso em: 4 jun. 2025.

THEOBALD, Oliver. Machine Learning For Absolute Beginners: A Plain English Introduction (Second Edition) (Machine Learning From Scratch Book 1). Scatterplot Press. Edição do Kindle.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL (UFRGS). Manual de trabalhos acadêmicos: normas da ABNT. Porto Alegre: UFRGS, 2009.

VAN DER MAATEN, L.; HINTON, G. Visualizing data using t-SNE. Journal of Machine Learning Research, v. 9, p. 2579–2605, 2008.

VERZELLA, F. et al. Dimensionality reduction techniques in natural language processing: A comparative study. Expert Systems with Applications, v. 183, p. 115376, 2021.

VIÉGAS, Fernanda B.; WATTENBERG, Martin; FEINBERG, Jonathan. Participatory visualization with Wordle. IEEE Transactions on Visualization and Computer Graphics, v. 15, n. 6, p. 1137–1144, 2009. DOI: 10.1109/TVCG.2009.171.

WATTENBERG, M.; VIEGAS, F.; JOHNSON, I. How to use t-SNE effectively. Distill, 2016.

XIA, Jiazhi et al. Revisiting Dimensionality Reduction Techniques for Visual Cluster Analysis: An Empirical Study. 2021. Disponível em: https://arxiv.org/abs/2110.02894. Acesso em: 4 jun. 2025.