UNIVERSIDADE DO VALE DO RIO DOS SINOS – UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA
NÍVEL DOUTORADO

EDILSON GUEDES DE ALMEIDA

EXPLORANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA EM TEXTOS LEGAIS ANTITRUSTE

PORTO ALEGRE - RS
2024

Edilson Guedes de Almeida

Explorando Algoritmos de Aprendizado de Máquina em Textos Legais Antitruste

Tese apresentada como requisito parcial para obtenção do título de Doutor em Economia, pelo Programa de Pós-Graduação em Economia da Universidade do Vale do Rio dos Sinos (UNISINOS).

Orientador: Dr. Magnus dos Reis

Coorientador: Dr. Rafael Kunst

Porto Alegre - RS

2024

A447e Almeida, Edilson Guedes de.
Explorando algoritmos de aprendizado de máquina em textos legais antitruste / por Edilson Guedes de Almeida. – Porto Alegre, 2024.

153 f. : il. (algumas color.) ; 30 cm.

Tese (doutorado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Economia, Porto Alegre, RS, 2024.
Orientação: Prof. Dr. Magnus dos Reis; Coorientação: Prof. Dr. Rafael Kunst, Escola de Gestão e Negócios.

1.Economia – Processamento de dados. 2.Algoritmos. 3.Processamento de linguagem natural (Computação). 4.Aprendizado do computador. 5.Conselho Administrativo de Defesa Econômica (Brasil). 6.Inteligência artificial. 7.Direito antitruste – Brasil. I.Reis, Magnus dos. II.Kunst, Rafael. III.Título.

CDU 330:004
330:004.421
347.733(81):004.421

Catalogação na publicação:
Bibliotecária Carla Maria Goulart de Moraes – CRB 10/1252

Edilson Guedes de Almeida

Explorando Algoritmos de Aprendizado de Máquina em Textos Legais Antitruste

Tese apresentada como requisito parcial para obtenção do título de Doutor em Economia, pelo Programa de Pós-Graduação em Economia da Universidade do Vale do Rio dos Sinos (UNISINOS).

Aprovado em 6 de março de 2024.

BANCA EXAMINADORA:

Prof. Dr. Magnus dos Reis
Orientador

Prof. Dr. Rafael Kunst
Coorientador

Prof. Dr. Alexsandro Marian Carvalho
Avaliador Externo

Prof. Dr. Marcos Tadeu Caputi Lélis
PPGEcon UNISINOS

Prof. Dr. Tiago Wickstrom Alves
PPGEcon UNISINOS

Visto e permitida a impressão
Porto Alegre - RS, em 6 de março de 2024.

Este trabalho é dedicado à criança, hoje um cinquentenário, que um dia ousou sonhar em se tornar professor.

AGRADECIMENTOS

Em memória amorosa de minha querida mãe, meu pai e meu irmão Antonio Guedes dos Santos (Toinho), que iniciaram precocemente sua jornada ao encontro do Pai Celestial. Sinto a presença deles a cada momento, enchendo meu coração de gratidão e amor eterno.

Às minhas irmãs Luísa (Lu), Cristina (Tina), Edna (Branca), Enilda (com seus "cachinhos de anjo"), ao Fernando, ao Gleibson (Kero) e a todos os meus sobrinhos e sobrinhos-netos. Meu amor por vocês é infinito e eterno. Aos meus filhos, Caíque e Sofia Maria, que são meu orgulho e constantemente me lembram da importância de viver com integridade, dignidade e justiça, valores que espero deixar como legado. Oro para que o Todo-Poderoso me conceda muitas primaveras ao lado de vocês. À minha esposa, Rosangela Maria – "Rosa Angelorum" –, amiga e companheira em todos os momentos da vida. O que seria de mim sem você? Que valor teria o sucesso se não pudesse compartilhá-lo contigo? Como pedir mais ao Senhor dos Senhores quando Ele já me abençoou com seu sorriso, seu apoio e sua paz? Meu amor por você foi, é e sempre será eterno.

Ao casal Adonias e Eliones, a sua filha Larissa, bem como a toda família tocantinense Araújo Almeida, que nos aceitou como iguais e nos honrou com a mais bela, inteligente e estimada nora, a Dra. Ana Allen.

Ainda, ao meu Grande Amigo Fábio Luiz Morais Reis, pessoa reta e extremamente legalista, cujos momentos de conversa sublimava meu ser em razão de sua sabedoria, história de vida e senso de justiça. Um dia quero me tornar um "fábio" tão "sábio" quanto e quiçá possa ofertar também um pouco de "maná intelectual" a outras pessoas. Abraços fraternos.

Aos amigos e colegas Oficiais PMTO e PMMG da turma de Aspirantes APMMG 2001, com quem compartilhei o banco acadêmico e ombreei a mesma farda em tempos outros. Sou forte, porque vocês são fortes; a minha honra se reflete na sua. Grato por me acolherem como um dos seus.

Aos meus colegas e alunos do IFTO. Aos meus amigos Vinícius Braga Rodrigues Duarte (ex GRH) e ao estimado Prof. Dr. Joseane Granja Júnior, que me orientaram nas tarefas administrativas e me apoiaram durante a jornada do doutorado, meu sincero agradecimento. A todos colaboradores do PPGEcon da Unisinos, estendo minha profunda gratidão por seu serviço inestimável.

Por último, mas não menos importante, minha gratidão e apreço aos meus orientadores doutorais: Dr. Magnus dos Reis, Dr. Rafael Kunst, Dra. Luciana Costa Andrade e Dr. Guilherme Stein. Houve momentos de dúvida e medo durante essa jornada desafiadora, mas foi através da competência, sabedoria e humildade de vocês que encontrei esperança renovada e forças para completar esta significativa fase da minha vida acadêmica. Muito obrigado e até breve!

“Os que se encantam com a prática sem a ciência
são como os timoneiros que entram no navio sem timão nem bússola,
nunca tendo certeza do seu destino."
(Leonardo da Vinci)

RESUMO

Esta pesquisa explora a integração de algoritmos de aprendizado de máquina e PLN na análise antitruste do CADE no Brasil, utilizando a modelagem de tópicos para quantificar como a prevalência de tópicos pode auxiliar na previsão de decisões em casos de cartéis. O foco é identificar a técnica mais eficiente para examinar textos jurídicos do CADE, concentrando-se no entendimento do processo decisório e na avaliação de algoritmos relevantes, incluindo a investigação de hipóteses como a superioridade do modelo BERT, particularmente através do BERTopic, em identificar tópicos em textos legais antitruste. A metodologia abrange a coleta e análise de dados processuais e biográficos das autoridades do CADE, empregando várias ferramentas de modelagem, como NMF, LDA, CTM, Top2Vec e BERTopic, e métricas como NPMI, UMass Coherence, diversidade de tópicos e tempo de processamento, levando em conta considerações éticas. Os resultados mostram que o modelo BERTopic, especialmente nas configurações BERTimbau e DistilUSE, é notável em coerência, diversidade temática e eficiência temporal, tornando-se uma opção promissora para análises no contexto do CADE; a pesquisa enfatiza a importância da seleção criteriosa de modelos de PLN, variando desde o LDA, ideal para alta coerência e eficiência, até modelos baseados em *embeddings*, mais adequados para diversidade temática, e destaca as limitações encontradas, como os valores negativos de NPMI, sugerindo a necessidade de aperfeiçoamento na coerência dos tópicos e na precisão das configurações dos modelos. Além disso, a pesquisa explora o desempenho variado de diferentes técnicas de modelagem de tópicos e a inter-relação entre a sofisticação das técnicas e a necessidade de recursos computacionais, destacando a relevância dessas abordagens para as áreas da Economia e do Direito e sublinhando o valor da aplicação de métodos computacionais avançados nestes campos. Ao concluir, a tese ressalta a importância do pré-processamento de dados e do equilíbrio entre as técnicas de PLN e a disponibilidade de recursos computacionais, confirmando a eficácia do BERTopic na modelagem de tópicos em contextos jurídicos, apesar da necessidade de ajustes na coerência e nas configurações; e finalmente sugere a necessidade de futuras investigações para aprimorar as técnicas de PLN e modelagem de tópicos, visando ampliar sua aplicabilidade e relevância.

Palavras-chaves: Modelagem de Tópicos. Processamento de Linguagem Natural (PLN). Análise Antitruste. Conselho Administrativo de Defesa da Concorrência (CADE). Aprendizado de Máquina.

ABSTRACT

This research explores the integration of machine learning algorithms and NLP in antitrust analysis by CADE in Brazil, utilizing topic modeling to quantify how the prevalence of topics can assist in predicting decisions in cartel cases. The focus is on identifying the most efficient technique for examining legal texts from CADE, concentrating on understanding the decision-making process and evaluating relevant algorithms, including investigating hypotheses such as the potential superiority of the BERT model, particularly through BERTopic, in identifying topics in antitrust legal texts. The methodology encompasses the collection and analysis of procedural and biographical data from CADE authorities, employing various modeling tools like NMF, LDA, CTM, Top2Vec, and BERTopic, as well as metrics like NPMI, UMass Coherence, topic diversity, and processing time, taking ethical considerations into account. The results show that the BERTopic model, especially in BERTimbau and DistilUSE configurations, is notable in coherence, thematic diversity, and temporal efficiency, becoming a promising option for analyses in the context of CADE; the research emphasizes the importance of a careful selection of NLP models, ranging from LDA, ideal for high coherence and efficiency, to embedding-based models, more suitable for thematic diversity, and highlights limitations encountered, such as negative NPMI values, suggesting a need for improvement in topic coherence and precision of model settings. Moreover, the research explores the varied performance of different topic modeling techniques and the interplay between the sophistication of the techniques and the need for computational resources, highlighting the relevance of these approaches for the fields of Economics and Law and underscoring the value of applying advanced computational methods in these fields. In conclusion, the thesis emphasizes the importance of data preprocessing and the balance between NLP techniques and the availability of computational resources, confirming the effectiveness of BERTopic in topic modeling in legal contexts, despite the need for adjustments in coherence and configurations; and finally suggests the need for future investigations to enhance NLP and topic modeling techniques, aiming to expand their applicability and relevance.

Key-words: Topic Modeling. Natural Language Processing (NLP). Antitrust Analysis. Administrative Council for Economic Defense (CADE). Machine Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE SIGLAS

AED Análise Econômica do Direito

AM Aprendizado de Máquina

AI Artificial Intelligence

API Application Programming Interface

BERT *Bidirectional Encoder Representations for Transformers*

BoW *Bag of Words*

CADE Conselho Administrativo de Defesa Econômica

CNJ Conselho Nacional de Justiça

docvars *document variables*

DEE Departamento de Estudos Econômicos

DTM *Document-Term Matrix*

DFM *Document-Features Matrix*

ETM *Embedded Topic Model*

FTC *Federal Trade Comission*

GLM Modelo Linear Generalizado

IA Inteligência Artificial

KWIC *Keywords in context,*

LDA *Latent Dirichlet Allocation*

LSA *Latent Semantic Analysis*

ML *Machine Learning*

MPF Ministério Público Federal

NLP *Natural Language Processing*

p., pág. página

pp., págs. páginas

PL Projeto de lei

PLN Processamento de Linguagem Natural

PLSA *Probabilistic Latent Semantic Analysis*

PO Pesquisa Operacional

ProCADE Procuradoria Federal Especializada

SBDC Sistema Brasileiro de Defesa da Concorrência

SDE Secretaria de Direito Econômico

SEAE Secretaria de Acompanhamento Econômico

SG Superintendência-Geral do CADE

STM *Structural Topic Model*

TADE Tribunal Administrativo de Defesa Econômica

TF-IDF *Term Frequency-Inverse Document Frequency*

TM *Topic Model*

LISTA DE SÍMBOLOS

α Letra grega minúscula *alfa*

β Letra grega minúscula *beta*

ϵ Letra grega maiúscula *epsilon*

χ Letra grega minúscula *khi*

ϕ Letra grega minúscula *phi*

σ Letra grega maiúscula *sigma*

θ Letra grega minúscula *teta*

ζ Letra grega minúscula *zeta*

SUMÁRIO

1 INTRODUÇÃO

A evolução tecnológica, particularmente no que se refere aos algoritmos de aprendizado de máquina (AM) e Processamento de Linguagem Natural (PLN), tem aberto novas avenidas para a otimização do processo analítico, especialmente em contextos jurídico-regulatórios como o brasileiro. Esta evolução é crucial no campo da Organização Industrial, onde o Conselho Administrativo de Defesa da Concorrência (CADE) desempenha um papel vital na instrumentalização dos julgamentos antitruste. Neste cenário, a modelagem de tópicos, uma técnica avançada de PLN, surge como uma ferramenta transformadora, prometendo *insights* profundos na análise de documentos legais e facilitando a tomada de decisão em casos complexos de cartéis.

No contexto atual, com a crescente acumulação de dados em linguagem natural, a importância das ferramentas de análise e interpretação desses dados torna-se cada vez mais evidente. A vastidão de informações, exemplificada pelo índice colossal do Google, desafia os limites da compreensão humana, impulsionando o desenvolvimento de métodos que possam eficientemente ler, interpretar e sintetizar essas informações.

A análise textual, dentro deste universo de dados, tem se mostrado uma ferramenta inestimável, não apenas na Economia Digital, mas também na análise de decisões antitruste. Textos legais e pareceres de autoridades, como os do CADE, são fontes ricas para o entendimento de dinâmicas de mercado e tomadas de decisões estratégicas. Esta abordagem é reforçada pela pesquisa de Gentzkow, Kelly e Taddy (2019), que destacam a aplicabilidade da análise de texto em previsões econômicas e políticas.

Com o avanço da inteligência artificial (IA) e do AM, a análise econômica ganhou uma nova dimensão, especialmente na interpretação de decisões sequenciais e seus impactos. A contribuição de Herbert Simon neste campo ressalta a importância de entender essas sequências de ações, uma perspectiva que se alinha perfeitamente com o contexto dos julgamentos do CADE (Russell; Norvig, 2021).

A virada do século marcou um renovado interesse na aplicação de teorias de decisão na IA, estimulado por avanços no poder computacional e pela emergência do *big data*. Neste cenário, a interação entre economia e IA se tornou mais profunda, com a análise de grandes volumes de dados textuais assumindo um papel central.

O PLN, como um campo fundamental na transformação de texto em dados numéricos, possibilita a interpretação dessas informações por algoritmos de AM. Métodos como tokenização, *stemming* e lematização, embora essenciais, envolvem simplificações que podem impactar a interpretação dos resultados, como notado por Gentzkow, Kelly e Taddy (2019).

A economia digital, com sua transformação de atividades econômicas em formatos digitais, tem sido um motor de crescimento econômico. A expansão deste setor e a conse-

quente geração de grandes volumes de dados têm implicações diretas para a economia e a IA, especialmente no que diz respeito à otimização de processos analíticos.

Dessa forma, a intersecção da IA com o direito, conhecida como Legal AI, apresenta um potencial transformador, especialmente no campo da análise antitruste. A utilização de algoritmos de AM e técnicas de PLN, como a modelagem de tópicos, pode agilizar significativamente a administração e a tomada de decisões judiciais. Contudo, é essencial considerar as questões éticas e jurídicas[1], além da transparência e compreensão das decisões algorítmicas, para assegurar a equidade no processo administrativo antitruste. Este cenário abre caminho para uma investigação profunda sobre a aplicação de técnicas avançadas de PLN na análise de documentos do CADE, visando identificar o modelo mais eficaz para capturar e quantificar a influência dos pareceres do MPF sobre as decisões dos relatores, uma análise que se alinha com os princípios regulação antitruste brasileira e oferece contribuições significativas para a Organização Industrial.

1.1 Problema de pesquisa

Neste estudo, o problema de pesquisa é formulado para destacar a análise quantitativa na modelagem de tópicos, bem como a utilização posterior das probabilidades de tópicos prevalente mediante aplicação de um modelo estatístico apropriado, visando trazer a lume suas implicações nos julgamentos de cartéis pelo CADE. As questões de pesquisa propostas são:

a) Qual técnica de modelagem de tópicos, avaliada através de métricas quantitativas, é mais eficaz para analisar um conjunto de dados de documentos de texto do CADE relacionados a julgamentos de cartéis?

b) Utilizando as probabilidades dos tópicos prevalentes obtidas na modelo de modelagem de tópicos, é possível quantificar a influência dos pareceres do MPF e dos votos dos conselheiros nas decisões de mérito em casos de práticas colusivas, aplicando análise estatística?

As etapas críticas desta pesquisa são:

a) **Seleção do modelo de modelagem de tópicos:** a primeira fase concentra-se em identificar a técnica mais eficiente para a modelagem de tópicos no corpus do CADE, empregando uma abordagem quantitativa. Serão avaliados modelos de PLN, como o BERTopic, utilizando métricas específicas para assegurar a precisão na identificação de tópicos relevantes.

b) **Análise quantitativa de influência e padrões:** após a seleção do modelo de tópico mais apropriado, a pesquisa progride para uma análise estatística detalhada. Esta fase

[1] O adensamento de informações e necessário debate das questões ético-jurídicas atinentes ao emprego de AM e PLN no antitruste não faz parte do escopo desta pesquisa.

investiga, de maneira quantitativa mediante aplicação de um modelo empírico,o impacto dos pareceres do MPF e dos votos dos conselheiros sobre as decisões dos relatores em casos de condutas colusivas.

1.2 Objetivos

1.2.1 Objetivo geral

O objetivo geral desta pesquisa é realizar uma investigação quantitativa abrangente sobre a modelagem de tópicos aplicada na análise de documentos textuais do CADE. O foco será na identificação do(s) modelo(s) mais eficiente(s) para tal análise e na avaliação quantitativa do impacto dos pareceres do MPF e dos votos dos conselheiros relatores nas decisões sobre condutas colusivas, baseando-se em princípios da Teoria da Organização Industrial.

1.2.2 Objetivos específicos

Os objetivos específicos, que direcionam o alcance do objetivo geral, são:

a) Determinar o modelo de tópicos mais eficaz: identificar a técnica de modelagem de tópicos, incluindo abordagens como BERT e *embeddings* da OpenAI, que oferece maior precisão e eficiência, baseada em métricas quantitativas, para a análise de documentos do CADE em casos de cartéis;

b) Quantificar a influência dos pareceres e dos votos dos conselheiros relatores: utilizar o modelo selecionado para mensurar, de forma quantitativa, o grau de influência dos pareceres do MPF e votos dos relatores nas decisões sobre práticas colusivas;

c) Previsão de decisões futuras: avaliar a capacidade do modelo escolhido para prever futuras decisões dos relatores com base dados textuais de pareceres do MPF e votos dos conselheiros relatores;

d) Integração com conceitos de Organização Industrial: correlacionar os resultados obtidos com conceitos fundamentais da Organização Industrial, especialmente relacionados à regulação antitruste e ao comportamento racional dos agentes econômicos.

1.3 Justificativa

Esta tese é justificada pela crescente intersecção entre Economia e IA, com foco especial na economia digital e no uso de *big data* para análises quantitativas no ambiente econômico. A evolução dos algoritmos de Aprendizado de Máquina (AM), utilizando técnicas de PLN, proporciona uma abordagem quantitativa única para otimizar análises econômicas e jurídicas, especialmente no contexto do CADE no Brasil.

O aumento exponencial de dados em linguagem natural na era digital ressalta a necessidade de ferramentas analíticas quantitativas para interpretar esses grandes volumes de dados. Este cenário é reforçado pela expansão da economia digital e sua influência quantitativa no crescimento econômico.

O avanço da IA e do AM, apoiado pela pesquisa operacional e pelo conceito de agentes racionais, tem sido fundamental na análise econômica, evidenciando uma sinergia entre economia e IA. Esta convergência é caracterizada por uma abordagem quantitativa robusta, onde a análise de grandes volumes de dados textuais torna-se essencial.

No âmbito jurídico, a aplicação de algoritmos de AM, conhecida como Legal AI, tem potencial para revolucionar a análise de documentos legais. Especificamente no CADE, a modelagem de tópicos e técnicas de PLN são utilizadas para identificar e organizar temas prevalentes em documentos legais de maneira quantitativa, auxiliando na decisão de casos antitruste.

Esta tese visa investigar, através de uma abordagem quantitativa, como a modelagem de tópicos pode identificar padrões consistentes que refletem o comportamento do órgão regulador. A pesquisa focará na influência quantificável dos pareceres do MPF e dos votos dos conselheiros nas decisões sobre cartéis, integrando princípios da Organização Industrial. A aplicação de técnicas avançadas de PLN e AM, com um enfoque quantitativo, representa um avanço significativo na compreensão da análise legal e econômica no contexto dos julgamentos de cartéis pelo CADE.

1.4 Delimitações e estruturação do trabalho

Este trabalho está delimitado à análise quantitativa de documentos de julgamentos de cartéis de licitação pelo CADE, empregando técnicas de modelagem de tópicos e algoritmos de AM. Será analisada uma amostra quantificada de documentos emitidos entre janeiro de 2015 e julho de 2022, focando em casos de condutas anticompetitivas em licitações no Brasil. Os documentos incluem os emitidos pelo MPF junto ao CADE e os votos dos conselheiros relatores.

O estudo explorará como a modelagem de tópicos, apoiada por análises estatísticas, pode revelar padrões consistentes nos documentos, quantificando a influência dos pareceres do MPF e dos votos dos relatores nas decisões. Esta abordagem está alinhada com os princípios da regulação antitruste e da Organização Industrial, ampliando a análise para além das técnicas tradicionais.

A estrutura do trabalho é a seguinte:

- **Capítulo 2: Revisão da Literatura**, abordando conceitos básicos de modelagem de tópicos, transformação de texto em dados quantitativos, uso de PLN e AM em textos legais, aplicação da Teoria Econômica e análise antitruste;

- **Capítulo 3: Metodologia**, descrevendo a seleção do modelo de modelagem de tópicos e a quantificação da influência dos pareceres do MPF e votos dos relatores;
- **Capítulo 4: Resultados e Discussão**, apresentando os resultados alcançados pela modelagem de tópicos e da aplicação do modelo empírico, discutindo hipóteses com enfoque na intersecção entre Organização Industrial e análise antitruste;
- **Capítulo 5: Conclusões**, apresentando implicações práticas e teóricas, e sugestões para futuras pesquisas.

2 REVISÃO DA LITERATURA

Este capítulo de Revisão da Literatura da tese explora a intersecção entre a Economia e a IA, com ênfase particular na Economia Digital e na integração de grandes conjuntos de dados (*big data*) no contexto econômico. A relevância do aprendizado de máquina e da Inteligência Artificial aplicada ao Direito (Legal AI) é destacada, sublinhando como essas tecnologias estão revolucionando a análise de dados e a tomada de decisões no setor público antitruste brasileiro.

A transformação de texto em dados numéricos é um elemento crucial desta investigação, e o capítulo dedica uma atenção especial às técnicas de PLN, abordando desde a distribuição de palavras e leis empíricas até a tokenização e a conflação textual. Além disso, a modelagem de tópicos, que é o foco central da tese, é examinada em detalhes, cobrindo metodologias como *Latent Semantic Analysis*, *Non-negative Matrix Factorization*, *Probabilistic Latent Semantic Analysis*, *Latent Dirichlet Allocation*, entre outras.

O capítulo também discute as aplicações práticas da IA no setor público antitruste brasileiro, enfatizando a concorrência e práticas colusivas, com um olhar aprofundado sobre a conduta de cartéis, especialmente em licitações, e o papel do CADE neste contexto. A incorporação de novas tecnologias pelo órgão antitruste brasileiro é também um tópico crucial, alinhando-se com as atuais tendências digitais. Finalmente, o capítulo sintetiza o estado da arte e formula a tese, integrando as descobertas e as metodologias estudadas com a questão de pesquisa principal da tese.

2.1 Da Economia à Inteligência Artificial

São numerosos os exemplos da utilização da análise textual em modelagens econômicas. Gentzkow, Kelly e Taddy (2019, p. 535) afirma que em finanças, conhecimento extraído de notícias financeiras, mídias sociais e registros de empresas são utilizadas para previsão de movimentos de preços de ativos, bem como na investigação do impacto causal de novas informações no mercado. Esses autores, ainda, informam que em macroeconomia, o texto é usado para prever a variação da inflação e do desemprego, além de estimar os efeitos da incerteza política. Esclareça-se que na citada Economia Digital, o texto das notícias e das mídias sociais é usado para estudar os fatores e os efeitos da inclinação política, com aplicações na Economia Política, onde os discursos políticos são utilizados para estudar dinâmicas de agendas e o debate de políticas públicas. Por fim, no âmbito da Organização Industrial, o texto de anúncios e análises de produtos é usado para estudar os impulsionadores da tomada de decisão do consumidor Gentzkow, Kelly e Taddy (2019, p. 535).

No debate econômico, Russell e Norvig (2021, p. 19) lançam algumas indagações que se mostram pertinentes ao desenvolvimento desta pesquisa, por exemplo, como devemos tomar

decisões de acordo com nossas preferências? Como devemos fazer isso quando outros podem não concordar? Como devemos fazer isso quando a recompensa pode estar longe no futuro? Para responder indagações como essa, necessário realizar um breve escorço aos primórdios da ciência econômica. Russell e Norvig (2021, p. 19–63) realizam uma síntese seminal desde a publicação da obra "Uma investigação sobre sobre a natureza e as causas da riqueza das nações", por Adam Smith, em 1776[2], que se propôs analisar as economias a existência de diversos agentes individuais que atendem a seus próprios interesses, os economistas, com exceções, não abordaram a terceira questão listada acima: como tomar decisões racionais quando os retornos das ações não são imediatos, mas resultam de várias ações tomadas em sequência.

O advento da Segunda Grande Guerra deslocou o debate para o campo da pesquisa operacional (PO), a partir dos esforços na Grã-Bretanha para otimizar as instalações de radar e, posteriormente, de encontrar inúmeras aplicações civis. O trabalho de Richard Bellman (1957) formalizou uma classe de problemas de decisão sequenciais chamados processos de decisão de Markov, que que baseia o campo do "aprendizado por reforço"[3] que é uma área de estudo dentro do Aprendizado de Máquina (Russell; Norvig, 2021, p. 19–63).

Os esforços da ciência econômica e da pesquisa operacional (PO) foram fundamentais na formação da concepção moderna de agentes racionais. No entanto, a IA se desenvolveu inicialmente em um caminho distinto, em parte devido à complexidade envolvida na tomada de decisões racionais, como apontado por Russell e Norvig (2021, p. 19–63). Herbert Simon, um pioneiro da IA, ganhou o Prêmio Nobel de Economia em 1978 por demonstrar que modelos baseados em *satisficing*[4] representam mais precisamente o comportamento humano real, uma conceituação detalhada em Russell e Norvig (2021, p. 21).

Na virada do século, houve um renovado interesse na aplicação de teorias de decisão na IA. Avanços significativos no poder de computação e o advento da internet proporcionaram um volume sem precedentes de dados, conhecido como *big data*, que inclui trilhões de palavras de texto, bilhões de imagens e horas de áudio e vídeo, dados genômicos, de rastreamento de veículos, *clickstream*[5] e de redes sociais, conforme descrito por Russell e Norvig (2021, p. 19–63). Essa riqueza de informações tem permitido explorações mais profundas e complexas na interseção entre economia e IA, marcando uma era de integração e inovação entre estas disciplinas.

---

[2] Livre tradução para *"An Inquiry into the Nature and Causes of the Wealth of Nations"*.

[3] Russell e Norvig (2021, p. 1176) esclarecem que na aprendizagem por reforço o agente aprende a partir de uma série de reforços: recompensas e punições. Por exemplo, no final de um jogo de xadrez, o agente é informado de que ganhou (uma recompensa) ou perdeu (uma punição). Cabe ao agente decidir quais das ações anteriores ao reforço foram as maiores responsáveis por ele, e alterar suas ações visando mais recompensas no futuro. Os autores dedicam todo o capítulo 22 de sua obra para discutir esse tipo de aprendizado de máquina.

[4] Russell e Norvig (2021, p. 21) conceitua *satisficing* como sendo o ato de tomar decisões que são "boas o suficiente", em vez de calcular laboriosamente uma decisão ótima

[5] *Clickstream* é o registro dos caminhos seguidos pelos usuários ao clicar em *links on-line*, fornecendo insights valiosos sobre o comportamento do usuário, que ajudam empresas a otimizar experiências digitais, melhorar taxas de conversão e aumentar a receita.

Pelo exposto, verifica-se que a conexão entre economia e IA é profunda e multifacetada. Desde os primórdios da economia com Adam Smith, até os avanços contemporâneos em inteligência artificial, a busca por entender e otimizar a tomada de decisões tem sido um tema central. A evolução da PO e os desenvolvimentos na área de AM ressaltam a importância de sequências de decisões e suas consequências a longo prazo. A contribuição de Herbert Simon e a incorporação de conceitos econômicos na IA sublinham essa intersecção (Russell; Norvig, 2021). Atualmente, o uso extensivo de *big data*, advindo de múltiplas fontes, ilustra como a economia e a IA continuam a se entrelaçar, influenciando tanto a teoria quanto a prática em ambas as áreas.

Mas afinal, o que é inteligência artificial? é o mesmo que aprendizado de máquina? Para dirimir essas dúvidas e direcionar o debate, necessário realizar uma pequena imersão nos termos utilizados nesta pesquisa.

Em conceito amplo, pode-se afirmar que as técnicas de aprendizado de máquina (AM)[6] são algoritmos utilizados em computação para realizar tarefas e extrair conhecimento subjacente aos dados. Trata-se de um ramo da inteligência artificial (IA)[7], esta mais abrangente e incorpora campos de visão computacional, PLN, robótica e outros (Veras; Barreto, 2022, p. 4; Ferreira, 2018, p. 17; Brink; Richards; Fetherolf, 2017, p. 4–5). Outro conceito importante é o *Deep Learning*, uma ramificação do aprendizado de máquina que emprega algoritmos avançados, redes neurais profundas e treinamento intensivo com dados (Veras; Barreto, 2022, p. 4). Isso permite aprendizado, classificação e tomada de decisões, modelando o comportamento do cérebro humano de forma inspiradora. Para ter uma visão de como esses campos estão interconectados, vide a Figura 1 (p. 9).

PLN é o campo de *design* de métodos e/ou técnicas[8] que recebem como entrada, ou produzem como saída, dados de linguagem natural não estruturados, definindo a área própria de pesquisa e aplicações representadas por modelos estatísticos e algoritmos responsáveis pela análise e representação da linguagem natural, tanto fonética quanto escrita (Polo; Ciochetti; Bertolo, 2021, p. 1–2; Goldberg, 2017, p. 1). Assim, apreende-se que os métodos e técnicas de PLN são utilizados na transformação de texto em uma representação numérica visando a utilização em algum modelo. São exemplos de métodos de PLN os processos de tokenização, de *stemming*, de lematização, de *word embeddings* e de retirada das *stopwords*.

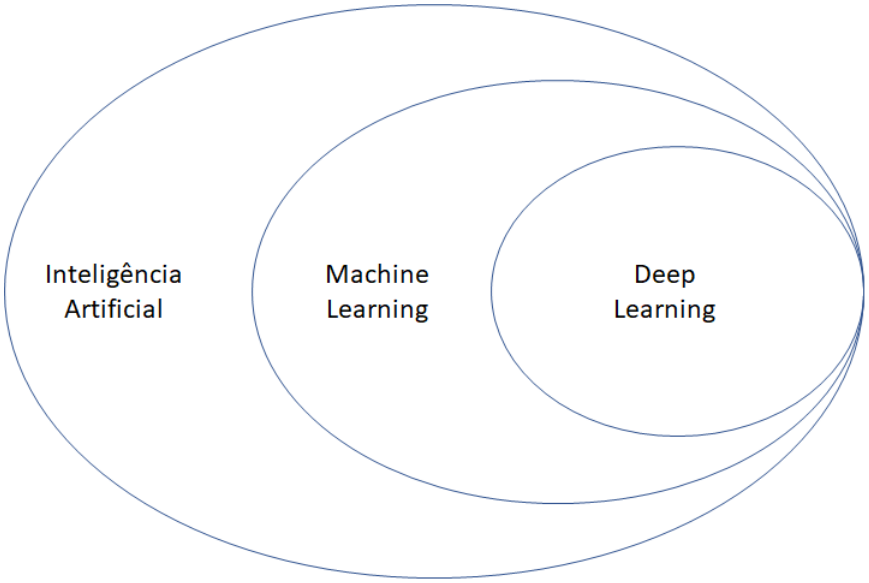
Com o volume crescente de informações textuais, os métodos de AM são convenientes

[6] Idem, *Machine Learning*.

[7] Idem, *Artificial Intelligence*.

[8] Calha esclarecer que, em geral, o termo "método"se refere a um conjunto de diretrizes, procedimentos e estratégias que guiam o desenvolvimento de uma determinada pesquisa ou projeto, visando atingir objetivos preestabelecidos. Já o termo "técnica"pode ser entendido como uma ferramenta, abordagem ou procedimento específico utilizado dentro desse método para a obtenção de resultados. Em resumo, o método é um conjunto mais amplo de orientações, enquanto a técnica é uma ação ou ferramenta específica utilizada dentro desse conjunto (Santos, 2018, p. 13, 16–17, 20). Nesta pesquisa, para fins de fluidez no texto, apesar de metodologicamente não serem sinônimos, as palavras “método”, “técnica”, “algoritmo”, “aplicação” e “ferramenta” serão muitas vezes empregadas como sinônimos próximos.

Figura 1 – Conexões entre IA, AM e *Deep Learning*



Fonte: adaptado de Chollet e Allaire (2018, p. 3).

para investigação da eficiência de modelos. Essa área da IA utiliza da modelagem estatística aliada a algoritmos de classificação e regressão com aplicações em PLN. Como consequência do continuado avanço do fluxo de dados textuais, houve mudança de paradigma no processamento de dados em larga escala, estabelecendo-se o campo das redes neurais, aplicações modernas que utilizam Métodos de aprendizado profundo[9], que requerem mais tempo e recursos, *v.g.* máquinas virtuais de processamento, mas potencializam a utilização PLN (Agerri *et al.*, 2015, p. 37–38). Calhar esclarecer que dados textuais não estruturados são volumosos, contrastando com a escassez de dados estruturados disponíveis para consulta, especialmente em idioma português brasileiro.

Pela amplitude de seu conceito, análise de texto pode ser entendida como um conjunto de técnicas e algoritmos aplicáveis a dados textuais, sendo bastante utilizada para a transformação de dados brutos e não estruturados, mas em linguagem natural, em dados estruturados que podem ser melhor analisados em abordagens quantitativas (Jockers; Thalken, 2020, p. 8–16). Seus métodos incluem a análise descritiva, análise de palavras-chave, análise de tópicos, medição e dimensionamento, agrupamento, comparações de texto e vocabulário ou análise de sentimento. Pode-se, também, incluir análise causal ou modelagem preditiva. Atualmente esses métodos são empregados a modelos de geração de linguagem natural que alimentam a mais nova geração de sistemas de inteligência artificial. Esses métodos serão melhor abordados na seção 2.2 e na subseção 2.2.2, páginas 12 e 21, respectivamente.

O PLN, pois, possibilita a transformação de texto em números visando emprego de recur-

[9] Idem, *Deep Learning*.

sos computacionais. Por conseguinte, torna-se possível a interpretação do dado por algoritmos de aprendizado de máquina. Alguns modelos de aprendizagem de máquina mais utilizados em PLN são os Modelos Lineares Regularizados (GLM), os de Máquinas Vetoriais de Suporte (SVM), o de *Naïve Bayes* e modelos baseados em árvores como *Random Forests*.

Finalmente, esse trabalho parte do conceito de que o texto deve ser organizado em documentos. Daí a necessidade do empréstimo de terminologia utilizada na Linguística, especialmente a ideia que *corpora* é o plural de *corpus*, sendo o último uma coleção de textos organizados em nível de documentos e metadados, que por sua vez são chamados de "variáveis de documentos", cuja abreviação utilizada nessa pesquisa é "docvars"[10]. No nível dos documentos, os dados são organizados em *stems* e *lemmas* – radicais e forma canônica de palavras, respectivamente – utilizados a normalização dos *tokens*.

2.1.1 Economia digital e *big data*

A economia digital é o conjunto de mercados com base em tecnologias digitais que facilitam o comércio de bens e serviços por meio do comércio eletrônico (Cramer; Hayes, 2013, p. 5). Uma interpretação atual possível desse conceito é que a economia digital resulta da transformação de atividades econômicas, produtos e serviços tradicionais em formato digital mediante o emprego da internet e com suporte em por mídias eletrônica diversas. A expansão do setor digital tem sido um dos principais impulsionadores do crescimento econômico nos últimos anos, e a mudança para um mundo digital teve efeitos na sociedade que se estendem muito além do contexto da tecnologia digital (Cramer; Hayes, 2013, p. 5). Observe-se que, em um cenário de cerca de oito bilhões de componentes conectados somente à internet, os dados tornam-se um "novo petróleo", remetendo à economia digital o desenvolvimento de novos modelos de negócios e gerando outra grande quantidade de dados nesse processo (Javornik; Nadoh; Lange, 2019, p. 295–296; Loi; Dehaye, 2017, p. 137). Assim como a denominação "economia digital", cunharam-se outros termos. Entra-se assim na seara do *big data*.

Segundo Javornik, Nadoh e Lange (2019, p. 306) e Japkowicz e Stefanowski (2016, p. 306), o termo *big data* não se refere apenas ao tamanho ou volume massivo. Monteiro (2017, p. 17) afirma que definições variadas para a expressão *big data* podem ser encontradas na doutrina, variando quanto a amplitude, a inclusão ou a restrição conceitual. Tais definições variam de acordo com a utilização destinada ao vernáculo, como por exemplo quando há apropriação do termo em razão da utilização de ferramentas de tecnologia da informação exploradas com determinado fim econômico. Hu *et al.* (2014, p. 652) explicam que o termo *big data* foi cunhado para capturar o significado de uma existência de um grande volume de dados, em grande parte desestruturados, e atrelados a outras características únicas. Os atributos são importantes para a construção de uma definição do *big data*, rotulados a partir de definições atributiva, comparativa e arquitetural (Monteiro, 2017, p. 24; Hu *et al.*, 2014, p. 654). Esse atributos, assim como outros

[10] Idem, *document variables*.

observados pela doutrina, somam-se em uma cadeia de valor derivada da transformação de dados brutos em informação, em conhecimento, tornando-se um ativo valioso para diversos fins e atividades econômicas, baseando melhores tomadas de decisão (Monteiro, 2017, p. 23–24; Japkowicz; Stefanowski, 2016, p. 306–307; Stucke; Grunes, 2015, p. 2–3). A conceituação do *big data* no decorrer desta pesquisa se constrói não somente em razão do volume massivo dos dados, mas também da cadeia de valor geradora de um *output* – informação ou conhecimento útil – a partir da coleta, extração, tratamento e processamento dos dados, estruturados ou não.

2.1.2 Aprendizado de Máquina e *Legal AI*

Wang (2022, p. 3) cita, de forma geral, apenas dos tipos de AM, os modelos supervisionados e os não supervisionados. Já Hsu (2020, p. 1–2) e Veras e Barreto (2022, p. 4) informa que AM consiste em três tipos diferentes: aprendizagem supervisionada, aprendizagem não-supervisionada e aprendizagem semi-supervisionada - uma combinação de métodos de aprendizagem supervisionados e não supervisionados. Entretanto, para essa pesquisa será adotada a divisão estabelecida por Park e Ko (2020, p. 7–12) que divide a aprendizagem em supervisionada, não-supervisionada e por reforço.

A aprendizagem supervisionada envolve o treinamento de um modelo em dados rotulados, onde os pares de entrada e saída são fornecidos. A aprendizagem não-supervisionada envolve o treinamento de um modelo em dados não rotulados, onde o algoritmo aprende padrões e relacionamentos sem receber resultados específicos. O aprendizado por reforço é um tipo de aprendizado de máquina em que um agente aprende a interagir com um ambiente e executa ações para maximizar um sinal de recompensa (Park; Ko, 2020, p. 7–12).

É amplo e incremental o uso de algoritmos de AM nas mais diversas em várias áreas do conhecimento humano, incluindo no campo jurídico. Esses algoritmos possuem a capacidade de analisar grandes volumes de dados e identificar padrões que podem ser úteis na tomada de decisões (Maranhão; Florêncio; Almada, 2021, p. 158; Schrepel, 2021, p. 3; Park; Ko, 2020, p. 1–2). No entanto, é importante ressaltar que a aplicação desses algoritmos no campo jurídico levanta questões éticas e jurídicas que precisam ser cuidadosamente consideradas. Ademais, a transparência e a compreensão da lógica por trás das decisões algorítmicas são fundamentais para garantir a equidade no processo judicial (Schrepel, 2021, p. 11–12, 14–15; Park; Ko, 2020, p. 17–18).

Mas o que é *Legal AI*? Autores têm defendido que a Inteligência Artificial Jurídica, também conhecida como *Legal AI*, é a aplicação de tecnologias avançadas de Inteligência Artificial no campo do direito. Essas tecnologias utilizam algoritmos de aprendizado de máquina, PLN e análise preditiva para automatizar e melhorar diversas tarefas do setor jurídico (Maranhão; Florêncio; Almada, 2021, p. 157, 169; Nitta; Satoh, 2020, p. 471–481).

O Ait Staff Writer (2023) destaca que a Legal AI como aquele onde aplicações de ferramentas baseadas em AI são utilizadas visando aumentar a eficiência, economizando tempo

em tarefas como revisão de documentos e pesquisa jurídica. Além disso, o artigo relata que essas ferramentas podem melhorar a pesquisa jurídica ao analisar grandes volumes de dados e fornecer *insights* relevantes. Por outro lado, o artigo reconhece limitações da AI, como a compreensão contextual limitada de conceitos legais complexos e preocupações éticas relacionadas à privacidade e viés. Também é mencionado que o aumento da dependência da AI pode reduzir a interação humana essencial na representação legal (Ait Staff Writer, 2023).

As vantagens da aplicação dessa abordagem envolve a utilização das tecnologias de IA para tarefas jurídicas, como revisão contratual, pesquisa legal e previsão de resultados com o potencial de agilizar e otimizar processos no campo jurídico, tornando-os mais eficientes e fornecendo resultados mais precisos(Veiga; Załucki, 2022, p. 361; Ulenaers, 2020, p. 15–17). Mas também há óbices a se considerar, posto que é crucial que se estabeleça uma regulamentação adequada para garantir a usabilidade ética e responsável da IA no campo jurídico. A regulamentação deve levar em consideração questões como transparência, prestação de contas e proteção dos direitos e interesses das partes envolvidas. Além disso, a interdisciplinaridade entre profissionais do direito, engenharia e computação se mostra essencial para o desenvolvimento e implementação eficaz de soluções de inteligência artificial no campo jurídico (Maranhão; Florêncio; Almada, 2021, p. 173; Park; Ko, 2020, p. 24).

Em suma, a aplicação de interdisciplinar de algoritmos de AM traz implicações no campo jurídico e introduz a denominada *Legal AI*. Isso se traduz em benefícios significativos tanto na parte administrativa, quanto na tomada de decisões judiciais, tornando mais ágeis, eficientes e precisas as tarefas jurídico-administrativas. Além disso, a utilização da aprendizagem automática em análises jurídicas e jurimétricas pode fornecer informações e previsões valiosas.

Por conseguinte, ao analisar processos e resultados jurídicos anteriores, os modelos de aprendizagem automática podem identificar padrões e tendências, permitindo previsões mais precisas do resultado de processos jurídicos futuros (Ulenaers, 2020, p. 6–7; Aletras *et al.*, 2016, p. 1). Por fim, técnicas de AM, como aprendizado supervisionado, não supervisionado e por reforço, juntamente com a integração do processamento de linguagem natural, têm imenso potencial para transformar o campo jurídico, melhorando a eficiência, a precisão e os processos de tomada de decisão.

2.2 Transformando Texto em Dados

Para transformar o texto em dados, técnicas de análise de texto são aplicadas. Isso inclui dividir o texto em documentos, selecionar elementos relevantes, e convertê-lo em representação numérica (Gentzkow; Kelly; Taddy, 2019). Na visão dos autores (Javornik; Nadoh; Lange, 2019, p. 296–297; Loi; Dehaye, 2017, p. 137–138), “os dados são um novo petróleo”. Por isso mesmo, uma aplicação de PLN, com foco em modelagem de tópicos, em língua portuguesa e na área da Economia Digital parece ser adequada para prospecção e revelação de ideias, subjacentes ou

explícitas em caso de análise textual.

Em sede de transformação do conhecimento, na etapa de pré-processamento, palavras comuns ou incomuns são excluídas, números e pontuação são removidos, e recursos linguísticos específicos são definidos. Após, é realizada alguma tarefa de representação numérica do texto, que pode ser binária (presença/ausência) ou contagem de frequência de palavras. Essa representação numérica, chamada de matriz de *features*, é então usada para análise estatística, como regressão, classificação ou modelagem de tópicos. É importante ressaltar que a transformação do texto em dados envolve simplificações e restrições para lidar com a complexidade do texto. Essas simplificações podem afetar a interpretação e a validade dos resultados, exigindo validação e interpretação cuidadosa dos resultados (Gentzkow; Kelly; Taddy, 2019, p. 555–556; Grimmer; Stewart, 2013, p. 271).

A partir desse ponto em diante, apresentaremos o estado da Arte sobre o processamento (pré e pós) de texto. Na Seção de Metodologia serão apresentados as ferramentas de validação e interpretação dos *outputs* dos modelos de tópicos.

2.2.1 Técnicas de PLN (pré-processamento)

Os passos básicos, ou estágios, para se analisar um texto como dado são descritos por Benoit (2020, p. 478–489). Para esse autor de início há que (*i*) transformar os textos em um objeto *corpus*; (*ii*) converter o texto em um formato eletrônico; (*iii*) selecionar os documentos para etapa de análise; (*iv*) definir e refinar as *features*[11]; (*v*) converter as *features* textuais em uma matriz de dados quantitativos; (*vi*) analisar a matriz de dados usando procedimentos estatísticos apropriados; por último, (*vii*) interpretar e reportar os resultados.

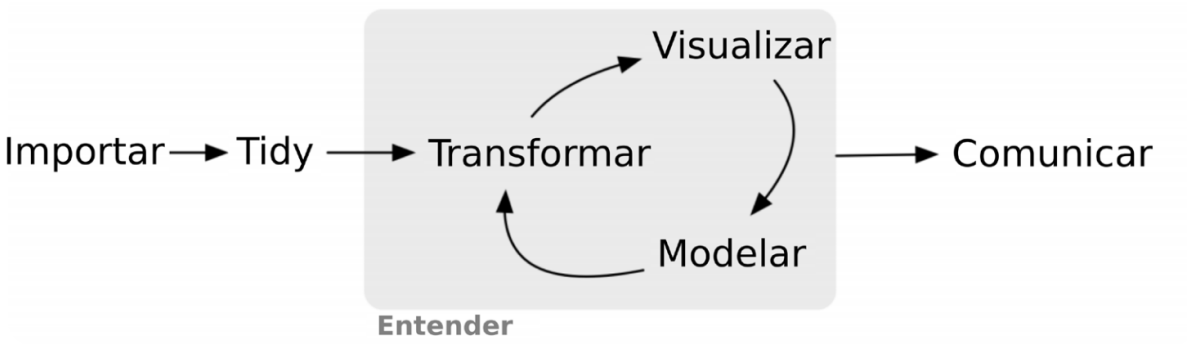
Wickham e Grolemund (2021, p. 1) e Silva (2022, p. 92) também descrevem, as suas maneiras, o fluxo de atividades necessárias à conversão de dados não estruturados – nessa pesquisa, serão os textos constantes dos pareceres e votos das autoridades do CADE –, em dados estruturados. Essa abordagem metodológica é sintetizada no que esses autores denominam como fluxo de trabalho em ciência de dados do *tidyverse* (Figura 2). Para esses autores, na fase de **importação** é realizada a coleta e seleção dos dados. Na etapa da **arrumação** (do inglês, *tidy* na linguagem do *tidyverse*), procede-se a limpeza, seleção e armazenagem dos dados relevantes sob a forma estruturada. Dessa forma, os dados passam de brutos para tratados e podem ser consumidos como recursos nas fases seguintes. Na fase de **transformação** elabora-se o *corpus*[12], cujo plural é *corpora* e pode ser conceituado como um construto de material, seja ele textual, gráfico, áudio e/ou vídeo, sobre o qual se baseia alguma análise (Desagulier, 2017, p. 3). É nessa fase que se agrega o conteúdo textual de todos os documentos, além de eventuais metadados associados ao conteúdo. Determina-se, ainda nessa fase, qual será a unidade de análise (*v.g.*,

[11] São os atributos, caracterizados na análise como variáveis preditoras ou covariáveis.

[12] Após transformações do texto em dado, é possível utilizar o *output* em modelos de regressão e classificação com dados de texto. Para esse mister é que se cria o *corpus*.

token, palavra, sentença) visando o processamento do texto para utilização nas fases seguintes. Por fim, as fases de **visualização** e de **modelagem** são autoexplicativas e necessárias para realização da *comunicação*.

Figura 2 – Fluxo de Ciência de Dados



Fonte: Silva (2022, p. 92).

Inobstante a abordagem de Wickham e Grolemund (2021, p. 1) e Silva (2022, p. 92), essa pesquisa explorou uma versão mais simples dos estágios apontados por Benoit (2020, p. 478–489), usando elementos do "tidyverse" [bastante explorados pelos primeiros autores] quando se fez necessário. Assim, o fluxo de dados desse trabalho consistiu em: (*i*) coletar textos em linguagem natural e dispostos de forma não estruturada; (*ii*) estruturar os dados textuais em forma matricial; e, por fim realizar uma análise exploratória de dados visando obtenção de *insights* para elaboração de uma modelagem mais rigorosa e robusta. Essa abordagem será melhor detalhada adiante, na Seção de Metodologia (página 64). De qualquer sorte, resta alertar ao leitor que os termos etapas, estágios ou fases serão aqui utilizados como sinônimos.

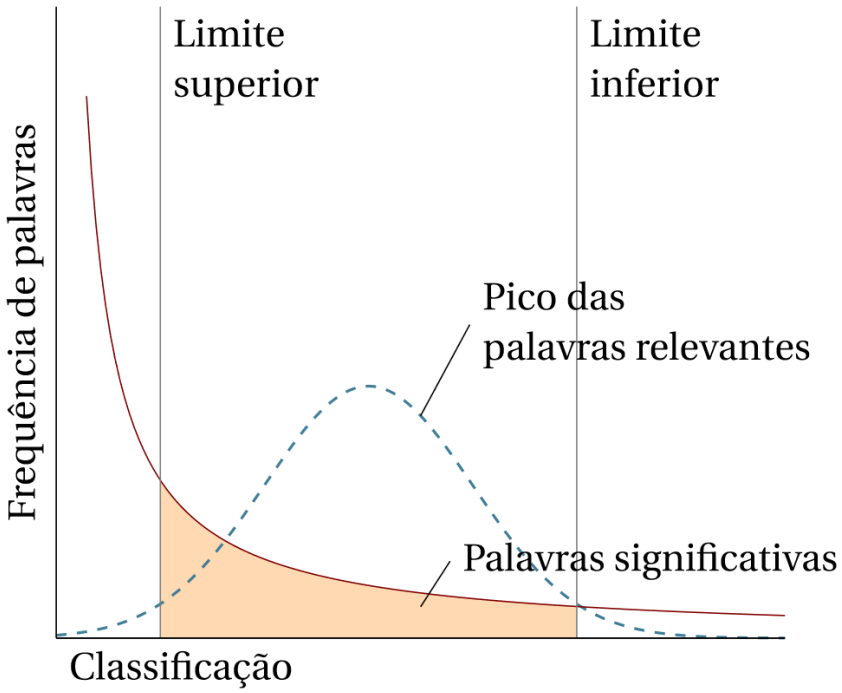
2.2.1.1 Distribuição de palavras e leis empíricas

Segundo Rosa (2010, p. 139–140), as leis empíricas seriam juízos universais sobre objetos empíricos que são pensados com rigorosa universalidade, de modo que nenhuma exceção se admite como possível. Em PLN frequentemente se recorre a duas leis empíricas: a de *Zipf* e a de *cortes de Luhn*.

De acordo com a *Lei de Zipf*, a frequência de qualquer palavra em um corpus é inversamente proporcional ao seu *rank* ou posição na distribuição geral de frequência (Jockers; Thalken, 2020, p. 28). Ou seja, a segunda palavra mais frequente ocorrerá cerca de metade da frequência da palavra mais frequente. Observa-se, pois, que o número de palavras que aparece em qualquer repetição faz o gráfico de ranqueamento de frequências aparecer com uma cauda longa à direita. Para facilitar a visualização, costuma-se utilizar os logaritmos dos valores do *rank* da frequência e a própria frequência. O gráfico resultante dessa conversão à potência é chamado de "Poder da Lei" (ou potência da lei).

Por seu turno, Santos (2019, p. 21–23) traz à lume pesquisas que apontam que a frequência de ocorrência das palavras em um texto pode fornecer uma medida útil sobre a expressividade das mesmas. Isto se dá porque o autor de um texto, como regra geral, é repetitivo em algumas palavras quando desenvolve argumentos ou elabora retóricas sobre assunto qualquer. Observou-se a existência de dois grandes grupos de palavras: (a) um com palavras muito frequentes e inexpressivas, e (b) outro formado com palavras especialmente raras, que ao contrário do primeiro grupo, apresenta frequências baixíssimas. Desse modo, Luhn (1958, p. 160–162) afirma que o gráfico de distribuição de frequências ranqueadas, denominado curva do grau de discriminação (ou "poder de resolução"), apresenta os grupos de palavras, infrequentes ou muito frequentes, nas caudas das distribuições. Desse modo, cabe ao pesquisador tomar divisões arbitrárias, de acordo com a experiência, visando extirpar esses dois grupos de termos na fase de pre-processamento textual. Essa afirmação ficou conhecida como *Lei de Cortes de Luhn* (Santos, 2019, p. 22).

Figura 3 – Distribuição hipotética da frequência de termos textuais



Fonte: Santos (2019, p. 21).

Importa esclarecer que a principal contribuição de Luhn foi a de associar a distribuição da frequência das palavras com a Lei de Zipf (Leite, 2010, p. 21–22). Na Figura 3 acima pode-se observar os limites empíricos de Luhn aplicados em um gráfico da Lei de Zipf. Em pontilhado temos uma distribuição normal. Finalmente, a parte hachurada do gráfico, após o estabelecimento do corte de Luhn, caracteriza as palavras significativas do discurso, devendo as análises recaírem sobre esse grupo.

Duas métricas comumente relatadas em análise de texto são a frequência do termo (TF) e o inverso da frequência do documento (IDF). A TF representa o número de vezes que um termo

t aparece em um documento *d*, significando que quanto maior a frequência de um determinado termo em um documento, mais importante esse termo é para o conteúdo desse documento. Já o inverso da frequência do documento (IDF) é uma medida que representa a raridade de um termo *t* no conjunto de documentos, ou seja, em nosso *corpus* de análise. Estas métricas são então formuladas como se segue:

$$tf_{(t,d)} = \frac{\text{Número de vezes que a palavra aparece no documento}}{\text{Número total de palavras no documento}} \tag{1}$$

$$idf_{(d)} = log_{10}\frac{\text{Número de documentos no } corpus}{\text{Número de documento que contém o termo } d} \tag{2}$$

Da conjunção dessas duas métricas nasceu a estatística TF-IDF[13], que representa a TF ponderada pela IDF, sob a premissa que as palavras mais importantes são aquelas mais usadas em um determinado documento, ao tempo que são relativamente raras no *corpus* geral. O cálculo da TF-IDF resulta da multiplicação da Equação 1 pela Equação 2, resultando na fórmula $TF - IDF = tf_{(t,d)} \cdot idf_{(d)}$. Em outras palavras, a TF-IDF considera a distribuição da frequência dos termos caracterizada pela Lei de Zipf, concentrando a análise nos termos mais relevantes para análise, conforme seria obtido aplicando-se os Cortes de Luhn. Abaixo, para fins de facilitar a visualização de TF-IDF, apresenta-se a Figura 4 obtido a partir de dois documentos de nosso *corpus* de análise.

No caso em análise, computamos a TF-IDF mediante a aplicação da função *dfm_tfidf()* do pacote *quanteda*. Resta observar que no cômputo da TF-IDF o *quanteda* utiliza o logaritmo decimal, mas essa base pode ser alterada pelo pesquisador. Outros pacotes utilizam padrões de bases logarítmicas diferentes, por exemplo, o *tidytext* usa a base neperiana e o *tm* usa a base binária. As palavras mais frequentes, conforme metodologia de TF-IDF, em documentos do *corpus* de análise pode ser observada nas figuras Figura 5 e Figura 6 (pp. 18 e 18).

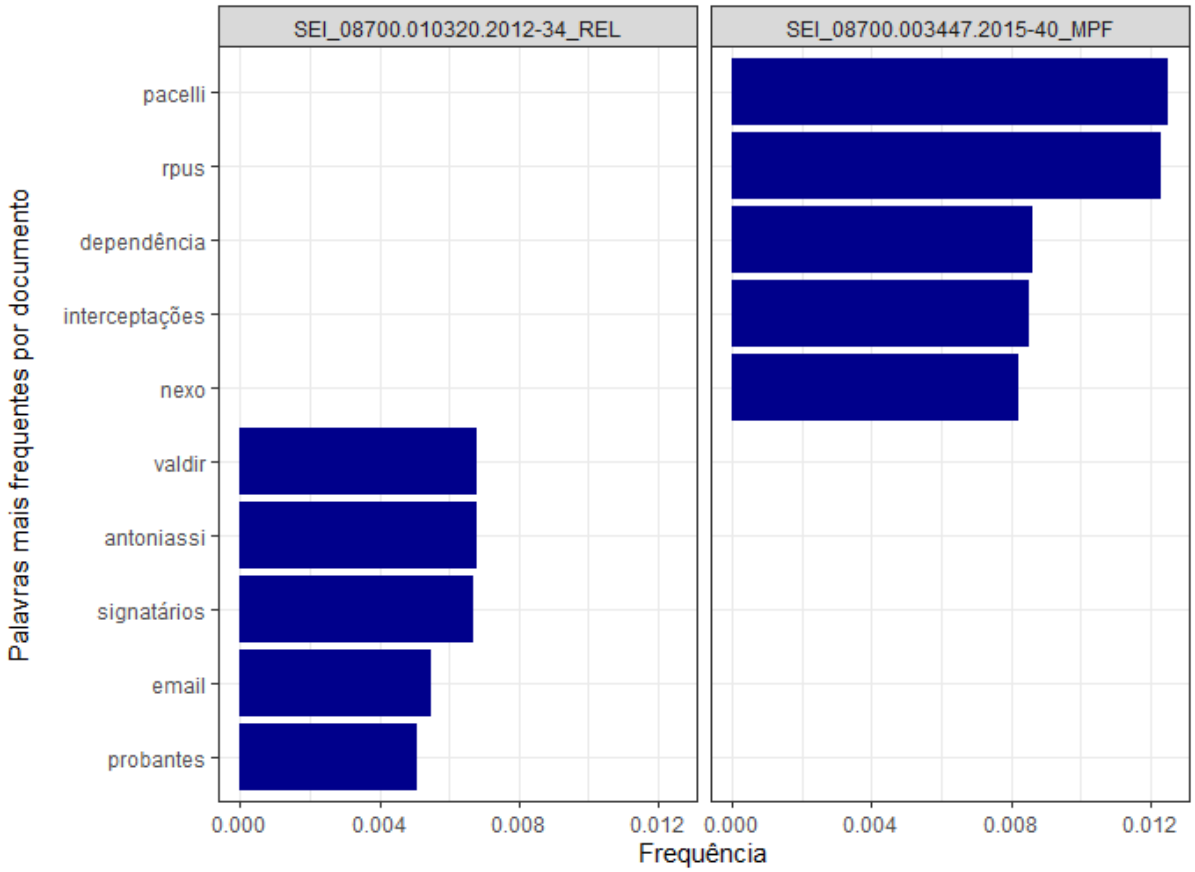
Palavras-chave em contexto[14] em um texto ou uma coleção de textos (em um objeto de *corpus quanteda*), retorna uma lista de uma palavra-chave fornecida pelo usuário em seu contexto imediato, identificando o texto fonte e o número de índice da palavra dentro do texto fonte. KWIC, ou concordâncias, são o método mais usado em linguística de *corpus*. A ideia é muito intuitiva: a palavra tem mais valor semântico se examinanda como ela está sendo usada em um contexto mais amplo. Na Tabela 1 se pode visualizar a utilização da palavras “cartel” em relação a seus vizinhos próximos.

Ainda empreendendo análise das palavras mais frequentes, o *quanteda* admite que seja realizada o agrupamento de variáveis contidas em *docvars*. Assim, promoveu-se o agrupamento

[13] *Term frequency - inverse document frequency*.

[14] *Keywords in context*, ou *kwic*.

Figura 4 – TF-IDF para dois documentos do corpus



Fonte: elaborado pelo autor.

Tabela 1 – Duas palavras anteriores e posteriores a “cartel”

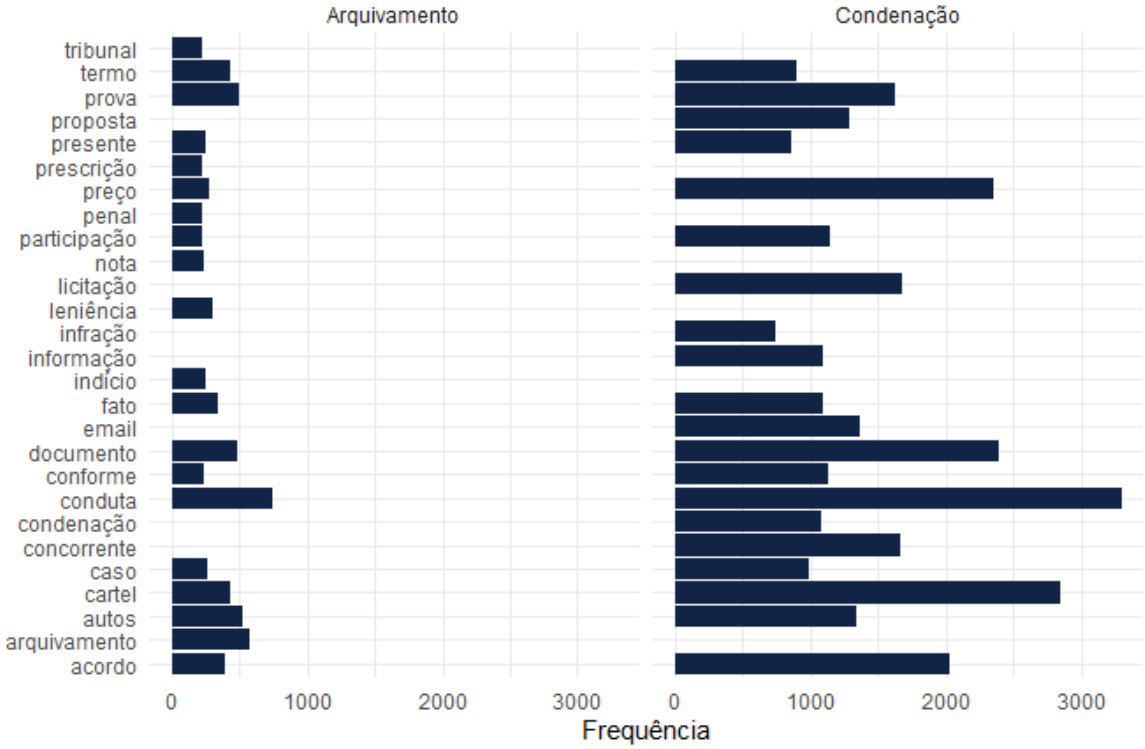
Documento	Anterior	Keyword	Posterior
SEI_08012.000030.2011-50_MPF	–	cartel	licitação pregão
SEI_08012.000030.2011-50_MPF	objeto conduta	cartel	licitação classifica
SEI_08012.000030.2011-50_MPF	noticia suposto	cartel	praticado veículo
SEI_08012.000030.2011-50_MPF	suposta configuração	cartel	base seguintes
SEI_08012.000030.2011-50_MPF	indício suposto	cartel	prestação manutenção
SEI_08012.000030.2011-50_MPF	configurarem prática	cartel	prevista artigo
SEI_08012.000030.2011-50_MPF	configurarem prática	cartel	prevista artigo
SEI_08012.000030.2011-50_MPF	ilícito objeto	cartel	licitação acompanha
SEI_08012.000030.2011-50_MPF	objeto trata	cartel	licitação presentes

Fonte: elaborado pelo autor.

da variável “merito” em suas componentes “Arquivamento” e “Condenação”. O resultado é apresentado na Figura 5:

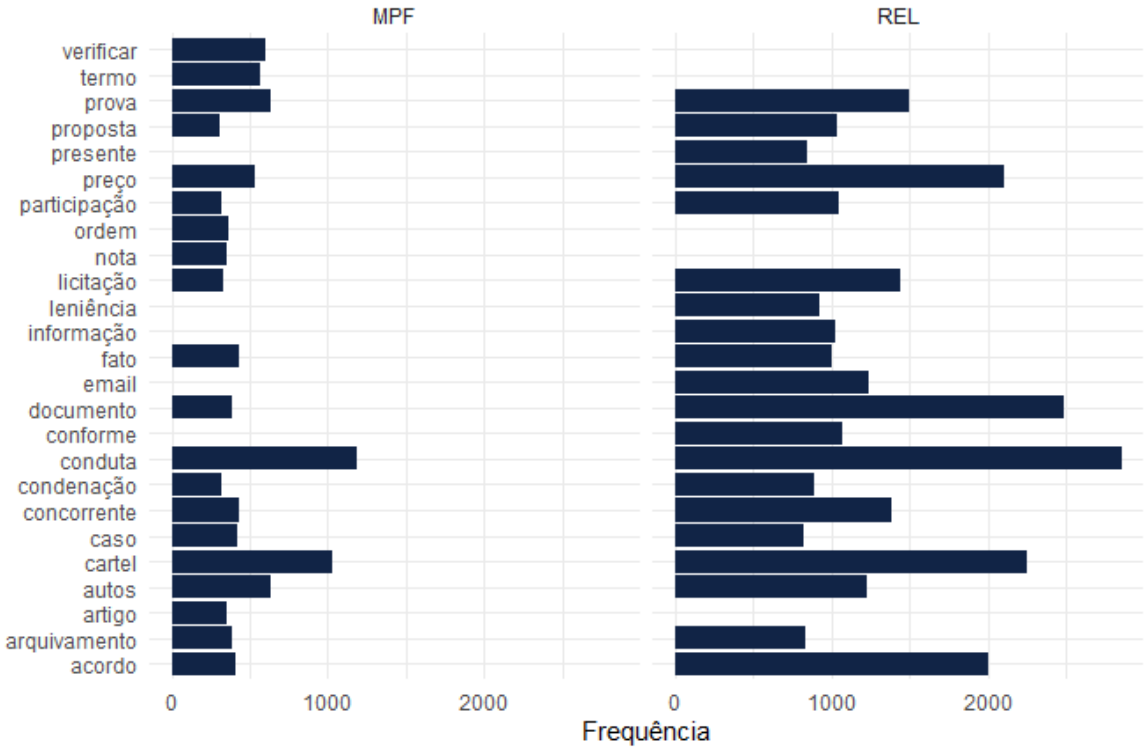
Verifica-se que os documentos referentes à “Arquivamento” tendem a ter menos *tokens* e menor frequência, fato observável no tamanho das barras. Mais uma vez é possível inferir que os documentos embasadores de uma condenação tendem a apresentar maior volume de palavras em

Figura 5 – Palavras mais frequentes segundo o juízo de mérito



Fonte: elaborado pelo autore.

Figura 6 – Palavras mais frequentes nos textos dos Relatores e do MPF



Fonte: elaborado pelo autor.

razão da necessária argumentação para evidenciar a conduta colusiva. Observe-se que o termo "email" não é citado nos documentos de "Arquivamento", enquanto é relativamente frequente nos documentos de "Condenação", o que ratifica a inferência de que esse tipo de prova é bastante utilizado na evidenciação de existência de cartel.

2.2.1.2 Tokenização e conflação textual

O processo de dividir o texto em elementos menores, mas significativos, é chamado de tokenização (Arnold; Tilton, 2015, p. 132). Os *tokens* podem se apresentar sob a forma de palavras, bigramas ou trigramas[15] (Faraco, 2020, p. 62)

A conflação é o ato de fusão ou combinação para igualar variantes morfológicas de palavras. Esta ação pode ser realizada mediante algum tipo de expressão regular, ou com a aplicação de processos computacionais chamados radicalizadores (*stemers*) ou lematizadores(*lemmatizers*) (Santos, 2019, p. 23). Conforme citado anteriormente, foi utilizado o pacote *spacyr* para fins de tokenização e conflação dos dados.

Tabela 2 – Exemplos de Conflação

Texto	Radicalização (stem)	Lematização (lemma)
cantando	cant	cantar
cantasse	cant	cantar
canto (substantivo)	cant	canto

Fonte: elaborado pelo autor.

A análise textual, *a priori*, tem por escopo indexar a unidade de análise em uma coleção de documentos. A representação mais trivial dessa indexação é conhecida por *bag of words* (BoW), que é um conjunto de palavras contidas em um documento, ou coleção de documentos, juntamente com uma contagem de quantas vezes cada uma aparece em no documento (Witten, 2004, p. 4). Esse modelo também é conhecido por sua tradução literal, "saco de palavras".

Muito embora o fato conhecido da BoW descartar a informação sequencial dada pela ordem das palavras, ela é aplicada juntamente à outras técnicas de PLN, tornando a análise mais eficaz e popularizando seu uso por pesquisadores. Witten (2004, p. 5–6) destaca alguns problemas práticos na aplicação da BoW, tais como definir uma determinada "palavra" dentro da de determinado documento, ou como se representar algarismos e numerais. O pesquisador esclarece que na prática, esses problemas são invariavelmente resolvidos por simples heurísticas *ad hoc*.

O modelo de saco de palavras (ou saco de *tokens* ou saco de termos) funciona a partir de uma matriz de termo de documento (DTM ou DFM, como será explicado na subseção 2.2.1.3,

[15] O conceito de n-grama leva em consideração a quantidade de *tokens* constituintes, se for dois será um bigrama, três formam um trigrama.

página 20), onde cada linha representa um documento e cada coluna representa um tipo (um "termo" no vocabulário) e as entradas são as contagens de *tokens* correspondentes ao termo no documento atual.

Figura 7 – Algoritmo de contagem de *tokens* e de *types*

```
library(quanteda)

# Elabora exemplo de tokenização
exemplo <- tokens(
  c(exemplo = 'Aqui digitamos as informações textuais'),
  remove_punct = TRUE)

# Computa o total de tokens
ntoken(exemplo)

# Computa o número de tokens únicos
ntype(exemplo)
```

Fonte: elaborado pelo autor.

Resta estabelecer a diferença estabelecida pelo *quanteda* para fins de tokenização de textos, conforme observado no algoritmo da Figura 7. Enquanto os *tokens* são agrupamentos de caracteres com valor semântico, que podem se repetir ao longo do texto, os *types* são *tokens* únicos no *corpus* segmentado. As funções *ndoc*, *ntoken*, *ntype* e *nsentence* retornam o número de documentos, *tokens*, tipos e sentenças. Essas estatísticas podem ser convenientemente geradas junto com metadados em nível de documento por meio da função *summary* do R base. Caso se queira acessar os metadados do *corpus* ou os alterar, poderá fazer isso a qualquer momento usando o comando *docvars*.

2.2.1.3 DFM e esparsidade dos dados

O objeto DFM resulta em uma matriz com algum nível de esparsidade. Matrizes esparsas são aquelas com muitos valores iguais a zero. Em razão de um corpus utilizar poucas palavras de um léxico, as matrizes resultantes (DFM ou DTM) sempre serão esparsas. Essa informação é importante porque muitos algoritmos de associação e classificação costumam utilizar matrizes esparsas, haja vista que seu armazenamento se dá de forma mais eficiente (Izbicki; Santos, 2020, p. 241).

Em termos práticos, o valor de esparsidade pode ser definido como o limite de frequência relativa do documento para um termo, acima do qual o termo será removido. A frequência relativa do documento aqui significa uma proporção que pode variar de 0 a 1. Há menor dispersão quando à medida de esparsidade se aproxima de um, nunca assumindo valores extremos, apenas apenas valores intermediários.

Por exemplo, definindo-se *sparse* = 0.99 como argumento da função *removeSparseTerms()* do *quanteda*, serão removidos apenas os termos mais esparsos que $0,99$. A interpretação

exata para *sparse* = 0.99 é que para o termo *j*, serão mantidos todos os termos para os quais $Corpus_j > N * (1 - 0,99)$, onde N é o número de documentos – neste caso, provavelmente, todos os termos serão retidos. Perto do outro extremo, se *sparse* = 0.01, somente os termos que aparecem em (quase) todos os documentos serão mantidos. Obviamente que isso depende do número de termos e do número de documentos e, em linguagem natural, palavras comuns como artigos e preposições provavelmente ocorrerão em todos os documentos e, portanto, nunca serão "escassas".

Do exposto, a vantagem de se trabalhar com algum nível de esparsidade se dá em razão da eficiência de uma DFM em armazenar apenas as entradas que têm frequência diferente de zero, o que permite aos algoritmos computacionais utilizar a representação esparsa para que matrizes de ordem elevadas possam ser alocadas eficientemente na memória do computador.

2.2.2 Modelagem de tópicos (pós-processamento)

Pode-se conceituar modelagem de tópicos como sendo um método de processamento de linguagem natural (PLN) que tem como objetivo identificar e extrair os tópicos principais de um conjunto de documentos, sendo amplamente utilizada em áreas como mineração de texto, recuperação de informação e análise de sentimento (Abdelrazek *et al.*, 2023, p. 1). Por conseguinte, ao aplicar a modelagem de tópicos é possível obter uma representação compacta dos documentos e facilitando, pois, a apreensão sobre o conteúdo abordado.

A modelagem de tópicos é realizada por meio de algoritmos que analisam o conteúdo dos documentos e identificam palavras-chave, frases ou conceitos relevantes que representam os temas abordados. Esses tópicos podem ser apresentados como categorias ou *clusters* que agrupam documentos relacionados. A Revisão de Literatura demonstrou que essa abordagem pode ser útil em diversas aplicações, como análise de textos das mais diversas áreas, seja social, financeira, da saúde, do setor jurídico, etc. Sua aplicação visa captar, por exemplo, preferências dos usuários, realizar agrupamento de documentos em coleções de grande volume para facilitar sua organização, recomendar conteúdo personalizado em redes sociais, entre outras aplicações possíveis. As seções que se seguem exploram diferentes técnicas e algoritmos utilizados na modelagem de tópicos, bem como apresenta algumas vantagens e limitações de sua aplicação de forma geral.

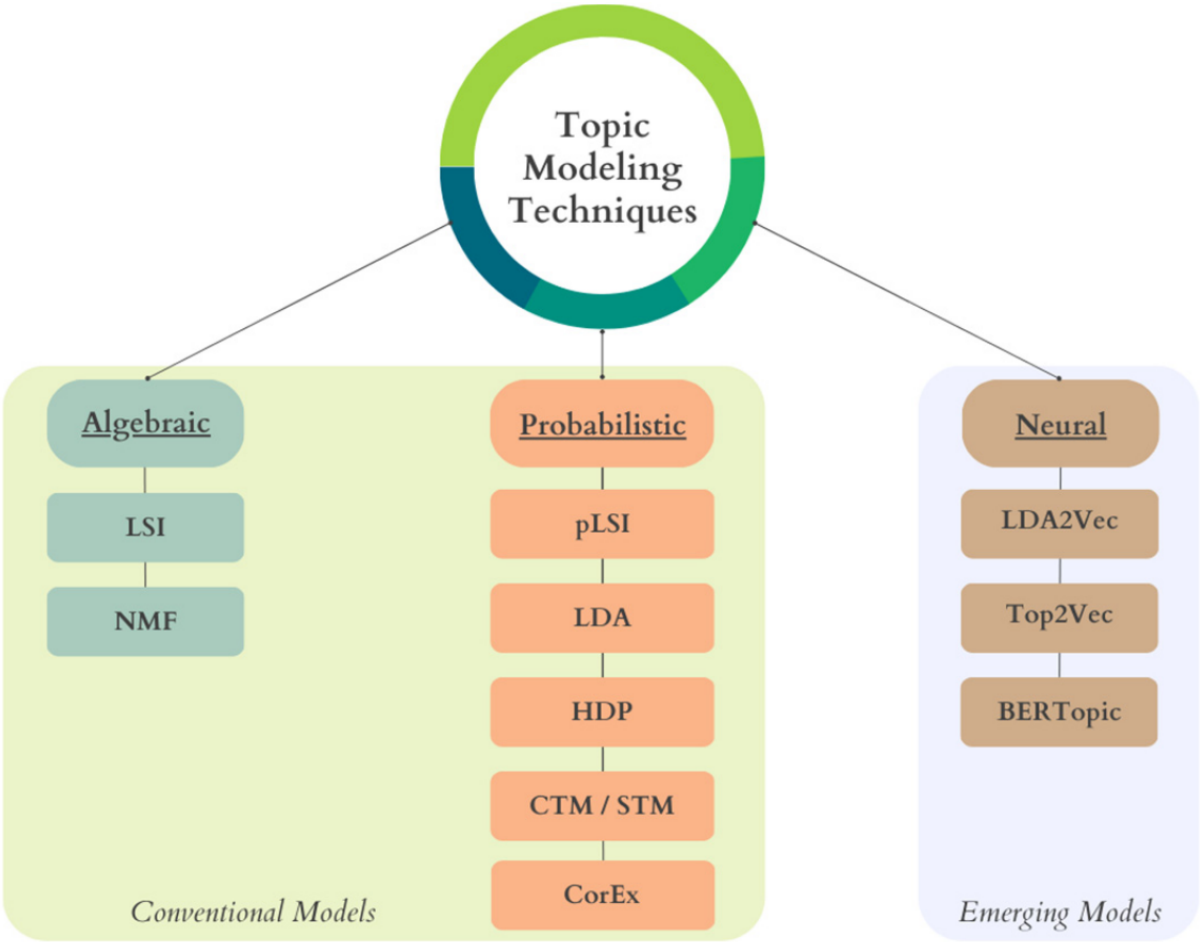
Krishnan (2023, p. 2–3) apresenta uma possível classificação para os modelos de tópicos, dividindo-os em três grandes grupos nominados como modelos algébricos, probabilísticos e neurais.

Modelos algébricos são abordagens estatísticas tradicionais usadas na modelagem de tópicos. Eles abrangem técnicas como indexação semântica latente (LSI) e fatoração de matriz não negativa (NMF). O LSI analisa as relações entre palavras e documentos para identificar tópicos latentes, enquanto o NMF utiliza uma abordagem de álgebra linear para descobrir tópicos latentes ao fatorar a matriz termo-documento.

Modelos probabilísticos usam métodos estatísticos para associar palavras e documentos a tópicos de maneira probabilística, ou seja, baseando-se na probabilidade de que uma palavra pertença a um tópico específico. O LDA é um modelo estatístico generativo que assume que cada documento é uma mistura de tópicos e cada palavra está associada a um tópico específico. Outros modelos probabilísticos incluem indexação semântica latente probabilística (PLSi), explicação da correlação ancorada (CoreX), processo hierárquico de Dirichlet (HDP), modelo de tópico correlacionado (CTM) e modelo de tópico estrutural (STM).

Os **modelos neurais** representam o mais recente avanço na modelagem de tópicos, aproveitando redes neurais artificiais e se beneficiando da ampla adoção do aprendizado profundo no Processamento de Linguagem Natural (PNL). Esses modelos utilizam técnicas como representações de codificadores bidirecionais de transformadores (BERT) e TF-IDF baseado em classe (c-TF-IDF) para obter um desempenho de última geração. O BerTopic, um modelo neural, surgiu como uma abordagem dominante no campo da modelagem de tópicos, apresentando resultados impressionantes em vários conjuntos de dados com requisitos mínimos de pré-processamento de dados.

Figura 8 – Classificação de modelos segundo Chen *et al.* (2023)



Fonte: Chen *et al.* (2023, p. 4).

Chen *et al.* (2023, p. 3–4) apresenta uma classificação das técnicas em dois grandes grupos: modelos convencionais e modelos emergentes. Os modelos convencionais se subdivide em algébricos e probabilísticos, já os modelos emergentes são aqueles do tipo neural. A Figura 8, na página 22 capta essa classificação.

Como se pode constatar, não há pacificação na classificação de modelos. Dentre os modelos nominados pela academia, os mais comumente mencionados nos documentos consultados, sem ordem de preferência, foram aqueles relacionados na Figura 9, na página 25. Tais algoritmos são amplamente utilizados para extrair tópicos de *corpora* para fins de realização de análises de texto em diferentes domínios. Calha ratificar o alerta que nossa classificação é apriorística, não fechada, derivada do corpo de trabalhos apresentados na seção de Referências, servindo tão somente para um breve entendimento sobre os diversos modelos revelados durante a fase Revisão de Literatura. Passa-se, pois, a elencá-los e exemplificá-los.

I - Modelos de Alocação Latente:

- *Latent Dirichlet Allocation* (LDA): modelo probabilístico que assume que documentos são misturas de tópicos e que cada tópico é uma mistura de palavras.
- *Latent Semantic Analysis* (LSA): usa decomposição em valores singulares para reduzir a dimensionalidade dos dados de texto e descobrir estruturas latentes.
- *Probabilistic Latent Semantic Analysis* (PLSA):o modelo se fundamenta na concepção de que cada documento consiste em uma mistura de tópicos, e cada tópico, por sua vez, em uma distribuição probabilística de palavras.
- *Structural Topic Model* (STM): algoritmo de inferência bayesiana, por meio da qual se estima os parâmetros do modelo. Essa abordagem permite a identificação de tópicos latentes, mas também abarca a capacidade de incorporar variáveis contextuais (covariáveis).

II - Modelos Baseados em *Embeddings*[16]:

- Word2Vec: aprende representações vetoriais de palavras que capturam contextos semânticos e sintáticos.
- Doc2Vec: extensão do Word2Vec que aprende representações vetoriais de documentos inteiros.
- *Global Vectors for Word Representation* (GloVe): combina as abordagens de matriz de co-ocorrência global e aprendizado local para produzir embeddings de palavras.

[16] *Embeddings* são representações numéricas de alta dimensão de dados textuais, como palavras, frases ou documentos inteiros, aprendidas de modo que entidades com significados semelhantes tenham *embeddings* próximos no espaço vetorial. Essas representações são fundamentais em várias tarefas de processamento de linguagem natural (NLP), como classificação de texto e análise de sentimentos, capturando o contexto semântico das palavras ou textos (Mikolov; Chen *et al.*, 2013).

- FastText: melhora o Word2Vec levando em conta subpalavras, o que é útil para lidar com palavras fora do vocabulário.
- *Bidirectional Encoder Representations from Transformers* (BERT): modelo de linguagem baseado em transformers que captura contextos bidirecionais.
- BERTopic: método que combina BERT com algoritmos de *clustering* para melhorar a modelagem de tópicos.

III - Modelos Híbridos:

- Top2Vec: algoritmo que combina *embeddings* de palavras e documentos para identificar tópicos de forma não-supervisionada.
- *Non-negative Matrix Factorization* (NMF): método de fatoração de matrizes que restringe os fatores a serem não-negativos, facilitando a interpretação dos tópicos.

IV - Modelos Baseados em Rede Neural:

- Neural Topic Models: modelos que utilizam redes neurais para descobrir tópicos em coleções de documentos.

V - Modelos de Tópico Temporal:

- *Dynamic Topic Models*: adaptam modelos como o LDA para lidar com mudanças nos tópicos ao longo do tempo.

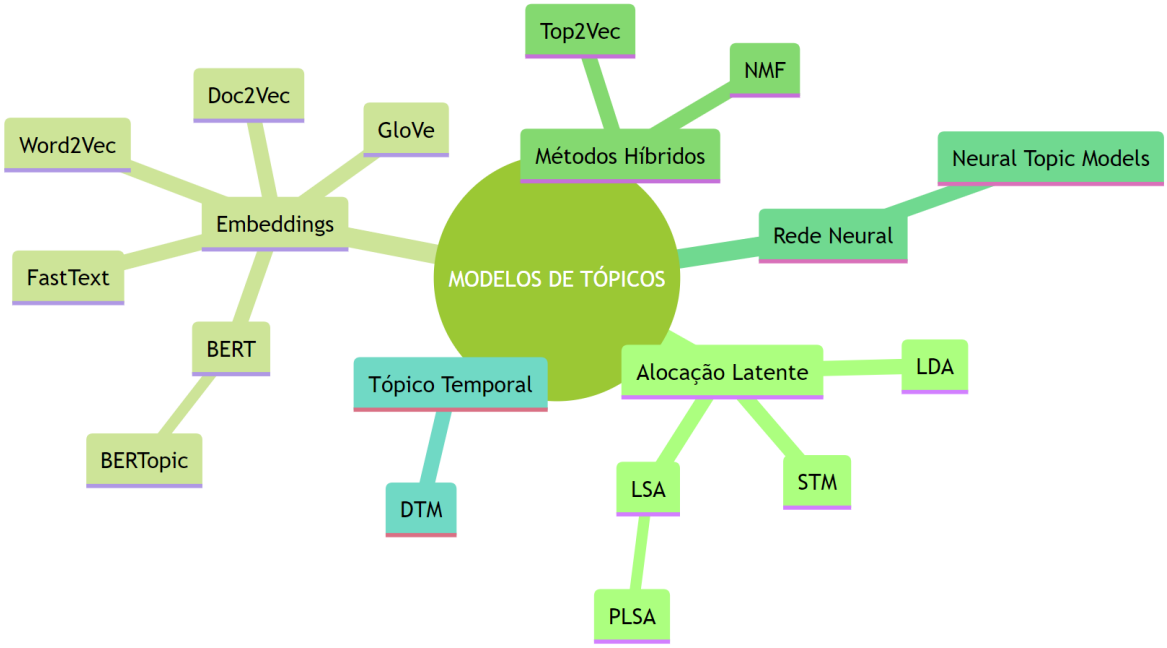
Dessa forma, sem pretensão de esgotar o tema, apresenta-se uma classificação de modelos tão somente para fins de padronização de terminologia nessa pesquisa, conforme a Figura 9 (página 25).

A classificação ora apresentada se constitui a partir dos métodos de modelagem de tópicos mais comuns identificado na Revisão de Literatura desta pesquisa (vide Apêndice B, p. 141). Cada um tem suas particularidades e pode ser mais adequado dependendo do tipo de dados e do objetivo da análise. Dada essas particularidades, cada modelo é aplicado de acordo com a necessidade e o contexto do problema em questão, ratificando que todos apresentam vantagens e limitações em suas aplicações. Nos tópicos que se seguem serão apresentadas informações coligidas sobre tais métodos de modelagem de tópicos.

2.2.2.1 Latent Semantic Analysis (LSA)

A Análise Semântica Latente (*Latent Semantic Analysis* - LSA), proposta por Deerwester *et al.* (1990), constitui um modelo que emprega métodos estatísticos e computacionais para

Figura 9 – Especies de Modelos de Tópicos



Fonte: elaborado pelo autor.

representar o significado semântico das palavras em textos. Baseando-se na premissa de que palavras em contextos semelhantes possuem significados semelhantes, o LSA representa palavras como vetores em um espaço multidimensional. A proximidade entre vetores neste espaço visa refletir a relação semântica entre as palavras. Esses autores oferecem uma explanação detalhada do algoritmo de LSA, suas aplicações em indexação documental e discutem as características e desafios do modelo, apoiando suas afirmações com experimentos comparativos.

Esse modelo é capaz de identificar relações semânticas entre palavras e documentos, facilitando a recuperação de informações em documentos que não compartilham termos exatos. O modelo também possibilita a redução da dimensionalidade do espaço vetorial de palavras, simplificando o processamento e análise de grandes conjuntos documentais. Além disso, o LSA considera o contexto no qual as palavras são empregadas, contribuindo para a análise semântica de documentos. A capacidade do LSA em facilitar a recuperação de informações é evidenciada pela identificação de termos semanticamente relacionados em extensos conjuntos documentais (Deerwester *et al.*, 1990).

Contudo, o LSA apresenta desafios. A interpretação dos resultados gerados pode ser complexa, uma vez que os vetores resultantes nem sempre são diretamente traduzíveis em conceitos compreensíveis. Adicionalmente, o modelo pode ser vulnerável a termos irrelevantes ou ruidosos nos documentos, o que pode afetar a qualidade das análises e exigir um pré-processamento rigoroso dos dados. A eficácia do LSA também é influenciada pela quantidade e qualidade dos dados disponíveis. Conjuntos de dados menores ou documentos de baixa qualidade podem impactar negativamente a precisão das análises (Deerwester *et al.*, 1990).

Em resumo, o LSA apresenta capacidades relevantes para a identificação de relações semânticas e recuperação de informações, entretanto, enfrenta desafios associados à interpretabilidade dos resultados e à sensibilidade ao ruído nos dados.

2.2.2.2 Non-negative Matrix Factorization (NMF)

A Fatoração de Matriz Não Negativa (NMF), conforme delineada por Lee *et al.* (1999) e também discutida em trabalhos posteriores como o de Krishnan (2023), é uma técnica de aprendizado de máquina que opera sob os princípios da álgebra linear para decompor uma matriz não negativa em duas outras matrizes igualmente não negativas.

A decomposição da matriz pode ser expressa como $V \approx WH$, onde: V é a matriz original que desejamos decompor; W e H são as matrizes de fatores não-negativas; a matriz W contém as características de baixo nível que são extraídas dos dados e a matriz H contém a informação da combinação dessas características para reconstruir os dados originais. O objetivo do NMF é encontrar as melhores matrizes W e H que minimizam o erro de reconstrução entre a matriz original V e a sua reconstrução pleo produto WH.

Amplamente empregada em diversos campos, incluindo processamento de imagens e mineração de dados, a NMF é particularmente notável na modelagem de tópicos, onde é utilizada para identificar temas em matrizes termo-documento (Krishnan, 2023; Lee *et al.*, 1999). A NMF foi comparado a outras técnicas de modelagem de tópicos, como Latent Dirichlet Allocation (LDA), e se descobriu que produz tópicos mais coerentes em certos casos (O'Callaghan *et al.*, 2015; Adhikary; Murty, 2012).

Contudo, apesar de suas vantagens em interpretabilidade, extração de características e redução de dimensionalidade, a NMF apresenta limitações, incluindo sensibilidade a ruídos, dependência de inicialização, problemas de escalabilidade e restrição de não-negatividade. Além disso, sua propensão a valores atípicos pode influenciar a coerência dos termos associados, exigindo atenção e, potencialmente, a utilização da mediana das avaliações como uma solução para esses desafios (Krishnan, 2023; Lee *et al.*, 1999).

Em resumo, a NMF é uma técnica robusta para a identificação de tópicos latentes, oferecendo precisão temática superior em conteúdos especializados quando comparada a métodos como o LDA. No entanto, suas limitações, incluindo sensibilidade a valores atípicos e restrições de não-negatividade, requerem consideração cuidadosa em sua aplicação.

2.2.2.3 Probalistic Latent Semantic Analysis (PLSA)

A Análise Semântica Latente Probabilística (PLSA), introduzida por Hofmann (2001, 1999) emerge como uma metodologia estatística robusta para a exploração e interpretação de extensos conjuntos de dados textuais. Pautando-se na identificação de padrões semânticos subjacentes em documentos, o modelo se fundamenta na concepção de que cada documento

consiste em uma mistura de tópicos, e cada tópico, por sua vez, em uma distribuição probabilística de palavras. Através desta abordagem, o PLSA busca desvendar as relações ocultas entre termos e documentos, considerando a incerteza inerente à seleção lexical em contextos variados. O PLSA é um método de modelagem de tópicos amplamente usado que tem sido aplicado no campo da análise de mídias sociais (Krishnan, 2023).

No processo de análise, o PLSA emprega técnicas de inferência probabilística para estimar as distribuições de tópicos por documento e de palavras por tópico, visando maximizar a verossimilhança dos dados observados. Esta metodologia não apenas facilita a identificação dos tópicos dominantes em documentos específicos, mas também permite a execução de operações complexas, como a clusterização de documentos com base nas distribuições temáticas. Sua aplicação estende-se por diversas áreas, incluindo recuperação de informações, classificação textual e análise de sentimentos, oferecendo uma perspectiva avançada na análise de grandes volumes de dados textuais e proporcionando *insights* semânticos significativos (Hofmann, 2001, 1999).

No entanto, apesar das vantagens destacadas, como a representação probabilística que permite interpretações mais precisas das relações entre termos e tópicos e a redução de ambiguidades textuais, o PLSA não está isento de limitações. A sensibilidade do modelo aos dados de treinamento, as dificuldades no tratamento de palavras raras e os desafios na interpretabilidade dos tópicos extraídos são alguns dos obstáculos enfrentados pelos pesquisadores ao empregar este método (Hofmann, 2001, 1999).

Concluindo, a Análise Semântica Latente Probabilística (PLSA) representa uma abordagem valiosa na análise semântica de grandes corpora textuais, apesar das limitações inerentes ao modelo. Suas capacidades de revelar tópicos latentes e atenuar ambiguidades linguísticas a posicionam como uma ferramenta relevante em variadas aplicações analíticas, reforçando seu papel como uma metodologia de referência no campo da análise de dados textuais.

2.2.2.4 Latent Dirichlet Allocation (LDA)

A *Latent Dirichlet Allocation* (LDA), conforme descrito por Blei (2011) e Blei, Ng e Jordan (2003), é um modelo estatístico não supervisionado amplamente utilizado em aprendizado de máquina para análise de tópicos e modelagem de documentos. Baseando-se na "mistura de tópicos" e na "distribuição de palavras", o LDA postula que cada documento é uma combinação de vários tópicos e que cada tópico pode ser representado por uma distribuição probabilística de palavras. Este modelo é configurado por hiperparâmetros de Dirichlet e pelo número de tópicos, e visa deduzir os tópicos subjacentes em coleções documentais, empregando probabilidade condicional para estimar distribuições.

Além disso, o LDA permite descobrir tópicos ocultos em grandes conjuntos de dados e é flexível para uma variedade de aplicações, como análise de sentimentos e classificação de documentos. Sua eficiência na inferência torna-o adequado para cenários de *big data*, e sua

aplicabilidade estende-se a dados de diferentes tipos, incluindo texto, imagem e redes sociais, permitindo uma interpretação compacta e facilitada dos dados (Blei, 2011; Blei; Ng; Jordan, 2003).

Contudo, o modelo apresenta limitações, como a simplificação na representação de documentos e tópicos, dificuldades em distinguir tópicos similares e a necessidade de definição prévia do número de tópicos. Ademais, a suposição de independência condicional entre as palavras pode não capturar a complexidade dos dados reais, e a interpretação dos tópicos gerados pode ser subjetiva, requerendo análise adicional (Blei, 2011; Blei; Ng; Jordan, 2003).

Em suma, o LDA é uma ferramenta fundamental para a compreensão semântica e identificação de tópicos predominantes em grandes volumes de dados, embora apresente desafios que requerem atenção na sua aplicação e interpretação dos resultados.

2.2.2.5 Dynamic Topic Model (DTM)

Os Modelos Dinâmicos de Tópicos (DTM, do inglês) representam uma abordagem estatística inovadora no campo da modelagem de tópicos em conjuntos de dados textuais. Este método distingue-se por incorporar a dimensão temporal na análise de tópicos, permitindo a observação da evolução destes ao longo do tempo. Originados como uma extensão do modelo LDA, os DTMs superam uma limitação fundamental do LDA, que não contempla a variabilidade temporal na distribuição dos tópicos (Blei; Lafferty, 2006). No modelo LDA, postula-se que documentos são compostos por uma mescla de tópicos, com palavras sendo geradas a partir desses tópicos. A inovação trazida pelo DTM consiste em permitir que os tópicos se modifiquem com o passar do tempo. Esta característica é particularmente relevante para a análise de coleções de documentos que cobrem extensos períodos, como arquivos históricos e fluxos contínuos de notícias.

Os DTMs empregam o Processo de Dirichlet Hierárquico (*Hierarchical Dirichlet Process* - HDP[17]) para modelar a distribuição de tópicos em cada período temporal, facultando a inclusão de novos tópicos à medida que o tempo avança. A evolução dos tópicos é inferida a partir das palavras presentes em diferentes períodos, oferecendo assim um meio para identificar transformações temáticas ao longo do tempo (Blei; Lafferty, 2006).

A aplicação dos DTMs pode revelar percepções valiosas sobre a dinâmica dos tópicos em dados textuais, facilitando a análise de tendências, mudanças de opinião e a evolução de conceitos. Tais análises são de grande utilidade em domínios como a análise de sentimentos, mídias sociais, análise de tendências e estudos históricos (Blei; Lafferty, 2006).

Em suma, os Modelos Dinâmicos de Tópicos constituem uma ferramenta poderosa para a modelagem de tópicos em textos, capaz de capturar e analisar a evolução temporal dos tópicos.

[17] O HDP é um modelo estatístico que permite a identificação de um número variável de categorias em análises bayesianas, ajustando-se aos dados sem a necessidade de definir previamente a quantidade de categorias. Foi introduzido por Teh *et al.* (2006).

Contudo, apesar de suas vantagens, como a detecção de mudanças temáticas e uma representação temporal aprimorada, enfrentam desafios, como a complexidade na interpretação de tópicos dinâmicos e demandas elevadas de processamento computacional. A compreensão plena das vantagens e limitações do DTM, contudo, requer um exame mais detalhado que ultrapassa o escopo deste resumo.

2.2.2.6 Structural Topic Model (STM)

O modelo estrutural de tópicos (STM), concebido por Roberts, Stewart e Tingley (2019, 2016), Roberts, Stewart, Tingley *et al.* (2014) e Roberts, Tingley *et al.* (2013) representa um método estatístico avançado, aplicado primordialmente ao exame de respostas abertas em pesquisas. Este modelo distingue-se de alternativas como o LDA pela sua capacidade de incorporar as particularidades estruturais dos dados, contemplando as características das perguntas formuladas e as respectivas respostas dos entrevistados (Roberts; Tingley *et al.*, 2013).

A fundamentação teórica do STM reside na inferência bayesiana, por meio da qual estima os parâmetros do modelo. Essa abordagem bifurca a variabilidade em duas vertentes principais: a variação dos tópicos e a variação dos termos presentes em cada tópico. Este método utiliza uma matriz de termos e outra de atributos contextuais, combinando-as em um modelo unificado que estima a distribuição tópica em cada documento e a frequência das palavras em cada tópico (Roberts; Stewart; Tingley *et al.*, 2014).

A relevância do STM não se restringe à identificação de tópicos latentes, mas também abarca a capacidade de incorporar variáveis contextuais. Essa característica possibilita uma compreensão mais aprofundada de como fatores contextuais influenciam a distribuição dos tópicos nos documentos analisados (Roberts; Stewart; Tingley *et al.*, 2014; Roberts; Tingley *et al.*, 2013).

Em contrapartida, há desafios associados ao uso do STM, como a exigência de pré-processamento dos dados, a sensibilidade à escolha de hiperparâmetros e a necessidade de expertise técnica para sua implementação e interpretação. Ademais, o modelo pode enfrentar limitações na identificação de tópicos menos frequentes e é mais apto para análises de dados textuais não estruturados (Roberts; Stewart; Tingley, 2019; Roberts; Stewart; Tingley *et al.*, 2014).

Em suma, o STM manifesta-se como uma ferramenta robusta e flexível, com aplicações que transcendem a análise de dados textuais, abrangendo a modelagem de variações temporais e a incorporação de covariáveis. No entanto, é imperativo reconhecer suas limitações e desafios para otimizar sua utilização em contextos de pesquisa.

2.2.2.7 Word2Vec

O Word2Vec, delineado por Mikolov, Chen *et al.* (2013), é um modelo de representação de palavras que opera por meio de vetores distribuídos, buscando apreender relações semânticas e sintáticas em um espaço vetorial contínuo. Utilizando uma arquitetura neural, especificamente modelos CBOW ou *skip-gram*, o Word2Vec é treinado para prever a probabilidade de uma palavra com base em seu contexto imediato, aprendendo padrões de coocorrência de palavras em grandes corpora textuais. Posteriormente ao treinamento, o modelo atribui a cada palavra um vetor que encapsula aspectos semânticos e sintáticos. Tais vetores permitem operações que revelam analogias, enriquecendo aplicações práticas em diversas tarefas de processamento de linguagem natural.

As características principais desse modelo envolvem sua eficiência computacional, decorrente de seu método de aprendizado não supervisionado; a habilidade de generalizar informações semânticas entre palavras, conservando similaridades; e sua escalabilidade, permitindo a aplicação em volumosos conjuntos de texto. Entretanto, o modelo possui limitações, como a focalização apenas no contexto local de palavras, dificuldades em lidar com polissemia e na interpretação de frases compostas ou expressões idiomáticas (Mikolov; Chen *et al.*, 2013).

Portanto, o Word2Vec é uma abordagem robusta na representação de palavras em um espaço vetorial contínuo, impactando significativamente o campo do processamento de linguagem natural. Sua arquitetura neural e capacidade de gerar representações vetoriais densas contribuem para a eficácia em diversas aplicações práticas, apesar das limitações inerentes ao modelo.

2.2.2.8 Doc2Vec

O Doc2Vec, um algoritmo de aprendizado de máquina proposto por Le e Mikolov (2014), é empregado para representar documentos e palavras em um espaço vetorial distribuído. Distinto das representações de palavras individuais fornecidas pelo modelo Word2Vec, o Doc2Vec possibilita a representação de documentos inteiros como vetores em um espaço contínuo.

O algoritmo visa capturar o significado das palavras e o contexto em que são utilizadas dentro de um documento. Para tanto, atribui-se um vetor exclusivo a cada palavra, combinando-os posteriormente para formar uma representação singular do documento. Esta representação se mostra útil em diversas tarefas de processamento de linguagem natural, como classificação de documentos, recuperação de informações e análise de sentimentos. Utilizando um modelo de aprendizado não supervisionado, o Doc2Vec é treinado por meio de um algoritmo denominado *skip-gram with negative sampling*, que tem como objetivo prever palavras em um contexto dado. Durante o treinamento, as representações de palavras e documentos são ajustadas para minimizar a discrepância entre as palavras previstas e as reais (Le; Mikolov, 2014).

Em comparação com outras abordagens, como o modelo de vetor global (GloVe), o Doc2Vec se destaca por oferecer uma maneira eficiente de representar documentos inteiros,

explorando relações entre palavras e incorporando-as em um espaço vetorial contínuo. Entre as vantagens do Doc2Vec, destacam-se a capacidade de capturar o contexto semântico de documentos, a possibilidade de realizar operações matemáticas com documentos e a economia de recursos de armazenamento, decorrente do uso de representações vetoriais compactas. Por outro lado, o Doc2Vec enfrenta desafios como a necessidade de um grande volume de dados para treinamento, a dependência de hiperparâmetros e a dificuldade em representar documentos curtos, que podem oferecer contexto insuficiente para uma boa representação (Le; Mikolov, 2014).

Conclui-se que, apesar de suas vantagens, como a habilidade de capturar contexto semântico e a realização de operações matemáticas com documentos, é fundamental considerar as limitações do Doc2Vec, especialmente a exigência de grandes volumes de dados de treinamento e as dificuldades na representação de documentos breves.

2.2.2.9 Global Vectors for Word Representation (GloVe)

O *Global Vectors for Word Representation* (GloVe), conforme apresentado por Pennington, Socher e Manning (2014), é um modelo de representação de palavras em espaço vetorial contínuo, projetado para capturar relações semânticas e sintáticas baseadas em coocorrência estatística em amplos conjuntos de documentos. O GloVe visa gerar representações vetoriais que refletem as características semânticas das palavras, permitindo que termos semanticamente similares estejam proximamente situados no espaço vetorial. Tal objetivo é alcançado modelando a distribuição vetorial das palavras por meio de informações contextuais derivadas de suas coocorrências.

O método emprega uma matriz de coocorrência, cujas entradas refletem a frequência de coocorrência de palavras em contextos próximos. A partir dessa matriz, uma função de custo é estabelecida para aprender vetores de palavras que minimizem a discrepância entre os produtos internos desses vetores e os logaritmos das contagens de coocorrências. Uma vantagem distintiva do GloVe é sua capacidade de gerar vetores de palavras que encapsulam tanto relações de similaridade quanto de analogia. Isso é possível pela incorporação de informações de coocorrência global no processo de treinamento, diferentemente de modelos que se concentram exclusivamente em coocorrências locais. E outras palavras, o modelo permite uma representação vetorial proximal de palavras semanticamente similares, beneficiando diversas tarefas de processamento de linguagem natural (Pennington; Socher; Manning, 2014).

As vantagens do GloVe incluem sua eficiência computacional comparativa e sua habilidade em capturar relações semânticas e representar diferentes sentidos de uma palavra. Entretanto, o modelo enfrenta limitações decorrentes de sua dependência de informações estatísticas, negligência da ordem das palavras e dificuldades em lidar com palavras polissêmicas (Pennington; Socher; Manning, 2014).

Concluindo, o GloVe emerge como um método eficiente e avançado na representação

vetorial de palavras, apesar de suas limitações inerentes, oferecendo contribuições significativas ao campo do processamento de linguagem natural.

2.2.2.10 FastText

O modelo FastText, proposto por Joulin, Grave, Bojanowski e Mikolov (2016) e Joulin, Grave, Bojanowski, Douze *et al.* (2016), representa uma abordagem de aprendizado de máquina para o processamento de texto, destacando-se pela eficiência em lidar com grandes volumes de texto e sua aplicabilidade em problemas de classificação de texto. Diferenciando-se de outros modelos de processamento de texto, o FastText emprega representações de palavras baseadas em *n-grams*, capturando a estrutura interna das palavras e seu contexto em sentenças.

Além disso, o FastText é reconhecido pela rapidez e eficiência de seu treinamento, atributos decorrentes da implementação da técnica *Hierarchical Softmax*[18], que otimiza o tempo de treinamento ao estruturar palavras em uma árvore hierárquica e aplicar probabilidades condicionais. A arquitetura do modelo inclui ainda a *Bag-of-Tricks*[19] para classificação, capaz de tratar palavras desconhecidas ou fora do vocabulário, uma funcionalidade crucial para cenários onde o modelo é exposto a palavras não presentes no treinamento (Joulin; Grave; Bojanowski; Mikolov, 2016; Joulin; Grave; Bojanowski; Douze *et al.*, 2016).

O FastText sobressai por sua eficiência computacional, permitindo processamento rápido de extensos conjuntos textuais. Suas representações de palavras capturam nuances semânticas e estruturais, incluindo subpalavras[20] e afixos, facilitando diversas tarefas de processamento de texto. Na classificação de texto, o modelo demonstra eficiência e precisão notáveis. Adicionalmente, sua capacidade de tratar palavras fora do vocabulário permite generalizações para termos desconhecidos. No entanto, esse modelo exige conjuntos de dados volumosos para desempenho ótimo, podendo ter seu desempenho afetado em conjuntos menores. Embora seja efetivo em classificação de texto e previsão de palavras, o modelo não é tão adequado para tarefas que demandam compreensão de sequências mais longas. A qualidade do FastText também está atrelada à tokenização adequada dos textos de entrada (Joulin; Grave; Bojanowski; Mikolov,

[18] *Hierarchical Softmax* é uma técnica que acelera a computação de probabilidades em modelos de linguagem com vocabulários amplos, como o Word2Vec. Em vez de calcular a probabilidade de uma palavra em relação a todo o vocabulário, o que é computacionalmente custoso, o *Hierarchical Softmax* usa uma árvore binária onde cada palavra é uma folha. Calcula-se a probabilidade seguindo o caminho da raiz até a folha da palavra desejada, reduzindo drasticamente as operações necessárias. Para mais detalhes, pode-se consultar o trabalho de Mikolov, Sutskever *et al.* (2013) entitulado *Distributed Representations of Words and Phrases and their Compositionality*.

[19] *Bag-of-Tricks* é uma técnica usada no processamento de linguagem natural (NLP) para representar um texto como uma coleção desordenada de palavras, sem levar em conta a estrutura gramatical ou a ordem das palavras. Essa abordagem trata o texto como um "saco" (bag) de palavras e ignora a sequência de palavras, considerando apenas a frequência de palavras individuais. Ao criar essa representação, são contabilizadas as ocorrências de cada palavra no texto, formando um vetor de características. Essa técnica é frequentemente utilizada como uma forma de representação inicial dos dados para alimentar algoritmos de aprendizado de máquina, como modelos de classificação (Joulin; Grave; Bojanowski; Mikolov, 2016).

[20] Uma subpalavra de uma palavra é qualquer sequencia de símbolos que compõem a palavra (Centro De Informática Da Ufpe, 2023).

2016; Joulin; Grave; Bojanowski; Douze *et al.*, 2016).

Em síntese, o FastText se destaca como um modelo eficiente e veloz no processamento de texto, com aptidão para resolver problemas de classificação de texto e manipular palavras desconhecidas. Suas vantagens incluem eficiência computacional e representações de palavras detalhadas, embora enfrentem desafios com conjuntos de dados menores e compreensão de sequências mais longas.

2.2.2.11 Neural Topic Model (NTM)

As Modelagens Neurais de Tópicos (NTM, do inglês) representam um avanço significativo na análise de dados textuais. Utilizando redes neurais, esses modelos buscam extrair temas subjacentes de vastos conjuntos de texto não rotulado. Diferentemente das abordagens convencionais, como a Alocação Latente de Dirichlet (LDA), as Modelagens Neurais de Tópicos empregam redes neurais para aprender representações de palavras e documentos, captando as associações semânticas entre eles. Uma instância notável de tal modelo é o TopicRNN, introduzido por Dieng, Wang *et al.* (2017), que se vale de uma rede neural recorrente para discernir dependências semânticas de longo alcance.

O desenvolvimento do TopicRNN exemplifica a habilidade dos Modelos Neurais de Tópicos de aprender representações distribuídas das palavras, que, ao serem combinadas, formam tópicos. Tais modelos encontram aplicação em uma série de contextos, incluindo análise de sentimentos, classificação de documentos e recomendação de conteúdo, promovendo a descoberta automática de tópicos sem a necessidade de rotulação manual (Dieng; Wang *et al.*, 2017).

Os benefícios das Modelagens Neurais de Tópicos são diversos. Eles incluem a habilidade de capturar relações semânticas complexas e proporcionar tópicos mais coesos e interpretáveis, além de lidar eficientemente com dados não estruturados. Entretanto, desafios permanecem, como a necessidade de grandes volumes de dados para obter resultados satisfatórios, limitações na interpretabilidade de modelos de grande escala e dificuldades em lidar com ambiguidade e polissemia (Dieng; Wang *et al.*, 2017).

Em suma, as NTM oferecem vantagens substanciais na análise semântica de textos. Contudo, é fundamental reconhecer e endereçar suas limitações, as quais incluem a dependência de vastos conjuntos de dados, desafios interpretativos em modelos complexos e a gestão de ambiguidades linguísticas. Uma avaliação criteriosa das capacidades e restrições desses modelos é essencial antes de sua implementação em contextos específicos.

2.2.2.12 Bidirectional Encoder Representations from Transformers

O *Bidirectional Encoder Representations from Transformers* (BERT) é um modelo de linguagem construído sob o domínio do *deep learning*, proposto por Devlin *et al.* (2019) com o propósito de aprimorar a compreensão da linguagem humana por meio de uma representação

bidirecional aprimorada de sentenças. Baseando-se na arquitetura de *Transformers*[21], que permitem a consideração das relações de dependência entre as palavras em uma frase, o BERT é capaz de capturar o contexto das palavras tanto retrospectiva quanto prospectivamente.

O treinamento prévio do BERT é realizado com um volume substancial de dados não rotulados, empregando a técnica de predição de palavras mascaradas e o mecanismo de *next sentence prediction*, que treina o modelo para prever se uma frase é sequencial a outra. Esse processo resulta em um modelo capaz de capturar informações contextuais complexas, aplicáveis a uma diversidade de tarefas de processamento de linguagem natural, como classificação de texto e extração de informações, nas quais o BERT tem demonstrado desempenho de vanguarda (Devlin *et al.*, 2019).

As vantagens do BERT incluem sua capacidade de processar a linguagem de maneira bidirecional, atribuir representações vetoriais ricas a palavras em uma frase, ser pré-treinado em vastos conjuntos de dados e sua aplicabilidade transversal em diferentes tarefas e domínios por meio do transferência de aprendizado. Contudo, o modelo exige recursos computacionais significativos, enfrenta desafios ao lidar com textos longos e depende da disponibilidade de dados rotulados para treinamento e ajuste fino (Devlin *et al.*, 2019).

Em síntese, o BERT representa um avanço significativo na modelagem de linguagem, capacitando a captura de nuances contextuais de forma bidirecional e a aplicação em múltiplas tarefas de processamento de linguagem natural. Não obstante, o modelo impõe desafios relacionados a recursos computacionais e tratamento de textos extensos, além da dependência de dados rotulados para treinamento.

2.2.2.13 Top2Vec

Top2Vec é um algoritmo de representação de tópicos distribuídos proposto por introduzido por Angelov (2020). A abordagem do modelo visa extrair e representar eficientemente tópicos em extensas coleções de documentos, diferenciando-se de métodos tradicionais de modelagem de tópicos, como Latent Dirichlet Allocation (LDA) ou Non-Negative Matrix Factorization (NMF), que geram representações de tópicos baseadas em estatísticas. Em vez disso, o Top2Vec emprega uma abordagem baseada em vetores distribuídos.

No que concerne à sua metodologia, o algoritmo opera em dois passos distintos: o agrupamento e a análise dos documentos. Inicialmente, o Top2Vec emprega um algoritmo de agrupamento para criar *clusters* de documentos com características semelhantes. Após essa etapa,

[21] A arquitetura de *Transformers* é um tipo de modelo de aprendizado profundo que foi introduzido por Vaswani *et al.* (2017) no artigo *Attention Is All You Need*. Essencialmente, os *Transformers* são baseados em mecanismos de atenção que permitem ponderar a importância relativa de diferentes partes da entrada de dados. Em vez de processar sequencialmente cada elemento de entrada, os *Transformers* operam em conjunto sobre todos os elementos de entrada, permitindo que o modelo aprenda as dependências entre eles, independentemente da distância nas sequências. Isso os torna particularmente eficazes para tarefas de processamento de linguagem natural, como tradução automática e compreensão de texto, onde a compreensão do contexto e das relações entre as palavras é crucial.

um modelo de palavras vetoriais é treinado com base nos documentos de cada cluster, resultando na geração de uma representação vetorial para cada tópico (Angelov, 2020).

Uma característica notável do Top2Vec é sua capacidade de identificar tópicos emergentes que não estavam evidentes durante o treinamento do modelo. Tal característica decorre do fato de que o algoritmo atribui vetores tanto para os documentos quanto para os tópicos, proporcionando uma adaptabilidade a novas informações. Ademais, o Top2Vec exibe vantagens significativas em termos de escalabilidade e eficiência computacional, pois, ao distribuir o processamento em múltiplas unidades, possibilita o processamento de grandes volumes de dados de maneira mais célere e eficaz (Angelov, 2020).

Em suma, o Top2Vec emerge como um algoritmo inovador no campo da representação de tópicos distribuídos, oferecendo vantagens em termos de flexibilidade, escalabilidade e eficiência computacional. Entretanto, o algoritmo não está isento de limitações, enfrentando desafios relacionados à escalabilidade em grandes conjuntos de dados, dependência de ajustes de hiperparâmetros e sensibilidade à qualidade dos dados de entrada. Tais desafios exigem atenção e abordagens meticulosas para garantir a acurácia e eficácia do Top2Vec em aplicações práticas.

2.2.2.14 BERTopic

O BERTopic é um modelo de aprendizado profundo que incorpora a tecnologia BERT (subseção 2.2.2.12) para a modelagem de tópicos em documentos. Foi idealizado por Grootendorst (2022) que descreve o método como uma modelagem de tópicos neural que utiliza uma variante de *Frequency-Inverse Document Frequency* (TF-IDF)[22] baseada em classes.

A tecnologia BERT, fundamental para o BERTopic, consiste em um modelo de linguagem natural pré-treinado que adota uma arquitetura de rede neural de *transformers*. Esta arquitetura possibilita que a codificação de palavras considere o contexto bilateral, melhorando a compreensão e contextualização do significado das palavras em um texto. O modelo utiliza as representações codificadas fornecidas pelo BERT, aplicando um procedimento TF-IDF baseado em classes para designar tópicos aos documentos, que o autor denominou de c-TF-IDF (Grootendorst, 2022).

A metodologia de classes do BERTopic implica no agrupamento prévio dos documentos em categorias semelhantes antes da aplicação do TF-IDF. Tal estratégia permite que o modelo considere a similaridade entre os documentos de uma mesma classe ao atribuir tópicos, baseando-se nessas relações. Esse modelo apresenta vantagens como a eficiência na extração de tópicos, a capacidade de representar semanticamente os textos por meio de vetores pré-treinados do

[22] O TF-IDF é uma técnica estatística utilizada para avaliar a importância de uma palavra em um documento, que faz parte de um *corpus*. Calcula-se multiplicando duas métricas: a frequência do termo (TF), que é o número de vezes que uma palavra aparece em um documento, e a frequência inversa do documento (IDF), que é o logaritmo do número de documentos no corpus dividido pelo número de documentos que contêm a palavra (Grootendorst, 2022, p. 3). Este método destaca palavras que são mais relevantes, atribuindo a elas pesos mais altos enquanto palavras comuns, que aparecem em muitos documentos, recebem pesos mais baixos.

BERT, e a flexibilidade na configuração dos hiperparâmetros. Todavia, enfrenta limitações como a dependência de recursos computacionais, a sensibilidade ao tamanho do *corpus* e a necessidade de ajuste fino dos hiperparâmetros (Grootendorst, 2022).

Generalizando, o BERTopic é uma abordagem avançada de aprendizado de máquina que alia a capacidade de codificação contextual do BERT com uma técnica de TF-IDF baseada em classes, visando uma modelagem de tópicos mais precisa e contextualizada em corpora documentais. Apesar das vantagens consideráveis, suas limitações demandam atenção no que tange aos recursos computacionais e ao ajuste de hiperparâmetros para uma performance otimizada.

2.3 Teoria da Organização Industrial e o antitruste brasileiro

Na sequência deste documento, serão apresentados alguns elementos fundamentais da Organização Industrial. Esse introito é crucial para uma compreensão abrangente da disciplina e serve como base para o desenvolvimento de análises mais avançadas. Serão abordados temas essenciais da Organização Industrial, que incluem as estruturas de mercado, o comportamento de práticas colusivas e a atuação da autoridade antitruste brasileira. Além disso, haverá uma ênfase especial na interação entre inovação e dinâmicas econômicas. A subseção final deste trabalho apresentará as aplicações práticas da inteligência artificial no contexto do setor público antitruste, fornecendo uma visão holística e aplicada da Organização Industrial, significativa tanto para o âmbito acadêmico quanto para as práticas regulatórias no setor público.

2.3.1 Generalidades sobre Organização Industrial

Segundo Tirole (2019, p. 361), no capítulo que discute sobre os desafios da indústria, a Organização Industrial tem como foco o estudo de como o poder de mercado é exercido e regulado. Para esse autor, isso é feito através do desenvolvimento de modelos que capturam os aspectos cruciais de diferentes cenários. Ainda, as previsões geradas por esses modelos são submetidas a testes econômicos, seja em ambientes controlados como laboratórios ou em situações reais. Para serem eficazes, esses modelos devem se basear em suposições plausíveis e gerar previsões consistentes que sejam corroboradas por dados empíricos. Com base nisso, os economistas estão em posição de oferecer recomendações políticas ou estratégias de negócios com maior segurança (Tirole, 2019, p. 361).

Para Tirole (2019, p. 361–362), a Organização Industrial tem suas raízes na França do século XIX através dos trabalhos de Cournot e Dupuit, e se desenvolveu ao longo do tempo, abrangendo desde questões de preços e valoração de serviços até teorias complexas de segmentação de mercado. Dupuit, por exemplo, introduziu o conceito de "excedente do consumidor", analisando as escolhas dos consumidores em relação à qualidade e preço dos serviços, como ilustrado em sua análise sobre a qualidade dos vagões de trem. Essas ideias iniciais evoluíram para aplicativos em vários campos, desde transportes até software (Tirole,

2019, p. 361–362).

Para o multicitado autor, no final do século XIX, com a introdução da Lei *Antitrust Sherman* nos EUA, a Organização Industrial começou a influenciar as políticas públicas, focando em limitar práticas anticompetitivas. Esse enfoque foi fortalecido pela Escola de Harvard, que defendeu a intervenção pública no mercado. Posteriormente, nas décadas de 1960 e 1970, a Escola de Chicago questionou a base teórica de muitas políticas antitruste, iniciando uma contra-revolução na área. Essa crítica levou a uma reavaliação das ideias subjacentes à legislação e regulamentação da concorrência, culminando em um trabalho que fortaleceu o argumento para a intervenção pública, principalmente nos anos 70 e 80 (Tirole, 2019, p. 361–362).

Dessa forma, assume-se que a Organização Industrial é um ramo da economia que estuda a estrutura e o funcionamento dos mercados, com foco nas interações entre as empresas e o comportamento estratégico que adotam. Alguns dos principais conceitos da Organização Industrial incluem:

a) Estrutura de mercado: refere-se à organização e características dos mercados, como o número de empresas, a existência de barreiras à entrada, a diferenciação de produtos e a presença de poder de mercado;

b) Monopólio: é uma estrutura de mercado em que uma única empresa detém o controle total sobre a oferta de um determinado produto ou serviço, o que lhe confere poder de mercado significativo;

c) Oligopólio: é uma estrutura de mercado em que um pequeno número de empresas domina a oferta de um produto ou serviço. Essas empresas podem interagir estrategicamente, levando em consideração as ações das concorrentes ao tomar decisões;

d) Concorrência perfeita: é uma estrutura de mercado em que há um grande número de empresas que produzem produtos homogêneos, não havendo poder de mercado significativo. Nesse tipo de mercado, as empresas são tomadoras de preço;

e) Comportamento estratégico das empresas: refere-se às ações e decisões tomadas pelas empresas para maximizar seus lucros em um ambiente competitivo. Isso inclui a definição de preços, a diferenciação de produtos, a inovação, a entrada em novos mercados, entre outros aspectos.

Leal e Figueiredo (2018) examina as estratégias empresariais, focando em como as empresas procuram otimizar seus lucros e ampliar sua participação no mercado. A discussão engloba teorias relevantes como oligopólio e jogos estratégicos, destacando especialmente a influência crítica da informação e inovação na dinâmica competitiva industrial. O autor analisa o impacto da oligopolização nos investimentos em Pesquisa e Desenvolvimento (P&D) a nível empresarial, conforme ilustrado em um quadro específico (Leal; Figueiredo, 2018, p. 30).

Adicionalmente, aborda a relevância das estratégias nacionais de inovação, ressaltando que as dinâmicas de mercado por si só podem não ser suficientes para promover a implementação de inovações nas empresas (Leal; Figueiredo, 2018, p. 21).

Casos de sucesso em inovação são mencionados, incluindo setores como pesquisa agrícola e exploração de petróleo, bem como empresas como Embraer e Vale (Leal; Figueiredo, 2018, p. 6). No entanto, o autor também aponta a excessiva oligopolização, parcialmente causada pela estrutura tributária e desincentivos aos investimentos em inovação, como um obstáculo (Leal; Figueiredo, 2018, p. 8). A importância das capacidades tecnológicas para a inovação, tanto em nível empresarial quanto nacional, é também enfatizada (Leal; Figueiredo, 2018, p. 27, 29), assim como a necessidade de políticas de inovação que abordem tanto a oferta quanto a demanda (Leal; Figueiredo, 2018, p. 22). Além disso, a colaboração entre empresas e parceiros diversos é destacada como crucial para o processo inovativo (Leal; Figueiredo, 2018, p. 27, 29).

Leal e Figueiredo (2018) aborda principalmente a inovação tecnológica e suas implicações para o desenvolvimento econômico brasileiro. O foco recai sobre a importância das capacidades tecnológicas das empresas e países para impulsionar a produtividade e agregar valor econômico (Leal; Figueiredo, 2018, p. 29). O autor discute a necessidade de políticas públicas voltadas para a inovação tecnológica, buscando aumentar o retorno social dos investimentos (Leal; Figueiredo, 2018, p. 28).

Contrariando a percepção de que a inovação tecnológica ocorre primordialmente nas universidades, Leal destaca o papel fundamental das empresas neste processo (Leal; Figueiredo, 2018, p. 12). A inovação é apresentada não apenas como aplicação do conhecimento científico, mas também como agregação de valor e solução de problemas sociais (Leal; Figueiredo, 2018, p. 21). Embora o artigo não trate especificamente da relação entre inovação tecnológica e Organização Industrial, ele ressalta a importância de políticas de inovação orientadas para a demanda, enfatizando as aplicações comerciais das inovações (Leal; Figueiredo, 2018, p. 22).

Quanto à propriedade intelectual e à regulação econômica em ambientes de alta tecnologia, embora o artigo não aborde diretamente esses aspectos, enfatiza-se a necessidade de investimentos governamentais eficazes e de uma compreensão comum entre os formuladores de políticas sobre os papéis dos atores no processo de inovação (Leal; Figueiredo, 2018, p. 28). Esses elementos podem ser indiretamente relacionados à criação de um ambiente regulatório favorável e à proteção dos direitos de propriedade intelectual para fomentar a inovação tecnológica.

Em síntese, a Organização Industrial é uma área de estudo que analisa o comportamento das empresas e a estrutura dos mercados, buscando entender como esses fatores afetam a concorrência e o desempenho econômico. No setor público, ela é relevante para a análise de políticas e regulação, visando promover a eficiência e o bem-estar social.

2.3.2 Sobre a concorrência

para Tirole (2019, p. 362–364), nem sempre a concorrência é benéfica para o mercado. O autor justifica a afirmação explicando que a concorrência pode, de fato, levar à duplicação desnecessária de custos, especialmente em setores com altos custos fixos e efeitos de rede, como infraestruturas de transporte e comunicação. Por exemplo, a construção de múltiplas redes de distribuição elétrica ou linhas ferroviárias paralelas pode não ser prática ou econômica. De outro lado, ainda segundo o autor, a existência de um monopólio natural em um ponto da cadeia de valor não deve transformar todo o setor em monopólio. Em certos casos, como nas redes de telecomunicações ou ferrovias, é crucial que o proprietário da infraestrutura forneça acesso não discriminatório a operadores concorrentes (Tirole, 2019, p. 362).

Em termos de políticas públicas, Tirole (2019, p. 363) aponta que foi observada uma tendência de separar a infraestrutura do serviço para promover a concorrência nos segmentos do mercado que são potencialmente competitivos. Para o autor, exemplos históricos incluem o desmantelamento da AT&T em 1984 e a separação de aeroportos das companhias aéreas. No entanto, também houve exemplos de introdução de concorrência por razões ideológicas que não tiveram resultados positivos, como as "guerras de ônibus"no Reino Unido (Tirole, 2019, p. 363).

Por fim, Tirole (2019, p. 364) afirma que a concorrência pelo mercado (licitação para um contrato de concessão) pode substituir a concorrência diária em certos serviços públicos, como abastecimento de água ou transporte público. Por fim, é crucial que a concorrência no mercado beneficie os usuários e não seja distorcida por práticas desleais. A supervisão do comportamento no mercado é uma parte essencial do direito da concorrência, garantindo que a competição seja justa e beneficie os consumidores (Tirole, 2019, p. 364).

Já para Silva Filho e Haro (2013, p. 1), a competição de mercado é um fenômeno que ocorre quando diversas empresas de um mesmo setor disputam as mesmas demandas do mercado. É importante ressaltar que a competição não é um objetivo primário para as empresas; ela surge como um subproduto da busca por seus próprios interesses. Nesse contexto, a presença de concorrência varia conforme o cenário econômico específico. Por exemplo, em um pequeno município com apenas uma oficina mecânica detendo um monopólio legal, a concorrência é praticamente inexistente. No entanto, a legislação proíbe a obtenção de monopólio por meios anticoncorrenciais, o que requer a intervenção do CADE (Silva Filho; Haro, 2013, p. 1).

Historicamente, conforme observado por Salomão Filho (2002, p. 58–60), Silva Filho e Haro (2013, p. 2) e Aguillar (2006, p. 224), o mercado dos Estados Unidos no século XIX era marcado por auto-regulação, resultando em práticas abusivas dos capitalistas contra os consumidores. A alta concentração industrial no nordeste, o monopólio das ferrovias e o descontentamento do setor agrário foram fatores que contribuíram para a criação do *Sherman Act* em 1890, uma legislação pioneira e influente no direito antitruste.

A Constituição de 1934 foi a primeira no Brasil a incluir um capítulo dedicado ao direito

econômico empresarial, apesar de não focar em normativas de concorrência (Silva Filho; Haro, 2013, p. 3; Salomão Filho, 2002, p. 73). Posteriormente, a legislação de 1946, que pretendia orientar o Direito Antitruste, não teve eficácia prática. A lei nº 4.137 de 1962, entretanto, estabeleceu o CADE, evidenciando o papel do Estado nas relações comerciais e visando prevenir o abuso de poder de mercado e práticas anticoncorrenciais. Adicionalmente, a Constituição ressalta a proteção ao livre comércio e concorrência no artigo 170, garantindo o livre exercício de atividades econômicas, exceto em casos especificados por lei (Silva Filho; Haro, 2013, p. 4).

Conforme destacado por Carvalho (2015b, p. 98), a promoção da competitividade é crucial para a sustentabilidade dos mercados. As políticas antitruste são reconhecidas globalmente como essenciais na política econômica, assegurando a competitividade econômica. Ademais, a defesa da concorrência é vital para o bom funcionamento do mercado, não apenas proporcionando preços mais baixos, mas também incentivando a qualidade, diversidade e inovação dos produtos. Isso beneficia o consumidor e promove o desenvolvimento econômico, sendo um princípio fundamental da ordem econômica brasileira, conforme estabelecido na Constituição (Carvalho, 2015b, p. 98).

2.3.2.1 Legislação nacional correlata

Nos termos defendidos por Carvalho (2015b, p. 99), no ano de 1990 foram implementadas uma série de reformas legislativas, incluindo a privatização, liberação dos preços e regulação do mercado; em 1994, adotou-se o plano Real, além de se aprovar a Lei n° 8.884/1994, denominada de Lei da Defesa da Concorrência, que fez o órgão antitruste nacional ganhar reputação, visto que suas decisões refletiram-se na política da concorrência.

Na abordagem de Carvalho (2015b), é destacado que os esforços anticoncorrenciais no Brasil eram divididos entre três entidades distintas: a Secretaria de Acompanhamento Econômico do Ministério da Fazenda (SEAE), a Secretaria de Direito Econômico do Ministério da Justiça (SDE) e o Conselho Administrativo de Defesa Econômica (CADE), este último uma autarquia associada ao Ministério da Justiça. Essas instituições, conforme apontado pelo autor, geriam a política antitruste do país de maneira pouco eficaz. O autor ressalta ainda que, naquela época, havia uma ênfase maior nos atos de concentração, que não eram considerados uma ameaça significativa à concorrência. Por outro lado, havia uma menor atenção ao combate aos cartéis, apesar de serem práticas com um impacto mais profundo no mercado (Carvalho, 2015b, p. 99).

A atual estrutura legal brasileira definidora do campo de atuação das autoridades na defesa da concorrência é a seguinte:

a) artigos 170 e 173 da CRFB/1988;

b) Lei Delegada nº 4, de 16/09/1962;

c) Lei n. 12.529/2011: Lei Antitruste e controle do poder econômico;

d) Lei n. 8.176, de 8/02/1991: Crimes contra a ordem econômica;

e) Lei n. 8.137, de 27/12/1990: Crimes contra a ordem econômica.

Observa-se a existência de alguma preocupação no legislador originário quando erigiu a defesa da ordem econômica a categoria de norma constitucional. Isto se justifica na *mens legis* [23] normativa que insere o arcabouço cogente já na Carta Magna, pautando-se na necessidade de intervenção estatal para regular o mercado quando da ocorrência de alguma falha.

2.3.3 Estruturas de Mercado

Uma das pedras angulares da Organização Industrial é o estudo das estruturas de mercado. Em mercados de concorrência perfeita, muitas empresas vendem produtos homogêneos, enquanto em oligopólios, poucas empresas dominam o mercado, e em monopólios, uma única empresa controla todo o mercado. A análise dessas estruturas ajuda a entender como a concentração de mercado afeta preços, produção e eficiência econômica.

Estas estruturas se apresentam como configurações de mercado que mostram como os diferentes mercados são organizados, destacando aspectos essenciais da interação entre oferta e demanda a partir de determinadas hipóteses teóricas de comportamento das firmas e dos consumidores e de características encontradas nos mercados no mundo real, tais como poder de mercado, políticas de preço, interdependência entre as firmas, condições de maximização de lucros e quantidade produzida. A análise dessas características permite determinar suas implicações quanto ao nível de bem-estar social, à qualidade da distribuição de renda e à maior ou menor necessidade de regulação estatal do mercado em questão

A defesa da concorrência deve ser uma prática constante de Estado. As práticas anticoncorrenciais dever ser submetidas ao controle estatal para não turbar o mercado ou a economia do país. Essas práticas podem ser classificadas nas suas formas mais conhecidas. Esta pesquisa tomará as estruturas de referência citadas por Rossetti (1997, p. 406), como: concorrência perfeita, concorrência monopolística, oligopólio e monopólio. Cada uma tem características próprias que nos permitem as diferenciar uma das outras conforme será pormenorizado nos próximos subtópicos.

2.3.3.1 Mercados em concorrência perfeita

O termo "concorrência" pode ser definida como o processo de rivalidade entre os agentes econômicos, a competição do mercado. Expressa-se em termos de preço, qualidade, diversidade ou qualquer outra característica economicamente relevante. Nesse instituto, as forças de mercado atuam livremente para mitigar a escassez. Ou seja, o objetivo da concorrência e garantir que os recursos limitados sejam usados o mais eficientemente possível. De outro modo, o objetivo é maximizar o bem-estar social (Martinez, 2014, p. 5).

[23] Segundo Conselho Nacional do Ministério Público (2021), essa expressão latina pode ser interpretada como o espírito ou finalidade da lei.

Em um mercado no estado de concorrência perfeita, a diferença entre os preços praticados pelo produtor e seus custos marginais será a menor possível. Em termos técnicos, em um mercado altamente competitivo, o preço é fixado no ponto em que iguala ao custo marginal de produção, incluída aí a remuneração de seu custo de capital. Ou seja, em um ambiente de concorrência perfeita, os preços cobrados dos consumidores são aqueles necessários à remuneração do capital empregado na produção.

Segundo Martinez (2014, p. 5–6), pode-se elencar cinco pressupostos para a existência de um mercado em concorrência perfeita:

a) ausência de informação assimétrica;

b) ausência de economias de escala no longo prazo;

c) maximização da utilidade pelos consumidores e de do lucro pelos produtores;

d) produtores atuando como tomadores de preço;

e) correspondência entre o preço praticado e custo marginal da produção.

A mesma autora esclarece que com ausência de custos de entrada e saída significativos, pode ocorrer concorrência perfeita mesmo em mercados em que apenas um agente oferte produtos e/ou serviços. A explicação é que a entrada de novas empresas concorrentes pode ser interpretada pelo agente econômico como passível de ocorrer. Sendo assim, Martinez (2014, p. 5) exemplifica que "caso a empresa pratique preços de monopólio, funcionando como pressão para garantir um mercado com características competitivas. Nesse ambiente de concorrência perfeita (ainda que potencial), os recursos disponíveis na economia são alocados aos usos que melhor refletem as preferências dos consumidores".

Segundo Aguiar, Daher e Tabak (2018, p. 191), a esta lista se pode acrescentar a existência de bem homogêneo, ou padronizado. Ademais, Esses autores esclarecem que a concorrência perfeita é uma estrutura ideal, de forma a, caso implementada, garantir um maior bem-estar e impelindo o Estado a agir caso não haja estruturas de mercado mais concentradas (Aguiar; Daher; Tabak, 2018, p. 192).

2.3.3.2 Mercados em concorrência imperfeita

Tendo a concorrência perfeita como uma abstração teórica, passaremos a abordar os mercados com estruturas de concorrência imperfeita.

I - O oligopólio

Oligopólio a situação onde existe um pequeno número de vendedores no mercado, ou onde pode existir um maior variedade de vendedores, mas uma pequena parcela destes domina o mercado. Embora não haja barreiras, de forma explícita, o poder de mercado[24] das empresas

[24] O poder de mercado é comumente definido como a diferença entre o preço cobrado por uma firma e seu custo

oligopolistas é desestimulante à entrada de novas empresas neste mercado. Por existirem poucas firmas nessa estrutura, os oligopolistas podem se unir para evitar a concorrência entre eles, além de poder impor um preço ao mercado. A assimetria de informações nesse cenário podem acessoa às mesmas informações sobre o processo produtivo, níveis de oferta e outros, mas em razão da rivalidade existente, não há completude nas informações.

Montella (2012, p. 104) afirma que existem três tipos de oligopólio:

a) oligopólio concentrado: caracterizado pela alta concentração e homogeneidade do produto;

b) oligopólio diferenciado: tem concentração mais baixa e elevada diferenciação entre os produtos;

c) oligopólio misto: uma combinação de "a"e "b".

Segundo Samuelson e Marks (2012, p. 349–392), na esfera dos oligopólios, a tomada de decisões ótimas é fundamentalmente ancorada na capacidade de antecipar as ações dos concorrentes. Isso se reflete no modelo de empresa dominante, onde empresas menores operam sob a premissa de preços dados, ajustando suas quantidades produzidas com base nessa percepção. A empresa dominante, por sua vez, aproveita essa dinâmica ao maximizar seu lucro através da determinação estratégica da quantidade e do preço, seguindo a lógica de que a Receita Marginal deve igualar o Custo Marginal $(R_M = C_M)$, ao longo de sua curva de demanda líquida.

No contexto de uma competição simétrica entre oligopolistas, como proposto no modelo de Cournot[25], cada empresa busca maximizar seus lucros através da previsão das quantidades ótimas estabelecidas pelos rivais. Esta dinâmica ilustra a complexidade inerente à competição oligopolista, onde a interdependência das decisões é uma constante (Samuelson; Marks, 2012, p. 349–392).

A competição por preços em tal ambiente pode rapidamente se transformar em um dilema similar ao do prisioneiro, onde a busca pelo ideal individual leva a reduções de preços e, consequentemente, a uma diminuição dos lucros coletivos. Nesse sentido, a publicidade emerge como um componente crucial, devendo ser perseguida até que o lucro marginal obtido com o aumento das vendas se iguale ao custo marginal do último dólar investido em publicidade (Samuelson; Marks, 2012, p. 349–392).

O oligopólio é definido por um mercado controlado por um número limitado de empresas, cujos lucros estão intrinsecamente ligados às ações de seus concorrentes. Medidas como a razão de concentração e o Índice Herfindahl-Hirschman [26] (HHI) fornecem *insights* sobre a estrutura

marginal de produção. Para Castro (2017, p. 14), também é denominada poder de monopólio, termos que são usados indistintamente e de forma intercambiável.

[25] O modelo desenvolvido pelo economista francês Augustin Cournot nos anos 1838 ao observar a concorrência em um duopólio. Mas seus resultados também se aplicam a oligopólios (Tirole, 1988).

[26] O índice Herfindahl-Hirschman, representa uma métrica utilizada para avaliar a magnitude das empresas em

do mercado, indicando o grau de concentração e, por extensão, o potencial impacto sobre preços e lucros. Altos índices refletem um mercado mais concentrado, onde é razoável esperar que aumentos na concentração estejam correlacionados com elevações nos preços e nos lucros (Samuelson; Marks, 2012, p. 349–392).

A dinâmica de competição quantitativa é delineada por dois modelos principais: a competição com uma empresa dominante e a competição entre iguais. Em ambos, as quantidades de equilíbrio são ajustadas de tal forma que nenhuma empresa teria vantagem ao alterar sua produção planejada. Especificamente, no modelo de fixação de quantidade, o equilíbrio tende ao resultado perfeitamente competitivo conforme o número de empresas idênticas aumenta (Samuelson; Marks, 2012, p. 349–392).

Por fim, Samuelson e Marks (2012) afirmam que a expectativa de que reduções de preços sejam correspondidas pelos rivais, mas não os aumentos, introduz uma distorção na curva de demanda da empresa. Isso resulta em uma relativa estabilidade de preços, uma vez que alterações tendem a ser não lucrativas, evidenciando a complexidade e a sutileza das estratégias em contextos oligopolistas.

II - O monopólio

O monopólio é uma estrutura de mercado que se encontra no extremo oposto à concorrência perfeita. A sua característica principal é a existência de uma única firma vendedora de um produto que não tenha substitutos próximos. Assim, esse vendedor tem o poder de fixar o preço que melhor lhe aprouver. Outra característica é a existência de barreiras que impedem o surgimento de competidores. Tais barreiras, segundo Montella (2012, p. 103), dizem respeito:

a) à existência de economias de escala;

b) ao controle sobre o fornecimento de matérias-primas;

c) à posse de patentes;

d) concessão de monopólio legal.

Segundo Samuelson e Marks (2012, p. 320–321), um monopólio puro é caracterizado por ser um mercado com apenas um vendedor, ou seja, uma única empresa. Ao analisar o mercado estadunidense, esse autor afirma que que monopólios puros são extremamente raros, representando menos de 3% do Produto Interno Bruto (PIB) dos Estados Unidos, conforme definido por mercados nos quais uma única empresa detém 90% ou mais da participação de mercado. Não obstante, apesar de sua raridade, o estudo do monopólio puro é significativo tanto por sua própria natureza quanto pela sua aplicabilidade em situações de quase monopólios, onde poucas empresas dominam o mercado. O modelo de monopólio também é útil para compreender

relação ao conjunto de sua indústria, servindo como um indicativo do nível de competição existente entre elas. Nomeado em homenagem aos economistas Orris C. Herfindahl e Albert O. Hirschman, esse índice é um instrumento econômico de grande relevância no contexto da implementação de políticas de defesa da concorrência, na regulamentação antitruste e na administração da inovação tecnológica.

o comportamento de cartéis, que são grupos de produtores que determinam preços e produção de forma conjunta (Samuelson; Marks, 2012, p. 321).

Ao analisar o monopólio, dois aspectos principais devem ser considerados: o comportamento monopolista, que diz respeito a como um monopolista maximizador de lucros define preço e produção, e a existência de barreiras de entrada, que são fatores que impedem a entrada de novas empresas no mercado, garantindo que o monopolista não enfrente concorrência direta. No que tange à decisão sobre preço e produção, o monopolista, por ser o único produtor, tem a liberdade de aumentar o preço sem preocupar-se com a perda de vendas para concorrentes, pois não existem. Entretanto, isso não significa que o preço possa ser elevado indefinidamente, pois a política ótima de preço e produção depende da demanda de mercado (Samuelson; Marks, 2012, p. 321).

Como o monopolista representa a própria indústria, sua curva de demanda é a curva de demanda do mercado. A partir das informações sobre demanda e custo, é possível prever o preço e a produção monopolistas. O monopolista deve ajustar sua produção de modo que a receita marginal, derivada da curva de demanda do mercado, iguale-se ao custo marginal de produção. O lucro total excessivo do monopolista é representado pela área de um retângulo específico no gráfico, sendo o produto do lucro por unidade e da produção total (Samuelson; Marks, 2012, p. 321).

Ressalta-se que o monopólio permite à empresa obter lucros superiores aos que teria em um mercado competitivo, onde os lucros econômicos tendem a ser nulos no longo prazo. Contudo, o montante do lucro excessivo depende diretamente da relação entre a demanda de mercado e os custos. Por exemplo, se existirem substitutos próximos para o produto do monopolista, a demanda pode ser relativamente elástica, oferecendo pouca margem para lucros excessivos. Assim, para aumentar significativamente seus lucros, a empresa monopolista precisa reduzir seus custos médios de produção ou aumentar a demanda de mercado, embora seja possível que não exista demanda alguma para um produto único do monopolista. Isso destaca que, embora o monopólio puro permita a obtenção de lucro excessivo, o tamanho real desse lucro varia conforme a comparação entre demanda e custo.

III - A concorrência monopolística

A concorrência monopolística, por sua vez, é um *mix* dos conceitos de monopólio e concorrência perfeita. Dessa forma, apresenta um grande número de ofertantes vendedores com alguma diferenciação de produtos. Explicando melhor, existe um grande número de vendedores que se comportam, em razão dos produtos serem diferenciados, como monopolistas (Montella, 2012, p. 105–104).

Na Tabela 3 (p. 46), pode-se visualizar as figuras de mercado referenciadas nessa pesquisa, até o momento.

Dada as características de coordenação existente nos monopólios, estas estruturas de

Tabela 3 – Estruturas de mercado abordadas

Características	**Concorrência**		**Oligopólio**	**Monopólio**
	Perfeita	**Monopolista**		
Qtde. vendedores	Muitos pequenos	Muitos pequenos	Poucos grandes	Um grande
Qtde. compradores	Muitos pequenos	Muitos pequenos	Muitos pequenos	Muitos pequenos
Tipo do produto	Homogêneos	Diferenciado	Semelhante	Único
Controle de preços	Não há	Variável	Muito do vendedor	Total do vendedor
Barreiras à entrada	Não há	Variável	Várias	Todas
Fluxo de informação	Total	Boa	Limitada	Ausente

Fonte: Adaptado de Rossetti (1997, p. 407).

mercado são pontos fulcrais para o estabelecimento das chamadas condutas colusivas, tratadas com mais pormenor nas subseções que se seguem.

2.3.4 Sobre a concorrência e práticas colusivas

A concorrência, um pilar fundamental do mercado econômico, refere-se à interação competitiva entre empresas no mesmo segmento de mercado. Este fenômeno, longe de ser um objetivo empresarial direto, emerge como um resultado natural da busca por objetivos corporativos individuais. Este tópico explora a natureza multifacetada da concorrência, suas implicações legais e históricas, e sua relevância na ordem econômica contemporânea (Silva Filho; Haro, 2013, p. 1).

Em primeiro lugar, a compreensão da concorrência exige uma análise contextualizada. Em cenários onde a concorrência é inexistente, como em pequenas localidades com empresas monopolistas legais, surge a necessidade de regulação antitruste para assegurar a competitividade do mercado, um papel exercido pelo CADE no Brasil (Silva Filho; Haro, 2013, p. 1). A história do direito antitruste remonta a 1889 com a legislação canadense, marcando o início de uma era de regulamentações destinadas a coibir práticas comerciais restritivas (Silva Filho; Haro, 2013, p. 2).

Nos Estados Unidos, o *Sherman Act* de 1890 representa um marco na legislação antitruste, influenciado por uma combinação de auto-regulação de mercado, concentração industrial e monopólios em setores-chave. Este ato legislativo surgiu como resposta às dinâmicas econômicas e sociais do século XIX e moldou significativamente o direito econômico (Silva Filho; Haro, 2013, p. 2). No Brasil, a evolução do direito econômico empresarial começou com a Constituição de 1934, seguida pela Lei n. 4.137 de 1962, que estabeleceu o CADE, crucial para a prevenção de abusos de posição dominante e para a proteção da livre concorrência (Silva Filho; Haro, 2013, p. 3).

A Constituição Federal brasileira fortalece ainda mais este arcabouço ao garantir a proteção ao livre comércio e à concorrência. As políticas antitruste, reconhecidas mundialmente, são essenciais para a sustentabilidade e competitividade dos mercados, enfatizando a importância

da livre concorrência para o desenvolvimento econômico e o bem-estar do consumidor (Silva Filho; Haro, 2013, p. 4; Carvalho, 2015a, p. 98).

Sintetizando, a concorrência desempenha um papel vital na manutenção da saúde e eficiência dos mercados econômicos. As leis e regulamentos antitruste, tanto em contextos históricos quanto contemporâneos, são fundamentais para garantir um ambiente de mercado justo e equilibrado, promovendo inovação, diversidade e preços acessíveis para os consumidores. Este equilíbrio, entre a liberdade de mercado e a regulamentação, é essencial para o desenvolvimento econômico sustentável e para a proteção dos interesses dos consumidores, conforme estabelecido na ordem econômica brasileira (Carvalho, 2015a, p. 98).

2.3.4.1 Conceito e tipologia da conduta de cartel

Dentre as práticas anticompetitivas, destacam-se a formação de cartel, condutas uniformes e unilaterais. Estas práticas incluem a combinação de preços, adoção de tabelas de preços por associações e sindicatos, e ações abusivas por agentes dominantes no mercado, como a criação de barreiras à entrada e imposição de preços predatórios.

O conceito de cartel, entendido como uma das mais graves lesões à concorrência, é amplamente discutido na literatura e jurisprudência antitruste. Esta conduta, que constitui uma infração à Ordem Econômica, está tipificada na Lei n. 12.529, em seus diversos incisos (Cade, 2009, p. 6). Este tópico visa elucidar a natureza e as implicações da conduta de cartel, destacando sua tipologia e efeitos no mercado.

De acordo com a legislação vigente, a formação de cartel não se restringe apenas ao ajuste de preços entre concorrentes, mas se estende a outras variáveis de mercado, como quantidades ofertadas, clientes e áreas geográficas. Essa conduta tem como objetivo a simulação de uma entidade única no mercado, limitando a concorrência e manipulando variáveis para obter vantagens ilícitas (Braga, 2015, p. 111).

O cartel se caracteriza pela criação de um pacto de cooperação entre empresas, estabelecendo um controle sobre determinado mercado. Essa prática resulta na fixação de preços, imposição de barreiras à entrada de novas empresas e, consequentemente, na diminuição do bem-estar do consumidor devido à restrição da oferta de bens e serviços a preços competitivos.

As infrações à ordem econômica decorrentes de condutas de cartel, independentemente de culpa, incluem a limitação da livre concorrência, domínio de mercado relevante, aumento arbitrário de lucros e exercício abusivo de posição dominante. A posição dominante é presumida quando uma empresa ou grupo de empresas controla 20% ou mais do mercado relevante, com possibilidade de ajuste desse percentual pelo CADE para setores específicos (Cade, 2016, p. 14, 2019a, p. 15).

Na lição de Martinez (2014, p. 7–8), existem três tipos de ineficiências econômicas promovidas pelos cartéis: alocativa, produtiva e dinâmica. Ao sintetizar o pensamento da citada

autora, Aguiar, Daher e Tabak (2018, p. 195) explicam que **ineficiência alocativa** é a situação em onde há má alocação de recursos - como ambiente está cartelizado, esclarecem que parte do que seria excedente do consumidor vai para o produtor, e outra se transmuda em "peso morto"; **ineficiência produtiva** acontece quando empresas operam com custos mais altos do que no caso em que não haja o cartel; por fim a **ineficiência dinâmica** se traduz ao fato de que existe redução da inovação. Assim, o cartel diminui o interesse dos participantes de se aprimorar em seus processos produtos.

Importante, ainda, registrar que cartéis criam óbices para entrada de novas firmas no mercado (Aguiar; Daher; Tabak, 2018, p. 195). Ademais, o relatório publicado pelo OCDE (2002, p. 72) apresenta pesquisa realizada para levantar e avaliar os números sobre o tamanho do comércio afetado por cartéis em seu países membros, no período compreendido entre os anos de 1996 e 2000, inclusive. Estimou-se que a quantidade de comércio afetado em 16 casos de grande magnitude, que superaram US $ 55 bilhões. A conclusão foi que os preços maiores em torno de 15% a 20% do que em um mercado competitivo.

Ressalte-se que um estudo mais abrangente foi realizado por Connor (2014, p. 86), que examinou mais de 700 estudos econômicos publicados e decisões judiciais sobre 2.041 casos envolvendo cartéis. Considerando cartéis atuantes nos Estados Unidos e em âmbito internacional, entre os anos de 1790 e 2004, constatou uma média de sobrepreço de 23% em todos os períodos analisados (Aguiar; Daher; Tabak, 2018, p. 196).

Portanto, a conduta de cartel representa uma grave ameaça à livre concorrência e ao bem-estar social. As legislações e regulamentações antitruste desempenham um papel crucial na prevenção e punição dessas práticas, visando preservar a integridade e a competitividade do mercado. A análise detalhada das diversas formas de cartel, bem como a compreensão de suas consequências, são essenciais para o desenvolvimento de estratégias eficazes de combate a essas práticas ilícitas no âmbito das licitações e no mercado em geral.

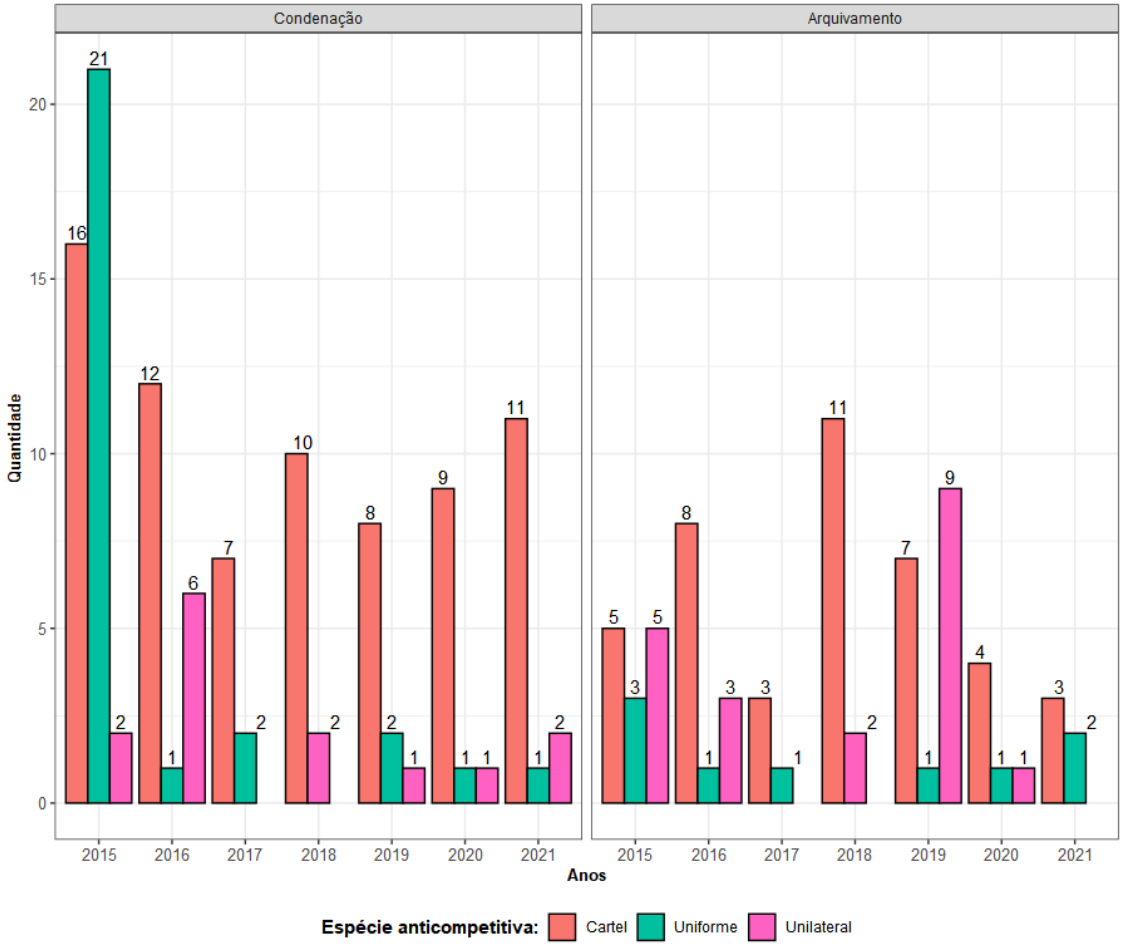
Ainda na fase de qualificação dessa pesquisa, realizada em 2022, mediante acesso à API[27] do CADE, denominada “CADE em Números” (Cade, 2021) foi possível prospectar e obter dados sobre o número de julgamentos desse Órgão Regulador. Uma amostra desses dados foram então consolidados no gráfico de barras contido no Figura 10.

De forma expedita, verificou-se que de 2015 até agosto de 2021 foram protocolados 185 processos referente a atos anticompetitivos distribuídos da seguinte forma: (*i*) 114 de cartel, (*ii*) 37 de conduta uniforme e (*iii*) 34 referentes à conduta comercial unilateral. Nesse mesmo período, foram proferidas 115 condenações e determinado o arquivamento de 70 processos administrativos.

Constatou-se, ainda da leitura do Figura 10, a superioridade numérica dos processos

[27] API, via de regra, é conceituada como um rol de definições e protocolos utilizados no desenvolvimento e na integração de software de aplicações.

Figura 10 – Número de condutas anticompetitivas informadas pelo CADE (2015-2021)



Fonte: elaborado pelo autor com dados da API do Cade.

refentes à carteis. Este grupo somente tem seu quantitativo anual superado no ano de 2015 por quantidade expressiva de processos referentes a conduta uniforme. Não é razoável emitir juízo que justifique esse fato, muito devido ao reduzido interregno abordado (seis anos), bem como pela falta de maior contextualização e cotejamento com outras informações (operações policiais, nova estratégia de atuação etc.). De qualquer sorte, nos períodos posteriores a 2015 vemos o número de processos sobre conduta uniforme oscilar no intervalo [1,2], que parece demonstrar sua verdadeira média de processamento de novos casos.

Verificou-se, pois, a existência de quantidade significativa de processos refentes à cartéis. Isso possibilitou a realização da pesquisa, que foi autorizada a se ater nas decisões de julgamento de conduta de cartéis em licitação, tanto nos casos resultantes de condenação, bem como naqueles processos cujo arquivamento fora determinado. Dessa forma a elaboração da tese oportunizou criar uma robusta base de dados que por si só já é uma contribuição da pesquisa, haja vista a aparência hermética de dados referentes aos processos resultantes de julgamentos do CADE.

Observou-se que os dados públicos da API, passíveis de manipulação e consolidação, são abundantes, mas que no bojo da pesquisa serão tomados a partir do mês de janeiro de 2015. Isto se deve ao fato de não haver fontes de dados digitalizados anteriores a essa data e disponíveis para consulta via API.

2.3.4.2 Formulação do conceito de cartel em licitação

A prática de cartel em licitações representa uma forma específica de conluio no âmbito das aquisições públicas. O CADE define esta prática como uma colaboração entre agentes econômicos que visa eliminar ou restringir a concorrência em processos de contratação de bens e serviços pela Administração Pública (Cade, 2019c, p. 11). Este tópico explora as características, metodologias e implicações deste tipo de cartel.

A estimativa da OCDE (2016) é que 13% do PIB dos países membros da organização sejam gastos em licitações públicas, percentual esse que pode ser bem superior em se tratando de países em desenvolvimento. Considerando o sobrepreço médio em torno de 15% a 20% causado pelos cartéis, segundo estudo da mesma Organização, conclui-se que a cartelização das licitações pode provocar altíssimos prejuízos ao erário.

Países em desenvolvimento, incluindo o Brasil, sofrem consequências ainda mais danosas. Não somente porque estão mais à corrupção e a colusão, mas também conta dos elevados investimentos em infraestrutura (Martinez, 2014, p. 4). Conforme estudos, é razoável admitir que em cartéis em licitações a elevação dos preços seja ainda maior do que nos cartéis ordinários, causando graves danos sociais (Aguiar; Daher; Tabak, 2018, p. 197).

É cediço que as aquisições públicas tem importante papel estratégico na economia do país, por conseguinte, na qualidade e eficiência dos serviços que o Estado fornece aos seus cidadãos. Em 2020, o Brasil gastou cerca de R$ 35,5 bilhões em bens, serviços e obras. Em 2017, as compras públicas representaram cerca de 13,5% dos gastos totais do governo brasileiro e aproximadamente 6,5% do PIB. Devido ao tamanho dos fluxos financeiros envolvidos, as compras públicas estão expostas a riscos de conluio entre fornecedores, assim como fraude e corrupção (OECD, 2021, p. 11).

As licitações proporcionam um terreno fértil para a formação de cartéis, onde as práticas colusivas podem manifestar-se de diversas maneiras. Estas incluem a fixação de preços entre concorrentes, a determinação de vencedores de licitações específicas, divisão de licitações entre membros do cartel, apresentação de propostas com valores inflacionados, e a subcontratação dos participantes desistentes pelo vencedor (Carvalho, 2015a, p. 84).

Essas licitações públicas se tornam locais onde as partes envolvidas praticam condutas colusivas para obter vantagens indevidas. Tal prática se configura na apresentação de propostas em conluio, onde os concorrentes decidem não competir e acordam para elevar os preços ou reduzir a qualidade dos bens e serviços, conhecida internacionalmente como *bid rigging* (Aguiar;

Daher; Tabak, 2018, p. 197; Colacino, 2016, p. 20).

O CADE, em sua Nota Técnica n. 29/2019, identifica várias estratégias associadas a cartéis em licitações. Estas incluem propostas de cobertura, supressão de propostas, rodízio de propostas, divisão de mercado, subcontratação, retirada de propostas e apresentação de propostas pró-forma ou em desconformidade (Cade, 2019b, p. 25).

Em resumo, o cartel em licitação representa uma grave ameaça à integridade e eficácia dos processos de aquisição pública. Esta prática não apenas contraria os esforços da Administração Pública em utilizar seus recursos de maneira eficiente e eficaz, mas também prejudica a sociedade como um todo, comprometendo a oferta de bens e serviços de qualidade e a promoção do desenvolvimento nacional.

2.3.4.3 Consequências econômicas da formação de cartel em licitação

Rememorando a lição de Martinez (2014, p. 7–8), existem três tipos de ineficiências econômicas promovidas pelos cartéis: alocativa, produtiva e dinâmica. Ao sintetizar o pensamento da citada autora, Aguiar, Daher e Tabak (2018, p. 195) explicam que **ineficiência alocativa** é a situação em onde há má alocação de recursos - como ambiente está cartelizado, esclarecem que parte do que seria excedente do consumidor vai para o produtor, e outra se transmuda em "peso morto"; **ineficiência produtiva** acontece quando empresas operam com custos mais altos do que no caso em que não haja o cartel; por fim a **ineficiência dinâmica** se traduz ao fato de que existe redução da inovação. Assim, o cartel diminui o interesse dos participantes de se aprimorar em seus processos produtos.

Importante, ainda, registrar que cartéis criam óbices para entrada de novas firmas no mercado (Aguiar; Daher; Tabak, 2018, p. 195). Os autores citam o relatório publicado pelo OCDE (2002, p. 72) em 2002, onde um trecho que retrata a pesquisa realizada para conhecer e avaliar o tamanho do comércio afetado por cartéis em seu países membros, isto nos anos de 1996 e 2000. Estimou-se que a quantidade de comércio afetado em 16 casos de grande magnitude, que superaram US $ 55 bilhões. A conclusão foi que os preços maiores em torno de 15% a 20% do que em um mercado competitivo.

Ressalte-se que um estudo mais abrangente foi realizado por Connor (2014, p. 86), que examinou mais de 700 estudos econômicos publicados e decisões judiciais sobre 2.041 casos envolvendo cartéis. Considerando cartéis atuantes nos Estados Unidos e em âmbito internacional, entre os anos de 1790 e 2004, constatou uma média de sobrepreço de 23% em todos os períodos analisados (Aguiar; Daher; Tabak, 2018, p. 196).

2.3.5 Investigação de cartel em licitação pelo CADE

O cartel em licitação ocorre quando empresas que deveriam competir de forma legítima entre si conspiram secretamente para aumentar os preços ou diminuir a qualidade dos bens ou

serviços ofertados no curso do processo licitatório. Essa prática prejudica a concorrência dos processos de contratação e a otimização do uso dos recursos pelo setor público, trazendo efeitos negativos aos serviços públicos ofertados e às economias nacionais. O cartel em licitação é considerado ilegal entre todas as jurisdições da OCDE e é tipificado como crime em 29 dos 37 países-membros. No Brasil, os cartéis em licitação constituem tanto uma infração administrativa quanto um crime (OECD, 2021, p. 15).

OECD (2021) elaborou uma revisão das compras públicas federais no Brasil e que teve por foco o combate aos cartéis em licitações, tendo em conta as recomendações do Conselho dessa Instituição sobre o tema. O relatório traz uma série de recomendações sobre o quadro normativo e as práticas referentes às compras públicas federais, buscando aprimorá-las e intensificar o combate aos cartéis.

Observa-se que OCDE tem elaborado Diretrizes para Combater o Conluio entre Concorrentes em Contratações Públicas, que se tornaram referências globais e têm auxiliado os países a avaliar o seu quadro legal de contratações públicas e a implementar reformas e aprimoramentos favoráveis à concorrência. Ambas serviram de base para o desenvolvimento de estratégias nacionais de combate a cartéis em licitações, orientação quanto ao desenho de processos licitatórios pró-competitivos e estruturação de programas de advocacia da concorrência e treinamento para servidores públicos quanto aos riscos de conluio nas contratações (OECD, 2021, p. 18).

As políticas de compras públicas no Brasil têm como objetivo garantir a eficiência na prestação de serviços pelo governo aos seus cidadãos, além de desempenhar um papel estratégico na economia do país. Em 2020, as compras públicas do governo brasileiro totalizaram cerca de R$ 35,5 bilhões, e políticas de compras públicas robustas que visam combater os cartéis em licitações podem gerar economias significativas e contribuir para que os governos obtenham melhores relações de custo-benefício nas contratações (OECD, 2021, p. 5).

O relatório OCDE (2021) apresenta recomendações quanto à prevenção de conluio entre concorrentes nas contratações públicas por meio do desenho de processos licitatórios competitivos e efetivos e aperfeiçoamento da detecção de esquemas colusivos. Estas recomendações se baseiam na Recomendação e Diretrizes da OCDE entre outras boas práticas internacionais. O relatório também propõe a identificação de padrões de preços suspeitos e comportamentos dos licitantes, assim como informações que possam chamar a atenção dos servidores públicos encarregados pela contratação para possíveis manipulações do processo de compra pública (OECD, 2021, p. 17–18).

Outrossim, conforme exposto nos parágrafos anteriores, a integração do aprendizado de máquina está transformando a maneira como o CADE realiza análises de fusões e aquisições. Algoritmos avançados possibilitam uma avaliação mais detalhada do impacto potencial dessas operações no mercado, considerando uma variedade de cenários e variáveis. Isso resulta em decisões mais informadas e precisas, essenciais para prevenir a formação de estruturas de mercado prejudiciais à concorrência e aos consumidores.

A Lei nº 12.529 de 2011 especifica as infrações contra a ordem econômica dividindo-as em duas categorias: infrações verticais e horizontais. As infrações verticais englobam práticas como estabelecimento de preços de revenda, limitação de acesso a clientes, contratos de exclusividade, recusa em comercializar, condicionamento de venda e diferenciação de preços. As infrações horizontais, por outro lado, incluem formações de cartéis, acordos entre concorrentes, condutas impróprias de entidades profissionais e estratégias de preços predatórios. Com a evolução tecnológica e do mercado, é importante observar que as concentrações horizontais podem agora oferecer riscos maiores à ordem econômica e práticas antes não observadas em mercados concentrados podem ser consideradas ilegais (Silva Filho; Haro, 2013, p. 5).

A presente pesquisa se concentrou nas infrações horizontais, com ênfase nas condutas colusivas, particularmente a prática anticoncorrencial denominada "cartel em licitações", conforme pode se observar a partir desse ponto.

Lima (2022), como exemplo de aplicação de AM no antitruste, aponta o projeto Cérebro é uma iniciativa do CADE que utiliza ferramentas de análise de dados e aprendizado de máquina para monitorar mercados, detectar cartéis e outras condutas anticompetitivas. Desde 2013, o projeto tem sido responsável pela detecção de práticas anticompetitivas e tem sido utilizado em investigações antitruste. As ferramentas desenvolvidas no âmbito do projeto também são usadas na investigação de condutas já detectadas (Lima, 2022, p. 4–5).

Além disso, o CADE tem se beneficiado da capacidade do aprendizado de máquina de processar e interpretar grandes volumes de dados para identificar tendências de mercado e comportamentos emergentes. Esta capacidade é particularmente valiosa no contexto de mercados digitais, onde os padrões de consumo, competição e interações empresariais são complexos e em constante evolução.

2.3.5.1 Estrutura do SBDC e funções do CADE

Observa-se que o principal objetivo de um sistema de proteção à concorrência se baseia na promoção da competição econômica mediante atuação preventiva e repressiva, estabelecendo ações que possam coibir a limitação, ou prejuízo, “de uma disputa saudável ou transparente por parcelas de mercado relevante, com base na legislação vigente” (Figueiredo, 2014, p. 226). Nesse mesmo diapasão, Silva Filho e Haro (2013, p. 5) esclarece que a função primeira do sistema de proteção "é proteger a coletividade prevenindo e reprimindo as infrações a ordem econômica, preservando uma concorrência livre e competitiva". No Brasil, essa competência funcional é atribuída ao SBDC por força da Lei 12.529/2011.

A Lei Antitruste estrutura o SBDC, determinando sua composição e instituindo a SEAE e o CADE como órgãos de instância máxima. A SEAE, antes detentora da atribuição de produzir pareceres sobre atos de concentração e investigação de condutas anticompetitivas, na nova legislação tem como atribuição opinar nos procedimentos em trâmite no CADE, que por sua vez é quem emite uma decisão administrativa definitiva sob os casos que lhe são apresentados.

Pertencia a estrutura do extinto Ministério da Fazenda, e agora faz parte da estrutura do Ministério da Economia. Exerce a função de promover e divulgar a concorrência nos órgãos governamentais e na sociedade, chamada de advocacia da concorrência ou *advocacy* pela Doutrina (Taufick, 2012, p. 115). O CADE se apresenta como uma autarquia federal vinculada ao Ministério da Justiça. Importa registrar que o CADE se divide em três órgãos básicos: (*i*) o Tribunal de Administrativo de Defesa Econômica; (*ii*) a Superintendência-Geral; e, (*iii*) o Departamento de Estudos Econômicos. Estes órgãos serão exaustivamente examinados em tópico específico. De mais a mais, oficia junto ao Conselho uma instância do Ministério Público Federal, cujo o membro é designado pelo Procurador Geral da República (Ferreira, 2013, p. 27).

A nova estrutura do SBDC, instituída pela Lei 12.529/2011, apresenta similitudes entre o CADE e desenho da autoridade antitruste americana chamada *Federal Trade Comission* (FTC). Essa entidade é composta pelo (*i*) escritório de economia, (*ii*) escritório concorrencial e o (*iii*) Tribunal[28]. Tal desenho foi reproduzido na estrutura do órgão antitruste brasileiro quando da nova reestruturação, modelando-se no Departamento de Estudos Econômicos, na Superintendência-Geral e no Tribunal Administrativo de Defesa Econômica, respectivamente, ganhando as feições de um órgão com mais atribuições, sendo recepcionado pela doutrina como sendo um "Super CADE" (Taufick, 2012, p. 2). Outrossim, a Doutrina faz alerta que a fusão dos órgãos de investigação de condutas e de concentração ao mesmo órgão julgador também aconteceu recentemente na Espanha, França e Reino Unido (Pereira Neto; Casagrande, 2016, p. 34–38).

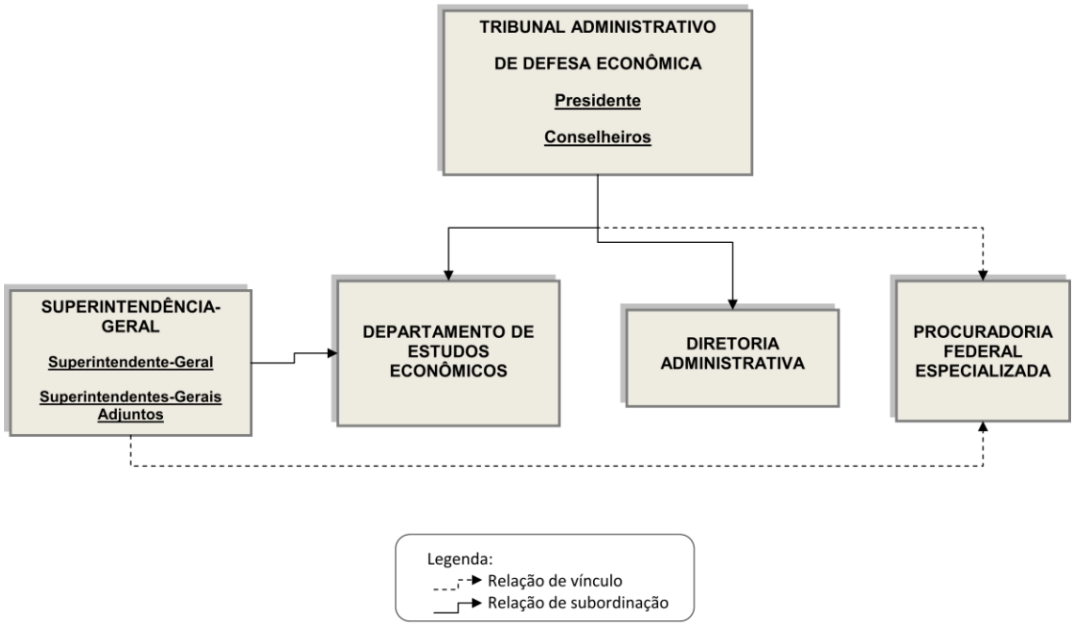
Em adição, esclarece-se que a Autoridade Antitruste brasileira opera como uma autarquia[29]. Ela se baseia no Direito Administrativo, já que é uma autarquia sujeita ao regime jurídico do Direito Público. Importante também esclarecer que, normalmente, existe separação entre as instâncias investigatória e acusatória nos processos administrativos sancionadores de atos anticompetitivos, sendo a Superintendência Geral e o Tribunal do CADE, respectivamente. Esta condição não se mantém apenas na situação de avocação do procedimento pelo Tribunal, aspecto que será detalhado em tópico específico sobre o fluxo do processo administrativo sancionador. A interação dessas instâncias pode ser melhor compreendida ao observar o organograma apresentado na Figura 11.

A Superintendência-Geral (SG) tem atribuições para a investigação das condutas anticompetitivas. É gerida pelo Superintendente Geral com o auxílio de dois Superintendentes Adjuntos. Seus mandatos tem duração de dois anos, sendo permitida uma única recondução. A SG também tem a atribuição de (*i*) investigar as condutas anticompetitivas; (*ii*) propor acordos e medidas preventivas; (*iii*) analisar atos de concentração sumários; e (*iv*) instruir atos de concentração, expedindo parecer opinativo. Tem competência, ainda, de negociar e celebrar acordos de

[28] Respectivamente *Bureal of Economics*, *Bureau of Competition* e *Comission*.

[29] Seguindo recomendação da OCDE, houve a inclusão do CADE na Lei Geral das Agência Reguladoras (Lei n. 13.848/2019. Com essa inclusão, a autarquia federal adquiriu autonomia administrativa, orçamentária e financeira.

Figura 11 – Estrutura do Cade



Fonte: Cade (2016, p. 21).

leniência (Brasil, 2011). Calha registrar que a Secretaria de Direito Econômico (SDE), existente no ordenamento anterior, fora extinta. Em consequência, a função de instruir e investigar os atos de concentração de empresas, além de outras competências, foram remetidas à SG/CADE (Silva Filho; Haro, 2013, p. 6).

Segundo Magalhães Júnior (2018, p. 32), o Tribunal Administrativo de Defesa Econômica (TADE)[30] é o órgão judicante da Autarquia. É o órgão colegiado, o Plenário, cuja composição se perfaz com seis funções de conselheiros e uma de presidente. Os mandatos de seus componentes são de quatro anos, não coincidentes e com recondução proibida. O Plenário tem competência para adotar medidas preventivas e representa a última instância de julgamento sobre atos de concentração e condutas anticoncorrenciais no âmbito do Poder Executivo, sem prejuízo à revisão judicial destas decisões (Magalhães Júnior, 2018, p. 32; Brasil, 2011).

Por fim, tem-se o Departamento de Estudos Econômicos (DEE). Este órgão interno é gerido por um economista-chefe, que tem a atribuição de elaborar estudos e parecer econômicos de ofício ou a requerimento de algum outro componente do CADE. Em razão da excelência técnica em assuntos econômicos atinentes ao antitruste, esse órgão tem a importante missão de

[30] Segundo Cardoso (2012), a citada lei padece do mesmo vício existente em normas e estatutos brasileiros, pois denominou como "Tribunal" um ente administrativo em razão de sua atribuição de julgar em flagrante colisão com o fato de não existir coisa julgada fora do processo judicial em razão do disposto no Princípio da Inafastabilidade do Judiciário a (art. 5º, XV da CRFB/1988). Exemplos desta prática são diversos: Tribunais de Contas da União e dos Estados, Tribunais Administrativos Tributários e Tribunais de Justiça Desportiva existentes no país. Assim, apesar da denominação de "Tribunal", o TADE é um órgão administrativo, que não integra o Poder Judiciário nacional.

elaborar estudos econômicos e econométricos visando a análise do impacto das operações já aprovadas sobre a economia (Brasil, 2011).

Magalhães Júnior (2018, p. 32–33) esclarece que, para auxiliar no mister de promover suas atribuições, o CADE é assessorada pela Procuradoria Federal Especializada (ProCADE), órgão que tem a competência de prestar consultoria e assessoramento jurídico, inclusive representando a Autarquia judicial e extrajudicialmente visando a prevalência dos interesses da autoridade antitruste (Taufick, 2012, p. 107). Ademais, como o autor deixa claro, o CADE conta com ação de um representante do Ministério Público Federal, responsável por emitir parecer nos processos administrativos para imposição de sanções administrativas por infrações à ordem econômica, de ofício ou por provação do Conselheiro-Relator (Magalhães Júnior, 2018, p. 33).

Recentemente, houve a inclusão do CADE na Lei Geral das Agências Reguladoras, mediante a promulgação da Lei Federal n. 13.848/2019. Com essa inclusão, a autarquia federal passa a ter autonomia administrativa, orçamentária e financeira. Calha, ainda, esclarecer que nos termos do art. 4º da Lei Antitruste, o Cade tem jurisdição em todo o território nacional, com sede e foro no Distrito Federal. Outrossim, cabe enfatizar que os órgãos do SBDC se apoiam em três frentes que conforme estipulado por Monteiro (2002, p. 13) assim se desenvolvem:

> A atuação dos órgãos do sistema subdivide-se em três tipos: preventiva, através do controle de estruturas de mercado, via apreciação dos atos de concentração (fusões, aquisições e incorporações de empresas); repressiva, através do controle de condutas ou práticas anticoncorrenciais, que busca verificar a existência de infrações à ordem econômica, das quais são exemplos as vendas casadas, os acordos de exclusividade e a formação de cartel; e educacional, que corresponde ao papel de difusão da cultura da concorrência, via parceria com instituições para a realização de seminários, palestras, cursos e publicações de relatórios e matérias em revistas especializadas, visando um maior interesse acadêmico pela área, o incremento da qualidade técnica e da credibilidade das decisões emitidas e a consolidação das regras antitruste junto à sociedade. (Monteiro, 2002, p. 13)

Do exposto, assume-se que os princípios constitucionais de defesa de uma economia de livre mercado e da concorrência e a chamada Lei Antitruste (Lei n. 12.529/2011) são a principal estrutura jurídica em que o SBDC irá atuar, apoiando-se nas frentes citadas acima, mediante as ações da SEAE e do CADE, buscando atingir um dos objetivos específicos desta pesquisa, passa-se a considerar os modos de atuação as funções do CADE.

O CADE foi criado pela Lei 4.137/1962, mas fora com a edição da Lei 8.158/1981 que seus procedimentos de análise das práticas anticoncorrências foi concertada (Salomão Filho, 2002, p. 73). Na mesma década fora elevado a condição de uma autarquia federal pela lei Lei 8.884/1994, e tem a sua atual estrutura funcional moldada pela lei 12.529/2011[31] (Silva Filho; Haro, 2013, p. 3). Tem como objetivo primeiro prevenir e reprimir as infrações a ordem

[31] Importa destacar que nos termos do art. 4º da Lei 12.529/2011, o Cade é uma entidade judicante com jurisdição em todo o território nacional, constituída na forma de autarquia federal e vinculada ao Ministério da Justiça. Possui sede e foro no Distrito Federal, tendo suas competências previstas nesta Lei.

econômica, tais quais aquelas condutas que resultem em um prejuízo a livre concorrência e à ordem econômica enquanto princípios constitucionais (Carvalho, 2015a, p. 97).

Destaque-se que o CADE, no âmbito de suas competências institucionais, tem as atribuições para investigar condutas prejudiciais à livre concorrência e, se for o caso, aplicar punições aos infratores, além de disseminar a cultura da livre concorrência. Condutas anticompetitivas são todas as práticas de um agente econômico que, de forma concreta ou potencial, causar danos à livre concorrência, independentemente do autor ter tido intenção de prejudicar o mercado (Cade, 2016, p. 11).

Outrossim, o princípio da livre concorrência[32] se baseia no pressuposto de que a concorrência não pode ser restringida por agentes econômicos com poder de mercado[33]. Cabendo ao Cade a investigação das condutas de tais agentes econômicos.

Ademais, Com a reestruturação, extinção de órgãos (v.g., a SDE) e convergência para um único órgão controlador de atividades antes dispersas, as atribuições do CADE – de forma não exaustiva: a elaboração de estudos e pareceres econômicos, apurações das infrações a ordem econômica, analise e pareceres nos processos de atos de concentração e a função de julgar estes processos – ganharam mais visibilidade e alcance, sendo recepcionada como um "Super"CADE (Taufick, 2012, p. 29).

Dessa forma, ante os preceitos acima debatidos, pode-se sintetizar as funções do Cade em três grupos: (*i*) preventiva, que diz respeito à análise e decisão sobre atos de concentração econômica; (*ii*) repressiva, concernente à investigação e julgamento de cartéis e outras infrações à ordem econômica; e (*iii*) educativa, a partir da divulgação da cultura da concorrência.

2.3.6 Aplicações de IA no setor público antitruste brasileiro

Tirole (2019, p. 378) afirma que a digitalização está transformando radicalmente a sociedade e a economia no século XXI, influenciando praticamente todas as atividades humanas. Atividades cotidianas, como compras, transações bancárias, leitura de notícias, e serviços de transporte e hospedagem, estão sendo cada vez mais realizadas online, através de plataformas como Uber, Airbnb e BlaBlaCar. A mudança não se limita ao comércio ou ao turismo, estendendo-se a setores como mídia, seguros, saúde, energia e educação. Por exemplo, a *National Public Radio* (NPR) dos EUA adaptou-se ao declínio da mídia tradicional ao oferecer programação personalizada através do seu aplicativo NPR *One* (Tirole, 2019, p. 378).

Ainda segundo Tirole (2019)A digitalização também está remodelando as relações pessoais, a vida cívica e a política, além de desafiar as estruturas industriais tradicionais e a natureza do trabalho. Questões como cibersegurança, direitos de propriedade intelectual e

[32] Previsto no artigo 170, inciso IV da Constituição Federal

[33] Para o Cade, uma empresa (ou um grupo de empresas) possui poder de mercado se for capaz de manter seus preços sistematicamente acima do nível competitivo de mercado sem com isso perder todos os seus clientes (Cade, 2016, p. 8)

regulamentação são de crescente preocupação. Com a ascensão de plataformas bilaterais, como Apple, Google e Amazon, que conectam diferentes lados do mercado, a economia digital promete progresso tecnológico significativo, mas também apresenta novos riscos. Este capítulo e os seguintes visam explorar os maiores desafios da digitalização para preparar a sociedade para esta profunda transformação (Tirole, 2019, p. 378–379).

A teoria clássica do crescimento econômico, que se baseia na acumulação de capital e no aumento da força de trabalho, foi desafiada pelo trabalho de Robert Solow em 1956. Solow identificou que esses fatores, por si só, não explicavam completamente o crescimento econômico, destacando a importância de outros elementos, como o progresso tecnológico (Tirole, 2019, p. 433). Atualmente, a inovação tecnológica é ainda mais central para o crescimento econômico, especialmente na economia do século XXI, frequentemente descrita como uma economia do conhecimento. Esta economia é caracterizada por uma mudança tecnológica abrangente, indicando um deslocamento do paradigma clássico de crescimento para um modelo onde a inovação desempenha um papel fundamental (Tirole, 2019, p. 433).

Assim, estamos testemunhando uma transformação sem precedentes nos mercados globais, impulsionada pela ascensão da economia digital e as inovações tecnológicas. Neste contexto, a política antitruste e o aprendizado de máquina emergem como elementos cruciais no entendimento e na regulação dessas mudanças dinâmicas. Esta seção e as seguinte subseções se propõem a explorar a intersecção entre a política antitruste, o avanço do AM e as funções do CADE no Brasil, oferecendo uma análise integrada de como esses elementos interagem e influenciam o cenário econômico antitruste atual.

Nessa toada da inovação tecnológica, Araújo Jr. (1999) explica o papel do antitruste na dinâmica da concorrência econômica, guiado pela noção schumpeteriana de destruição criativa. Nesse contexto, a concorrência é vista como um processo dinâmico e evolutivo, onde inovações e estratégias diferenciadas são essenciais para o sucesso empresarial. O artigo destaca que o antitruste enfrenta desafios significativos, principalmente na identificação e intervenção em práticas anticompetitivas que distorcem o mercado. Argumenta-se que a inovação deve ser o principal meio de criação de assimetrias informacionais, enquanto práticas prejudiciais, como a busca de aluguéis e o crime organizado, devem ser combatidas (Araújo Jr., 1999).

Além disso, o papel do antitruste é enfatizado como um regulador fundamental na manutenção de um mercado eficiente e justo, orientando a competição para a eficiência e equidade, e protegendo os interesses nacionais contra práticas anticompetitivas internacionais (Araújo Jr., 1999). O artigo também aborda os desafios específicos enfrentados pelas agências antitruste na América Latina, ressaltando a importância da cooperação e harmonização das políticas antitruste em vários níveis. Em suma, o estudo coloca a destruição criativa de Schumpeter no centro da competição econômica, apontando o antitruste como um elemento vital para assegurar um ambiente competitivo saudável e baseado na inovação (Araújo Jr., 1999).

Nesse cenário, a política antitruste, essencial para garantir a concorrência leal e prevenir

práticas monopolísticas, enfrenta desafios inéditos em um ambiente cada vez mais digital e orientado por dados. A introdução de métodos baseados em aprendizado de máquina oferece novas ferramentas e abordagens para analisar e responder a essas mudanças. O aprendizado de máquina, definido como o estudo e desenvolvimento de algoritmos que se aprimoram automaticamente através da experiência e dos dados, está remodelando a maneira como dados são interpretados e decisões são tomadas, impactando diretamente a eficácia das políticas antitruste.

Na seara da inovação tecnológica no setor público brasileiro, como contexto inicial, o artigo de Veras e Barreto (2022) aborda a utilização da IA no sistema judiciário brasileiro, com foco na celeridade processual. É mencionado o Projeto Victor[34], uma IA desenvolvida em parceria com a Universidade de Brasília (UnB) e o Supremo Tribunal Federal (STF), que tem como objetivo reduzir o tempo de tramitação dos processos e lidar com o congestionamento dos tribunais. O citado projeto utiliza técnicas de AM e PLN para analisar e classificar os processos, identificar precedentes e auxiliar na tomada de decisões. A IA também é vista como uma ferramenta que pode melhorar a eficiência e a precisão das consultas processuais e jurisprudenciais. No entanto, o texto ressalta que a utilização de IA para criar juízes robôs ainda é improvável, devido à complexidade da ponderação de contextos e tomada de decisões em casos de dúvida. No geral, a IA é vista como uma tecnologia promissora para agilizar e aprimorar o sistema judiciário, mas também são mencionados desafios relacionados à regulação e proteção de dados (Veras; Barreto, 2022, p. 1–4).

Em sintonia com o Projeto Victor e outros esforços do Conselho Nacional de Justiça (CNJ), o Brasil está propondo o Marco Legal da Inteligência Artificial[35]. Esse marco estabelece princípios como respeito aos direitos humanos, igualdade e transparência. Ele introduz a figura do agente de IA, responsável legalmente pelas decisões do sistema, estabelece obrigações em conformidade com a LGPD (2018). Além disso, exige relatórios de impacto de IA e consulta pública para intervenção estatal, promovendo a adoção de IA no setor público, capacitação e apoio à pesquisa (Veras; Barreto, 2022, p. 6).

Neste mesmo diapasão, o papel do CADE, como o órgão responsável pela defesa da concorrência no Brasil, se torna ainda mais vital. A integração de ferramentas de aprendizado de máquina, como demonstrado pelo projeto Cérebro do CADE, que utiliza estas tecnologias para monitoramento de mercados e detecção de práticas anticompetitivas desde 2013, exemplifica a evolução do papel regulatório em resposta às demandas de um mercado em constante evolução (Lima, 2022, p. 3). Esse autor afirma que Departamento de Estudos Econômicos (DEE) do Cade utiliza diversas técnicas, incluindo ciência de dados, para a análise de condutas e avaliação de fusões e aquisições. Isso envolve a estimação de demanda, definição de mercado relevante e

[34] O projeto homenageia Victor Nunes Leal, que faleceu em 1985 e serviu como ministro do Supremo Tribunal Federal (STF) de 1960 a 1969. Durante seu mandato, Leal foi fundamental na organização da jurisprudência do STF em súmulas, uma inovação que desde então tem simplificado a utilização de precedentes judiciais em recursos.

[35] Projeto de Lei do Senado Brasileiro n° 2338, de 2023, de iniciativa Senador Rodrigo Pacheco (PSD/MG).

outros procedimentos antitruste, com a colaboração de profissionais experientes em ciência de dados.

Esta pesquisa apresenta uma visão abrangente e multidisciplinar sobre como a política antitruste, impulsionada pelo aprendizado de máquina, está sendo adaptada e aplicada no Brasil, especialmente através das ações do CADE. Ao fazer isso, busca-se não apenas compreender os desafios atuais, mas também antecipar futuras direções e oportunidades para a a apropriação de técnicas de AM e PLN, especificamente a modelagem de tópicos, em julgamentos de lides antitruste em uma era marcada por rápidas inovações tecnológicas.

2.3.6.1 Incorporação de novas tecnologias pelo CADE

O CADE desempenha um papel fundamental na regulação e manutenção da concorrência justa no mercado brasileiro. Com a emergência de tecnologias inovadoras, especialmente o aprendizado de máquina, o CADE enfrenta o desafio de integrar essas novas ferramentas em suas operações para aprimorar sua eficácia e eficiência. Esta integração é uma resposta necessária às complexidades crescentes dos mercados contemporâneos, particularmente aqueles influenciados pela economia digital.

Lima (2022) apresenta alguns modelos e procedimentos de aprendizado de máquina que podem ser aplicados em diferentes etapas da análise antitruste. O objetivo do autor é mostrar como esses modelos podem ser utilizados para previsões, classificações e outras tarefas, sendo por vezes mais adequados que os modelos econométricos tradicionais. Além disso, o documento destaca a importância do emprego do aprendizado de máquina na análise antitruste, considerando as mudanças no cenário econômico e tecnológico recente.

As autoridades antitruste têm se adaptado ao cenário tecnológico e de dados dos últimos anos investindo em pessoal e equipamento para aprimorar seus métodos de investigação e monitoramento de mercados com o uso de tecnologia da informação. A utilização de modelos de aprendizado de máquina e algoritmos de *big data* permitem às autoridades detectar colusões, outras práticas anticompetitivas e prever comportamentos de ofertantes e consumidores ou de alterações de estruturas competitivas. Algumas autoridades, como a Autoridade antitruste grega, criaram unidades de ciência de dados que desenvolveram plataformas para coleta e processamento de dados econômicos em tempo real (Lima, 2022, p. 15–16).

A integração do aprendizado de máquina também está transformando a maneira como o CADE realiza análises de fusões e aquisições. Algoritmos avançados possibilitam uma avaliação mais detalhada do impacto potencial dessas operações no mercado, considerando uma variedade de cenários e variáveis. Isso pode resultar em decisões mais informadas e precisas, essenciais para prevenir a formação de estruturas de mercado prejudiciais à concorrência e aos consumidores.

Existem diversas técnicas de aprendizado de máquina que podem ser empregadas na análise antitruste, tais como modelos de regressão e agrupamento, redes neurais artificiais,

algoritmos de árvore de decisão, análise de componentes principais e máquinas de suporte vetorial. Essas técnicas podem ser aplicadas em diferentes etapas da análise antitruste, como a definição de mercado, detecção e prova de cartel, avaliação de predatória e exclusão, além de auxiliar na predição de impactos de operações de concentração e investigação de potenciais abusos de posição dominante (Lima, 2022, p. 13).

2.4 Síntese do Estado da Arte e formulação da tese

No cenário atual, caracterizado por um volume crescente de dados e pela necessidade de eficiência econômica, a Organização Industrial no Setor Público encontra-se no epicentro de uma transformação significativa. A IA, com destaque para o PLN, emerge como uma ferramenta poderosa, capaz de converter dados em conhecimento útil para a tomada de decisões informadas. No contexto jurídico-administrativo, onde o Conselho Administrativo de Defesa Econômica (CADE) atua como regulador das práticas antitruste, a otimização processual torna-se uma necessidade premente e evidente.

Dentro deste contexto, a modelagem de tópicos, uma técnica avançada de PLN, mostra-se promissora para auxiliar no julgamento de condutas colusivas. Este estudo tem como objetivo explorar a aplicabilidade de diversas técnicas de modelagem de tópicos em textos legais, com foco particular nos julgamentos de cartéis em licitações realizados pelo CADE.

Conforme pode ser observado no Apêndice B (p. 141), a literatura revisada indica uma variedade de aplicações da modelagem de tópicos em diferentes áreas, embora sua utilização em textos legais antitruste ainda seja pouco explorada. Assim, esta tese propõe uma série de hipóteses que norteiam a investigação. A Hipótese Principal sugere que a modelagem BERT, devido à sua modernidade, mostrará um desempenho superior na identificação de tópicos subjacentes. Também é discutido o desempenho variado das técnicas, a complexidade computacional, a correlação com decisões jurídicas, a interpretabilidade e a adequação do pré-processamento, detalhadas a seguir.

Desse modo, a tese divide-se em duas grandes etapas, cada uma com um conjunto específico de hipóteses para resolver o problema de pesquisa proposto.

Na primeira etapa, será realizada uma análise não supervisionada, que permitirá selecionar a técnica de modelagem de tópicos mais adequada para o *dataset* do CADE. As hipóteses a serem testadas nesta etapa são:

Hipótese de Eficiência em Modelagem Avançada: esta hipótese propõe que a modelagem BERT, destacando-se por sua capacidade de capturar contextos complexos em textos legais antitruste, superará outras técnicas como NMF, LDA, e Top2Vec em precisão e relevância na identificação de tópicos. Avaliará a efetividade do BERT contra outras metodologias, considerando tanto a qualidade analítica quanto os desafios computacionais.

Hipótese de Impacto Analítico nas Decisões do CADE: examina a relação entre a precisão na identificação de tópicos através de diferentes técnicas de PLN e o padrão das decisões jurídicas no CADE. Investiga se uma correlação significativa entre a análise de tópicos e as decisões pode contribuir para a previsibilidade e aprimoramento dos julgamentos antitruste, considerando também a interpretabilidade dos resultados para sua aplicação prática.

A segunda etapa envolve a realização de uma análise supervisionada de classificação no *dataset* do CADE, utilizando o modelo selecionado na etapa anterior. Para esta serão analisadas duas hipóteses como se segue:

Hipótese de Previsibilidade das Decisões: Esta hipótese afirma sobre a viabilidade de antecipar tendências de condenação em processos de cartel através de análises probabilísticas de tópicos dominantes em pareceres do MPF e votos dos conselheiros relatores, com a consideração de variáveis profissiográficas (gênero, formação acadêmica, histórico profissional) destes últimos *players* como moderadoras dessas previsões. Propõe-se que padrões preditivos e influências profissiográficas podem ser conjuntamente modelados, utilizando-se de técnicas estatísticas avançadas, para prever decisões judiciais de forma mais abrangente e fundamentada.

A integração da Teoria da Organização Industrial com técnicas avançadas de Processamento de Linguagem Natural na análise de práticas antitruste oferece uma abordagem inovadora para responder às hipóteses formuladas. Esta teoria proporciona um arcabouço teórico para entender as dinâmicas de mercado e o impacto das estruturas empresariais sobre a concorrência e eficiência econômica. Ao aplicá-la em conjunto com análises de dados e linguagem, espera-se desvendar padrões complexos de comportamento anticompetitivo, permitindo uma avaliação mais precisa e fundamentada das práticas de mercado. Essa abordagem promete não apenas enriquecer a compreensão teórica, mas também aprimorar a eficácia das intervenções regulatórias.

A integração da IA nos campos da Economia e do Direito apresenta desafios técnicos, éticos e sociais. As técnicas de modelagem de tópicos devem ser precisas, interpretáveis e adaptáveis ao contexto jurídico. A transparência e a explicabilidade das técnicas de IA são fundamentais, dada a relevância das decisões legais para indivíduos e a sociedade. Além disso, a adequação e a qualidade dos dados utilizados são cruciais, uma vez que o pré-processamento impacta significativamente os resultados da modelagem de tópicos.

Com a conclusão deste estudo, espera-se contribuir significativamente para o campo acadêmico e para a prática jurídico-administrativa. Os *insights* adquiridos podem ajudar a otimizar os processos decisórios do CADE, melhorar a eficiência e a transparência das decisões antitruste e promover uma compreensão mais aprofundada das técnicas de IA aplicáveis à Economia e ao Direito. Ademais, ao elucidar as interações entre IA e processos jurídicos de controle antitruste, esta tese visa estimular um debate contínuo sobre as melhores práticas para a integração ética e eficaz da tecnologia no âmbito do direito antitruste.

O capítulo seguinte detalha os procedimentos metodológicos adotados para o teste das hipóteses apresentadas nesta seção, respondendo, assim, aos objetivos específicos deste estudo.

3 METODOLOGIA

Esta pesquisa, classificada como exploratória e aplicada, incursionou no universo dos julgamentos do Conselho Administrativo de Defesa Econômica (CADE), com enfoque nas práticas anticoncorrenciais de formação de cartel em licitações. Utilizando métodos avançados de PLN, o estudo extraiu *insights* relevantes e identificou padrões nas decisões do CADE, contribuindo para a ampliação do conhecimento em Economia e Direito. O trabalho busca contribuir para o campo de estudos econômicos e jurídicos ao aplicar métodos de PLN para analisar decisões do CADE, oferecendo uma nova perspectiva sobre como dados textuais podem ser explorados para entender melhor as práticas anticoncorrenciais e a tomada de decisão em julgamentos relacionados a cartéis em licitações.

O desenho da pesquisa, sob a ótica da resolução do problema, foi elaborado conforme contido no Quadro 3 (p. 140). Os procedimentos metodológicos englobaram a seleção de uma base de dados abrangente, coleta e pré-processamento dos dados textuais, e uma série de análises quantitativas e qualitativas. As subseções a seguir detalham cada etapa do processo, enfatizando a integridade, a precisão, e a relevância acadêmica do trabalho realizado.

3.1 Classificação da pesquisa

Esta pesquisa pode ser classificada como exploratória e aplicada, com foco na análise de dados textuais provenientes de julgamentos do Conselho Administrativo de Defesa Econômica (CADE), especificamente relacionados a práticas anticoncorrenciais de formação de cartel em licitações. A abordagem adotada envolve a utilização de técnicas avançadas de PLN, com o intuito de extrair *insights* relevantes e identificar padrões nas decisões do CADE.

A pesquisa é exploratória, pois busca investigar um campo ainda pouco explorado: a aplicação de PLN em julgamentos de cartéis em licitações no contexto brasileiro. Este tipo de análise permite uma compreensão mais aprofundada das decisões judiciais e contribui para o desenvolvimento de modelos analíticos que podem ser aplicados em estudos futuros.

Do ponto de vista metodológico, a pesquisa é aplicada, pois utiliza dados reais obtidos a partir de consultas ao Sistema Eletrônico de Informações (SEI) do CADE. Os dados coletados foram transformados em um corpus para análise, composto por documentos relevantes que incluem Pareceres do Ministério Público junto ao CADE e Votos dos Conselheiros-Relatores. Este corpus foi submetido a um processo de pré-processamento, que incluiu a tokenização, remoção de *stopwords* e lematização, seguido da transformação em uma matriz de recursos (*Document-Feature Matrix* - DFM).

A pesquisa também emprega uma abordagem quantitativa, com a utilização de software estatístico R e pacotes específicos para análise de texto, como o *quanteda* e *tidytext*. Estas

ferramentas possibilitam a realização de análises estatísticas detalhadas, incluindo a frequência de termos, análise de TF-IDF (*Term Frequency-Inverse Document Frequency*) e visualizações de dados, que auxiliam na interpretação dos resultados e na formulação de conclusões.

3.2 Base de dados e amostra

A pesquisa foca nos julgamentos realizados pelo Conselho Administrativo de Defesa Econômica (CADE) e presentes na base de dados obtida por meio das suas APIs (Cade, 2023, 2021). O escopo do estudo inclui processos administrativos referentes a cartéis em licitações, com um intervalo temporal que abrange de 2011 até o primeiro quadrimestre de 2023. A escolha desse período se justifica pela promulgação da Lei 12.529/2011, um marco regulatório fundamental para a compreensão dos impactos dessa legislação sobre as decisões do Órgão Regulador.

A amostra foi selecionada com base em critérios específicos para capturar a essência do impacto da Lei 12.529/2011 nos julgamentos do CADE relacionados a cartéis em licitações. A partir da população total de julgamentos disponíveis na base de dados do CADE, a amostra foi refinada para incluir apenas os processos administrativos pertinentes ao tema de cartéis em licitações. Esse recorte temporal e temático permite uma análise mais direcionada e significativa dos efeitos da legislação no comportamento regulatório do CADE.

A base de dados utilizada para a pesquisa é composta por registros detalhados dos processos administrativos julgados pelo CADE. Estes registros incluem informações essenciais como datas de julgamento, partes envolvidas, natureza das infrações, decisões tomadas, e as razões por trás dessas decisões. Esses dados foram extraídos utilizando as APIs do CADE, que fornecem acesso a um vasto repositório de informações processuais e decisórias.

3.2.1 Dados processuais do CADE

Consultou-se na API de dados “CADE em Números” (Cade, 2021), selecionando-se a aba “Faça você mesmo!” e aplicando os filtros conforme contido na Tabela 4.

A extração permitiu filtrar a informação sobre o número do processo referentes à “cartel”. Esse passo foi importante, pois possibilitou consultar e realizar a extração manual dos dados na API de consulta processual do CADE (Cade, 2023). Não se utilizou técnicas de *webscraping* pois os dados, apesar de constantes em planilhas de informação, carecerem de estrutura rígida de rotulação dos documentos, o que poderia levar um algoritmo automatizado de extração de dados não encontrar o documento desejado, ou fazer o *download* de um documento indesejado. De qualquer forma, caso se optasse por uma extração por *webscraping* e pelas razões expostas [ausência de rigidez de classificação], o algoritmo realizaria o *download* de todos os documentos de cada processo, sendo necessário estabelecer uma etapa extra para filtrar dentre os documentos baixados, quais seriam interessantes para a pesquisa.

Tabela 4 – Critérios de consulta na API "CADE em Números"

Filtro	Valor inserido
Classificação	Processo Administrativo
Instância (apenas AC)	Tribunal, tribunal, Ttibunal [sic]
Mérito	Arquivamento, Condenação
Conduta	Cartel, Conduta Comercial Uniforme, Conduta Unilateral, Conduta unilateral, abuso de posição dominante e dominação de mercado relevante, NA
_dimension	NOT Representante, Requerente
_mesure	Total de Atos de Concentração Julgados, Total de Processos Administrativos Julgados, Outros Procedimentos Julgados, Total de Multas Aplicadas (R$), Total de Contribuições Pecuniárias (R$)

Fonte: elaborado pelo autor.

Desse modo, como afirmado alhures, preferiu-se realizar a consulta *on-line* a cada processo de interesse e apenas realizar o *download* dos documentos referentes ao Parecer do Ministério Público junto ao CADE, bem como o Voto do Conselheiro-Relator dos autos. Esse procedimento operacional levou mais tempo que o desejado, mas possibilitou salto de qualidade na documentação extraída da API. Torna-se óbvio que dada a escalabilidade e volume de transações, onde seja necessário obter uma visão mais ampla dos processos autuados e documentos produzidos *ex officio* pelo CADE, a escolha deverá recair em um processo automatizado de *webscraping*.

3.2.2 Dados biográficos das autoridades do CADE

Diante da insuficiência de informações biográficas sobre os Conselheiros no Portal do CADE, realizou-se coleta de dados públicos constantes na Plataforma Lattes[36] e Diário Oficial da União. Assim, coletou-se dados dos Conselheiros referentes à formação acadêmica, exercício de função ou cargo público anteriormente a nomeação, mandato no CADE e processos de cartel julgados. Embora os nomes dos Conselheiros sejam associados aos atos que eles realizam, bem como todos os processos tratados na pesquisa serem de natureza pública, constantes em bancos de dados disponibilizados na rede mundial de computadores, será adotado um rótulo para nominar as Autoridades do CADE.

Não se investigou dados pessoais dos Membros do MPF que oficiaram nos processos de interesse da pesquisa. Isto se deu em razão dos princípios da Unidade, da Independência Funcional e da Indivisibilidade do Ministério Público[37], bem como decorre de sua atuação no processo como fiscal da Lei (*custos legis*). Se decidíssemos de outra maneira, com respeito aos

[36] https://lattes.cnpq.br/

[37] Os princípios institucionais estabelecem, em apertada síntese, que o do Ministério Público é único e sua divisão é meramente funcional (Unidade); que o membro do *Parquet* tem autonomia de convicção (Independência Funcional); e que seus membros agem em nome da Instituição e não por eles mesmos, por isso a possibilidade de um membro substituir o outro, dentro da mesma função, sem consequências processuais de mérito (Indivisibilidade).

princípios anunciados, seria preciso levantar informações de todos os membros do MPF, não se adequando aos objetivos desta pesquisa.

3.2.3 Limitações e considerações

É importante notar que a base de dados final, que posteriormente constituiu-se no *corpus*, depende da disponibilidade e da qualidade das informações fornecidas pelas APIs do CADE. Além disso, a análise está sujeita às limitações inerentes às técnicas de PLN e aos métodos estatísticos empregados. Apesar dessas limitações, a pesquisa oferece *insights* valiosos sobre o impacto da legislação antitruste nos julgamentos do CADE, particularmente no contexto de cartéis em licitações.

Ademais, assumiu-se como resultado de mérito a primeira decisão colegiada exarada nos autos, independentemente da decisão haja sido reformada em instância administrativa (novo julgamento no CADE) ou por instância judicial. Tal medida se mostra necessária pois o julgamento do primeiro mérito do processo se deu exatamente com o conhecimento pleno do contido nos documentos de parecer do MPF e do voto do conselheiro relator acostado aos autos.

3.3 Procedimento de coleta de dados

Na pesquisa acadêmica, a coleta de dados é um processo fundamental, especialmente em estudos que envolvem a análise de práticas anticompetitivas. Este estudo emprega a API "Cade em Números"como instrumento primordial na coleta de informações sobre o Conselho Administrativo de Defesa Econômica (CADE) até o ano de 2016. A escolha dessa ferramenta e o período determinado para a coleta refletem uma abordagem metódica e estratégica para garantir a relevância e a integridade dos dados.

O procedimento de coleta de dados iniciou-se com a seleção criteriosa da API "Cade em Números", reconhecida por sua extensa e organizada base de informações sobre condutas anticompetitivas. Esta plataforma foi escolhida devido à sua capacidade de proporcionar acesso facilitado e estruturado a um conjunto substancial de dados. A definição do recorte temporal até o ano de 2016 fundamentou-se na análise preliminar que revelou uma maior disponibilidade e confiabilidade de dados estruturados neste período, tornando-o o mais apropriado para a pesquisa.

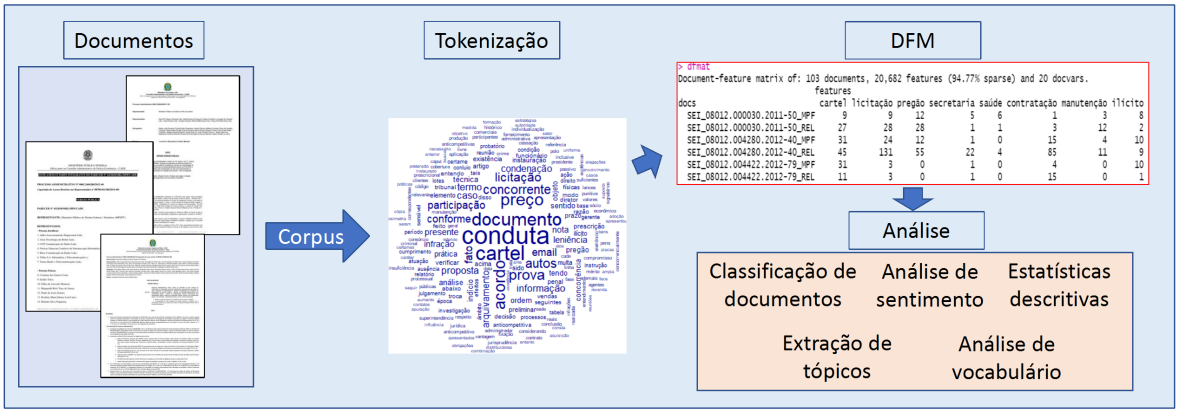
A coleta foi conduzida através de consultas diretas à API do CADE, uma abordagem que permitiu a identificação rápida e eficiente de dados estruturados relevantes. Este processo não apenas otimizou a coleta, mas também estabeleceu uma base sólida para as investigações subsequentes. Após a coleta, os dados passaram por um meticuloso processo de verificação e organização. Este tratamento inicial foi crucial para assegurar que os dados estivessem em um formato compatível com as ferramentas de análise e preparados para o processamento e interpretação subsequentes.

Uma decisão significativa tomada durante este processo foi a de não empregar técnicas de *webscraping*. A ausência de uma estrutura rígida de rotulação nos documentos disponíveis na API apresentou um desafio, o qual poderia comprometer a precisão e a relevância das informações coletadas. Por isso, optou-se por uma abordagem manual na seleção dos dados, priorizando a precisão e a qualidade das informações coletadas, apesar do aumento no tempo despendido nesta fase.

3.4 Procedimentos de pré e pós processamento do *corpus*

A análise de textos em PLN é um processo complexo que se inicia com uma etapa crucial: o pré-processamento dos dados. Esta fase é dedicada à preparação do *corpus* de texto para análise, envolvendo uma série de operações essenciais para limpar e organizar os dados. Inicialmente, ocorre a limpeza de dados, que implica na remoção de caracteres indesejados, como símbolos de pontuação e números, além da correção de erros de digitação. Segue-se a tokenização, um processo de divisão do texto em unidades menores, como palavras ou frases. Outra etapa importante é a remoção de *stopwords*, que são palavras comuns com pouca contribuição significativa para o significado do texto, como "e", "o", "em", etc. A fase de pré-processamento inclui também o *stemming* ou lematização, que reduz as palavras às suas raízes ou formas base, e a normalização, como a conversão de todo o texto para letras minúsculas.

Figura 12 – Desenho da Pesquisa



Fonte: elaborado pelo autor.

Conforme o *framework* ilustrado na Figura 12, uma visão simplificada daquela na Figura 1, o pacote*quanteda*, do R, é utilizado para realizar as tarefas de pré-processamento em um fluxo contínuo, empregando funções como *corpus()*, *tokens()* e *dfm()*. Este procedimento é cíclico, retornando sempre à fase anterior até que se obtenha um corpus ideal para a análise subsequente. A conversão do texto em caracteres minúsculos, a remoção de *stopwords* e pontuação, a tokenização e a elaboração da *Document-Feature Matrix* (DFM) são etapas integrantes deste processo. A DFM, oriunda da conversão do corpus *em* uma matriz de contabilização de *features*

por documento, representa as frequências das features, mas não traz informações sobre a posição das palavras nos documentos, devido à natureza não posicional do modelo de saco de palavras.

Posteriormente, o pós-processamento é realizado para refinar os resultados obtidos e extrair *insights* mais profundos. Inclui-se a análise de sentimentos, avaliando o tom emocional dos textos e classificando-os como positivos, negativos ou neutros. A identificação de temas e tópicos é realizada com técnicas de modelagem de tópicos, que descobre tópicos recorrentes nos textos. A visualização de dados, por meio de gráficos e mapas de calor, é outra etapa crucial para representar visualmente os resultados da análise e facilitar a interpretação. Além disso, agrupamento e classificação de textos, utilizando algoritmos de AM, são empregados para categorizar os textos em grupos semelhantes. A validação e ajuste de modelos também são fundamentais, avaliando a precisão dos modelos utilizados na análise e fazendo ajustes conforme necessário para aprimorar a acurácia.

Conforme explicado por Bartholomay (2021, p. 33), as técnicas de Topic Model são empregadas em problemas de aprendizado não supervisionado, em que o conteúdo dos tópicos do estudo é inferido ao invés de assumido. O autor esclarece que, ao utilizar essa técnica, o pesquisador não define os tópicos antecipadamente. Além disso, o citado autor informa que diversos métodos e suposições podem ser utilizados para validar e avaliar dados em texto, variando desde a leitura manual dos textos até modelos de aprendizado não supervisionado e Topic Model, sendo que cada método possui um custo envolvendo a análise dos textos.

3.4.1 Ferramentas de modelagem de tópicos

Foram utilizados os recursos conforme o Quadro 1. Foram usados modelos treinados em português do Brasil (pt-BR), mas para permitir comparações com trabalhos relacionados, também foram utilizados modelos multilíngue.

As diferenças entre encodes do tipo *transformer*, *sentence transformer BERT* e *sentence transformer DistilBERT* podem ser compreendidas ao explorar as características de cada um e o contexto em que são usados, especialmente no campo PLN.

A arquitetura *Transformer* foi introduzida por Vaswani *et al.* (2017, p. 2) e é fundamental no desenvolvimento dos novos modelos em PLN, servindo de base para modelos como BERT, GPT e DistilBERT. Seu principal avanço é o "mecanismo de atenção"[38], que avalia a importância de diferentes partes de um texto.

O *Sentence Transformer BERT* aplica a arquitetura *Transformer* no BERT para gerar

[38] O mecanismo de atenção permite que o modelo dê mais importância a certas palavras sobre outras ao processar uma frase. Para Vaswani *et al.* (2017, p. 3) uma função de atenção pode ser descrita como o mapeamento de uma consulta e um conjunto de pares de valores-chave para uma saída, onde a consulta, as chaves, os valores e a saída são todos vetores. A saída é calculada como uma soma ponderada dos valores, onde o peso atribuído a cada valor é calculado por uma função de compatibilidade da consulta com a chave correspondente. Isso ajuda o modelo a entender melhor o significado e a relação entre as palavras em um texto.

Quadro 1 – Modelos utilizados na pesquisa

Tipo de algoritmo	Encode	Recurso	Modelo
Embedding	Transformer BERT	neuralmind/bert-base-portuguese-cased	BERTopic-BERTimbau
Embedding	Sentence Transformer DistilBERT	distiluse-base-multilingual-cased	BERTopic-DistilUSE
Embedding	Sentence Transformer BERT	ulysses-camara/legal-bert-pt-br	BERTopic-Legal BERT
Embedding	Sentence Transformer BERT	rufimelo/Legal-BERTimbau-sts-large	BERTopic-Legal BERTimbau
Embedding	Transformer BERT	felipemaiapolo/legalnlp-bert	BERTopic-LegalNLP
Embedding	Sentence Transformer BERT	paraphrase-multilingual-MiniLM-L12-v2	BERTopic-MiniLM
Embedding	text-embedding-ada-002	OpenAI	BERTopic-OpenAI
Embedding	Sentence Transformer BERT	stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0	BERTopic-STJIris
Embedding	Transformer	universal-sentence-encoder-multilingua	BERTopic-USE Multi
Neural	Não usa	OCTIS	CTM
Tradicional	Não usa	OCTIS	LDA
Tradicional	Não usa	OCTIS	NMF
Híbrido	Sentence Transformer DistilBERT	distiluse-base-multilingual-cased	Top2Vec-DistilUSE
Híbrido	Sentence Transformer BERT	ulysses-camara/legal-bert-pt-br	Top2Vec-Legal BERT
Híbrido	Sentence Transformer BERT	paraphrase-multilingual-MiniLM-L12-v2	Top2Vec-MiniLM
Híbrido	Sentence Transformer BERT	stjiris/bert-large-portuguese-cased-legal-mlm-sts-v1.0	Top2Vec-STJIris
Híbrido	Transformer	universal-sentence-encoder-multilingual	Top2Vec-USE Multi

Fonte: elaborado pelo autor.

embeddings de sentenças. O BERT se diferencia pelo seu treinamento bidirecional[39], permitindo a compreensão contextual baseada em todas as palavras de uma sentença. Essa adaptação foca em *embeddings* de sentenças inteiras, ao invés de palavras individuais.

O "Sentence Transformer DistilBERT"utiliza o DistilBERT, uma versão mais eficiente do BERT desenvolvida pela *Hugging Face*. Ele oferece um equilíbrio entre precisão e eficiência, sendo adequado para ambientes com recursos limitados.

O "text-embedding-ada-002"da OpenAI é um modelo de *embedding* de texto que se diferencia dos modelos *Transformer* citados (BERT e DistilBERT). Embora possua algumas semelhanças em termos de tecnologia e aplicação, sendo projetado para gerar incorporações de texto, que na verdade são representações vetoriais de textos (frases, parágrafos, documentos). O modelo "ada"é parte de uma série de modelos com diferentes níveis de capacidade e eficiência, oferecendo um equilíbrio entre desempenho e uso de recursos.

ainda em relação ao modelo da OpenAI, enquanto BERT e DistilBERT são exemplos de modelos baseados na arquitetura *Transformer* original, com ênfase na compreensão de contexto e relações bidirecionais no texto, os modelos da OpenAI, como "text-embedding-ada-002", são otimizados para tarefas específicas como geração de embeddings.

Em síntese, enquanto *transformer* refere-se à arquitetura de modelo geral, *sentence transformer BERT* e *sentence transformer DistilBERT* são implementações específicas dessa arquitetura, otimizadas para gerar *embeddings* de sentenças completas com eficiência e precisão,

[39] O treinamento bidirecional do BERT permite que o modelo aprenda o significado das palavras considerando o contexto completo da frase, analisando tanto as palavras antes quanto depois de uma palavra específica. Isso difere de métodos mais antigos que focavam apenas em palavras anteriores ou posteriores, proporcionando uma compreensão mais profunda do contexto e significado das palavras.

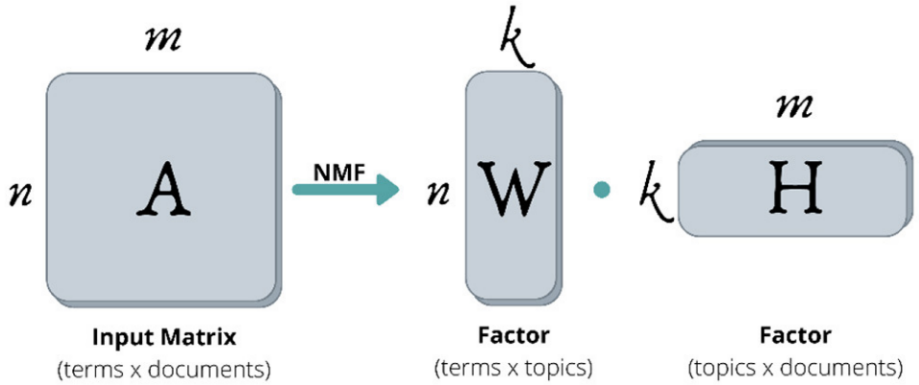
com o DistilBERT oferecendo uma alternativa mais leve e rápida ao BERT.

3.4.1.1 NMF

NMF (Non-negative Matrix Factorization) é um método de aprendizado de máquina que visa encontrar uma representação de partes em dados observáveis, como imagens ou documentos de texto. Ele é usado para decompor uma matriz não negativa em duas matrizes de fatores não negativos, W e H, de forma que sua multiplicação se aproxime da matriz original.

A ideia principal por trás do NMF é que cada variável oculta (representada por uma coluna em H) coativa um subconjunto de variáveis observáveis (representadas por colunas em V), chamadas de "partes". A ativação de uma constelação de variáveis ocultas combina essas partes de forma aditiva para gerar um todo. O NMF utiliza restrições de não-negatividade para permitir apenas combinações aditivas, o que resulta em uma representação baseada em partes.

Figura 13 – Representação gráfica do modelo NMF



Fonte: Egger e Yu (2022, p. 5).

A Figura 13 ilustra o processo de fatoração de matrizes não negativas para a modelagem de tópicos, cujo detalhamento dos componentes é o seguinte:

- **Matriz de Entrada** A: Com dimensões $n \times m$, onde n representa o número de termos e m o número de documentos. As entradas da matriz indicam a frequência ou importância dos termos nos documentos.
- **NMF**: O algoritmo que decompõe a matriz A em duas matrizes de fatores W e H, sujeitas à condição de não negatividade dos elementos.
- **Matriz de Fatores** W: De dimensões $n \times k$, onde k é o número de tópicos. As colunas representam tópicos como distribuições de palavras, e as linhas correspondem às palavras.
- **Matriz de Fatores** H: Com dimensões $k \times m$, representa a distribuição de tópicos nos documentos. As colunas são documentos e as linhas são tópicos.

- **Produto das Matrizes de Fatores**: Indicado pelo ponto entre W e H, representa a multiplicação das duas matrizes para reconstruir aproximadamente a matriz de entrada A.

A finalidade do NMF na modelagem de tópicos é identificar padrões latentes nos textos. Documentos são expressos como combinações de tópicos, e os tópicos são caracterizados por palavras frequentemente co-ocorrentes. Sua formulação matemática pode ser descrita como a busca por uma aproximação da matriz original *A* por meio da multiplicação de duas matrizes não negativas, *W* e *H*:

$$A \approx W \times H \tag{3}$$

Onde:

- A é a matriz original de dados observáveis;
- W é a matriz de partes, onde cada coluna representa uma parte e cada elemento é não negativo;
- H é a matriz de ativações, onde cada coluna representa uma combinação de partes e cada elemento é não negativo.

O objetivo do NMF é encontrar as matrizes W e H que melhor aproximam a matriz original A. Isso é feito por meio de iterações de atualização dos valores de W e H, utilizando regras de atualização específicas. O algoritmo do NMF converge para um máximo local da função objetivo, que é determinada pela diferença entre a matriz original A e a matriz aproximada *W* x *H*.

O NMF tem sido aplicado em diversas áreas, como análise de imagens, análise de texto, processamento de sinais e reconhecimento de padrões. Ele é especialmente útil quando se deseja encontrar uma representação baseada em partes dos dados, em oposição a uma representação holística fornecida por métodos como a Análise de Componentes Principais (PCA) ou a Quantização Vetorial (VQ).

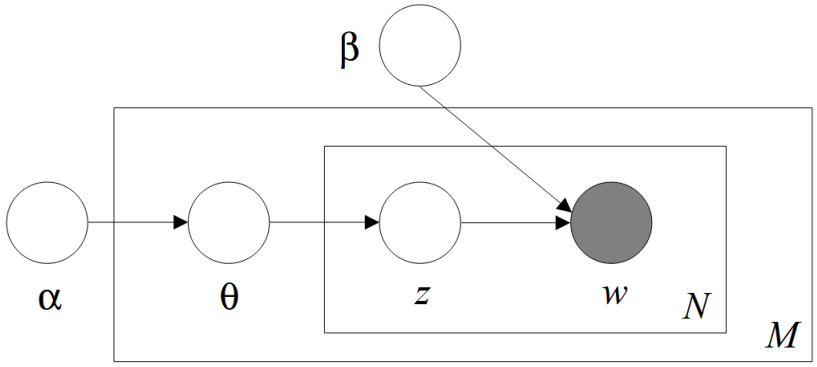
3.4.1.2 LDA

Latent Dirichlet Allocation (LDA) é um modelo probabilístico generativo para coleções de dados discretos, como textos. Utilizado para descobrir tópicos latentes em documentos, o LDA baseia-se na suposição de troca simples para palavras e tópicos, aplicando o teorema de representação de *de Finetti*[40].

[40] Segundo Blei, Ng e Jordan (2003, p. 994), um teorema de representação clássico devido a *de Finetti* estabelece que qualquer coleção de variáveis aleatórias trocáveis tem uma representação como uma distribuição de mistura – em geral uma mistura infinita. Assim, se quisermos considerar representações intercambiáveis para documentos e palavras, precisamos considerar modelos de mistura que capturem a permutabilidade de palavras e documentos.

Como uma técnica de redução de dimensionalidade, o LDA pode ser considerado uma extensão do modelo de indexação semântica latente (LSI), oferecendo uma semântica probabilística subjacente adequada aos dados modelados. É um modelo hierárquico bayesiano de três níveis, onde cada item de uma coleção é modelado como uma mistura finita de tópicos, que por sua vez são misturas infinitas de probabilidades de tópicos subjacentes.

Figura 14 – Representação gráfica do modelo de LDA



Fonte: Blei, Ng e Jordan (2003, p. 997).

A Figura 14 representa um diagrama de notação de placa para o modelo descrito Blei, Ng e Jordan (2003, p. 997), cuja descrição de cada componente é:

- **Círculos**: Representam variáveis probabilísticas, onde círculos vazios são variáveis latentes a serem aprendidas e círculos preenchidos são variáveis observadas nos dados.
- **Placas**: As caixas retangulares indicam replicação. A placa externa com o rótulo M representa os documentos e a placa interna com o rótulo N representa as palavras em cada documento, demonstrando a repetição do processo para cada documento e palavra.
- α: Hiperparâmetro que influencia a distribuição Dirichlet das distribuições de tópicos nos documentos, onde valores maiores indicam uma mistura mais uniforme de tópicos em cada documento.
- β: Hiperparâmetro que afeta a distribuição Dirichlet das palavras nos tópicos, com valores maiores sugerindo que cada tópico contém uma mistura mais uniforme de palavras.
- θ_d: As distribuições de tópicos para os documentos, com cada documento d tendo sua própria distribuição θ_d.
- z_{dn}: Variável de alocação de tópicos que designa um tópico para cada palavra n em um documento d.
- w_{dn}: Variável observada, representando a palavra n no documento d.

Este diagrama encapsula a descrição concisa do modelo generativo que o LDA utiliza para explicar a geração de documentos em um corpus. O processo generativo descrito pelo modelo é como segue: para cada documento d, uma distribuição de tópicos θ_d é escolhida a partir de uma distribuição Dirichlet com parâmetro α. Para cada palavra n no documento d, um tópico z_{dn} é escolhido a partir de θ_d, e uma palavra w_{dn} é observada, selecionada a partir da distribuição de palavras do tópico z_{dn}, parametrizada por β.

Para Blei, Ng e Jordan (2003, p. 997), a formulação matemática do LDA envolve distribuições de Dirichlet e é representada como um modelo gráfico probabilístico de três níveis. Os parâmetros α e β são parâmetros de nível de corpus, enquanto θ_d são variáveis de nível de documento, e z_{dn} e w_{dn} são variáveis de nível de palavra.

Para cada documento d no corpus:

$$\theta_d \sim \text{Dir}(\alpha) \tag{4}$$

Para cada palavra n no documento d:

$$\begin{aligned} z_{dn} &\sim \text{Multinomial}(\theta_d) \\ w_{dn} &\sim \text{Multinomial}(\beta_{z_{dn}}) \end{aligned} \tag{5}$$

A formulação geral do modelo é dada por:

$$P(w, z, \theta, \phi|\alpha, \beta) = P(\theta|\alpha) \cdot P(\phi|\beta) \cdot P(z|\theta) \cdot P(w|\phi, z) \tag{6}$$

Essa equação descreve a distribuição conjunta completa do LDA, utilizada para inferir as distribuições latentes de tópicos e palavras a partir dos documentos. Embora a inferência exata seja intratável, abordagens aproximadas como a aproximação variacional e o método de Monte Carlo de cadeias de Markov são comuns.

O LDA se destaca por sua modularidade e extensibilidade, podendo ser adaptado para diferentes tipos de dados e aplicações, incluindo modelagem de texto, classificação de texto e filtragem colaborativa.

3.4.1.3 Combined Topic Model (CTM)

Bianchi, Terragni e Hovy (2021) e Bianchi, Terragni, Hovy *et al.* (2021) não definem explicitamente o que é CTM (Combined Topic Model). No entanto, podemos inferir que o CTM é um modelo de tópicos que combina representações contextualizadas de documentos com o modelo de tópicos ProdLDA. O CTM utiliza representações de documentos geradas pelo SBERT (Sentence-BERT), que é uma extensão do modelo BERT para geração rápida de *embeddings* de sentenças. Essas representações são combinadas com a representação do tipo BoW do ProdLDA para melhorar a coerência e a interpretabilidade dos tópicos descobertos. O CTM demonstrou

produzir tópicos mais significativos e coerentes em comparação com modelos tradicionais de tópicos baseados em BoW e modelos de tópicos neurais recentes.

Importa registrar que o modelo propõe um método inovador e prático para a criação de tópicos coerentes utilizando modelos de tópicos neurais. Os estudos destacam a eficácia do uso de incorporações de documentos contextualizados, demonstrando que eles produzem tópicos significativamente mais coerentes em comparação com as representações tradicionais baseadas em um saco de palavras (BoW) (Papadia *et al.*, 2023, p. 11–13). As pesquisas evidenciam que a incorporação de informações contextuais latentes é benéfica para a modelagem de tópicos, abordando uma questão central nesta área. Como parte de sua contribuição, os autores disponibilizaram uma implementação do modelo, conhecida como *Contextualized Topic Model*[41] Bianchi, Terragni e Hovy (2021).

3.4.1.4 Top2Vec

No Top2Vec, desenvolvido por Dimo Angelov Angelov (2020) em seu artigo *Top2Vec: Distributed Representations of Topics*, o número de tópicos não é diretamente controlável da mesma maneira que os modelos apresentados anteriormente. O Top2Vec utiliza aprendizado não supervisionado e dimensionalidade reduzida para identificar tópicos de forma orgânica, com base na estrutura intrínseca dos dados.

Alguns autores apresentam formas de influenciar o número de tópicos. Por exemplo, através do ajuste de parâmetros em algoritmos de redução de dimensionalidade como o UMAP. Conforme McInnes, Healy e Melville (2020) discutem em seu artigo sobre UMAP, parâmetros como $n_neighbors$ e min_dist são cruciais para determinar como os dados são agrupados. Alterações nesses parâmetros podem afetar significativamente a estrutura dos tópicos identificados. Outros autores defendem alterações no processo de clusterização, um componente chave do Top2Vec, que pode ser influenciado ajustando-se os parâmetros do HDBSCAN, conforme descrito por Campello, Moulavi e Sander (2013). Os parâmetros como $min_cluster_size$ e $min_samples$ são essenciais para definir a densidade e o tamanho dos *clusters*, impactando diretamente o número de tópicos descobertos.

Ao ajustar esses parâmetros, é importante considerar o equilíbrio entre a granularidade dos tópicos e a relevância/coerência deles, como discutido por Blei, Ng e Jordan (2003). Esses autores destacam a importância da coerência e relevância dos tópicos em qualquer modelo de modelagem de tópicos. Assim, ajustes nos parâmetros podem levar à formação de tópicos mais genéricos ou específicos, dependendo da natureza da alteração. Nessa pesquisa preferiu-se utilizar a parametrização padrão utilizada por Grootendorst (2022), uniformizando os resultado visando a comparabilidade das métricas.

[41] Apesar de ter a mesma abreviatura, a biblioteca *Contextualized Topic Model* é uma das ferramentas computacionais para se modelar o *Combined Topic Model*. A biblioteca Python está disponível no seguinte link: `https://github.com/MilaNLProc/contextualized-topic-models`.

3.4.1.5 BERTopic

O BERTopic, desenvolvido por Grootendorst (2022), representa uma abordagem inovadora na modelagem de tópicos, destacando-se por sua similaridade estrutural com o Top2Vec, mas com características distintas e avançadas. Enquanto ambos os modelos empregam o BERT como incorporador de linguagem, o BERTopic se diferencia pela sua capacidade de extrair incorporações de documentos em mais de 50 idiomas, utilizando um sofisticado modelo de transformadores de frases. Este aspecto do BERTopic é crucial, pois permite uma análise mais profunda e abrangente dos documentos, abrindo caminho para interpretações mais ricas e diversificadas dos dados textuais.

Um dos aspectos centrais do BERTopic é a implementação do UMAP (Uniform Manifold Approximation and Projection) para a redução de dimensionalidade. O UMAP é uma técnica poderosa que permite uma representação mais eficiente dos dados em um espaço de menor dimensão, mantendo ao mesmo tempo as propriedades essenciais dos dados originais (McInnes; Healy; Melville, 2020). Essa técnica é particularmente eficaz na preservação das relações locais e globais entre os pontos de dados, facilitando a identificação de estruturas e padrões subjacentes nos dados textuais. A redução de dimensionalidade é um passo crítico na modelagem de tópicos, pois lida com a complexidade inerente aos dados de alta dimensão e melhora a eficiência computacional do processo de modelagem.

Além do UMAP, o BERTopic também incorpora o HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) para o agrupamento de documentos. O HDBSCAN é um algoritmo avançado que identifica clusters baseados em densidade, tratando eficientemente os dados dispersos como ruído (Campello; Moulavi; Sander, 2013). Esta abordagem é especialmente útil em conjuntos de dados textuais, onde a identificação de grupos coesos de documentos pode ser desafiadora devido à variabilidade e à natureza esparsa dos dados. O HDBSCAN fornece uma maneira robusta e flexível de agrupar documentos, o que é essencial para a extração significativa de tópicos.

A principal inovação do BERTopic, conforme destacado por reside na aplicação do algoritmo c-TF-IDF, que compara a importância dos termos dentro de um *cluster* para criar representações de termos. Este método difere significativamente das abordagens tradicionais, pois leva em conta a frequência dos termos dentro de um *cluster* específico em relação à sua frequência em outros *clusters* (Sánchez-Franco; Rey-Moreno, 2022, p. 8–10). Assim, o c-TF-IDF proporciona uma medida mais precisa da relevância dos termos, permitindo que o BERTopic identifique os termos mais representativos de cada tópico.

Em contraste com o LDA, o BERTopic oferece modelagem de tópicos de forma contínua, não discreta. Esta natureza contínua e estocástica do modelo resulta em variações nos resultados a cada execução, o que implica em uma rica diversidade na interpretação dos tópicos. Após a conclusão do modelo, os pesquisadores podem identificar os tópicos mais pertinentes, notando-se

que o Tópico 0, com uma contagem de -1, representa valores discrepantes e, portanto, não deve ser considerado em análises subsequentes. O BERTopic também oferece funcionalidades para a busca de palavras-chave e a visualização dos tópicos mais relevantes com base em sua pontuação de similaridade, além da possibilidade de inspecionar tópicos individuais através de suas palavras-chave.

Finalmente, para uma análise mais abrangente da potencialmente grande variedade de tópicos, o BERTopic proporciona um mapa interativo de distância intertópico, como ilustrado na Figura 3. Esta ferramenta visual permite aos pesquisadores uma visão geral inicial dos tópicos e facilita a realização de uma redução automática de tópicos, conforme necessário. Esta capacidade de visualização e interação com os dados é um aspecto notável do BERTopic, que enriquece significativamente o processo de análise e interpretação dos tópicos gerados pelo modelo (Grootendorst, 2022). Em suma, o BERTopic emerge como uma ferramenta poderosa e flexível na modelagem de tópicos, oferecendo *insights* valiosos e profundos sobre conjuntos de dados textuais complexos.

3.4.2 Métricas de avaliação

Serão utilizadas métricas distintas conforme a etapa de resolução do problema de pesquisa e modelo de *machine learning* utilizado, conforme detalhamento nas subseções abaixo.

3.4.2.1 Modelo não supervisionado

Para realizar a avaliação dos modelos utilizou-se do OCTIS para treinar, analisar e comparar modelos de tópicos. O OCTIS é uma biblioteca escrita em *Python* que permite aos pesquisadores treinar modelos existentes, integrar novos algoritmos de treinamento e inferência, e comparar justamente os modelos de tópicos de interesse. Além disso, essa ferramenta utiliza a otimização bayesiana para ajustar os hiperparâmetros dos modelos de tópicos. Essa otimização busca encontrar as melhores configurações de hiperparâmetros para um determinado conjunto de dados e métrica de avaliação (Terragni; Harrando *et al.*, 2022, p. 328; Terragni; Fersini *et al.*, 2021, p. 263; Terragni; Fersini, 2021, p. 1409–1411).

As métricas NPMI, UMass e *Topic Diversity* são comumente usadas para avaliar modelos de tópicos, enquanto o Tempo de Computação é uma métrica de eficiência. Em uma apertada síntese, temos que:

a) **NPMI**: mede o grau de associação entre as principais palavras em um tópico. É uma medida de coerência[42];

b) **UMass**: mede a frequência em que duas palavras aparecem juntas. Também é uma medida de coerência;

[42] Importa registrar que esta medida de coerência emula o julgamento humano com desempenho razoável (Grootendorst, 2022, p. 5; Lau; Newman; Baldwin, 2014, p. 531, 532, 536)

c) ***Topic Diversity***: mostra quão distintos os tópicos são um do outro;

d) **Tempo de computação**: aponta o tempo decorrido para se calcular o modelo.

Nas subseções a seguir são elaboradas explicações mais detalhadas sobre cada uma dessas medidas utilizadas nessa pesquisa, além daquelas obtidas na seção de Revisão de Literatura, subseção 2.2.2 (p. 21).

Normalized Pointwise Mutual Information (NPMI):

Esta métrica avalia a qualidade dos tópicos gerados pelo modelo. O NPMI mede a associação entre palavras dentro de um tópico, normalizando o valor de Informação Mútua Pontual (PMI). Valores mais altos indicam que as palavras em um tópico estão mais fortemente relacionadas, sugerindo tópicos mais coesos e significativos.

Para Bouma (2009, p. 36–38), a métrica NPMI é fundamental na avaliação da qualidade de modelos de tópicos, especialmente na determinação da coerência e relevância dos tópicos gerados. Esta métrica tem sido aplicada em diversos estudos acadêmicos para aprimorar a interpretação e eficácia de modelos de tópicos. Esse autor explora a aplicação do NPMI na análise de colocações linguísticas, demonstrando como essa métrica pode ser utilizada para medir a força da associação entre palavras.

Além disso, Rosner *et al.* (2014), no artigo *Evaluating Topic Coherence Measures*, investigam diferentes medidas de coerência de tópicos, incluindo o NPMI. Eles avaliam a eficácia de várias métricas na identificação de tópicos coerentes, ressaltando a relevância do NPMI nesse contexto. Nesse trabalho, os autores discutem a métrica NPMI como uma medida de coerência em conjuntos de palavras. Ainda segundo esses autores, a métrica é usada para avaliar a interpretabilidade de tópicos em modelos de tópicos e que essa métrica tem sido amplamente estudada e comparada com outras medidas de coerência (Rosner *et al.*, 2014, p. 401–402).

A fórmula para o cálculo do NPMI é dada por:

$$NPMI(w_1, w_2) = \log\left(\frac{P(w_1, w_2)}{P(w_1) \cdot P(w_2)}\right) / -\log(P(w_1, w_2)) \tag{7}$$

onde w_1 e w_2 são as palavras que estão sendo comparadas, $P(w_1, w_2)$ é a probabilidade conjunta das palavras w_1 e w_2 ocorrerem juntas, $P(w_1)$ é a probabilidade da palavra w_1 ocorrer e $P(w_2)$ é a probabilidade da palavra w_2 ocorrer.

É uma uma medida de associação que opera em um intervalo de valores compreendido entre -1 e 1, oferecendo uma escala padronizada e interpretável para a avaliação da relação entre termos em diversos contextos de análise textual. Um valor de -1 na escala NPMI indica uma associação negativa perfeita, sugerindo que a ocorrência de um termo é inversamente relacionada à presença do outro. Por outro lado, um valor de 0 representa a independência estatística, onde a presença ou ausência de um termo não influencia a ocorrência do outro. Em contraste, um valor de 1 denota uma associação positiva perfeita, implicando que a presença de um termo garante a presença do outro.

Ainda, Newman *et al.* (2010) empregam o NPMI para avaliar a coerência de tópicos em modelos de tópicos. Este estudo demonstra a aplicabilidade do NPMI na avaliação automática da qualidade dos tópicos, enfatizando sua importância em pesquisas que lidam com grandes volumes de dados textuais.

Por fim, A NPMI é amplamente reconhecida por sua eficácia na medição da força da associação entre palavras ou conceitos, especialmente no âmbito do processamento de linguagem natural e análise de tópicos. Sua normalização permite a comparação entre diferentes conjuntos de dados e contextos, independentemente das frequências absolutas dos termos, tornando-a uma ferramenta valiosa para pesquisadores e analistas que buscam *insights* qualitativos e quantitativos em textos.

University of Massachusetts Coherence (UMass):

Similar ao NPMI, a métrica UMass mede a coerência de um tópico, baseando-se na força da associação entre as palavras de um tópico. Diferentemente do NPMI, ela utiliza contagens de documentos ao invés de informações pontuais. Valores mais altos de UMass indicam tópicos mais coerentes. Uma referência importante para esta métrica é o trabalho é o artigo *Optimizing Semantic Coherence in Topic Model* (Mimno *et al.*, 2011, p. 265–266).

A formulação matemática da UMass é dada pela seguinte expressão:

$$\text{UMass} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \log \left(\frac{P(w_i, w_j) + \epsilon}{P(w_j)} \right) \tag{8}$$

onde:

- N é o número total de palavras no conjunto de palavras;
- $P(w_i, w_j)$ é a probabilidade conjunta das palavras w_i e w_j;
- $P(w_j)$ é a probabilidade marginal da palavra w_j;
- ϵ é um valor pequeno adicionado ao numerador para evitar a divisão por zero.

Essa formulação é utilizada para calcular a coerência de um conjunto de palavras com base na probabilidade conjunta e marginal das palavras. É fundamentada na premissa de que palavras semanticamente relacionadas tendem a aparecer conjuntamente em maior frequência, serve como um indicativo significativo da coerência de um tópico.

Operando em uma escala predominantemente negativa, a UMass varia de valores próximos a 0, denotando alta coerência, até valores extremamente negativos, que indicam uma coerência mais baixa. Valores que se aproximam de 0 refletem uma forte relação semântica entre as palavras dentro de um tópico, sugerindo que o tópico é bem definido e consistente. Por outro lado, valores mais negativos indicam uma falta de relação significativa entre as palavras, apontando para tópicos menos coerentes ou até aleatórios. Por ser baseada em dados de frequência

de palavras, a métrica UMass é particularmente sensível ao tamanho e à natureza do corpus de documentos.

Calculada a partir da frequência e co-ocorrência de palavras nos documentos, a UMass avalia a probabilidade de co-ocorrência de pares de palavras em comparação com suas ocorrências individuais, identificando padrões que revelam relações significativas entre as palavras de um tópico. A sensibilidade desta métrica ao tamanho e características do *corpus* documental a torna uma ferramenta valiosa para pesquisadores e cientistas de dados. Ela é especialmente útil na análise e otimização de modelos de tópicos em grandes conjuntos de dados textuais, oferecendo *insights* quantitativos e qualitativos para aprimorar a interpretação e eficácia desses modelos.

Topic diversity:

Esta métrica avalia a diversidade dos tópicos gerados por um modelo. Uma maior diversidade indica que o modelo é capaz de produzir uma variedade mais ampla de tópicos, o que pode ser desejável em certas aplicações para capturar uma gama mais ampla de temas ou ideias. Não há uma referência padrão para esta métrica, pois ela pode ser calculada de várias maneiras, dependendo da abordagem específica do modelo.

No âmbito acadêmico, a métrica de diversidade em modelos de tópicos tem sido amplamente explorada e valorizada, como evidenciado por vários estudos significativos. Dieng, Ruiz e Blei (2020), por exemplo, em seu trabalho intitulado *Topic Modeling in Embedding Spaces*, propuseram o *Embedded Topic Model* (ETM), uma abordagem inovadora que integra incorporações de palavras na modelagem de tópicos. Neste estudo, a diversidade de tópicos foi empregada como uma métrica chave para avaliar a eficácia do modelo proposto, ressaltando sua importância na avaliação de modelos de tópicos. Os autores Definiram diversidade de tópicos (*Topic Diversity*) como a porcentagem de palavras únicas nas 25 principais palavras de todos os tópicos (Dieng; Ruiz; Blei, 2020, p. 448). Diversidade próxima de 0 indica temas redundantes; diversidade próxima de 1 indica temas mais variados.

Chuang *et al.* (2012) não apresentam uma definição explícita para “diversity” ou “topic diversity”. No entanto, o artigo discute a importância de identificar áreas de sobreposição temática em diferentes departamentos acadêmicos e a necessidade de medir a similaridade entre essas áreas. Isso sugere que “diversity” pode se referir à variedade de tópicos abordados em diferentes departamentos ou à presença de interdisciplinaridade na pesquisa universitária. No contexto do artigo, “topic diversity” pode se referir à diversidade de tópicos de pesquisa abordados em diferentes departamentos. enfatizaram a relevância da diversidade dos tópicos. Os autores se concentraram em desenvolver visualizações para análise de texto baseadas em modelos, utilizando a diversidade dos tópicos como um meio para avaliar a interpretabilidade dos tópicos gerados, o que é crucial para a análise qualitativa em diversas aplicações.

Da mesma forma, embora o artigo não forneça uma definição formal ou uma fórmula matemática específica para “diversity” ou “topic diversity”, Peinelt, Nguyen e Liakata (2020)

sugerem que a incorporação de modelos de tópicos pode contribuir para a diversidade de tópicos em modelos de similaridade semântica. Esses autores combinaram modelos de tópicos com BERT para detecção de similaridade semântica. A diversidade de tópicos, neste contexto, foi utilizada para demonstrar a capacidade do modelo em capturar um amplo espectro de tópicos, evidenciando a aplicabilidade e eficiência dessa métrica.

Esses exemplos ilustram a relevância da métrica de diversidade em diferentes áreas de pesquisa, sublinhando sua utilidade não apenas para assegurar a precisão na identificação de tópicos, mas também para garantir uma cobertura abrangente de temas e assuntos nos modelos de tópicos.

Tempo de computação:

Tempo de Computação é uma medida prática que indica quanto tempo o modelo leva para treinar ou inferir tópicos a partir de um conjunto de dados. Essas métricas são fundamentais para avaliar e comparar modelos de tópicos, cada uma focando em um aspecto diferente da qualidade ou eficiência do modelo. Um tempo de computação menor é geralmente preferível, especialmente em aplicações que requerem processamento em tempo real ou que lidam com grandes conjuntos de dados.

Essa métrica é utilizada por Grootendorst (2022) para avaliar diferentes modelos de tópicos entre si, sendo relevante pois aponta a eficiência em se adotar uma ou outra ferramenta de modelagem dentro de um contexto de escassez de recursos computacionais.

Outrossim, cabe afirmar que o artigo de Grootendorst (2022) não conceitua ou define explicitamente o termo *computational time*. No entanto, menciona o uso de otimização bayesiana para ajustar os hiperparâmetros dos modelos de tópicos, o que implica em encontrar as melhores configurações de hiperparâmetros para um determinado conjunto de dados e métrica de avaliação. Esse processo de otimização pode exigir um tempo computacional significativo, dependendo da complexidade do modelo e do tamanho do conjunto de dados. Desse modo, pode-se inferir que tempo de computação se refere ao interregno decorrido para realizar cálculos e processamentos computacionais relacionados à otimização dos modelos de tópicos.

3.4.2.2 Modelo supervisionado (classificação)

Explorou-se diversas métricas utilizadas na avaliação de modelos de classificação em estatística e aprendizado de máquina, destacando sua importância e aplicabilidade em contextos diversos (vide a Tabela 5).

A acurácia, como aponta Japkowicz e Shah (2011), é uma medida fundamental para avaliar o desempenho geral de um modelo, indicando a proporção de previsões corretas realizadas. Entretanto, em cenários onde as classes são desbalanceadas, métricas como a ROC e a AUC, conforme discutido por Fawcett (2006) e Bradley (1997), tornam-se mais relevantes, oferecendo uma visão mais detalhada do comportamento do modelo em diferentes limiares de decisão.

Tabela 5 – Métricas de Classificação e suas Referências

Classificação	**Métrica**	**Autor-Data**
Desempenho	Acurácia	Japkowicz e Shah (2011)
	ROC	Fawcett (2006)
	AUC	Bradley (1997)
	Precision	Davis e Goadrich (2006)
	Recall	Powers (2011)
	F1-Score	Van Rijsbergen (1979)
Ajuste e Consistência	Kappa de Cohen	Cohen (1960)
	AIC	Akaike (1974)
	BIC	Schwarz (1978)
	Pseudo R^2 de McFadden	McFadden (1974)

Fonte: Elaborado pelo autor.

Além disso, a *precision* e o *recall*, descritos por Davis e Goadrich (2006) e Powers (2020), respectivamente, são métricas cruciais quando se deseja focar na qualidade das previsões positivas (*precision*) ou na capacidade do modelo de identificar todos os casos positivos (*recall*). O F1-Score, conforme Van Rijsbergen (1979) sugere, é uma métrica que combina harmoniosamente precisão e *recall*, sendo especialmente útil em situações onde é preciso um equilíbrio entre essas duas métricas.

No que se refere à consistência e ao ajuste dos modelos, o Kappa de Cohen, como Cohen (1960) descreve, oferece uma medida de concordância que leva em conta a concordância ocorrida por acaso, sendo particularmente útil em problemas de classificação com múltiplos *raters*. Por outro lado, critérios de informação como AIC e BIC, discutidos respectivamente por Akaike (1974) e Schwarz (1978), são essenciais para comparar modelos, equilibrando a complexidade do modelo com o seu desempenho. Enquanto o AIC foca na previsão do modelo, o BIC incorpora uma penalidade mais forte para modelos com maior número de parâmetros.

Por fim, o Pseudo R^2 de McFadden, conforme explorado por McFadden *et al.* (1973), é uma métrica específica para modelos logísticos, avaliando o quão bem o modelo ajusta-se aos dados em comparação com um modelo nulo. A utilização dessas métricas permite uma avaliação mais robusta e detalhada dos modelos de classificação, assegurando escolhas mais informadas na prática de modelagem estatística e de aprendizado de máquina.

3.5 Modelo empírico

Este estudo adota um modelo de regressão logística para investigar a influência dos discursos contidos nos pareceres exarados pelo MPF e dos conselheiros relatores do CADE sobre os resultados dos julgamentos. O método emprega a modelagem de tópicos para extrair probabilidades de tópicos prevalentes nos documentos relacionados a cada entidade, integrando-

as como variáveis independentes em um modelo preditivo. Esta abordagem permite uma análise quantitativa das variáveis discursivas, proporcionando uma nova perspectiva sobre as dinâmicas subjacentes às decisões judiciais.

A aplicação da modelagem de tópicos em documentos jurídicos e administrativos oferece uma compreensão mais detalhada das estratégias de comunicação e argumentação. Além disso, a integração destas probabilidades no modelo de regressão logística facilita a identificação de padrões e a previsão de resultados de julgamentos com base nas informações extraídas dos textos. Este modelo contribui para o campo da Organização Industrial e análise antitruste, fornecendo uma ferramenta analítica para entender melhor as influências nas decisões regulatórias.

3.5.1 Estrutura do modelo

Como revelado anteriormente, modelo empírico proposto é formulado como uma regressão logística, onde a variável dependente é binária representa o resultado do primeiro julgamento de mérito(*condenação* ou*não condenação*) em casos de prática anticompetitiva de cartel. Os vetores de probabilidades de tópicos extraídos dos documentos do MPF e dos relatores do CADE constituem as variáveis independentes.

Antes de apresentar o modelo, calha esclarecer que, embora a regressão logística forneça uma estimativa da probabilidade de um evento, a interpretação dos resultados deve considerar a distinção entre probabilidade, *odds* e *odds ratio*. Enquanto a probabilidade se refere à frequência esperada de um evento, os *odds* representam a razão dessa probabilidade pela probabilidade do evento não ocorrer. O *odds ratio*, por sua vez, indica o efeito de uma variável independente sobre os odds de ocorrência do evento.

Feita essa breve digressão, o modelo empírico é matematicamente expresso pela seguinte relação:

$$\log\left(\frac{p_M}{1-p_M}\right) = \beta_0 + \beta_1 P_{MPF1} + \beta_2 P_{MPF2} + \ldots + \beta_n P_{MPFn} + \beta_{n+1} P_{Rel1} + \ldots + \beta_{n+m} P_{Relm} \tag{9}$$

onde:

- p_M representa a probabilidade de mérito na condenação;
- $\beta_0, \beta_1, \ldots, \beta_{n+m}$ são os coeficientes do modelo a serem estimados;
- P_{MPFi} são as probabilidades dos tópicos prevalentes nos pareceres do MPF;
- P_{Relj} são as probabilidades dos tópicos prevalentes nos votos dos conselheiros relatores.

De outra forma, mais sucinta e clara, a fórmula da regressão logística pode incorporar somatórios para os vetores de probabilidades dos tópicos do MPF e do conselheiro relator,

resultando na seguinte formulação:

$$\log\left(\frac{p_M}{1-p_M}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i P_{MPFi} + \sum_{j=1}^{m} \beta_{n+j} P_{Relj} \tag{10}$$

Nesta equação:

- p_M representa a probabilidade de mérito na condenação;
- β_0 é o intercepto do modelo;
- Os termos $\beta_i P_{MPFi}$ representam os produtos dos coeficientes β e as probabilidades dos tópicos do MPF;
- Os termos $\beta_{n+j} P_{Relj}$ representam os produtos dos coeficientes β e as probabilidades dos tópicos do Relator;
- n é o número de tópicos do MPF, e m é o número de tópicos do Relator.

A interpretação dos coeficientes, especialmente em termos de *odds ratio*, é crucial para a compreensão do impacto de cada variável independente no resultado. A *odds ratio*, dado por e^{β_i} (a exponencial do logito) para cada coeficiente β_i, indica o efeito multiplicativo nas *odds* de condenação para uma mudança unitária na probabilidade do tópico correspondente. Este modelo proporciona uma base robusta para a análise de eventos binários e a extração de *insights* significativos de dados complexos.

3.5.2 Metodologia de estimação

A metodologia adotada para a estimação dos coeficientes β no modelo proposto é fundamentada no método de máxima verossimilhança. Este método é aplicado utilizando um conjunto de dados históricos abrangentes, que incluem não apenas os resultados dos julgamentos, mas também as distribuições de tópicos derivadas da modelagem de tópicos aplicada às decisões exaradas pelo órgão antitruste nacional. A análise destes dados proporciona uma base sólida para a estimação dos coeficientes, garantindo que o modelo reflita com precisão as complexidades inerentes aos processos de tomada de decisão.

A interpretação dos coeficientes β, resultante do processo de estimação, desempenha um papel crucial na compreensão da dinâmica subjacente às decisões de julgamento. Cada coeficiente oferece *insights* significativos sobre o impacto de um tópico específico na probabilidade de um resultado de condenação. A análise permite identificar quais tópicos são mais influentes nas decisões e como eles se correlacionam com os resultados dos julgamentos. Do exposto, o modelo não apenas fornece uma previsão quantitativa dos resultados, mas também enriquece a compreensão dos fatores que influenciam esses resultados, contribuindo significativamente para o campo da análise antitruste e da tomada de decisão jurídica.

3.5.3 Considerações metodológicas sobre a estimação

Na fase de validação do modelo proposto, foram adotadas práticas rigorosas para assegurar a precisão e confiabilidade dos resultados. Esta etapa é crucial para garantir que o modelo não apenas se ajuste bem aos dados utilizados para sua calibração, mas também seja capaz de fazer previsões acuradas em novos contextos. Duas considerações principais nortearam este processo:

- A multicolinearidade entre as variáveis independentes foi cuidadosamente examinada para assegurar a estabilidade e a interpretabilidade dos coeficientes estimados.
- A validade do modelo foi verificada através de técnicas de validação cruzada e testes de significância estatística dos coeficientes, garantindo a robustez e generalização do modelo para novos dados.

As estratégias empregadas na validação do modelo são fundamentais para reforçar a confiança nos resultados obtidos. A análise detalhada da multicolinearidade contribui para a interpretação correta dos coeficientes, evitando inferências enganosas decorrentes de relações excessivamente interdependentes entre as variáveis. Ademais, a aplicação de validação cruzada e testes de significância estatística fornece uma medida da robustez do modelo, assegurando que as conclusões derivadas sejam aplicáveis em uma variedade de contextos, além daqueles representados nos dados de treinamento. Estas práticas reforçam a aplicabilidade prática do modelo em futuras análises e tomadas de decisão no campo da análise antitruste.

3.5.4 Contribuições para a Organização Industrial

A análise antitruste, como um campo importante na Organização Industrial, tem como objetivo salvaguardar a concorrência no mercado e prevenir práticas monopolistas que podem prejudicar consumidores e a economia em geral. Este estudo propõe um modelo inovador de regressão logística que se integra com a modelagem de tópicos com fins de para enriquecer a análise antitruste. Nos parágrafos e subseções seguintes ser discutido como esse modelo impacta o campo da Organização Industrial e preenche lacunas importantes na análise antitruste.

De maneira geral, pode-se afirmar que a Organização Industrial, ao examinar estruturas, funcionamentos e desempenho dos mercados, pode ser beneficiada por abordagens que oferecem *insights* sobre a tomada de decisão de entidades reguladoras como o CADE. O modelo proposto oferece uma ferramenta quantitativa para analisar a influência do discurso do Ministério Público Federal e dos Relatores do CADE nos resultados de julgamentos. Esta abordagem possibilita uma análise mais detalhada das dinâmicas subjacentes às decisões antitruste, que são cruciais para compreender e prever mudanças na política e prática regulatória.

3.5.5 Preenchimento de lacunas na análise antitruste

A análise antitruste tradicionalmente foca em métricas quantitativas, como preços, quotas de mercado e concentração. No entanto, existe uma lacuna significativa no entendimento das motivações e influências qualitativas subjacentes às decisões regulatórias. O modelo desenvolvido aborda esta lacuna ao quantificar a influência de discursos e documentos textuais, proporcionando uma perspectiva mais holística e informada sobre as decisões antitruste.

3.5.5.1 Perspectiva comportamental

A concepção do modelo empírico surge em um contexto onde, conforme o levantamento realizado em sede de Revisão da Literatura, não se constatou a existência de trabalhos anteriores que abordem a integração da análise de texto com a teoria econômica no âmbito da análise antitruste brasileira e com fins de identificação de padrões no âmbito decisório. Esta lacuna na literatura ressalta a originalidade e relevância do presente estudo. O modelo proposto permite uma avaliação comportamental de entidades reguladoras, ultrapassando os limites da análise baseada exclusivamente em dados econômicos, ao incorporar também os aspectos qualitativos e discursivos.

Assim, a pesquisa apresenta esta abordagem inovadora, que combina a análise de texto com a teoria econômica, oferece uma nova dimensão para a compreensão das decisões no âmbito antitruste. Ao fazer isso, o modelo transcende os métodos convencionais, proporcionando uma visão mais holística e aprofundada das dinâmicas subjacentes às decisões regulatórias. Isso não só amplia o entendimento do campo da análise antitruste, mas também abre caminhos para futuras pesquisas que podem explorar outras intersecções entre a linguagem discursiva e as teorias econômicas.

3.5.5.2 Implicações para a formulação de políticas

Ao proporcionar uma análise mais detalhada das influências sobre os decisores antitruste, o modelo oferece *insights* valiosos para a formulação de políticas. Os responsáveis pela formulação de políticas podem utilizar esses *insights* para desenvolver regulamentos mais eficazes e direcionados, que refletem melhor a realidade das práticas de mercado e as preocupações regulatórias.

Em suma, o modelo proposto tem o potencial de impactar significativamente a Organização Industrial, fornecendo uma ferramenta robusta para análise antitruste. Ao preencher uma lacuna crítica na compreensão das motivações e influências nas decisões antitruste, o modelo oferece uma contribuição valiosa para a teoria e prática no campo da Organização Industrial.

3.6 Aspectos éticos

Em conformidade com as diretrizes éticas para pesquisas acadêmicas, este estudo adotou procedimentos rigorosos para garantir a confidencialidade e a privacidade dos dados. Especificamente, os dados coletados e manipulados eram exclusivamente aqueles publicizados pelo CADE em suas APIs, os quais foram submetidos a um processo cuidadoso de anonimização. Essa medida visava eliminar quaisquer informações que pudessem identificar indivíduos ou entidades envolvidas nos processos analisados. Tal abordagem estava alinhada com as normativas da Lei Geral de Proteção de Dados (LGPD), assegurando que a integridade, a privacidade e a segurança das informações fossem rigorosamente mantidas ao longo da pesquisa.

O estudo aderiu firmemente aos princípios de uso responsável de dados, assegurando que todos os dados coletados e analisados fossem utilizados exclusivamente para fins acadêmicos e de pesquisa. A integridade acadêmica foi uma pedra angular deste trabalho, refletida na precisão e na transparência com que os métodos de pesquisa foram apresentados. Todas as fontes de dados foram meticulosamente citadas, e as técnicas de processamento de linguagem natural e análise de dados foram descritas com detalhes, promovendo uma clareza metodológica essencial.

Além disso, a pesquisa reconheceu a importância de considerar as implicações sociais dos resultados encontrados. As análises e conclusões foram conduzidas e apresentadas com consciência de seu potencial impacto na área da Organização Industrial e nas práticas regulatórias do CADE. Uma atenção especial foi dada à não exposição de dados sensíveis relacionados às autoridades do CADE, respeitando sua privacidade e posição, conforme delineado pela LGPD.

Em suma, o estudo esteve comprometido com os mais altos padrões de ética em pesquisa, garantindo a conformidade com a LGPD e a proteção rigorosa de dados sensíveis, enquanto buscava contribuições significativas para o campo de estudo.

4 RESULTADOS E DISCUSSÃO

As análises foram realizadas tanto intra quanto intermodelos, levando em consideração os grupos específicos identificados durante a Revisão da Literatura e selecionados com base na heurística descrita na seção de Metodologia desta pesquisa. Para identificar os modelos mais eficazes, compararam-se as métricas de desempenho (NPMI, UMass, Diversidade de tópicos) e de eficiência (Tempo de computação). De maneira geral, um modelo é considerado mais eficiente quando demonstra um equilíbrio entre uma alta diversidade de tópicos e boas pontuações nas métricas NPMI e UMass, que são indicativos de coesão e qualidade dos tópicos. O tempo de computação foi adotado como uma métrica de penalização, uma vez que um período prolongado de processamento pode tornar impraticável o uso de um modelo específico em contextos práticos.

4.1 Aspectos descritivos

Esta seção oferece uma análise descritiva das estatísticas e outras métricas textuais oriundas dos resultados alcançados pela aplicação de modelos de tópicos, como o BERTopic, nas decisões no âmbito do CADE. As subseções subsequentes abordarão detalhes sobre o *corpus*, as comparações intra e intermodelos e, finalmente, apresentarão uma visão geral sobre os dados quantitativos compilados.

4.1.1 Sobre o *corpus*

Inicialmente, foi criado um *corpus* composto por 103 documentos relativos a decisões de mérito em processos de cartel em licitações. A etapa de tokenização revelou-se crucial para a redução da dimensionalidade dos dados. Notou-se que pouco mais de 25% do conjunto de palavras do *corpus* original foram mantidas no *corpus* tokenizado. Esse fato é evidenciado na Tabela 6, que apresenta um sumário dos principais objetos desenvolvidos a partir dos dados coletados.

Tabela 6 – Sumarização dos principais objetos criados

Objeto	Função	Documentos	Tokens	Types	Features
Função	–	ndoc()	ntoken()	ntype()	nfeat()
Corpus inicial	corpus()	103	1.199.682	46.028	–
Corpus tokenizado	spacy_tokenize()	103	303.366	111.511	–
DFM	dfm()	103	303.366	111.511	20.682

Fonte: elaborado pelo autor.

A conversão da base de dados em um objeto DFM resultou em uma matriz com vinte *docvars* e uma esparsidade de 94,77%. As palavras mais frequentes no *corpus* são ilustradas na

Figura 15. Observa-se que os termos "conduta", "documento", "cartel", "licitação", "acordo", "prova" e "*email*" são proeminentes na nuvem de palavras. Uma inferência possível é que as trocas de *emails* servem como provas da conduta de cartel em licitações, embora seja reconhecido que tais práticas são notoriamente difíceis de comprovar[43].

Figura 15 – Duzentas palavras mais frequentes

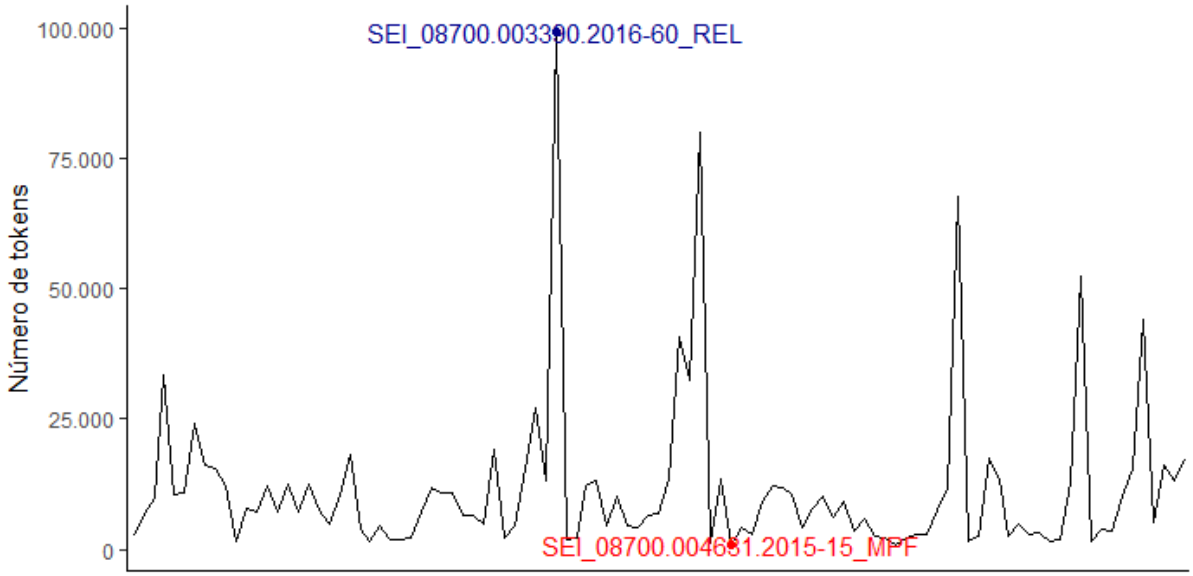
Fonte: elaborado pelo autor.

Quanto à utilização do objeto DFM, é importante destacar que, no design do *quanteda*, um texto torna-se um "recurso" apenas após o processo de tokenização. Essa observação é relevante, especialmente se há um conjunto de recursos que se deseja excluir. Para isso, é preciso

[43] Conforme Brasil (2019, p. 16), "a questão da prova do cartel é bastante sensível, visto que se, de um lado, os envolvidos na prática buscam eliminar os rastros do conluio, de outro, o mero paralelismo de preços não é suficiente para caracterizar a conduta, sendo necessário demonstrar a existência de um elemento adicional (o chamado fator plus) que indique que o comportamento dos agentes econômicos não foi espontâneo".

tokenizar o texto; caso contrário, não será possível excluí-los do *corpus* inicial. Diferentemente de outros pacotes de texto para R, essas etapas são aplicadas após a criação de um *corpus*, garantindo que este permaneça como um conjunto não processado de textos aos quais as manipulações serão posteriormente aplicadas. Segundo os autores do projeto *quanteda*, o objetivo é preservar a generalidade e, ao mesmo tempo, promover a reprodutibilidade e a transparência na análise de texto[44].

Figura 16 – Número de *tokens* por documento



Fonte: elaborado pelo autor.

Os documentos analisados no Processo SEI 08700.03390.2016-60 e no Processo SEI 08700.004651.2015-15 representam, respectivamente, o maior e o menor número de *tokens*, conforme ilustrado no gráfico (Figura 16). O primeiro, um processo administrativo iniciado em 9 de maio de 2016, investigou a formação de cartel no mercado brasileiro de tubos e conexões de PVC para obras de saneamento e construção civil. Concluído na 180ª Sessão do CADE em 30 de junho de 2021, resultou na condenação dos envolvidos e na aplicação de multas que totalizaram 193,8 milhões de reais. O segundo processo, iniciado em 6 de julho de 2015 e julgado na 110ª Sessão Ordinária do CADE em 6 de setembro de 2017, abordou alegadas práticas de cartel no fornecimento de módulos de *airbag*, cintos de segurança e volantes de direção, onde o Ministério Público Federal recomendou o arquivamento, decisão apoiada pelo Conselheiro-Relator.

Essa análise permite inferir que processos que culminam em condenações tendem a ser mais extensos, enquanto aqueles que resultam em arquivamento geralmente têm instruções mais simplificadas. Isso está alinhado à lógica de que processos que levam a penalidades requerem diligências mais aprofundadas e a coleta de evidências mais substanciais, prolongando o tempo

[44] Veja a explicação do autor do pacote *quanteda*, Ken Benoit, disponível em: https://stackoverflow.com/questions/38931507/create-dfm-step-by-step-with-quanteda.

de instrução e julgamento. A diferença na complexidade dos processos se reflete no volume de documentação produzida, como demonstrado pela variação no número de *tokens* entre os dois casos.

4.2 Modelagem por Aprendizagem Não Supervisionada

Essa análise foi conduzida visando obter evidências para aprimorar a análise de documentos do CADE em casos de cartéis, adotando uma metodologia tripartite. Inicialmente, identificamos a técnica de modelagem de tópicos mais eficaz, utilizando abordagens avançadas como BERT e *embeddings* da OpenAI, selecionadas por sua precisão e eficiência com base em métricas quantitativas. Posteriormente, aplicamos o modelo escolhido para quantificar a influência dos pareceres do MPF e votos dos conselheiros relatores nas decisões sobre práticas colusivas. Concluímos com uma análise quantitativa de padrões e comportamentos em decisões antitruste, proporcionando *insights* valiosos sob a perspectiva da Organização Industrial.

4.2.1 Sobre os Modelos Clássicos (LDA e NMF)

Neste estudo, realizou-se uma análise detalhada e quantitativa dos modelos de tópicos LDA e NMF, empregando dois conjuntos de dados - um geral e outro que é um subconjunto do primeiro, incluindo os modelos com as melhores métricas em seus respectivos grupos. O objetivo foi avaliar as características e a eficácia desses modelos populares de modelagem de tópicos em diferentes contextos.

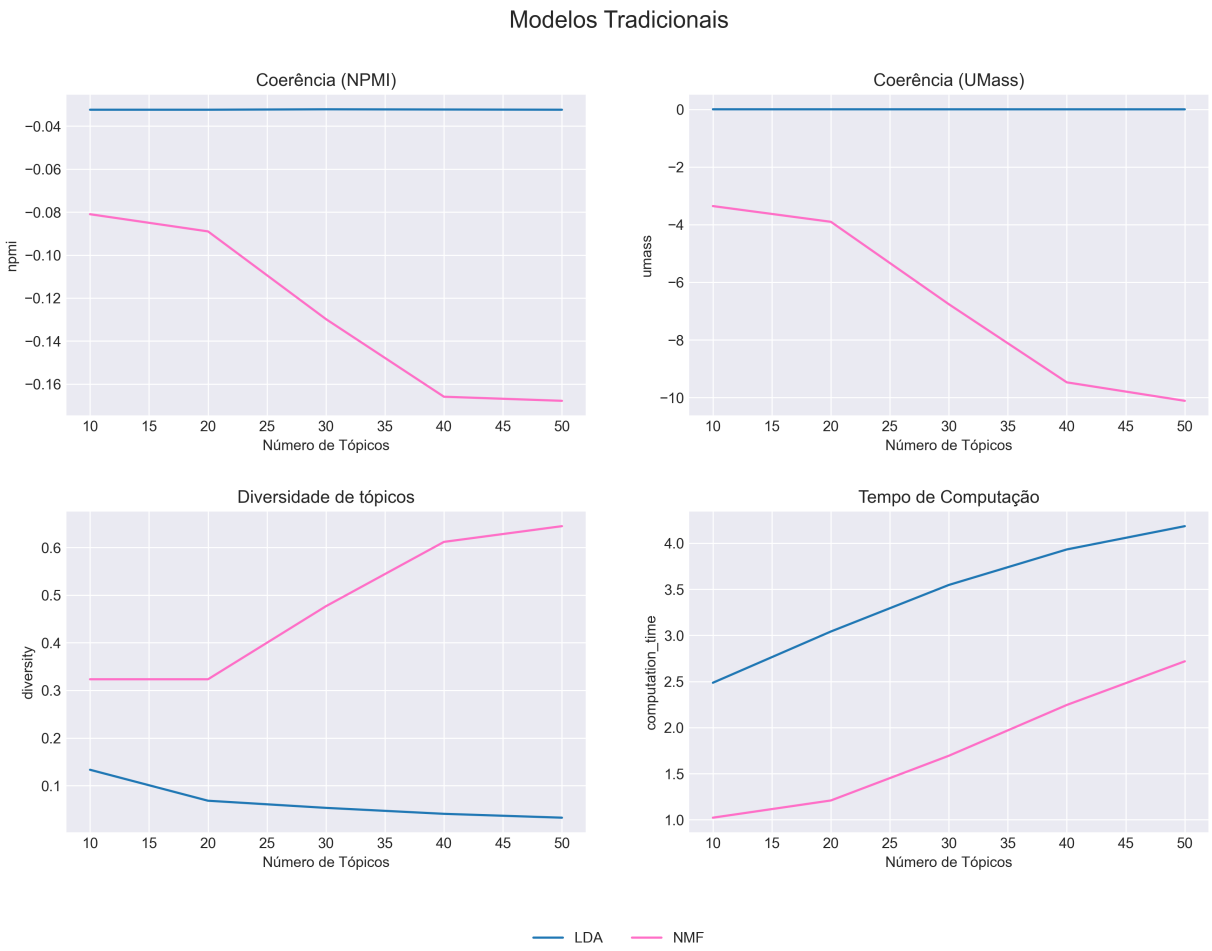
Identificaram-se diferenças significativas em termos de coerência, diversidade e eficiência computacional, como ilustrado no Figura 17 a seguir:

Pela tendência das linhas no gráfico, observa-se que, em termos de coerência, as métricas de NPMI e UMass apresentam trajetórias semelhantes, com melhores resultados para o LDA. Por outro lado, a diversidade de tópicos deste modelo é inferior à do NMF. Ambos os modelos apresentam um comportamento idêntico quanto ao uso de recursos computacionais, consumindo mais ao gerar novos tópicos.

Na análise do conjunto de dados geral, o LDA obteve um NPMI médio de -0.032331 e um UMass médio de -0.000675, com uma diversidade de tópicos média de 0.065700. Seu tempo de computação médio foi de aproximadamente $3.44\,\mathrm{s}$. Em contrapartida, o NMF registrou um NPMI médio de -0.126705 e um UMass médio de -6.721646, com uma diversidade de tópicos média de 0.475933 e um tempo de computação médio de cerca de $1.78\,\mathrm{s}$. Estes resultados sugerem que, enquanto o LDA oferece uma coerência temática ligeiramente melhor, o NMF destaca-se na diversidade de tópicos e na eficiência computacional.

Nos modelos selecionados, o LDA e o NMF mostraram um padrão similar. O LDA manteve um NPMI médio de -0.032355 e um tempo de computação médio de $2.49\,\mathrm{s}$, enquanto o

Figura 17 – Comparação entre os Modelos Tradicionais (LDA e NMF)



Fonte: elaborado pelo autor.

NMF teve um NPMI médio de -0.080980 e um tempo de computação médio de $1.02\,\mathrm{s}$. Esses resultados reforçam a tendência observada nos dados gerais.

Embora ambos os modelos sejam amplamente utilizados na modelagem de tópicos, os resultados indicam algumas limitações. O LDA, apesar de sua popularidade, exibe uma diversidade de tópicos relativamente baixa, sugerindo que pode não ser a melhor escolha para aplicações que requerem uma ampla exploração temática. Em contraste, apesar de o NMF exibir maior diversidade e eficiência computacional, sua coerência temática é inferior à do LDA, o que pode ser uma desvantagem em cenários que exigem alta precisão temática.

Egger e Yu (2022, p. 6) identificam que, embora o LDA e o NMF sejam eficazes na identificação de temas comuns, como respostas do governo e restrições de viagem, o NMF tende a produzir tópicos mais distintos e alinhados com o julgamento humano. Isso sugere que, para o contexto do CADE, a utilização do NMF pode ser mais adequada para capturar nuances específicas em documentos legais relacionados a julgamentos de cartéis em licitações.

Os modelos LDA e NMF oferecem abordagens valiosas para a modelagem de tópicos, cada um com seus pontos fortes e limitações. A escolha entre LDA e NMF deve basear-se em um

equilíbrio entre precisão na captura de tópicos, diversidade temática e eficiência computacional, conforme as necessidades específicas da aplicação. Considerando as métricas de NPMI, UMass, diversidade de tópicos e tempo de computação, fica evidente que uma seleção cuidadosa do modelo é crucial para o sucesso da modelagem de tópicos.

4.2.2 Sobre o Modelo Neural (CTM)

Neste estudo, realizou-se uma análise detalhada e quantitativa dos modelos de tópicos que empregam o CTM, utilizando dois conjuntos de dados: um geral e outro de modelos selecionados. O objetivo desta análise é compreender as características e a eficácia dos modelos CTM em diferentes contextos de aplicação.

Na análise do conjunto de dados geral, observou-se que o modelo CTM apresentou um NPMI médio de -0.250073 e um UMass médio de -1.346014, com uma diversidade de tópicos média de 0.616344. O tempo de computação médio foi de aproximadamente $1955.31\,\mathrm{s}$. Estes resultados indicam que, apesar de o CTM possuir uma capacidade razoável de capturar a coerência temática, ele requer um tempo de computação consideravelmente longo.

No conjunto de modelos selecionados, o CTM mostrou um NPMI médio de -0.219016 e um UMass médio de -0.928798, com uma diversidade de tópicos média de 0.641667. O tempo de computação médio foi de cerca de $1885.53\,\mathrm{s}$. Esta análise sugere que o CTM mantém um desempenho consistente em termos de coerência e diversidade de tópicos, embora ainda exija um tempo de computação significativo.

A análise dos modelos CTM revela algumas limitações importantes. Primeiramente, o tempo de computação elevado para ambos os conjuntos de dados sugere que o CTM pode não ser a escolha ideal para aplicações que exigem processamento rápido ou para grandes volumes de dados. Além disso, os valores de NPMI e UMass indicam que, embora o CTM seja capaz de capturar tópicos com uma coerência razoável, há espaço para melhorias na precisão e relevância dos tópicos gerados. A diversidade de tópicos, apesar de ser relativamente alta, também sugere a necessidade de uma seleção cuidadosa dos parâmetros para garantir a abrangência temática desejada.

Os modelos CTM oferecem uma abordagem inovadora para a modelagem de tópicos, com resultados promissores em termos de coerência e diversidade temática. Contudo, a eficiência computacional e a precisão na captura de tópicos relevantes variam consideravelmente. Portanto, a escolha do modelo CTM deve ser guiada por um equilíbrio entre precisão temática, diversidade de tópicos e eficiência computacional, conforme as necessidades específicas da aplicação.

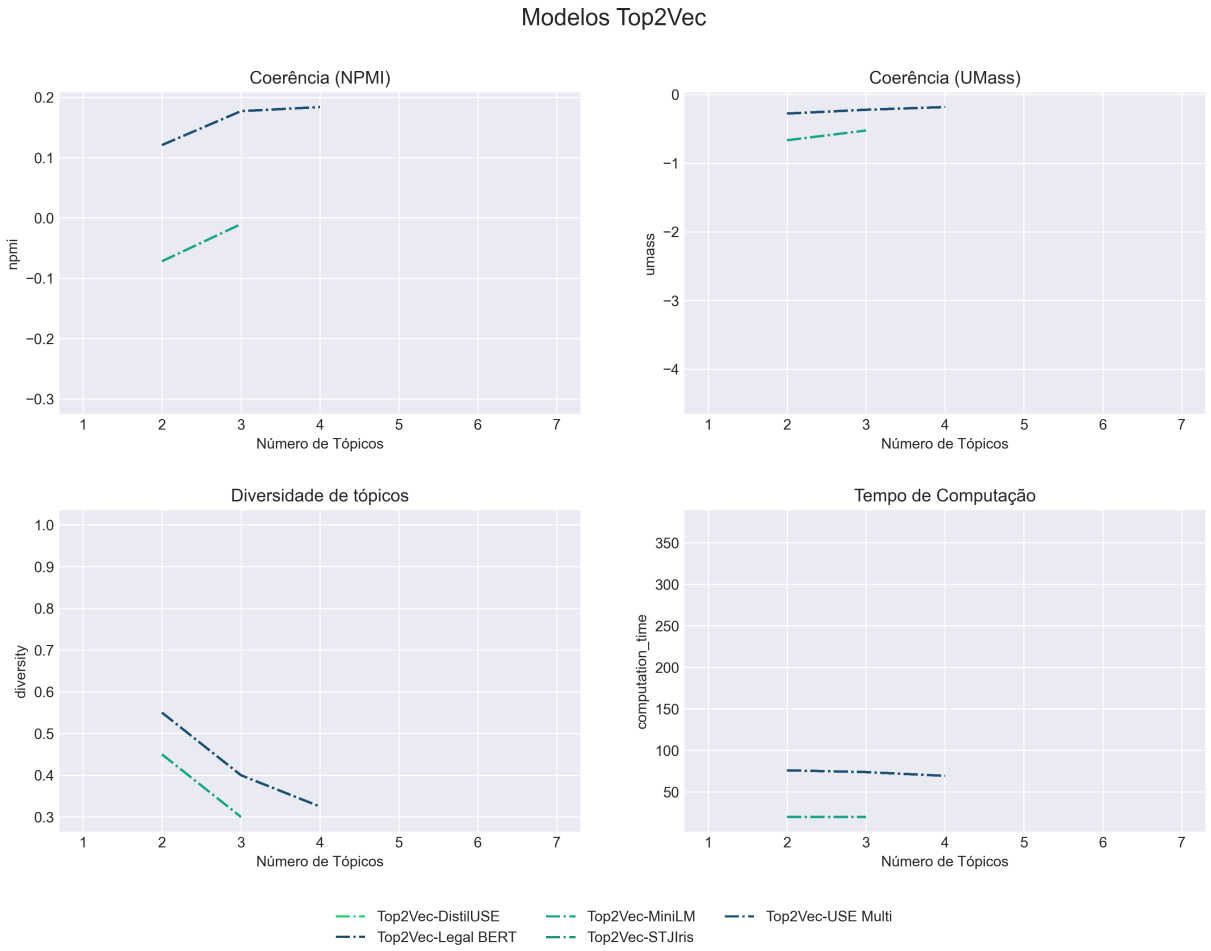
4.2.3 Sobre o Modelo Híbrido (Top2Vec)

Nesta subseção, exploram-se os modelos de tópicos Top2Vec, visando avaliar quantitativamente as métricas de desempenho desses modelos para proporcionar uma compreensão mais

aprofundada de suas características e eficácia.

Este modelo não gerou tópicos com a definição inicial de 15 unigramas por tópico. Após várias tentativas e ajustes, definiu-se o número ideal de cinco unigramas por tópico para o dataset do CADE. Observa-se no Figura 18 (p. 94) que, dentre todos os modelos, o Top2Vec apresentou o menor número de tópicos.

Figura 18 – Comparação entre os Modelos Top2Vec com 5 Unigramas por Tópico



Fonte: elaborado pelo autor.

Nos dados gerais, diversas configurações do modelo Top2Vec foram observadas, cada uma com suas métricas específicas. Por exemplo, o modelo "DistilUSE" registrou um NPMI médio de -0.116244 e um tempo de computação médio de aproximadamente $23.43\,\mathrm{s}$, indicando um equilíbrio entre coerência temática e eficiência computacional. Em contraste, o "STJIris" apresentou um NPMI médio de -0.263855, mas um tempo de computação significativamente mais alto, cerca de $371.61\,\mathrm{s}$.

No conjunto de modelos selecionados, a análise focou principalmente no "USE Multi", que exibiu um NPMI médio de 0.18374 e um tempo de computação médio de $69.21\,\mathrm{s}$. Este modelo demonstrou um desempenho superior em termos de coerência, mas com uma eficiência computacional reduzida.

Apesar de o Top2Vec mostrar uma capacidade notável de capturar tópicos relevantes, as variações significativas no tempo de computação levantam preocupações sobre sua aplicabilidade em grandes conjuntos de dados ou em cenários que exigem processamento rápido. Além disso, a diversidade de tópicos, embora variada, sugere que a escolha da configuração do Top2Vec deve ser cuidadosamente ponderada para garantir que os tópicos capturados sejam tanto relevantes quanto abrangentes. A variação no NPMI também indica que nem todas as configurações são igualmente eficazes na captura de tópicos coerentes.

Os modelos Top2Vec oferecem uma abordagem robusta para a modelagem de tópicos, com capacidade de capturar uma ampla gama de tópicos. No entanto, a eficiência computacional e a coerência dos tópicos variam significativamente entre as diferentes configurações. Assim, a seleção do modelo Top2Vec adequado deve ser baseada em um equilíbrio entre precisão temática, diversidade de tópicos e eficiência computacional, dependendo das necessidades específicas da aplicação.

4.2.4 Sobre os Modelos de *Embeddings* (BERTopic)

Esta análise foca nos modelos de tópicos que utilizam o "BERTopic", empregando dois conjuntos de dados distintos: um geral e outro de modelos selecionados. O objetivo é realizar uma avaliação quantitativa, examinando as métricas de desempenho desses modelos para compreender suas características e eficácia.

Observaram-se diferenças significativas em termos de coerência, diversidade e eficiência computacional ao variar o número de unigramas por tópico, conforme ilustrado nos Figura 19 e Figura 20 (pp. 96 e 97, respectivamente).

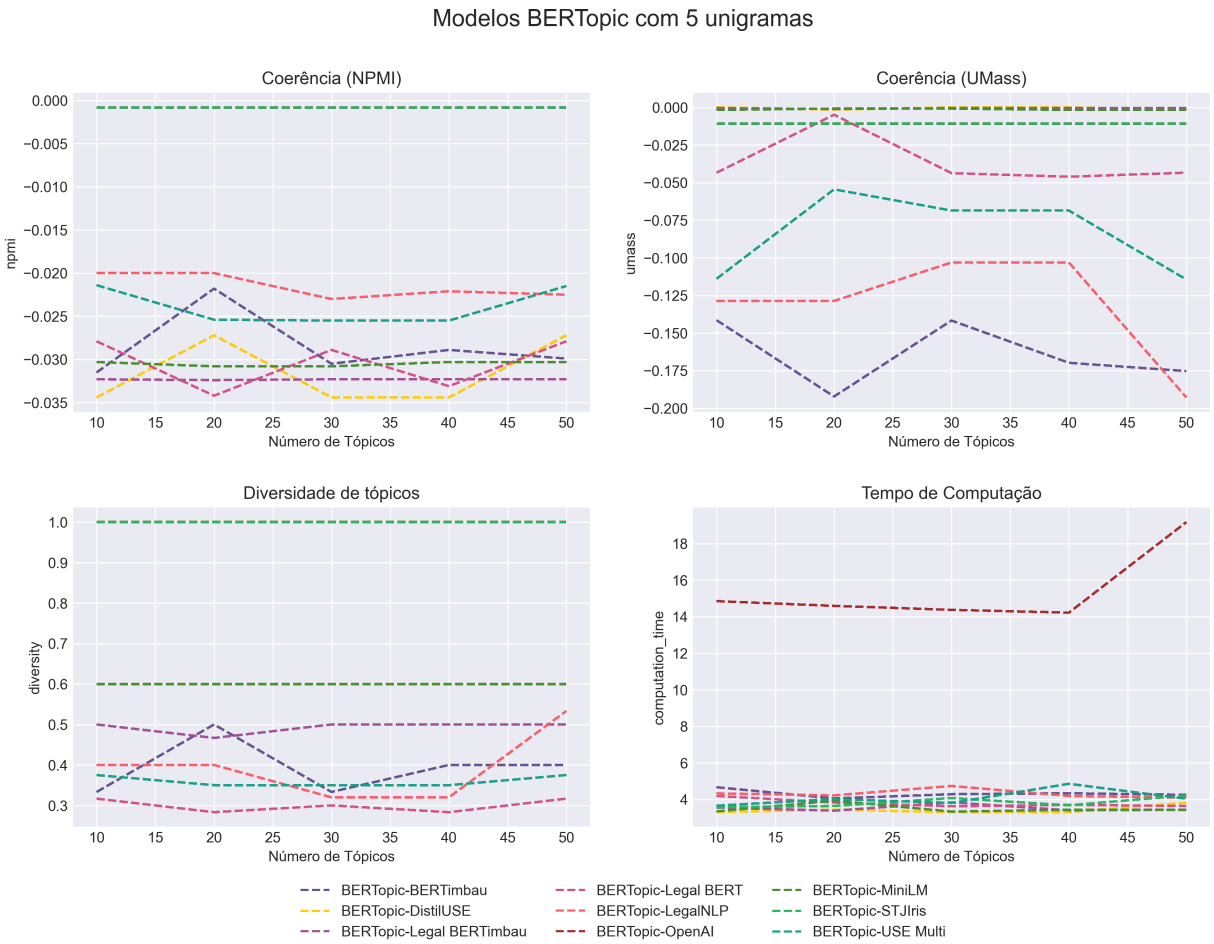
Verifica-se que, para as métricas de coerência, os gráficos tornam-se mais erráticos com o aumento do número de unigramas por tópico e do número de tópicos, observando-se também uma queda na diversidade de tópicos em alguns modelos analisados (BERTimbau e USE Multi).

Cabe ressaltar que, embora *embeddings* tenham sido utilizados na modelagem Top2Vec, foi necessária a alteração no código padrão da biblioteca Python para a utilização de *embeddings* não nativamente reconhecidos pelo algoritmo Top2Vec. Por essa razão, as análises dos modelos Top2Vec e BERTopic foram separadas, classificando os primeiros como híbridos e os últimos como de *embeddings*.

Na análise do conjunto de dados geral, observou-se uma diversidade nas configurações do "BERTopic", cada uma com suas métricas específicas. Por exemplo, o "BERTopic-BERTimbau" registrou um NPMI médio de -0.030503 e um tempo de computação de aproximadamente $4.33\,\mathrm{s}$. Outras variações, como o "BERTopic-DistilUSE" e "BERTopic-OpenAI", mostraram diferenças no NPMI e no tempo de computação, refletindo diferentes equilíbrios entre coerência temática e eficiência computacional.

No conjunto de modelos selecionados, focou-se principalmente nos modelos "BERTimb

Figura 19 – Comparação entre os Modelos BERTopic com 5 Unigramas por Tópico



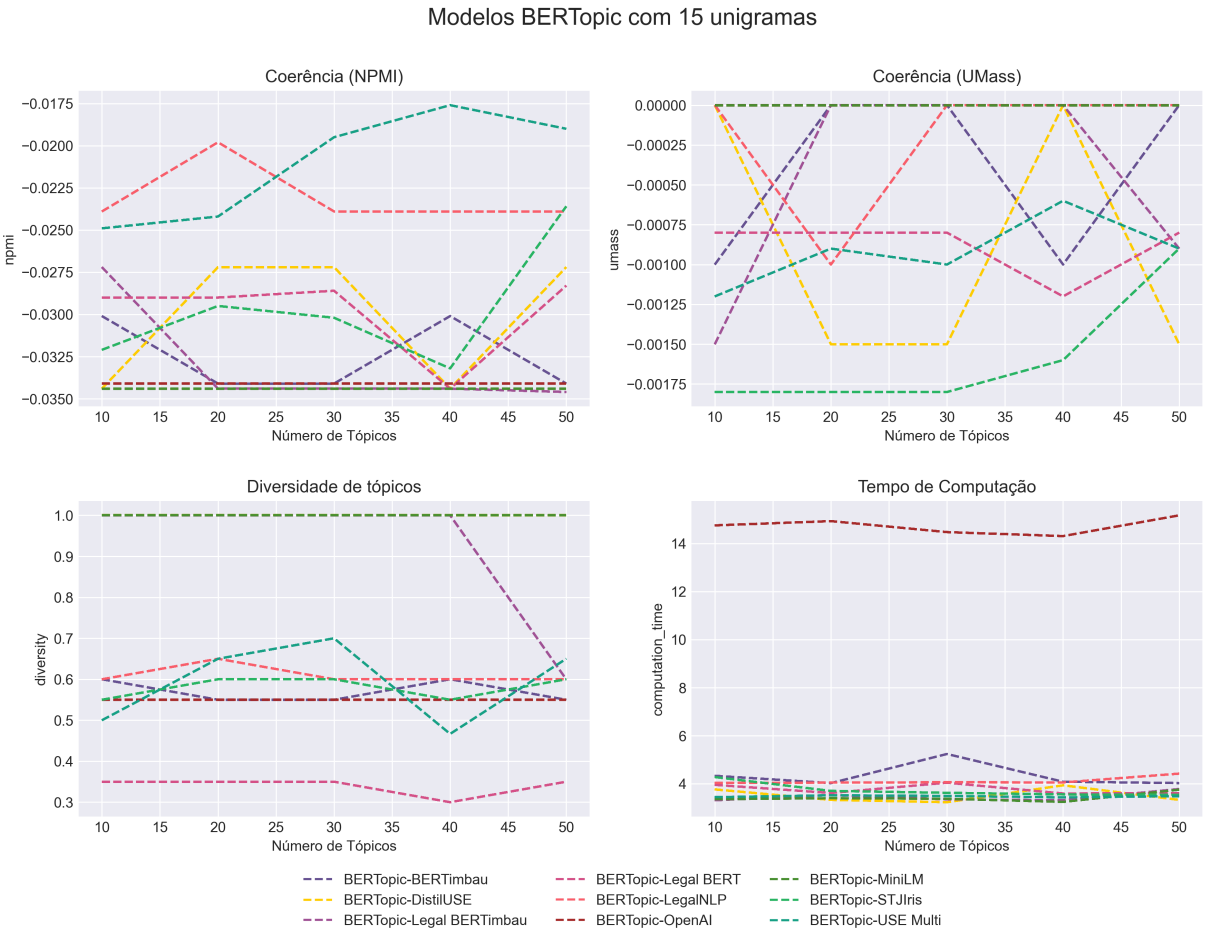
Fonte: elaborado pelo autor.

au" e "USE Multi". O primeiro apresentou um NPMI médio de -0.027946 e um tempo de computação médio de cerca de $4.04\,\mathrm{s}$, indicando um desempenho consistente em termos de coerência, mas com uma eficiência computacional ligeiramente reduzida.

Embora os modelos "BERTopic" demonstrem uma habilidade notável em capturar tópicos relevantes, há variações significativas tanto na coerência dos tópicos quanto na eficiência computacional entre as diferentes configurações. Por exemplo, modelos com maior tempo de computação podem não ser adequados para aplicações que exigem processamento rápido ou lidam com grandes volumes de dados. Além disso, a diversidade de tópicos, embora variada, sugere que a escolha da configuração do "BERTopic" deve ser cuidadosamente considerada para garantir que os tópicos capturados sejam relevantes e abrangentes.

Os modelos "BERTopic" oferecem uma abordagem robusta e flexível para a modelagem de tópicos, capaz de capturar uma ampla gama de tópicos. No entanto, a variação no desempenho entre as diferentes configurações destaca a importância de selecionar cuidadosamente a configuração apropriada do "BERTopic", baseando-se em um equilíbrio entre precisão temática, diversidade de tópicos e eficiência computacional, conforme as necessidades específicas da

Figura 20 – Comparação entre os Modelos BERTopic com 15 Unigramas por Tópico



Fonte: elaborado pelo autor.

aplicação.

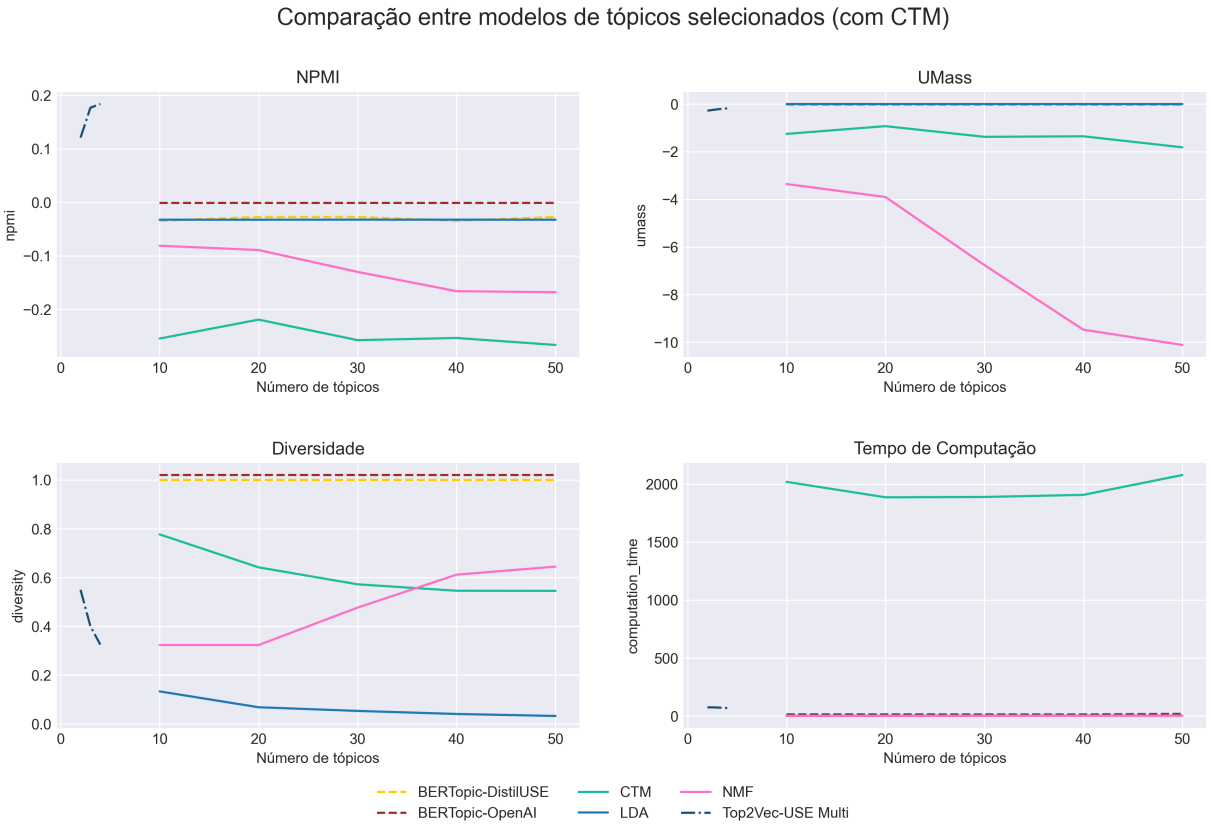
4.2.5 Considerações Gerais sobre os Modelos Analisados

Até o momento, foi realizada uma análise comparativa abrangente de diferentes modelos de modelagem de tópicos aplicados a um conjunto de dados, contemplando modelos tradicionais, baseados em *embeddings*, híbridos e neurais. A análise concentrou-se intramodelos, avaliando as métricas de coerência, diversidade dos tópicos e tempo de computação, considerando uma gama de tópicos de 10 a 50 e tamanhos mínimos de tópicos de 5 a 15 palavras[45].

Para a heurística de seleção do modelo de modelagem de tópicos ideal para textos do CADE, adotou-se uma abordagem metódica e criteriosa, focando em métricas específicas. As métricas selecionadas incluíam a coerência dos tópicos (NPMI e UMass), indicando a clareza e relevância dos tópicos gerados, e a diversidade dos tópicos, refletindo a capacidade do modelo de capturar uma ampla gama de temas. Adicionalmente, o tempo de computação foi considerado

[45] Para os modelos baseados no algoritmo Top2Vec, o range de tópicos variou entre 1 e 7, e o de unigramas foi definido como 5.

Figura 21 – Comparação entre os Melhores Modelos de Cada Grupo



Fonte: elaborado pelo autor.

como um fator decisivo em casos de empate técnico, priorizando a eficiência operacional em contextos de processamento rápido. A síntese qualitativa dessas métricas pode ser verificada no Quadro 2.

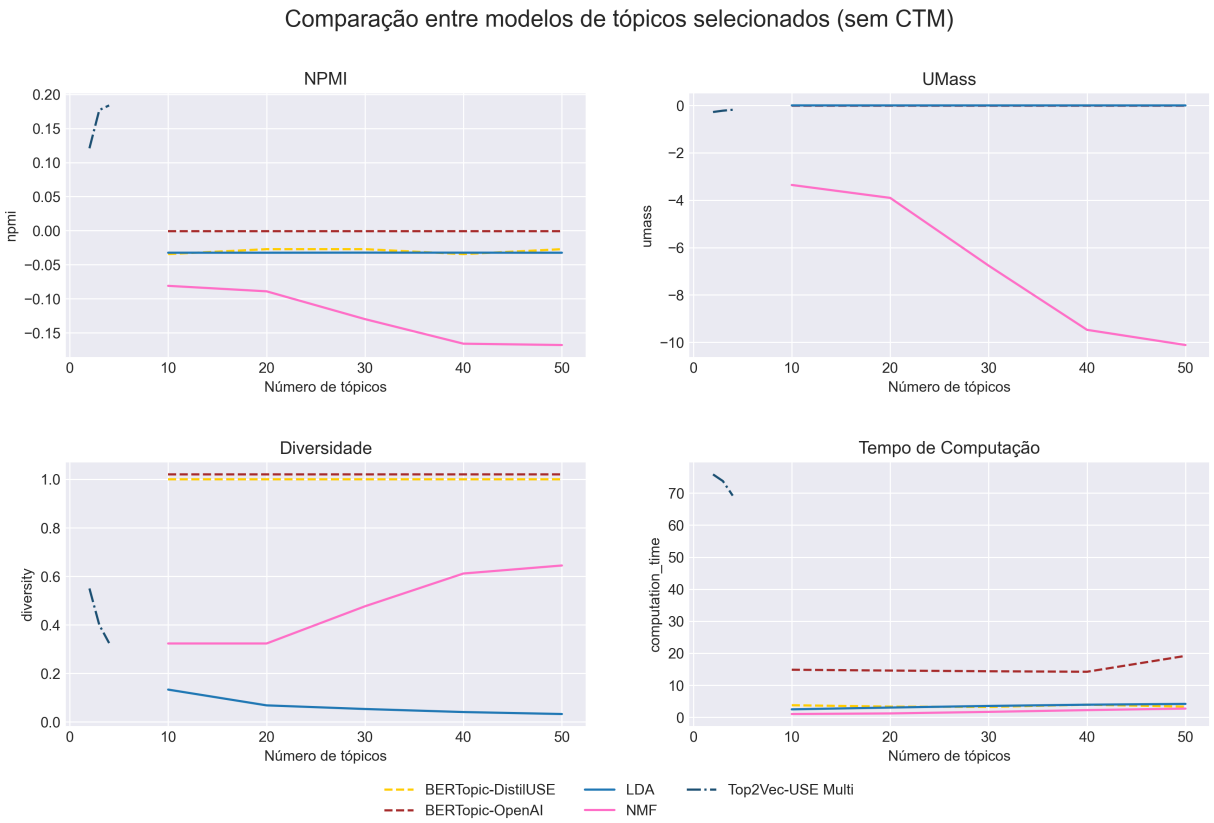
a) Quanto à coerência dos tópicos (NPMI e UMass), modelos com alta coerência são preferíveis, pois indicam que os tópicos gerados são significativos e interpretáveis, crucial para refletir efetivamente as nuances dos textos do CADE.

b) Quanto à diversidade dos tópicos, uma diversidade adequada é importante para assegurar que o modelo capte uma ampla gama de temas, evitando concentração excessiva em poucos tópicos.

c) Quanto ao tempo de computação, este critério é utilizado em desempate técnico, enfatizando a importância da eficiência computacional em contextos operacionais limitados.

O viés potencial desta heurística está na possibilidade de favorecer modelos que oferecem alta coerência e diversidade, mas que podem não ser os mais apropriados para capturar todas as nuances dos textos do CADE. Além disso, o uso do tempo de computação como critério de desempate pode levar à escolha de modelos mais rápidos, mas com qualidade inferior de resultados.

Figura 22 – Comparação entre os Melhores Modelos de Cada Grupo (sem CTM)



Fonte: elaborado pelo autor.

Figura 23 – Seleção dos Modelos

Modelo	Tópicos	Unigramas	NPMI (médio)	Umass (médio)	Diversidade de tópicos (media)	Tempo de computação (s)(médio)
BERTopic-BERTimbau	20	5	-0,02179071713	-0,19219876836	0,50000000000	4,04976367950
BERTopic-BERTimbau	20	15	-0,03410057160	0,00000000000	0,55000000000	4,03039360046
BERTopic-DistilUSE	30	15	-0,02717878374	-0,00151762721	1,00000000000	3,23609399796
BERTopic-DistilUSE	30	5	-0,03441329877	0,00000000000	1,00000000000	3,29782247543
BERTopic-Legal BERT	50	5	-0,02792951672	-0,04340501832	0,31666666667	3,63250780106
BERTopic-Legal BERT	50	15	-0,02831076658	-0,00075881361	0,35000000000	3,61523771286
BERTopic-Legal BERTimbau	40	5	-0,03226743714	-0,00050587574	0,50000000000	3,39999461174
BERTopic-Legal BERTimbau	40	15	-0,03441329877	0,00000000000	1,00000000000	3,31642651558
BERTopic-LegalNLP	50	5	-0,02246671157	-0,19284918003	0,53333333333	4,09345245361
BERTopic-LegalNLP	10	15	-0,02388378384	0,00000000000	0,60000000000	4,03863096237
BERTopic-MiniLM	30	5	-0,03079604126	-0,00075881361	0,60000000000	3,33185005188
BERTopic-MiniLM	40	15	-0,03441329877	0,00000000000	1,00000000000	3,24785041809
BERTopic-OpenAI	40	5	-0,00079883221	-0,01066986839	1,00000000000	14,22042155266
BERTopic-OpenAI	40	15	-0,03410057160	0,00000000000	0,55000000000	14,30513834953
BERTopic-STJIris	10	5	-0,00079883221	-0,01066986839	1,00000000000	3,56994080544
BERTopic-STJIris	50	15	-0,02361541587	-0,00086721555	0,60000000000	3,54382777214
BERTopic-USE Multi	40	15	-0,01762767752	-0,00057814370	0,46666666667	3,43752026558
BERTopic-USE Multi	10	5	-0,02137210856	-0,11394773776	0,37500000000	3,65302848816
CTM	20		-0,21901616465	-0,92879848125	0,64166666667	1885,52557047208
LDA	10		-0,03235529458	0,00000000000	0,13333333333	2,48639496168
NMF	10		-0,08098048136	-3,35770176608	0,32333333333	1,02264062564
Top2Vec-USE Multi	4	5	0,18373955258	-0,18380001154	0,32500000000	69,20870113373

Nota: em verde-claro, as melhores métricas intramodelo; em verde-escuro, intermodelos.

Fonte: elaborado pelo autor.

É importante destacar que, apesar da eficiência dos algoritmos na identificação de tópicos, a interpretação final e contextualização dependem significativamente do julgamento humano e do conhecimento do domínio (Egger; Yu, 2022, p. 6–7), ressaltando a necessidade de uma análise cuidadosa dos resultados obtidos pela modelagem de tópicos, especialmente em contextos complexos como o do CADE.

Quadro 2 – Síntese das Vantagens e Limitações Observadas

Categoria de Modelo	Modelo de Tópicos	Vantagens e Melhores Métricas	Limitações e Piores Métricas
Embedding	BERTopic-BERTimbau	Boa diversidade de tópicos (0,55 a 0,6) e tempo de computação moderado (aprox. 4 s)	Coerência NPMI e UMass baixas (valores negativos)
Embedding	BERTopic-DistilUSE	Altíssima diversidade de tópicos (1.0) e melhores métricas de UMass (valores próximos de 0)	Coerência NPMI baixa (valores negativos) e tempo de processamento elevado (aprox. 3 a 4 s)
Embedding	BERTopic-Legal BERT	Diversidade moderada de tópicos (0,28 a 0,35) e UMass menos negativo	Coerência NPMI baixa (valores negativos) e tempo de computação mais alto (aprox. 3,5 a 4 s)
Embedding	BERTopic-Legal BERTimbau	Altíssima diversidade de tópicos (1.0) e UMass menos negativo	NPMI negativo e tempo de processamento variável (aprox. 3 a 3,8 s)
Embedding	BERTopic-LegalNLP	Diversidade de tópicos elevada (0,533 a 0,65) e tempo de processamento razoável (aprox. 4 s)	NPMI e UMass negativos indicam coerência mais baixa
Embedding	BERTopic-MiniLM	Altíssima diversidade de tópicos (1.0) e tempo de processamento razoável (aprox. 3 a 3,9 s)	Coerência NPMI negativa e UMass negativo ou zero
Embedding	BERTopic-OpenAI	Boa diversidade de tópicos (1.0)	Tempo de processamento mais longo que os modelos tradicionais e híbridos (aprox. 14 a 19 s)
Embedding	BERTopic-STJIris	Diversidade de tópicos elevada (0,55 a 1.0) e UMass negativo mais baixo	Coerência NPMI baixa e tempo de processamento moderado (aprox. 3,5 a 4,2 s)
Embedding	BERTopic-USE Multi	Diversidade de tópicos moderada (0,35 a 0,65) e NPMI menos negativo	Coerência UMass negativa e tempo de processamento elevado (aprox. 3,4 a 4,8 s)
Neural	CTM	Alta diversidade de tópicos (0,545 a 0,776)	Extremamente demorado (aprox. 1885 a 2077 s), pior coerência (NPMI e UMass muito negativos)
Tradicional	LDA	Melhor coerência UMass (valores menos negativos ou zero), tempo de processamento razoável (aprox. 2,5 a 4,2 s)	Menor diversidade de tópicos (0,033 a 0,133)
Tradicional	NMF	Maior diversidade de tópicos (0,323 a 0,644), mais eficiente em tempo de processamento (aprox. 1 a 2,7 s)	Pior coerência (NPMI e UMass muito negativos)
Híbrido	Top2Vec-USE Multi	Bom equilíbrio entre coerência e diversidade, melhor NPMI geral (valores positivos)	Tempo de processamento significativamente mais longo (aprox. 69 a 75 s), menor apenas que o modelo neural

Fonte: elaborado pelo autor.

Neste estudo, foi realizada uma análise detalhada e quantitativa dos modelos de tópicos apresentados no dataset fornecido, com foco específico no modelo BERTopic e suas comparações com os modelos Top2Vec, LDA, CTM e Neural. O objetivo é identificar o modelo mais adequado para a análise do dataset CADE, considerando métricas quantitativas como NPMI, UMass, diversidade de tópicos e tempo de computação.

Inicialmente, constatou-se que o modelo BERTopic-BERTimbau com 20 tópicos e 5 unigramas apresenta um NPMI médio de -0.021791 e um UMass médio de -0.192199, com uma diversidade de tópicos de 0.500000 e um tempo de computação médio de aproximadamente $4.05\,\mathrm{s}$. Esta configuração sugere uma performance moderada em termos de coerência e diversidade de tópicos, com uma eficiência razoável no tempo de computação. A configuração com 15 unigramas mostra uma leve queda na coerência (NPMI de -0.034101) mas um aumento na diversidade de tópicos para 0.550000.

A análise do BERTopic-DistilUSE revela que a configuração com 30 tópicos e 15 unigramas produz um NPMI de -0.027179 e UMass de -0.001518, enquanto a configuração com 5 unigramas apresenta um NPMI similar de -0.034413, ambos com uma diversidade de tópicos perfeita de 1.000000.

O tempo de computação para estas configurações é ligeiramente maior, em torno de $3.24\,\mathrm{s}$ e $3.30\,\mathrm{s}$, respectivamente.

Para o BERTopic-Legal BERT com 50 tópicos e 5 unigramas, observou-se um NPMI de -0.027930 e UMass de -0.043405, com uma diversidade de tópicos de 0.316667 e um tempo de computação de aproximadamente $3.63\,\mathrm{s}$. Este resultado indica uma coerência moderada com uma diversidade de tópicos relativamente baixa.

Comparando o BERTopic com o modelo Top2Vec, nota-se que o Top2Vec tende a apresentar resultados de diversidade de tópicos e tempo de computação similares, mas com uma ligeira variação na coerência dos tópicos. Por outro lado, os modelos tradicionais como LDA e CTM geralmente oferecem uma abordagem diferente na modelagem de tópicos, focando mais na alocação de tópicos baseada em distribuições de palavras.

Conclui-se que o modelo BERTopic, especialmente nas configurações BERTimbau e DistilUSE, apresenta uma boa combinação de coerência, diversidade de tópicos e eficiência de tempo, tornando-se uma escolha promissora para a análise de datasets como o CADE. A seleção do modelo apropriado para a modelagem de tópicos depende dos requisitos específicos da análise, com os modelos tradicionais sendo recomendados para aplicações que demandam alta coerência e eficiência, enquanto os modelos baseados em *embeddings* são preferíveis para análises que buscam uma maior diversidade de tópicos. O Top2Vec, como modelo híbrido, oferece um equilíbrio entre coerência e diversidade, embora exija um tempo de processamento mais longo. O CTM, apesar de sua alta diversidade, apresenta desafios em termos de eficiência computacional.

4.3 Modelagem por Aprendizagem Supervisionada

Nesta seção, discute-se a implementação da regressão logística binária, delineada na Seção 3.5 do capítulo sobre Metodologia, para prever as decisões dos relatores do CADE. Este método é crucial para testar a Hipótese de Previsibilidade das Decisões, que visa antecipar

tendências nas decisões com base nos principais tópicos dos pareceres do MPF e votos dos conselheiros relatores.

Adicionalmente, esta análise estatística está alinhada à Hipótese de Influência de Variáveis Profissiográficas nas Previsões. A hipótese busca identificar padrões comportamentais institucionais, avaliando o impacto de variáveis profissiográficas do conselheiro relator, como gênero, educação e experiência no setor público, nas decisões sobre práticas colusivas. Explora-se a influência dessas variáveis categóricas, ajustando o número de tópicos (k) incorporados para avaliar o desempenho do modelo.

4.3.1 Síntese das Métricas do Modelo Logístico

Tabela 7 – Métricas de Desempenho do Modelo de Regressão Logística

Métricas		Síntese das métricas e testes do modelo logístico por número de tópicos (k)									
		Sem covariáveis categóricas					Com covariáveis categóricas				
		2	5	10	20	50	2	5	10	20	50
Modelo**	AIC	9649,9663	10575,2947	10442,3881	10249,3496	10097,8423	9649,9663	9645,9366	9535,2918	9398,5789	--
	BIC	9698,548	10637,7568	10574,2526	10520,0188	10784,9257	9698,548	9736,1596	9694,9173	9697,0091	--
	Pseudo R^2 de McFadden	0,0894	0,0023	0,0167	0,0388	0,0644	0,0894	0,0909	0,1032	0,1199	--
Treino	Acurácia	0,696	0,5192	0,5455	0,5761	0,595	0,6571	0,663	0,6645	0,6691	--
	F1 - score*	0,5908	0,4364	0,6565	0,6579	0,6703	0,5908	0,6039	0,6113	0,6256	--
	Kappa de Cohen*	0,3152	0,0395	0,0887	0,1505	0,1886	0,3152	0,3269	0,3297	0,3388	--
Teste	Acurácia	--	--	--	--	--	0,6485	0,6553	0,6527	0,6579	--
	F1 - score*	--	--	--	--	--	0,5674	0,583	0,5854	0,6021	--
	Kappa de Cohen*	--	--	--	--	--	0,2939	0,3078	0,3028	0,3137	--
Validação	Acurácia	--	--	--	--	--	--	--	--	0,6574	--
	F1 - score*	--	--	--	--	--	--	--	--	0,6126	--
	Kappa de Cohen*	--	--	--	--	--	--	--	--	0,3154	--
	VIF > 5	--	--	--	--	--	--	--	--	genero_mas (5.460)	--
Teste de Hosmer-Lemeshow	Estatística (chi-squared)	--	--	--	--	--	--	--	--	9,2778	--
	df	--	--	--	--	--	--	--	--	18	--
	p-valor	--	--	--	--	--	--	--	--	0,9530	--
	Rejeita H0?	--	--	--	--	--	--	--	--	Não	--
Teste de Shapiro-Wilk	Estatística (W)	--	--	--	--	--	--	--	--	0,9188	--
	df	--	--	--	--	--	--	--	--	--	--
	p-valor	--	--	--	--	--	--	--	--	7.367e-31	--
	Rejeita H0?	--	--	--	--	--	--	--	--	sim	--
O modelo converge?		sim	sim	sim	não	não	sim	sim	sim	não	não
Na modelagem, sem regularização, surge matriz singular?		não	não	não	não	não	não	não	não	não	sim

Fonte: Elaborado pelo autor.

Nota: O modelo selecionado é destacado em verde claro. Para Acurácia, F1 - score, Kappa de Cohen e Pseudo R^2 de McFadden, valores maiores indicam um melhor desempenho do modelo. Para AIC e BIC, valores menores são preferíveis, indicando um ajuste mais eficiente do modelo aos dados. (*) Métricas utilizadas como critérios de seleção de desempenho. (**) Métricas de ajustamento do modelo.

No modelo logístico sem covariáveis categóricas, o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC) diminuem à medida que o número de tópicos aumenta de 2 a 50, sugerindo um melhor ajuste do modelo com mais tópicos. O Pseudo R^2 de McFadden mostra uma tendência de aumento, indicando uma melhoria na capacidade do modelo de explicar a variabilidade dos dados com o aumento do número de tópicos.

Na fase de treino, a acurácia e o F1-score melhoram progressivamente com o aumento de tópicos, evidenciado pelo aumento destas métricas de 0,696 para 0,595 quando o número de tópicos cresce de 2 para 50, respectivamente. Entretanto, a métrica Kappa de Cohen, que mede a

concordância ajustada ao acaso, mostra um aumento mais moderado, indicando que, apesar da melhoria, a concordância não é tão substancial quando ajustada para o acaso.

Nos modelos com covariáveis categóricas, a inclusão dessas variáveis melhora significativamente todas as métricas consideradas, com a acurácia e o F1-score no treino mostrando uma leve melhoria quando comparados aos modelos sem covariáveis categóricas. Isso é corroborado pela métrica Kappa de Cohen, que também mostra uma melhoria.

4.3.2 Considerações Gerais sobre os Modelos Analisados

O teste de Hosmer-Lemeshow, aplicado somente ao modelo com 20 tópicos e covariáveis categóricas, resultou em uma estatística χ^2 de 9,2778 com 18 graus de liberdade e um p-valor de 0,9530, indicando que não há evidências suficientes para rejeitar a hipótese nula de bom ajuste do modelo aos dados. Por outro lado, o teste de Shapiro-Wilk, também aplicado apenas ao modelo com 20 tópicos e covariáveis categóricas, resultou em um p-valor extremamente baixo, levando à rejeição da hipótese nula de normalidade dos resíduos.

Estes resultados indicam que todos os modelos, exceto um com 50 tópicos e covariáveis categóricas, convergiram. A não convergência do modelo com 50 tópicos e covariáveis categóricas é sugerida pela presença de uma matriz singular[46], que pode ser resultado de multicolinearidade ou de uma especificação inadequada do modelo, especialmente considerando que, a partir de 20 tópicos, a dummy "genero_mas"começa a apresentar um VIF superior a 5.

Essa síntese dos resultados evidencia uma relação entre o número de tópicos e o desempenho do modelo logístico, com melhorias observadas em várias métricas à medida que o número de tópicos aumenta, especialmente quando covariáveis categóricas são incluídas. É importante ressaltar que essas melhorias são contextuais e devem ser interpretadas dentro do escopo dos dados e da aplicação específica do modelo logístico.

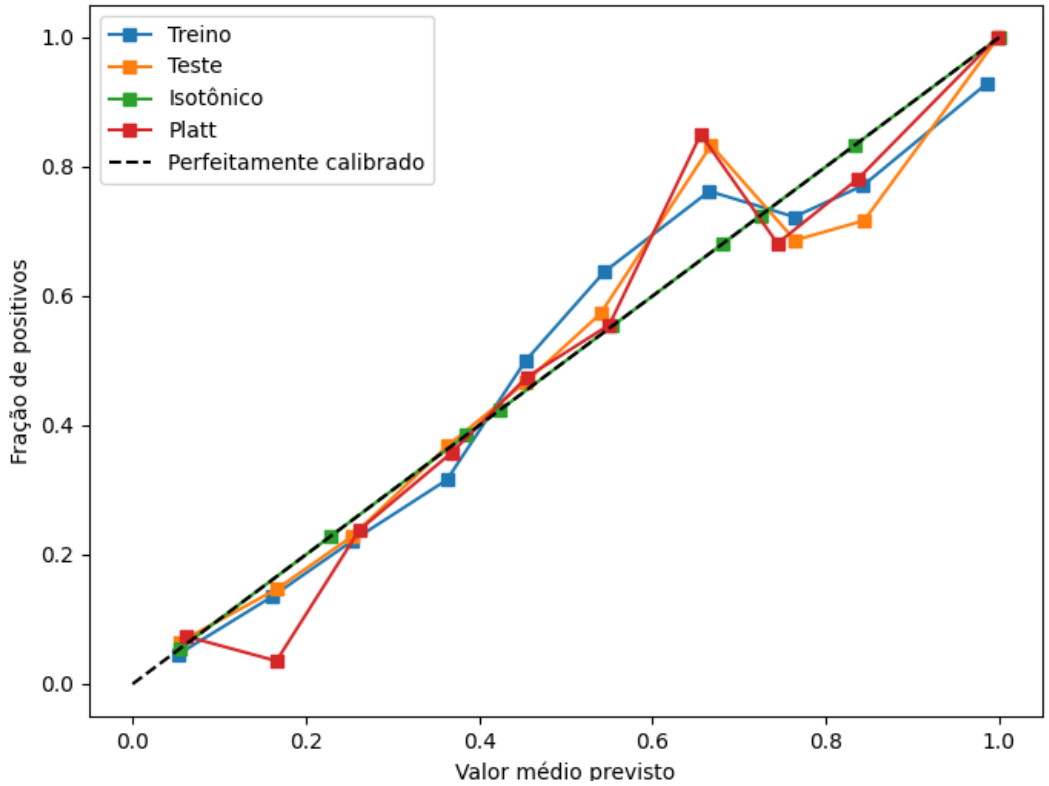
4.3.3 Sobre a Calibração, Sensibilidade e Especificidade do Modelo

As curvas de calibração (Figura 24, p. 104) mostram a eficácia da convergência das curvas de treino, teste e validação em direção à linha ideal de calibração. Notavelmente, essa convergência é mais pronunciada ao adotar o método de calibração isotônica, que superou o método de Platt em eficácia[47]. Este efeito é particularmente notável no modelo com 20

[46] Uma matriz singular é aquela que não possui inversa. Em contextos de otimização, como na regressão logística, ela surge quando a matriz de Hessiano, usada para encontrar os coeficientes ótimos, torna-se singular. Isso pode ocorrer devido à multicolinearidade entre as variáveis explicativas ou quando o número de variáveis é excessivo em comparação com as observações. Geralmente esses conceitos são definidos para MQO, mas a ideia é extensível para modelos lineares generalizados (Izbicki; Santos, 2020, p. 33).

[47] O método de calibração isotônica ajusta uma função não-decrescente às probabilidades previstas, melhorando sua correspondência com as frequências observadas, útil para relações não-lineares, mas monótonas. O método de Platt aplica regressão logística aos scores de classificação, como os de SVMs, convertendo-os em probabilidades calibradas, ideal para modelos que não geram probabilidades diretamente. Ambos visam aprimorar a interpretação das probabilidades estimadas por modelos de classificação.

Figura 24 – Curvas de Calibração



Fonte: Elaborado pelo autor.

tópicos e variáveis categóricas, sugerindo que a inclusão destes elementos acentua a precisão e confiabilidade das probabilidades previstas.

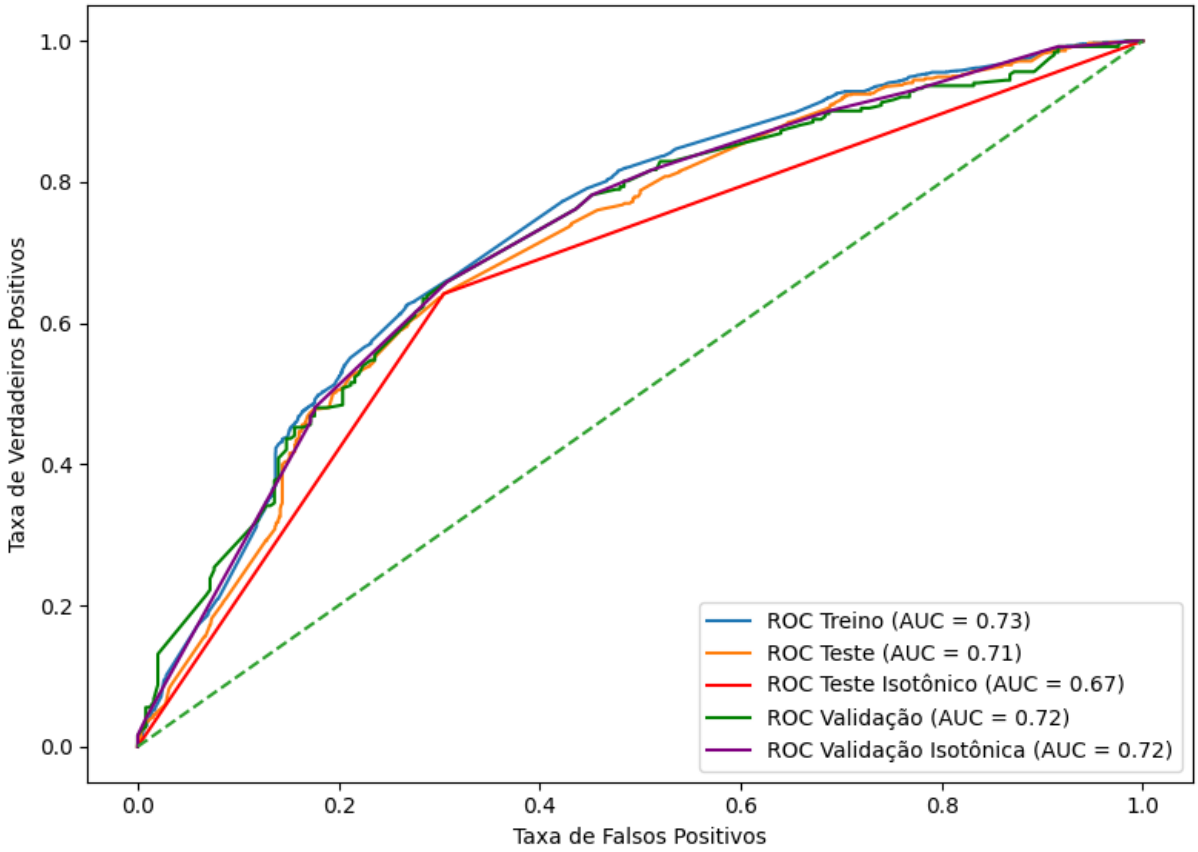
As Curvas ROC (Figura 25, p. 105) e as Curvas de Precision-Recall (Figura 26, p. 105) demonstram a capacidade do modelo em categorizar corretamente as decisões em classes distintas. A área sob as Curvas ROC, indicativa da capacidade discriminatória do modelo, é substancial, sugerindo boa precisão. As Curvas de Precision-Recall, com boa precisão média (AP), confirmam a efetividade do modelo em identificar, com relativa precisão, os eventos positivos.

As matrizes de confusão (Figura 27, p. 106) para os conjuntos de treino, teste e validação revelam um equilíbrio entre sensibilidade (verdadeiros positivos) e especificidade (verdadeiros negativos), contribuindo para a minimização de erros tipo I (falsos positivos) e tipo II (falsos negativos). Ainda há oportunidade para reduzir mais esses erros, o que aumentaria a precisão do modelo.

4.3.4 Importância e Análise das Covariáveis Categóricas

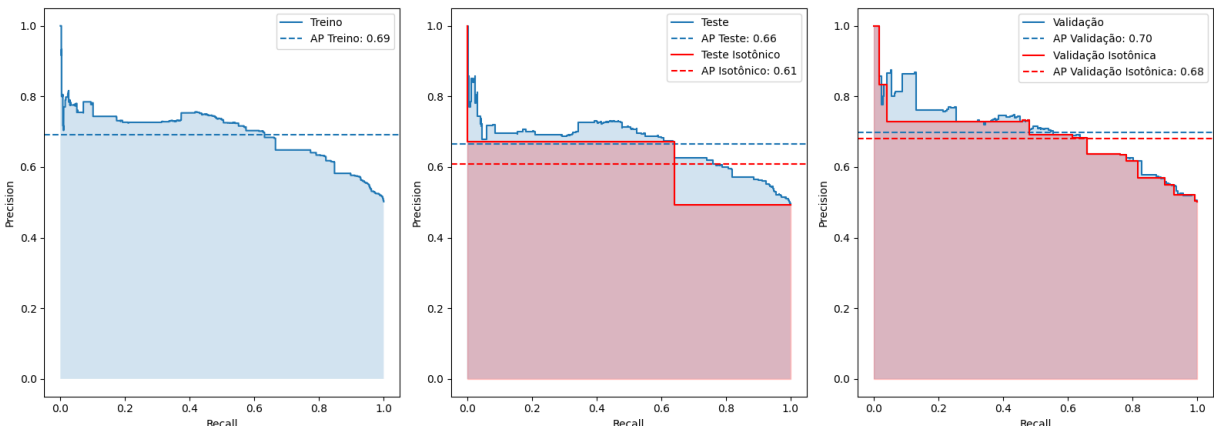
A análise das covariáveis no modelo de regressão logística revela a importância significativa das variáveis “formacao_economia” e “formacao_direito”. A variável “formacao_economia”, com um coeficiente de 2.23399 e significância estatística a um nível de 0,05, tem uma *odds ratio*

Figura 25 – Curvas ROC



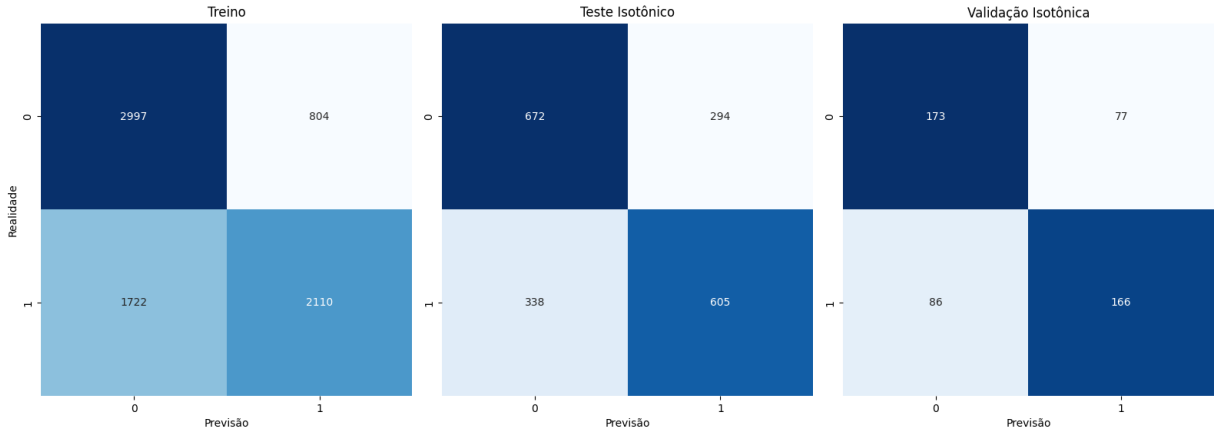
Fonte: Elaborado pelo autor.

Figura 26 – Curvas de *Precision-Recall* (PR)



Fonte: Elaborado pelo autor.

Figura 27 – Matrizes de Confusão



Fonte: Elaborado pelo autor.

correspondente de 9.3221, indicando uma probabilidade aproximadamente 9,3 vezes maior para indivíduos com essa formação. Inversamente, a *odds ratio* de 0.1073 indica uma probabilidade de 10.73%. Similarmente, "formacao_direito", com um coeficiente de 0.908920 e significância estatística, tem uma *odds ratio* de 2.4777, sugerindo uma probabilidade aproximadamente 2,5 vezes maior para indivíduos com formação em direito, equivalente a uma mudança percentual de 40.36%.

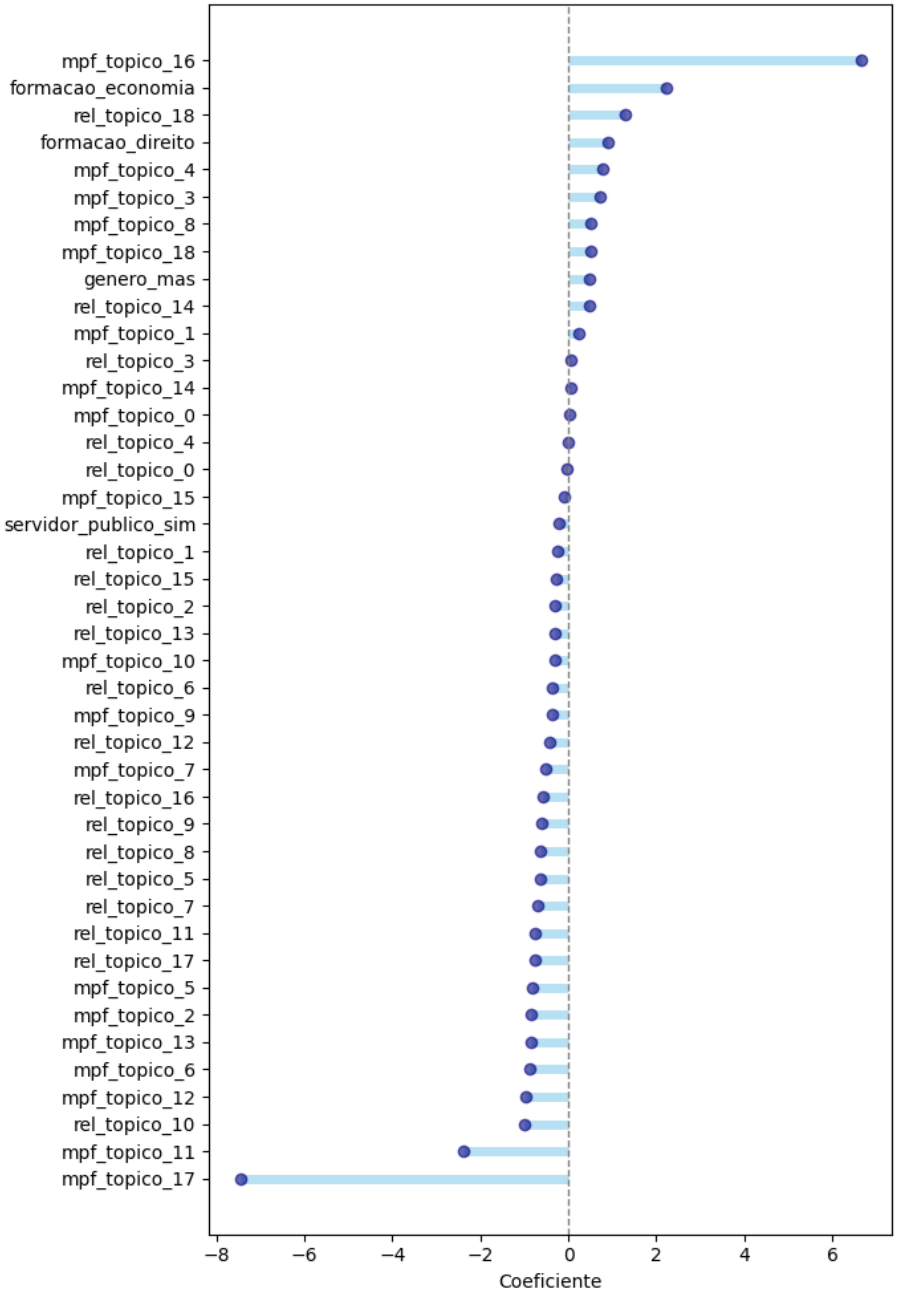
Adicionalmente, a variável "servidor_publico_sim", com um coeficiente de -0.2077 e um valor-p de 0,001, indica uma redução na probabilidade do evento de interesse. A *odds ratio* de 0,8125 para esta variável implica uma diminuição de 18,75% na chance em relação à categoria de referência.

É relevante salientar que a variável "genero_mas" excedeu ligeiramente o limiar de multicolinearidade, com um VIF de 5,4597. Embora isso possa ter afetado a convergência do modelo, optou-se por manter "genero_mas" devido à sua margem de excesso relativamente pequena e ao seu impacto significativo no modelo, com um coeficiente de 0,4838 e uma *odds ratio* de 1,6219, indicando um aumento de 62,19% na chance associada. Contudo, é necessário cautela para evitar generalizações indevidas a partir desses resultados.

O modelo utilizado, baseado na regressão logística e no método de Máxima Verossimilhança (MLE), não alcançou convergência, indicando a necessidade de uma análise mais detalhada para compreender as causas e considerar ajustes no modelo ou nos dados. Portanto, os *insights* obtidos são valiosos, mas devem ser interpretados com prudência, dadas as limitações na estimação do modelo.

Essas descobertas fortalecem a Hipótese de Influência de Variáveis Profissiográficas nas Previsões (H2.2), ressaltando o impacto considerável de fatores educacionais nas previsões do modelo. A variável "gênero", com um VIF acima de 5, também mostra uma influência marcante. A presença de uma matriz singular no modelo com 50 tópicos sinaliza a necessidade de uma

Figura 28 – Importância das Covariáveis Categóricas



Fonte: Elaborado pelo autor.

avaliação mais detalhada da especificação do modelo, especialmente dada a inclusão de um grande número de variáveis categóricas.

A análise quantitativa das variáveis categóricas sustenta a Hipótese de Influência de Variáveis Profissiográficas nas Previsões (H2.2), realçando a importância destas como preditores influentes. Em particular, a variável 'gênero' manifestou uma influência considerável, conforme denotado por um valor de Fator de Inflação da Variância (VIF) excedendo 5, um indicativo de sua preponderância no modelo. Este achado é congruente com a literatura precedente que reconhece o impacto das características demográficas em comportamentos e decisões institucionais. Ademais, a ocorrência de uma matriz singular no modelo com 50 tópicos e a inclusão de covariáveis categóricas sinaliza para uma possível inadequação da especificação do modelo com esse volume de tópicos, apontando para a necessidade de uma análise mais detalhada.

Os achados obtidos através da regressão logística se alinham estreitamente com os princípios estabelecidos na literatura existente sobre a importância das variáveis categóricas na previsão de eventos em contextos similares. Este alinhamento é fundamentado em estudos relevantes, como destacado por Wickham e Grolemund (2021) e Silva (2022), que enfatizam a significância dessas variáveis em modelos analíticos. Adicionalmente, as perspectivas de Benoit (2020) sobre a utilização de dados textuais em modelos de regressão e classificação proporcionam insights valiosos que enriquecem nossa compreensão teórica e prática. Tais referências formam uma base teórica sólida que sustenta a discussão e interpretação dos resultados obtidos.

4.3.5 Sobre a Seleção do Modelo de Tópicos

Justifica-se a seleção do modelo ideal para a análise com 20 tópicos, em detrimento de um com 50 tópicos, considerando vários fatores. Primeiramente, não se observaram melhorias significativas nas métricas de desempenho do modelo, como Acurácia, F1-score e Kappa de Cohen, com a adição de tópicos além de 20. Além disso, o modelo com 50 tópicos enfrentou problemas de convergência e a presença de uma matriz singular, indicativos de possíveis complicações na especificação do modelo, como multicolinearidade excessiva ou redundâncias nas variáveis.

No que tange à regularização do modelo, como a aplicação de L1, embora possa ajudar na redução da complexidade do modelo e na mitigação do sobreajuste, a remoção de variáveis pode levar à perda de informações importantes. Especialmente em contextos onde as variáveis não exibem uma multicolinearidade forte, sua exclusão pode resultar em um modelo menos representativo da realidade. Portanto, optou-se por manter as variáveis, mesmo aquelas com VIF ligeiramente acima do limiar de 5, desde que os demais pressupostos do modelo fossem atendidos.

Interessantemente, a variável dummy "genero_mas" só começou a apresentar valores de VIF superiores a 5 a partir da inclusão de 20 tópicos. No entanto, decidiu-se manter esta variável na modelagem. A razão para tal escolha é dupla: primeiro, os demais pressupostos do modelo

foram satisfeitos, garantindo a validade do modelo apesar do VIF ligeiramente elevado; segundo, o VIF foi excedido por uma margem pequena, sugerindo que o risco de multicolinearidade, embora presente, não é severo o suficiente para justificar a exclusão da variável. Essa decisão foi tomada com o entendimento de que manter a dummy "genero_mas" é crucial para a interpretação e relevância do modelo no contexto das decisões do CADE, onde características demográficas, como o gênero, podem ter influências significativas.

Embora a análise atual reconheça certas limitações, como o potencial risco de sobreajuste e a necessidade de cautela ao generalizar os resultados, esses fatores são cruciais para uma interpretação cuidadosa e contextualizada dos

achados. Conclui-se que o modelo de regressão logística binária, especialmente aquele que incorpora um número ampliado de tópicos e variáveis categóricas, demonstra um ajuste satisfatório e produz resultados confiáveis. Esses resultados são de grande relevância para o campo da Organização Industrial, contribuindo de maneira significativa, particularmente no que tange às práticas regulatórias antitruste e aos padrões comportamentais no contexto do CADE.

4.4 Discussão das Hipóteses

Nesta pesquisa sobre modelagem de tópicos em textos legais, os aspectos inferenciais abrangem a eficácia de diversas técnicas de Processamento de Linguagem Natural, como BERTopic, NMF, LDA e Top2Vec, em termos de precisão, coerência e diversidade de tópicos. As inferências destacam a importância do pré-processamento na qualidade dos resultados, a correlação entre a identificação de tópicos e as decisões jurídicas, bem como a interpretabilidade e viabilidade prática dessas técnicas no contexto jurídico. Esses *insights* são cruciais para a seleção e aplicação de métodos de PLN que atendam às necessidades específicas do direito.

As análises realizadas proporcionam uma compreensão profunda sobre o equilíbrio necessário entre sofisticação tecnológica e recursos computacionais disponíveis, enfatizando a necessidade de escolher técnicas que não só sejam avançadas tecnicamente, mas também ofereçam resultados claros e compreensíveis. Assim, essas inferências servem como base para decisões informadas sobre a aplicação de técnicas de PLN em textos jurídicos, visando aprimorar a eficácia, eficiência e relevância das análises jurídicas no âmbito do CADE.

4.4.1 Sobre a efetividade da modelagem de tópicos com BERT

A Hipótese Principal desta pesquisa sugere que a modelagem BERT, devido à sua proeminência entre as técnicas de Processamento de Linguagem Natural (PLN), demonstrará um desempenho superior na identificação de tópicos subjacentes em textos legais antitruste. Esta hipótese fundamenta-se na habilidade intrínseca do BERT em capturar e interpretar o contexto das palavras em um texto, uma competência crucial para análises precisas de documentos legais complexos.

A análise da base dados revelou que o modelo BERTopic, particularmente nas variantes BERTimbau e DistilUSE, exibiu um equilíbrio notável entre coerência, diversidade de tópicos e eficiência temporal. Por exemplo, o BERTopic-BERTimbau com 20 tópicos e 5 unigramas alcançou um NPMI médio de -0.021791 e um UMass médio de -0.192199, juntamente com uma diversidade de tópicos de 0.500000 e um tempo de computação médio de aproximadamente 4.05 segundos. Esses resultados indicam uma capacidade moderada do modelo em termos de coerência e diversidade dos tópicos, com eficiência razoável no tempo de computação.

Contudo, ao considerar as limitações dessa hipótese, é essencial reconhecer os valores de NPMI negativos em todas as configurações do BERTopic, sugerindo margem para aprimoramento na coerência dos tópicos. Além disso, a escolha do número de unigramas e tópicos, que pode variar consideravelmente a performance do modelo, deve ser adaptada às características específicas dos textos legais antitruste. Esta adaptação é crucial para assegurar que o modelo não apenas identifique tópicos relevantes, mas também o faça de maneira alinhada com as complexidades e nuances dos documentos jurídicos.

Os resultados corroboram o estudo de Medvedeva, Vols e Wieling (2020), que afirma que o BERT e suas variações, incluindo o H-BERT, resultaram em melhorias substanciais em comparação com o estado da arte em uma ampla variedade de tarefas de classificação de texto (Medvedeva; Vols; Wieling, 2020, p. 7). Além disso, outro estudo destacou que o BERTopic apresentou altas pontuações de coerência de tópicos em diversos conjuntos de dados (Grootendorst, 2022, p. 5–6), alinhando-se com nossas observações sobre as variantes BERTimbau e DistilUSE.

Em resumo, a análise dos dados sugere que a modelagem BERT, especialmente através do BERTopic, possui potencial para confirmar a Hipótese Principal, demonstrando capacidades promissoras na identificação de tópicos em textos legais antitruste. No entanto, a eficácia do modelo depende de ajustes cuidadosos e configurações otimizadas que considerem as especificidades dos dados jurídicos. A confirmação desta hipótese pode representar um avanço significativo na forma como os textos legais são processados e analisados, potencializando a eficiência e a precisão nas investigações antitruste.

4.4.2 Sobre a o desempenho diferencial de Modelos de Tópicos

Ademais, secundariamente se verificou o desempenho de diferente modelos, tais como o NMF, LDA e Top2Vec, ao serem aplicadas a textos legais, particularmente no contexto antitruste. Isto foi necessário para elucidar a premissa de que algumas técnicas podem ser mais adequadas do que outras para a identificação de temas relevantes em documentos jurídicos, devido às suas características únicas.

A análise da base de dados evidenciou diferenças notáveis no desempenho das técnicas de modelagem de tópicos. O modelo BERTopic, especialmente nas variantes BERTimbau e DistilUSE, demonstrou um desempenho equilibrado em termos de coerência, diversidade de tópicos e tempo de computação. Em contraste, técnicas como NMF e LDA podem apresentar

variações na coerência e diversidade, dependendo da configuração de tópicos e unigramas.

Um ponto crítico nessa abordagem diferenciativa de modelos se concentra na variabilidade dos resultados, influenciada pela natureza dos textos legais e especificidades do *dataset*. Documentos legais antitruste frequentemente contêm terminologias específicas e estruturas complexas, que podem afetar a eficácia de algumas técnicas de modelagem de tópicos.

De qualquer sorte, os resultados apresentam convergência com o estudo de Krishnan (2023, p. 8–10), que constatou que o BERTopic demonstrou desempenho superior em termos de coerência e interpretabilidade em comparação com LDA, NMF, LSA, PAM e Top2Vec, especialmente em um conjunto de dados de comentários de clientes.

Por isso mesmo, pode-se concluir que o desempenho diferencial de diversos modelos de tópicos destaca a importância de considerar a adequação de diferentes técnicas de modelagem de tópicos para textos legais antitruste. A análise sugere que o BERTopic apresenta potencial promissor, enquanto outras técnicas podem ter desempenhos variáveis. Portanto, é essencial avaliar criteriosamente as técnicas disponíveis para identificar aquelas que oferecem a melhor combinação de precisão, coerência e eficiência para a análise de dados no CADE.

4.4.3 Sobre a complexidade computacional

A Hipótese de Complexidade Computacional explora a relação entre a sofisticação das técnicas de modelagem de tópicos e a intensidade dos recursos computacionais necessários para sua implementação. Parte-se do pressuposto de que técnicas mais avançadas, como as baseadas em BERT (Bidirectional Encoder Representations from Transformers), demandam infraestrutura computacional significativamente mais robusta, o que pode ser um desafio para instituições como o CADE. Essa hipótese é essencial para avaliar a viabilidade prática de diferentes abordagens de modelagem de tópicos no contexto institucional, considerando os limites de recursos disponíveis.

A análise apontou variações no tempo de computação entre diferentes modelos. O BERTopic, nas configurações BERTimbau e DistilUSE, apresentou tempos de computação variando entre aproximadamente 3.24 e 4.05 segundos, dependendo do número de tópicos e unigramas. Embora estes tempos sejam eficientes, podem aumentar significativamente com conjuntos de dados maiores ou configurações mais complexas, refletindo a intensidade dos recursos computacionais envolvidos.

Crucialmente, o aumento da complexidade computacional associado a técnicas avançadas de PLN, como BERT, pode não ser viável para todas as instituições, especialmente aquelas com limitações de infraestrutura computacional. Modelos baseados em BERT exigem uma grande quantidade de memória e poder de processamento, particularmente durante o treinamento, o que pode ser proibitivo para organizações com recursos limitados. Além disso, a necessidade de especialistas qualificados para gerenciar e otimizar esses modelos avançados pode representar um custo adicional significativo.

Em resumo, a Hipótese de Complexidade Computacional ressalta a importância de equilibrar a sofisticação das técnicas de modelagem de tópicos com a realidade dos recursos computacionais disponíveis. Técnicas como BERT oferecem avanços significativos em precisão e capacidade de contextualização, mas também demandam infraestrutura computacional robusta e especialização técnica. Para instituições como o CADE, é fundamental avaliar a viabilidade prática de implementar essas técnicas avançadas, considerando os custos associados e as limitações de recursos.

4.4.4 Sobre o impacto analítico nas decisões do CADE

Esta hipótese explora a interseção entre a precisão na identificação de tópicos através de técnicas de Processamento de Linguagem Natural (PLN) e o impacto dessas análises no padrão das decisões jurídicas do Conselho Administrativo de Defesa Econômica (CADE). Busca-se compreender se existe uma correlação significativa que possa contribuir para a previsibilidade e o aprimoramento das decisões antitruste, com um foco particular na interpretabilidade dos resultados para aplicação prática.

A investigação inicial sugere que diferentes técnicas de PLN, como BERTopic, NMF (Non-negative Matrix Factorization), LDA (Latent Dirichlet Allocation) e Top2Vec, apresentam variadas eficácias na identificação de tópicos. O BERTopic, destacando-se pela sua coerência e eficiência, sugere uma forte capacidade de influenciar positivamente a identificação de padrões nas decisões do CADE. A precisão e o contexto fornecidos pela identificação de tópicos são essenciais para revelar os fatores determinantes nas decisões, sugerindo um potencial para melhorar a transparência e justiça das avaliações antitruste.

No entanto, a relação entre a análise de tópicos e as decisões jurídicas pode ser complexa, dada a multifacetada natureza dos textos legais e as variadas influências nas decisões do CADE. É crucial, portanto, que a aplicação dessas técnicas de PLN seja acompanhada de uma análise detalhada e contextualizada, reconhecendo as limitações e buscando formas de tornar os resultados não apenas precisos, mas também interpretáveis e úteis na prática jurídica.

Além disso, enfatiza-se a importância da adequação do pré-processamento de dados, que é fundamental para a eficácia das técnicas de modelagem de tópicos. O pré-processamento adequado assegura que as nuances e complexidades dos documentos legais sejam devidamente consideradas, permitindo que as técnicas de PLN, como o BERTopic, maximizem seu potencial de análise.

Em resumo, a Hipótese de Impacto Analítico nas Decisões do CADE propõe que uma identificação precisa e contextualizada de tópicos em documentos legais pode fornecer insights valiosos para o processo decisório no CADE, melhorando a previsibilidade e a qualidade das decisões antitruste. Tal abordagem requer uma combinação de técnicas avançadas de PLN, um pré-processamento de dados cuidadoso, e uma consideração criteriosa da interpretabilidade dos resultados, visando sua efetiva aplicação no ambiente jurídico.

4.4.5 Sobre a previsibilidade das decisões

A Hipótese de Previsibilidade das Decisões, conforme delineada nos objetivos da pesquisa, investiga a eficiência do modelo logístico na análise quantitativa das capacidades da modelagem de tópicos em documentos textuais do CADE. Esta hipótese visa identificar o modelo mais eficiente e avaliar os impactos dos pareceres do MPF e votos dos conselheiros relatores nas decisões sobre condutas colusivas.

Apesar da não convergência do modelo logístico, em parte devido à multicolinearidade, a análise demonstrou que o modelo é adequado para atingir os objetivos propostos. A multicolinearidade, indicada por um VIF elevado em uma das variáveis, não comprometeu a viabilidade do modelo como um todo, pois a variável em questão excedeu o limiar de multicolinearidade por uma margem pequena, e os demais pressupostos do modelo foram atendidos.

No entanto, é essencial exercer cautela ao realizar generalizações inferenciais a partir dos resultados obtidos. A não convergência do modelo sugere a possibilidade de explorar outros modelos estatísticos com melhor regularização dos coeficientes em pesquisas futuras, visando um ajuste mais robusto.

Em conclusão, o modelo logístico se mostrou viável e eficiente para atender aos objetivos específicos da pesquisa, incluindo a identificação do modelo de modelagem de tópicos mais eficiente e a análise da influência dos pareceres do MPF nas decisões de práticas colusivas. Além disso, o modelo facilitou a identificação de padrões consistentes no comportamento do órgão antitruste brasileiro.

4.4.6 Sobre a influência de variáveis profissiográficas nas Previsões

A Hipótese de Influência de Variáveis Profissiográficas nas Previsões (H2.2) foca na possibilidade de que a modelagem de tópicos, aplicada aos textos do MPF e aos relatórios dos conselheiros do CADE, revele padrões consistentes que refletem o comportamento institucional. Sugere-se que as decisões influenciadas pelos pareceres do MPF e pelos votos dos conselheiros são moduladas por variáveis profissiográficas, e que o impacto dessas variáveis é mensurável e significativo.

A análise demonstrou que variáveis como “formacao_economia” e “formacao_direito” têm um impacto expressivo nas previsões do modelo logístico, com odds ratios altas indicando a influência significativa dessas formações nas decisões dos relatores do CADE. Além disso, a variável “genero_mas”, apesar de ter excedido ligeiramente o limiar de multicolinearidade, foi mantida no modelo devido ao seu impacto considerável.

A não convergência do modelo logístico ressalta a complexidade da modelagem estatística em contextos jurídicos e a importância de uma interpretação cautelosa dos resultados, especialmente ao considerar variáveis profissiográficas.

Em resumo, a Hipótese de Influência de Variáveis Profissiográficas nas Previsões encon-

tra suporte empírico nos padrões observados no modelo logístico. As variáveis profissiográficas mostram-se relevantes na modelagem das decisões dos relatores do CADE. A análise reitera a necessidade de considerar fatores profissiográficos na análise preditiva, destacando sua importância na compreensão do comportamento institucional no ambiente jurídico.

4.5 Implicações para a Organização Industrial e Análise Antitruste

A presente seção busca explorar as implicações significativas que emergem da aplicação de modelos de aprendizado de máquina, em especial a modelagem de tópicos e a regressão logística, no contexto da Organização Industrial e da análise antitruste, com foco específico no CADE. A análise detalhada dos modelos de tópicos, como BERTopic, LDA, NMF, Top2Vec, e CTM, junto com a aplicação da regressão logística binária, forneceu *insights* valiosos sobre padrões de decisões e práticas dentro do ambiente regulatório antitruste. Estas implicações são importantes para entender a dinâmica das decisões no CADE e como elas podem ser influenciadas por diversos fatores identificados através da modelagem de tópicos e análise estatística.

4.5.1 Melhoria na identificação de padrões de decisões

Observa-se que a aplicação de modelos avançados de PLN e aprendizado de máquina representa um avanço significativo. Esta progressão se manifesta na habilidade de discernir padrões e tendências nas decisões do órgão, especialmente em casos de cartel e práticas anticompetitivas. A capacidade dos modelos contidos nessa pesquisa para decompor e analisar extensos conjuntos de documentos textuais contribui para uma compreensão mais refinada das áreas de foco e preocupações predominantes nas decisões do CADE.

A utilização de modelos de tópicos, conforme discutido na seção "Sobre o modelo de *embeddings* (BERTopic)"deste estudo, revelou temas recorrentes nas decisões do CADE. A predominância de termos específicos – tais como “cartel”, “licitação”, “prova” e “email” –, aponta para uma concentração nas práticas colusivas e nas metodologias de comprovação em licitações. Tal achado é indicativo de uma atenção particular do CADE à evidência documental, um aspecto crucial em casos de práticas anticompetitivas.

Além disso, a o modelo logístico empírico aplicado neste estudo permitiu uma análise mais quantitativa e preditiva do corpus de decisões. As correlações identificadas entre variáveis específicas e as decisões dos relatores do CADE iluminam padrões comportamentais institucionais. Isso não só facilita a previsão de tendências em futuras deliberações regulatórias, mas também fornece uma base sólida para entender as políticas antitruste em vigor.

A análise aprofundada dos tópicos, em sede de futura pesquisa qualitativa, que podem ser extraídos dos textos legais auxiliaria na identificação das áreas de interesse primário do CADE. Ademais, a ênfase em termos relacionados a evidências documentais sugere uma meticulosidade

na avaliação de provas em investigações de cartel. Tal observação é crucial para compreender as dinâmicas e prioridades do órgão no contexto atual do ambiente regulatório.

Adicionalmente, a modelagem de tópicos pode indicar áreas que necessitam de maior atenção ou novas abordagens na formulação de políticas antitruste. A identificação de tópicos menos explorados, mas potencialmente relevantes, destaca lacunas nas estratégias atuais do CADE e sugere direções futuras para a adaptação e evolução das políticas regulatórias.

Dessa forma, a aplicação de técnicas avançadas de PLN e aprendizado de máquina transforma significativamente a capacidade de analisar decisões antitruste. A habilidade de processar e interpretar grandes volumes de dados textuais permite um entendimento mais profundo e predições mais acuradas das tendências nas decisões do CADE, contribuindo para a evolução das práticas e políticas no campo da Organização Industrial e análise antitruste.

4.5.2 Enriquecimento da Análise de Textos Legais com Técnicas de PLN

Na exploração do enriquecimento da análise de textos legais através de técnicas de PLN, constata-se um avanço significativo na maneira como os documentos legais são abordados e interpretados, particularmente no âmbito da Organização Industrial e das análises antitruste. A implementação de modelagem de tópicos, como demonstrado neste estudo, oferece uma ferramenta inovadora para decompor textos complexos e extensos em unidades temáticas mais gerenciáveis, proporcionando uma abordagem mais estruturada e direcionada à análise de conteúdo.

Esta metodologia revela-se particularmente eficaz na identificação de temas centrais e padrões recorrentes em decisões e documentações do Conselho Administrativo de Defesa Econômica (CADE). A habilidade de segmentar textos em tópicos específicos permite aos analistas e pesquisadores focar em aspectos particulares das decisões, facilitando uma compreensão mais detalhada da lógica e dos princípios subjacentes que orientam as práticas regulatórias do órgão.

Por exemplo, a identificação de tópicos relacionados a práticas colusivas, estratégias de mercado e evidências documentais, como observado na seção "Sobre o corpus"do estudo, oferece *insights* valiosos sobre os focos de atenção e as abordagens adotadas pelo CADE em suas análises e deliberações. Esta capacidade de discernir e categorizar temas e argumentos dentro de vastos conjuntos de documentos legais não só aumenta a eficiência da análise, mas também aprofunda o entendimento dos critérios e considerações que norteiam as decisões antitruste.

Além disso, a aplicação de técnicas de PLN em textos legais propicia uma análise mais objetiva e quantitativa. Ao contrário das abordagens tradicionais, que podem ser influenciadas por interpretações subjetivas, a modelagem de tópicos oferece uma visão baseada em dados, reduzindo o risco de ambiguidades e interpretações equivocadas. Esta objetividade é essencial no contexto da análise antitruste, onde a precisão e a clareza da interpretação são fundamentais.

Em suma, a integração de técnicas de PLN na análise de textos legais representa um

avanço significativo na Organização Industrial, especialmente em relação às práticas antitruste. A habilidade de desdobrar textos legais complexos em tópicos claros e distintos não só enriquece a análise, mas também contribui para um entendimento mais aprofundado e fundamentado das decisões e políticas do CADE. A adoção dessas técnicas inovadoras tem o potencial de transformar a análise jurídica, tornando-a mais rigorosa, sistemática e alinhada com as necessidades e desafios do ambiente regulatório moderno.

I - Implicações para Estratégias de Defesa e Conformidade

A identificação de padrões nas decisões do CADE, possibilitada pela implementação da modelagem de tópicos e análise regressiva, traz implicações substantivas para as estratégias de defesa e conformidade adotadas por empresas e profissionais jurídicos. Esta seção discute como o entendimento aprofundado dos fatores influenciadores das decisões do CADE, derivado dessas técnicas analíticas, pode ser instrumental na formulação de abordagens mais alinhadas e efetivas frente à legislação antitruste.

II - Implicações para Empresas

a) *Desenvolvimento de Estratégias de Conformidade*: as empresas, ao compreenderem os tópicos e temas prevalentes nas decisões do CADE, podem ajustar suas práticas comerciais para garantir maior conformidade com as normativas antitruste. Esta compreensão, enraizada em uma análise objetiva e detalhada, permite identificar áreas de risco e oportunidades para alinhamento regulatório;

b) *Prevenção e Mitigação de Riscos*: a análise de tópicos frequentemente discutidos nas decisões do CADE permite que as empresas antecipem possíveis áreas de escrutínio e desenvolvam estratégias proativas para prevenir violações. Esta abordagem baseada em dados pode ser crucial na mitigação de riscos legais e na manutenção de práticas de negócios sustentáveis e éticas;

c) *Adaptação às Mudanças Regulatórias*: a capacidade de acompanhar as tendências e mudanças nas decisões do CADE auxilia as empresas a se adaptarem rapidamente às evoluções na legislação antitruste. Isso é particularmente relevante em um cenário econômico dinâmico, onde as práticas de mercado estão em constante evolução.

II - Implicações para Advogados

a) *Formulação de Defesas Jurídicas*: Advogados especializados em direito antitruste podem utilizar os *insights* obtidos pela modelagem de tópicos para construir argumentações e defesas mais robustas e fundamentadas. O conhecimento dos padrões nas decisões do CADE fornece uma base sólida para antecipar argumentos contrários e elaborar estratégias jurídicas mais eficazes.

b) *Abordagem Estratégica em Litígios*: a compreensão dos temas centrais e argumentos prevalentes nas decisões do CADE permite que advogados adotem uma abordagem

mais estratégica em litígios antitruste. Essa abordagem informada e orientada por dados pode aumentar significativamente as chances de um resultado favorável para seus clientes.

c) *Assessoria Jurídica Proativa*: além da representação em litígios, advogados podem oferecer assessoria proativa às empresas, orientando-as sobre como estruturar suas operações e transações de maneira a evitar infrações antitruste. A análise de dados do CADE pode ser um recurso valioso na identificação de práticas comerciais que potencialmente atrairiam a atenção do órgão regulador.

Em resumo, a aplicação de modelagem de tópicos e análise regressiva nas decisões do CADE tem implicações significativas tanto para as estratégias corporativas de conformidade quanto para a prática jurídica em casos antitruste. Esta abordagem analítica fornece uma compreensão mais profunda e matizada das tendências regulatórias, o que é crucial para o desenvolvimento de estratégias jurídicas e comerciais bem-sucedidas e alinhadas com as normativas antitruste.

4.5.3 Avanços em Direção a uma Abordagem Orientada por Dados na Organização Industrial

Javornik, Nadoh e Lange (2019, p. 296) esclarece que no mundo atual a transformação digital de produtos, serviços, empresas e indústrias inteiras é um processo sempre contínuo.

Por isso mesmo, a integração de técnicas de PLN e análises estatísticas avançadas na Organização Industrial sinaliza uma evolução significativa em direção a uma abordagem mais orientada por dados. Esta mudança representa um avanço notável na análise econômica e jurídica, contrastando com as metodologias tradicionais que tendem a ser mais subjetivas e qualitativas.

Historicamente, a Organização Industrial e o direito antitruste confiaram em interpretações qualitativas e análises baseadas em teorias e princípios estabelecidos, muitas vezes com limitações na objetividade e na capacidade de lidar com grandes volumes de dados. A adoção de PLN e técnicas estatísticas avançadas, como demonstrado nesta pesquisa, oferece uma alternativa notável, possibilitando uma abordagem sistemática e fundamentada em dados. Essas técnicas permitem a análise de extensos conjuntos de dados textuais, proporcionando uma visão mais abrangente e detalhada dos padrões e tendências subjacentes nas decisões econômicas e jurídicas.

A transição para uma análise orientada por dados traz maior objetividade para a tomada de decisões no campo da Organização Industrial. Com uma base quantitativa para análises, os formuladores de políticas e reguladores podem tomar decisões mais informadas e fundamentadas em evidências concretas e análises robustas. Isso é particularmente relevante no cenário atual, onde a complexidade e a quantidade de informações disponíveis estão em constante crescimento.

Além disso, a aplicação de PLN e análise estatística permite uma compreensão mais profunda e detalhada de questões complexas no direito e na economia. Questões que antes eram desafiadoras para serem abordadas devido à sua complexidade ou ao volume de informações

agora podem ser analisadas de maneira mais eficaz. Essa capacidade de processar e interpretar grandes quantidades de dados textuais abre novas possibilidades para a análise e a compreensão de temas econômicos e jurídicos.

A evolução para uma abordagem orientada por dados também favorece uma perspectiva mais multidisciplinar na análise econômica e jurídica. A combinação de *insights* obtidos através do PLN com teorias econômicas e princípios jurídicos enriquece a análise, oferecendo novas perspectivas e abordagens para questões tradicionais. Essa abordagem evolutiva está em consonância com o entendimento de que a dinâmica econômica mundial é caracterizada por ondas tecnológicas distintas, conhecidos por ciclos de Kondratieff[48] (Javornik; Nadoh; Lange, 2019, p. 295). A Organização Industrial, pois, encontra-se surfando a onda da Tecnologia de Informação – que no caso dessa pesquisa, traduz-se na obtenção de conhecimento útil a partir de dados textuais diversos –, dentro do quinto ciclo de Kondratieff, que teve início com a introdução da World Wide Web em 1993.

Em poucas palavras, a mudança para uma metodologia mais orientada por dados na Organização Industrial, facilitada pelo uso de técnicas de PLN e análise estatística, representa um avanço significativo na forma como as questões econômicas e jurídicas são abordadas e analisadas. Esta mudança não só permite uma análise mais objetiva e quantitativa, mas também proporciona uma compreensão mais rica e matizada de temas complexos, fundamentando decisões mais informadas e eficazes na esfera da política e regulação econômica.

4.5.4 Desafios e limitações

A aplicação de técnicas avançadas de PLN e análises estatísticas na Organização Industrial, embora traga uma série de benefícios, também enfrenta desafios e limitações significativas. Um dos principais desafios reside na interpretação dos dados. A capacidade de extrair significado preciso e relevante de grandes conjuntos de dados textuais não depende apenas da existência de ferramentas tecnológicas avançadas, mas também de uma compreensão profunda e contextualizada do ambiente legal e econômico no qual esses dados são situados.

A interpretação de dados em grande escala, especialmente em campos complexos como o direito e a economia, requer não só a habilidade técnica para processar informações, mas também a capacidade de discernir nuances e inferir significados que estão frequentemente entrelaçados com o jargão específico e as práticas institucionais. Este desafio é exacerbado pela

[48] Nikolai Kondratieff, economista russo que na década de 1920 analisou dados econômicos históricos disponíveis naquela época e identificou padrões que sugeriam uma natureza cíclica na economia. Ele observou intervalos regulares nos quais os indicadores econômicos apresentavam fases de aumentos graduais seguidos por períodos de declínio. Esses ciclos, conforme observados por Kondratieff, pareciam ter uma periodicidade de aproximadamente 50 anos. Ele documentou essas observações, detalhando as ondas longas específicas e suas fases. Pesquisas subsequentes sobre os ciclos de Kondratieff por outros estudiosos ampliaram a análise para incluir ondas longas adicionais identificadas durante e após o período pós-Primeira Guerra Mundial (Korotayev; Zinkina; Bogevolnov, 2011, p. 1280).

natureza muitas vezes ambígua e multifacetada da linguagem utilizada em documentos legais e econômicos, onde o contexto e a interpretação são cruciais para a compreensão adequada.

Além disso, há questões pertinentes relacionadas à acessibilidade e viabilidade da implementação dessas tecnologias avançadas. A realidade é que a adoção de PLN e análises estatísticas sofisticadas exige recursos computacionais substanciais e expertise técnica especializada. Esses requisitos podem ser um obstáculo significativo, especialmente para instituições com recursos limitados ou para pesquisadores que operam em ambientes com menos infraestrutura tecnológica. A necessidade de especialistas qualificados que possam gerenciar e interpretar esses sistemas complexos adiciona outra camada de desafio, tanto em termos de custo quanto de disponibilidade de mão-de-obra qualificada.

Portanto, enquanto as técnicas de PLN e análises estatísticas avançadas representam um avanço promissor na Organização Industrial e análise legal, é essencial abordar esses desafios para maximizar seu potencial e garantir sua aplicabilidade prática. A superação desses obstáculos exige não apenas o desenvolvimento contínuo de tecnologias mais acessíveis e fáceis de usar, mas também um investimento em educação e formação que permita aos profissionais das áreas jurídica e econômica aproveitar plenamente as oportunidades oferecidas por essas ferramentas analíticas.

Em suma, a aplicação de modelagem de tópicos e análise regressiva no contexto da Organização Industrial e análise antitruste oferece uma via promissora para enriquecer a compreensão e a eficácia das práticas regulatórias. As implicações desta abordagem são amplas, afetando a formulação de políticas, estratégias de defesa e conformidade, e contribuindo para uma abordagem mais orientada por dados na Organização Industrial. Contudo, é fundamental abordar os desafios associados a essas técnicas para maximizar seu potencial e garantir sua aplicabilidade prática no âmbito da análise antitruste.

5 CONCLUSÃO

Neste capítulo de conclusão da tese emerge uma síntese profunda das descobertas e do alcance dos objetivos propostos, refletindo a interseção entre a modelagem de tópicos e a análise quantitativa de textos jurídicos no contexto das decisões do CADE.

5.1 Resultados Alcançados

Inicialmente, a pesquisa iniciou com a composição e análise de um corpus substancial de documentos relacionados a cartéis em licitações, aplicando modelos de Processamento de Linguagem Natural (PLN), com destaque para o BERTopic. Durante a tokenização, etapa crucial no processamento de dados, observou-se uma redução significativa na dimensionalidade dos dados, restringindo o corpus a aproximadamente um quarto do volume original de palavras. Este processo permitiu identificar uma frequência elevada de termos específicos, ressaltando a relevância da comunicação por *email* como evidência em investigações de cartel.

Subsequentemente, a análise prosseguiu com uma avaliação detalhada dos modelos clássicos de modelagem de tópicos, LDA e NMF. Foi revelado um equilíbrio delicado entre a coerência temática e a eficiência computacional. Por outro lado, o modelo neural CTM, embora eficiente na captura da coerência temática, apresentou desafios relacionados à eficiência computacional.

Adicionalmente, a investigação também se concentrou nos modelos Top2Vec, nos quais se observaram variações notáveis em termos de coerência temática e eficiência computacional. Isso indicou a necessidade de uma seleção cuidadosa desses modelos, especialmente em grandes conjuntos de dados ou em contextos que demandam processamento rápido. Da mesma forma, a análise dos modelos de *embeddings*, particularmente usando o BERTopic, evidenciou variações na coerência e na eficiência computacional entre suas diferentes configurações.

De maneira complementar, a pesquisa realizada desvelou, ainda, aspectos cruciais sobre a aplicação das técnicas de PLN na modelagem de tópicos em textos legais. Salientou-se a vital importância do pré-processamento na determinação da qualidade dos resultados. Além disso, destacou-se a necessidade de um equilíbrio entre a sofisticação das técnicas de PLN e a disponibilidade de recursos computacionais.

Consequentemente, a investigação confirmou em grande parte a eficiência do modelo BERT, especificamente consumida no algoritmo BERTopic, na identificação de tópicos em textos legais antitruste. Os dados indicaram um equilíbrio entre coerência, diversidade de tópicos e eficiência temporal utilizando o BERTopic, ainda que seja necessária atenção ao aprimoramento na coerência dos tópicos e a uma configuração mais aprimorada.

Além disso, a pesquisa evidenciou variações nos resultados em razão da aplicação de técnicas diversas (algébricas, probabilísticas, híbridas e neurais) de modelagem de tópicos, destacando-se a inter-relação entre a sofisticação das técnicas e os recursos computacionais necessários.

No que tange à Hipótese de Previsibilidade das Decisões, esta foi investigada através da eficácia do modelo logístico na análise de tópicos em documentos do CADE, que focou em identificar modelo(s) mais eficiente(s) para prever decisões sobre condutas colusivas. Apesar de enfrentar desafios como a não convergência do modelo empírico logístico, principalmente devido à multicolinearidade, a análise confirmou a adequação do modelo para atingir os objetivos propostos, especialmente sobre a influência das variáveis profissiográficas nas decisões de condenação de cartelistas. A pesquisa ressalta a necessidade de cautela nas inferências, mas demonstra a adequabilidade do modelo logístico para revelar padrões consistentes no comportamento de *players* do CADE.

Ainda, em relação à influência de variáveis profissiográficas nas previsões, a análise explorou como a modelagem de tópicos em textos do MPF e relatórios do CADE poderia refletir padrões institucionais. A análise revelou que variáveis como formação em economia e direito impactam significativamente as previsões do modelo logístico, indicando sua influência nas decisões dos relatores do CADE. Apesar de desafios como a multicolinearidade ligeira da variável "gênero masculino"e a não convergência do modelo logístico, a pesquisa confirmou a importância dessas variáveis profissiográficas, destacando a necessidade de interpretação cuidadosa em contextos jurídicos e validando a hipótese com padrões observados no modelo.

Finalmente, a tese apresentada traz uma contribuição valiosa para a compreensão da modelagem de tópicos em contextos legais, com uma atenção especial voltada para os casos de antitruste e as decisões tomadas pelo CADE. Neste estudo, é dada ênfase à eficácia de diversas técnicas de PLN, incluindo BERTopic, NMF, LDA e Top2Vec. Um ponto central destacado é a importância do pré-processamento na determinação da qualidade dos resultados obtidos, ressaltando que os detalhes iniciais da análise podem ter um impacto significativo nos achados finais.

5.2 Contribuições Metodológicas e Teóricas

Inicialmente, a pesquisa destacada na tese coloca em evidência a importância crucial de uma escolha criteriosa na técnica de modelagem de tópicos. Esta seleção não é trivial, pois envolve a avaliação de fatores como a precisão temática, a diversidade de tópicos e a eficiência computacional. O estudo enfatiza que a escolha do modelo mais adequado para a modelagem de tópicos depende intrinsecamente dos requisitos específicos da análise em questão e das particularidades dos dados analisados. Esta abordagem cuidadosa assegura que os resultados obtidos sejam não apenas tecnicamente robustos, mas também relevantes e aplicáveis ao contexto

específico da pesquisa.

Ademais, a tese vai além e explora hipóteses específicas relacionadas ao campo da modelagem de tópicos. Uma das principais questões investigadas é a superioridade potencial da modelagem BERT, em particular através do uso do BERTopic, na identificação de tópicos em textos legais relacionados a questões antitruste. A pesquisa sugere que a capacidade do BERTopic de capturar e interpretar contextos complexos pode representar um avanço significativo na forma como os textos legais são processados e analisados. Esta evolução promete melhorar tanto a eficiência quanto a precisão das investigações antitruste, oferecendo novas ferramentas para enfrentar os desafios deste campo jurídico.

Além disso, a tese aborda outras hipóteses relevantes para a modelagem de tópicos. Isso inclui a Hipótese de Desempenho Diferencial, que considera a variação na eficácia de diferentes técnicas de modelagem de tópicos. Também examina a Hipótese de Complexidade Computacional, que explora a inter-relação entre a sofisticação das técnicas de modelagem e a intensidade dos recursos computacionais necessários para a sua implementação. Por fim, a pesquisa contempla a Hipótese de Correlação com Decisões Jurídicas, investigando a relação entre a precisão na identificação de tópicos e os padrões observáveis nas decisões jurídicas. Cada uma dessas hipóteses contribui para um entendimento mais profundo da modelagem de tópicos, reforçando a sua aplicabilidade e relevância em contextos jurídicos e de pesquisa.

5.3 Limitações da pesquisa

Contudo, embora esta pesquisa ofereça contribuições significativas para a compreensão do *corpus* de julgamentos de cartéis, ela não está isenta de limitações, comuns a estudos que empregam métodos complexos de PLN e modelagem de tópicos. Tais limitações residem principalmente na qualidade e representatividade dos dados utilizados, bem como nas características intrínsecas das técnicas de modelagem, que podem influenciar a precisão e a interpretabilidade dos tópicos gerados, consequentemente afetando a generalização dos resultados obtidos.

5.4 Recomendações para futuros trabalhos

Por conseguinte, o reconhecimento dessas limitações abre caminho para futuras investigações que visam aprimorar as técnicas de PLN e modelagem de tópicos, aumentando assim sua precisão e capacidade interpretativa. Há uma necessidade de explorar o uso de conjuntos de dados mais amplos e diversificados para fortalecer a generalização dos resultados. Paralelamente, estudos comparativos entre diferentes abordagens de modelagem de tópicos em contextos similares poderiam oferecer *insights* valiosos. Além disso, o estudo atual pode ser ampliado para incluir outras formas de análise textual, como a análise de sentimentos, e para explorar a modelagem de tópicos em diferentes contextos legais e de tomada de decisão. A aplicação de

modelos mais recentes de PLN, como os baseados em *Transformers*, apresenta-se como uma extensão promissora, capaz de proporcionar novas perspectivas e compreensões.

Em síntese, esta pesquisa ressalta o valor e a relevância da aplicação de métodos computacionais avançados em campos tradicionalmente não associados à tecnologia, como o Direito e a Economia. A integração dessas técnicas na análise de dados legais e econômicos é um passo significativo para uma tomada de decisão mais informada e para o avanço da pesquisa acadêmica. Por fim, a tese reflete o significado mais amplo desse estudo: a convergência progressiva entre tecnologia e ciências humanas, que está remodelando nossa abordagem e compreensão de questões complexas, como as decisões antitruste, sinalizando um futuro onde a inovação tecnológica se entrelaça com o aprofundamento do conhecimento humano.

REFERÊNCIAS

ABDELRAZEK, Aly; EID, Yomna; GAWISH, Eman; MEDHAT, Walaa; HASSAN, Ahmed. Topic Modeling Algorithms and Applications: A Survey. *Information Systems*, v. 112, fev. 2023. ISSN 03064379. DOI: 10.1016/j.is.2022.102131. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0306437922001090. Citado 1 vez na página 21.

ADHIKARY, Jyoti Ranjan; MURTY, M. Narasimha. Feature Selection for Unsupervised Learning. *null*, v. 7665, p. 382–389, 2012. DOI: 10.1007/978-3-642-34487-9_47. pmid: null. Disponível em: http://link.springer.com/10.1007/978-3-642-34487-9_47. Citado 1 vez na página 26.

AGERRI, Rodrigo; ARTOLA, Xabier; BELOKI, Zuhaitz; RIGAU, German; SOROA, Aitor. Big Data for Natural Language Processing: A Streaming Approach. *Knowledge-Based Systems*, v. 79, p. 36–42, maio 2015. ISSN 09507051. DOI: 10.1016/j.knosys.2014.11.007. Disponível em: https://www.sciencedirect.com/science/article/pii/S0950705114003992. Acesso em: 3 out. 2022. Citado 1 vez na página 9.

AGUIAR, Júlio Cesar; DAHER, Lenna; TABAK, Benjamin Miranda. Cartéis Em Licitações Públicas Sob o Enfoque Da Análise Econômica Do Direito. Os Incentivos Legais à Livre Concorrência São Suficientes Para Tornar o Custo de Um Cartel Superior Ao Seu Benefício? en-US. *Revista de Direito Econômico e Socioambiental*, v. 9, n. 3, p. 185, 2018. ISSN 2179-345X. DOI: 10.7213/rev.dir.econ.soc.v9i3.23437. Disponível em: https://periodicos.pucpr.br/index.php/direitoeconomico/article/view/23437. Citado 10 vezes nas páginas 42, 48, 50, 51.

AGUILLAR, Fernando Herren. *Direito econômico: do direito nacional ao direito supranacional*. São Paulo: Atlas, 2006. p. 407. Citado 1 vez na página 39.

AIT STAFF WRITER. *Transforming Legal Landscape: How AI Is Becoming the Ultimate Sidekick for Lawyers*. 5 set. 2023. Disponível em: https://aithority.com/ai-machine-learning-projects/transforming-legal-landscape-how-ai-is-becoming-the-ultimate-sidekick-for-lawyers/. Acesso em: 3 nov. 2023. Citado 2 vezes nas páginas 11, 12.

AKAIKE, Hirotugu. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado 1 vez na página 82.

ALETRAS, Nikolaos; TSARAPATSANIS, Dimitrios; PREOŢIUC-PIETRO, Daniel; LAMPOS, Vasileios. Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective. *PeerJ Computer Science*, v. 2, e93, 24 out. 2016. ISSN 2376-5992. DOI: 10.7717/peerj-cs.93. Disponível em: https://peerj.com/articles/cs-93. Citado 1 vez na página 12.

ANGELOV, Dimo. Top2Vec: Distributed Representations of Topics. *arXiv.org*, 2020. DOI: null. pmid: null. Citado 4 vezes nas páginas 34, 35, 75.

ARAÚJO JR., José Tavares De. Schumpeterian Competition and Its Policy Implications: The Latin American Case. *Brazilian Journal of Political Economy*, v. 19, n. 4, p. 569–580, out. 1999. ISSN 1809-4538, 0101-3157. DOI: 10.1590/0101-31571999-1025. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-31571999000400569&lang=en. Citado 4 vezes na página 58.

ARNOLD, Taylor; TILTON, Lauren. *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text*. Cham: Springer International Publishing, 2015. (Quantitative Methods in the Humanities and Social Sciences). ISBN 978-3-319-20701-8. DOI: 10.1007/978-3-319-20702-5. Disponível em: https://link.springer.com/10.1007/978-3-319-20702-5. Acesso em: 14 nov. 2023. Citado 1 vez na página 19.

BARTHOLOMAY, Eduardo Luís. *A Voz Do Povo é a Voz de Deus? A Influência Das Manifestações Na Comunicação Política*. 2021. 101 f. Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul, Porto Alegre. Citado 1 vez na página 69.

BENOIT, Ken. Text as Data: An Overview. *In:* CURINI, Luigi; FRANZESE, Robert. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd, 2020. p. 461–497. ISBN 978-1-5264-5993-0. DOI: 10.4135/9781526486387.n29. Disponível em: https://sk.sagepub.com/reference/the-sage-handbook-of-research-methods-in-political-science-and-ir/i4365.xml. Acesso em: 14 nov. 2023. Citado 3 vezes nas páginas 13, 14, 108.

BIANCHI, Federico; TERRAGNI, Silvia; HOVY, Dirk. *Pre-Training Is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence*. en-US. 17 jun. 2021. DOI: 10.48550/arXiv.2004.03974. arXiv: 2004.03974 [cs]. Disponível em: http://arxiv.org/abs/2004.03974. Acesso em: 22 nov. 2023. preprint. Citado 2 vezes nas páginas 74, 75.

BIANCHI, Federico; TERRAGNI, Silvia; HOVY, Dirk; NOZZA, Debora; FERSINI, Elisabetta. Cross-Lingual Contextualized Topic Models with Zero-Shot Learning. en-US. *In:* EACL 2021. Edição: Paola Merlo, Jorg Tiedemann e Reut Tsarfaty, p. 1676–1683. DOI: 10.18653/v1/2021.eacl-main.143. Disponível em: https://aclanthology.org/2021.eacl-main.143. Acesso em: 22 nov. 2023. Citado 1 vez na página 74.

BLEI, David M. Introduction to Probabilistic Topic Models. *P Communications of the ACM*, 2011. Citado 3 vezes nas páginas 27, 28.

BLEI, David M.; LAFFERTY, John D. Dynamic Topic Models. *In:* (ICML '06), p. 113–120. ISBN 978-1-59593-383-6. DOI: 10.1145/1143844.1143859. Disponível em: https://doi.org/10.1145/1143844.1143859. Acesso em: 4 nov. 2023. Citado 3 vezes na página 28.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2003. DOI: 10.5555/944919.944937. pmid: null. Disponível em: https://dl.acm.org/doi/10.5555/944919.944937. Citado 7 vezes nas páginas 27, 28, 72–75.

BOUMA, Gerlof. Normalized (Pointwise) Mutual Information in Collocation Extraction. en. *From Form to Meaning: Processing Texts Automatically. Proceedings of the the biennal GSCL Conference*, p. 31–40, 2009. Citado 1 vez na página 78.

BRADLEY, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, Elsevier, v. 30, n. 7, p. 1145–1159, 1997. Citado 1 vez na página 81.

BRAGA, Tereza Cristine Almeida. Cade, Cartéis e Licitações: Um Novo Nicho Da Política Antitruste Brasileira. en-US. *Revista de Defesa da Concorrência*, v. 3, n. 1, p. 108–132, 2015. Citado 1 vez na página 47.

BRASIL. Lei n. 12.529, de 30 de Novembro de 2011. Estrutura o Sistema Brasileiro de Defesa Da Concorrência; Dispõe Sobre a Prevenção e Repressão Às Infrações Contra a Ordem Econômica; Altera a Lei Nº 8.137, de 27 de Dezembro de 1990, o Decreto-Lei Nº 3.689, de 3 de Outubro de 1941 - Código de Processo Penal, e a Lei Nº 7.347, de 24 de Julho de 1985; Revoga Dispositivos Da Lei Nº 8.884, de 11 de Junho de 1994, e a Lei Nº 9.781, de 19 de Janeiro de 1999; e Dá Outras Providências. *Diário Oficial [da] República Federativa do Brasil*, Brasília, 2011. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12529.htm. Acesso em: 6 ago. 2021. Citado 3 vezes nas páginas 55, 56.

BRASIL. Lei n. 13.709, de 14 de Agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). en-US, 14 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 15 nov. 2023. Citado 1 vez na página 59.

BRASIL, Ministério Público Federal. *Combate a Cartéis*. en-US. 2019. Disponível em: https://www.mpf.mp.br/atuacao-tematica/ccr3/documentos-e-publicacoes/roteiros-de-atuacao/combate-a-carteis. Acesso em: 31 out. 2022. Citado 1 vez na página 89.

BRINK, Henrik; RICHARDS, Joseph W.; FETHEROLF, Mark. *Real-World Machine Learning*. Shelter Island-NY: Manning Publications Co., 2017. 242 p. ISBN 978-1-61729-192-0. Citado 1 vez na página 8.

CADE. *Anuário*. Brasília: Conselho Administrativo de Defesa Econômica, 2019a. Disponível em: https://www.gov.br/cade/pt-br/centrais-de-conteudo/publicacoes/anuarios-do-cade. Acesso em: 5 ago. 2021. Citado 1 vez na página 47.

CADE. *Cade em números*. pt-BR. Brasília, 2021. Disponível em: https://cadenumeros.cade.gov.br/QvAJAXZfc/opendoc.htm?document=Painel%2FCADE%20em%20N%C3%BAmeros.qvw&host=QVS%40srv004q6774&anonymous=true. Acesso em: 6 ago. 2021. Citado 3 vezes nas páginas 48, 65.

CADE. *Cartilha Do Cade*. Brasília: Conselho Administrativo de Defesa Econômica, 2016. Disponível em: https://cdn.cade.gov.br/Portal/acesso-a-informacao/perguntas-frequentes/cartilha-do-cade.pdf. Acesso em: 5 ago. 2021. Citado 3 vezes nas páginas 47, 55, 57.

CADE. *Combate a Cartéis e Programa de Leniência*. Brasília: Conselho Administrativo de Defesa Econômica, 2009. Disponível em: https://cdn.cade.gov.br/Portal/centrais-de-conteudo/publicacoes/guias-do-cade/guia-para-analise-de-atos-de-concentracao-horizontal.pdf. Acesso em: 5 ago. 2021. Citado 1 vez na página 47.

CADE. *Nota Técnica n. 29. Departamento de Estudos Econômicos*. en-US. Brasília, 2019b. p. 33. Citado 1 vez na página 51.

CADE. *Serviço Eletrônico de Informações (SEI/CADE)*. en-US. Brasília, 2023. Disponível em: https://sip.cade.gov.br/sip/login.php?sigla_orgao_sistema=CADE&sigla_sistema=SEI&infra_url=L3NlaS8=. Acesso em: 6 ago. 2023. Citado 2 vezes na página 65.

CADE, Conselho Administrativo De Defesa Econômica. *Guia de Combate a Cartéis Em Licitação*. Brasília: CADE, 2019c. 54 p. Disponível em: https://cdn.cade.gov.br/Portal/centrais-de-conteudo/publicacoes/guias-do-cade/guia-de-combate-a-carteis-em-licitacao-versao-final-1.pdf. Acesso em: 3 out. 2022. Citado 1 vez na página 50.

CAMPELLO, Ricardo J. G. B.; MOULAVI, Davoud; SANDER, Joerg. Density-Based Clustering Based on Hierarchical Density Estimates. en-US. *In:* PEI, Jian; TSENG, Vincent S.; CAO, Longbing; MOTODA, Hiroshi; XU, Guandong (ed.). v. 7819. (Lecture Notes in Computer Science), p. 160–172. ISBN 978-3-642-37456-2. DOI: 10.1007/978-3-642-37456-2_14. Disponível em: http://link.springer.com/10.1007/978-3-642-37456-2_14. Citado 2 vezes nas páginas 75, 76.

CARDOSO, Oscar Valente. *A Nova Estrutura Do CADE No Sistema Brasileiro de Defesa Da Concorrência Da Lei Nº 12.529/2011*. en-US. 2012. Disponível em: https://jus.com.br/artigos/22026/a-nova-estrutura-do-cade-no-sistema-brasileiro-de-defesa-da-concorrencia-da-lei-n-12-529-2011. Acesso em: 6 ago. 2021. Citado 1 vez na página 55.

CARVALHO, Erick Leonardo Freire. A Política Antitruste No Brasil e o Combate a Cartéis à Luz Do Novo Cade. en-US. *Revista da Faculdade de Direito da UERJ - RFD*, v. 0, n. 28, p. 97–117, 28 23 dez. 2015a. ISSN 2236-3475. DOI: 10.12957/rfd.2015.5252. Disponível

em: https://www.e-publicacoes.uerj.br/rfduerj/article/view/5252. Acesso em: 10 nov. 2023. Citado 4 vezes nas páginas 47, 50, 57.

CARVALHO, Erick Leonardo Freire. A política antitruste no Brasil e o combate a cartéis à luz do novo Cade. *Revista da Faculdade de Direito-UERJ*, Rio de Janeiro, n. 28, p. 97–117, 2015b. Citado 5 vezes na página 40.

CASTRO, Bruno Braz de. Preços exploratórios: por uma nova teoria da decisão. *Revista do Ibrac*, v. 23, n. 1, p. 11–69, 2017. Citado 1 vez na página 43.

CENTRO DE INFORMÁTICA DA UFPE. *Conceitos Básicos de Teoria Da Computação*. en-US. 2023. Disponível em: https://www.cin.ufpe.br/~gcb/tc/TC_CONCEITOS%20BASICOS.pdf. Acesso em: 5 nov. 2023. Citado 1 vez na página 32.

CHEN, Weisi; RABHI, Fethi; LIAO, Wenqi; AL-QUDAH, Islam. Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study. *Electronics*, Multidisciplinary Digital Publishing Institute, v. 12, n. 12, p. 2605, 12 jan. 2023. ISSN 2079-9292. DOI: 10.3390/electronics12122605. Disponível em: https://www.mdpi.com/2079-9292/12/12/2605. Acesso em: 4 nov. 2023. Citado 1 vez nas páginas 22, 23.

CHOLLET, François; ALLAIRE, Joseph J. *Deep Learning with R*. Shelter Island, NY: Manning Publications Co, 2018. 335 p. ISBN 978-1-61729-554-6. Citado 0 vez na página 9.

CHUANG, Jason; RAMAGE, Daniel; MANNING, Christopher; HEER, Jeffrey. Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis. en-US. *In:* (CHI '12), p. 443–452. ISBN 978-1-4503-1015-4. DOI: 10.1145/2207676.2207738. Disponível em: https://doi.org/10.1145/2207676.2207738. Acesso em: 28 nov. 2023. Citado 1 vez na página 80.

COHEN, Jacob. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960. Citado 1 vez na página 82.

COLACINO, Lucas D'Angelo. *Cartel Em Concorrências Públicas e Corrupção: Uma Abordagem Econômica*. 2016. Mestrado em economia – Universidade Federal do Rio de Janeiro. Citado 1 vez na página 51.

CONNOR, John M. Price-fixing overcharges: revised 3rd edition. *Available at SSRN 2400780*, 2014. Disponível em: https://www.researchgate.net/publication/272302307_Price-Fixing_Overcharges_Revised_3rd_Edition. Acesso em: 6 ago. 2021. Citado 2 vezes nas páginas 48, 51.

CONSELHO NACIONAL DO MINISTÉRIO PÚBLICO. *Glossário Institucional*. 2021. Disponível em: https://www.cnmp.mp.br/portal/institucional/476-glossario/7930-mens-legis. Acesso em: 30 nov. 2021. Citado 1 vez na página 41.

CRAMER, Meg; HAYES, Gillian R. The Digital Economy: A Case Study of Designing for Classrooms. en-US. *In:* IDC '13: INTERACTION DESIGN AND CHILDREN 2013, p. 431–434. ISBN 978-1-4503-1918-8. DOI: 10.1145/2485760.2485832. Disponível em: https://dl.acm.org/doi/10.1145/2485760.2485832. Acesso em: 17 nov. 2023. Citado 2 vezes na página 10.

DAVIS, Jesse; GOADRICH, Mark. The relationship between Precision-Recall and ROC curves. *In:* PROCEEDINGS of the 23rd international conference on Machine learning. 2006. p. 233–240. Citado 1 vez na página 82.

DEERWESTER, Scott; DUMAIS, Susan T.; FURNAS, George W.; LANDAUER, Thomas K.; HARSHMAN, Richard. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, v. 41, n. 6, p. 391–407, set. 1990. ISSN 0002-8231, 1097-4571. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9. Acesso em: 4 nov. 2023. Citado 3 vezes nas páginas 24, 25.

DESAGULIER, Guillaume. *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Cham: Springer International Publishing, 2017. (Quantitative Methods in the Humanities and Social Sciences). ISBN 978-3-319-64570-4. DOI: 10.1007/978-3-319-64572-8. Disponível em: http://link.springer.com/10.1007/978-3-319-64572-8. Acesso em: 14 nov. 2023. Citado 1 vez na página 13.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. 24 maio 2019. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805 [cs]. Disponível em: http://arxiv.org/abs/1810.04805. Acesso em: 25 out. 2023. Citado 3 vezes nas páginas 33, 34.

DIENG, Adji B.; RUIZ, Francisco J. R.; BLEI, David M. Topic Modeling in Embedding Spaces. en-US. *Transactions of the Association for Computational Linguistics*, v. 8, p. 439–453, 1 jul. 2020. ISSN 2307-387X. DOI: 10.1162/tacl_a_00325. Disponível em: https://doi.org/10.1162/tacl_a_00325. Acesso em: 28 nov. 2023. Citado 2 vezes na página 80.

DIENG, Adji B.; WANG, Chong; GAO, Jianfeng; PAISLEY, John. *TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency*. 26 fev. 2017. DOI: 10.48550/arXiv.1611.01702. arXiv: 1611.01702 [cs, stat]. Disponível em: http://arxiv.org/abs/1611.01702. Acesso em: 4 nov. 2023. preprint. Citado 3 vezes na página 33.

EGGER, Roman; YU, Joanne. A Topic Modeling Comparison between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, v. 7, p. 1–16, 2022. ISSN 2297-7775. Disponível em: https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498. Acesso em: 25 out. 2023. Citado 2 vezes nas páginas 71, 92, 100.

FARACO, Fernando Melo. *Modelo de Conhecimento Baseado Em Tópicos de Acórdãos Para Suporte à Análise de Petições Iniciais*. 2020. 131 f. Dissertação (Mestrado em Engenharia e Gestão do Conhecimento) – Universidade Federal de Santa Catarina, Florianopólis. Citado 1 vez na página 19.

FAWCETT, Tom. An introduction to ROC analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado 1 vez na página 81.

FERREIRA, Bráulio Cavalcanti. *A Repressão Dos Cartéis Pelo Sistema Brasileiro de Defesa Da Concorrência (SBDC) e Os Impactos Dos Conluios Em Licitações Na Administração Pública Brasileira: Uma Análise Do Processo Administrativo Nº 08012.001826/2003-10/CADE, Envolvendo as Empresas Prestadoras de Serviços de Segurança Privada No Rio Grande Do Sul*. 2013. Monografia (Trabalho de conclusão de curso) – Florianopólis. Citado 1 vez na página 54.

FERREIRA, Marcelo Herton Pereira. *Classificação de Peças Processuais Jurídicas: Inteligência Artificial No Direito*. 2018. 78 f. Monografia (Graduação em Engenharia de Software) – Universidade de Brasília, Brasília. Disponível em: https://bdm.unb.br/handle/10483/21570. Acesso em: 2 out. 2022. Citado 1 vez na página 8.

FIGUEIREDO, Leonardo Vizeu. *Lições de Direito Econômico*. 7. ed. Río de Janeiro: Forense, 2014. Citado 1 vez na página 53.

GENTZKOW, Matthew; KELLY, Bryan; TADDY, Matt. Text as Data. *Journal of Economic Literature*, v. 57, n. 3, p. 535–574, 1 set. 2019. ISSN 0022-0515. DOI: 10.1257/jel.20181020. Disponível em: https://pubs.aeaweb.org/doi/10.1257/jel.20181020. Acesso em: 6 out. 2022. Citado 6 vezes nas páginas 1, 6, 12, 13.

GOLDBERG, Yoav. *In:* NEURAL Network Methods for Natural Language Processing. Toronto, Canada: University of Toronto, 2017. (Synthesis Lectures on Human Language Technologies). ISBN 978-3-031-02165-7. DOI: 10.1007/978-3-031-02165-7. Disponível em: https://link.springer.com/book/10.1007/978-3-031-02165-7. Acesso em: 4 out. 2022. Citado 1 vez na página 8.

GRIMMER, Justin; STEWART, Brandon M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, Cambridge University Press, v. 21, n. 3, p. 267–297, 2013. ISSN 1047-1987, 1476-4989. DOI: 10.1093/pan/mps028. Disponível em: https://www.cambridge.org/core/journals/political-analysis/article/text-as-data-the-promise-and-pitfalls-of-automatic-content-analysis-methods-for-political-texts/F7AAC8B2909441603FEB25C156448F20. Acesso em: 11 out. 2022. Citado 1 vez na página 13.

GROOTENDORST, Maarten. *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure*. 2022. DOI: 10.48550/ARXIV.2203.05794. arXiv: 2203.05794 [cs]. Disponível em: http://arxiv.org/abs/2203.05794. Acesso em: 28 abr. 2023. preprint. Citado 11 vezes nas páginas 35, 36, 75–77, 81, 110.

HOFMANN, Thomas. Probabilistic Latent Semantic Indexing. *In:* (SIGIR '99), p. 50–57. ISBN 978-1-58113-096-6. DOI: 10.1145/312624.312649. Disponível em: https://dl.acm.org/doi/10.1145/312624.312649. Acesso em: 4 nov. 2023. Citado 3 vezes nas páginas 26, 27.

HOFMANN, Thomas. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, v. 42, n. 1, p. 177–196, 1 jan. 2001. ISSN 1573-0565. DOI: 10.1023/A:1007617005950. Disponível em: https://doi.org/10.1023/A:1007617005950. Acesso em: 5 nov. 2023. Citado 3 vezes nas páginas 26, 27.

HSU, Bi-Min. Comparison of Supervised Classification Models on Textual Data. *Mathematics*, v. 8, n. 5, p. 851, 24 maio 2020. ISSN 2227-7390. DOI: 10.3390/math8050851. Disponível em: https://www.mdpi.com/2227-7390/8/5/851. Citado 1 vez na página 11.

HU, Han; WEN, Yonggang; CHUA, Tat-Seng; LI, Xuelong. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*, IEEE, v. 2, p. 652–687, 2014. ISSN 2169-3536. DOI: 10.1109/ACCESS.2014.2332453. Disponível em: https://ieeexplore.ieee.org/document/6842585/. Citado 2 vezes na página 10.

IZBICKI, Rafael; SANTOS, Tiago Mendonça dos. *Aprendizado de Máquina: Uma Abordagem Estatística*. São Carlos, SP, 2020. 268 p. Citado 2 vezes nas páginas 20, 103.

JAPKOWICZ, Nathalie; SHAH, Mohak. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011. Citado 1 vez na página 81.

JAPKOWICZ, Nathalie; STEFANOWSKI, Jerzy (ed.). *Big Data Analysis: New Algorithms for a New Society*. Cham: Springer International Publishing, 2016. v. 16. (Studies in Big Data). ISBN 978-3-319-26987-0. DOI: 10.1007/978-3-319-26989-4. Disponível em: http://link.springer.com/10.1007/978-3-319-26989-4. Acesso em: 17 nov. 2023. Citado 2 vezes nas páginas 10, 11.

JAVORNIK, Marko; NADOH, Nives; LANGE, Dustin. Data Is the New Oil. *In:* redigido por Beate Müller e Gereon Meyer. Cham: Springer, 2019. (Lecture Notes in Mobility). p. 295–308. ISBN 978-3-319-99755-1. DOI: https://link.springer.com/chapter/10.1007/978-3-319-99756-8_19. Disponível em: http://link.springer.com/10.1007/978-3-319-99756-8_19. Citado 5 vezes nas páginas 10, 12, 117, 118.

JOCKERS, Matthew L.; THALKEN, Rosamond. *Text Analysis with R: For Students of Literature*. Cham: Springer International Publishing, 2020. 283 p. (Quantitative Methods in the Humanities and Social Sciences). ISBN 978-3-030-39642-8. DOI: 10.1007/978-3-030-39643-5. Disponível em: http://link.springer.com/10.1007/978-3-030-39643-5. Acesso em: 13 nov. 2023. Citado 2 vezes nas páginas 9, 14.

JOULIN, Armand; GRAVE, Edouard; BOJANOWSKI, Piotr; DOUZE, Matthijs *et al. FastText.Zip: Compressing Text Classification Models*. 12 dez. 2016. DOI: 10.48550/arXiv.

1612.03651. arXiv: 1612.03651 [cs]. Disponível em: http://arxiv.org/abs/1612.03651. Acesso em: 4 nov. 2023. preprint. Citado 3 vezes nas páginas 32, 33.

JOULIN, Armand; GRAVE, Edouard; BOJANOWSKI, Piotr; MIKOLOV, Tomas. *Bag of Tricks for Efficient Text Classification*. 9 ago. 2016. DOI: 10.48550/arXiv.1607.01759. arXiv: 1607.01759 [cs]. Disponível em: http://arxiv.org/abs/1607.01759. Acesso em: 5 nov. 2023. preprint. Citado 4 vezes na página 32.

KOROTAYEV, Andrey; ZINKINA, Julia; BOGEVOLNOV, Justislav. Kondratieff waves in global invention activity (1900–2008). lv. *Technological Forecasting and Social Change*, v. 78, n. 7, p. 1280–1284, set. 2011. ISSN 00401625. DOI: 10.1016/j.techfore.2011.02.011. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0040162511000503. Acesso em: 8 fev. 2024. Citado 1 vez na página 118.

KRISHNAN, Anusuya. Exploring the Power of Topic Modeling Techniques in Analyzing Customer Reviews: A Comparative Analysis. *arXiv.org*, 2023. DOI: 10.48550/arxiv.2308.11520. pmid: null. Disponível em: https://arxiv.org/abs/2308.11520. Citado 6 vezes nas páginas 21, 26, 27, 111.

LAU, Jey Han; NEWMAN, David; BALDWIN, Timothy. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. en-US. *In:* EACL 2014. Edição: Shuly Wintner, Sharon Goldwater e Stefan Riezler, p. 530–539. DOI: 10.3115/v1/E14-1056. Disponível em: https://aclanthology.org/E14-1056. Acesso em: 12 dez. 2023. Citado 1 vez na página 77.

LE, Quoc V.; MIKOLOV, Tomas. Distributed Representations of Sentences and Documents. *In:* INTERNATIONAL CONFERENCE ON MACHINE LEARNING. *Proceedings of the 31 St International Conference on Machine Learning*. Beijing, China: JMLR, 2014. DOI: 10.48550/ARXIV.1405.4053. Disponível em: https://proceedings.mlr.press/v32/le14.html. Acesso em: 4 nov. 2023. Citado 3 vezes nas páginas 30, 31.

LEAL, Carlos Ivan Simonsen; FIGUEIREDO, Paulo N. Inovação e tecnologia no Brasil: desafios e insumos para o desenvolvimento de políticas públicas. *Technological Learning and Industrial Innovation Working Paper Series*, n. 1, p. 1–32, 2018. Citado 15 vezes nas páginas 37, 38.

LEE, Daniel D.; LEE, Daniel D.; LEE, Daniel D.; SEUNG, H. Sebastian. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 1999. DOI: 10.1038/44565. pmid: 10548103. Citado 3 vezes na página 26.

LEITE, Daniel Saraiva. *Um Estudo Comparativo de Modelos Baseados Em Estatísticas Textuais, Grafos e Aprendizado de Máquina Para Sumarização Automática de Textos Em Português*. 2010. 213 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de São Carlos, São Carlos, SP. Disponível em: https://repositorio.ufscar.br/handle/ufscar/459. Acesso em: 30 set. 2022. Citado 1 vez na página 15.

LIMA, Tatiana de Macedo Nogueira. *Aprendizado de Máquina e Antitruste*. Brasília, DF, jul. 2022. Disponível em: https://cdn.cade.gov.br/Portal/centrais-de-conteudo/publicacoes/estudos-economicos/documentos-de-trabalho/2022/DOC_003-2022_Aprendizado-de-maquina-e-antitruste.pdf. Acesso em: 15 mar. 2023. Citado 6 vezes nas páginas 53, 59–61.

LOI, Michele; DEHAYE, Paul Olivier. If Data Is the New Oil, When Is the Extraction of Value from Data Unjust? Zurich, 2017. DOI: 10.5167/UZH-159945. Disponível em: http://fqp.luiss.it/files/2018/10/PPI_06_Loi-Dehaye_vol7_n2_2017def.pdf. Acesso em: 20 set. 2022. Citado 2 vezes nas páginas 10, 12.

LUHN, H. P. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, v. 2, n. 2, p. 159–165, abr. 1958. ISSN 0018-8646, 0018-8646. DOI: 10.1147/rd.22.0159. Disponível em: https://ieeexplore.ieee.org/abstract/document/5392672/authors#authors. Acesso em: 29 set. 2022. Citado 1 vez na página 15.

MAGALHÃES JÚNIOR, Danilo Brum De. *Arbitragem e Direito Concorrencial: A Arbitragem Como Método Para a Resolução de Disputas Privadas Que Envolvam Matéria Concorrencial No Direito Brasileiro*. 2018. Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos. Citado 4 vezes nas páginas 55, 56.

MARANHÃO, Juliano Souza De Albuquerque; FLORÊNCIO, Juliana Abrusio; ALMADA, Marco. Inteligência Artificial Aplicada Ao Direito e o Direito Da Inteligência Artificial. *Suprema - Revista de Estudos Constitucionais*, v. 1, n. 1, p. 154–180, 30 jun. 2021. ISSN 2763-7867, 2763-8839. DOI: 10.53798/suprema.2021.v1.n1.a20. Disponível em: https://suprema.stf.jus.br/index.php/suprema/article/view/20. Acesso em: 25 mar. 2023. Citado 3 vezes nas páginas 11, 12.

MARTINEZ, Ana Paula. *Aplicação do Direito da Concorrência a Licitações Públicas: Cartéis*. 2014. Disponível em: https://www.gov.br/fazenda/pt-br/centrais-de-conteudos/publicacoes/apostilas/advocacia-da-concorrencia/2-seae_aplicacao_direito_concorrencia_licitacoes_publicas_carteis.pdf. Acesso em: 6 ago. 2021. Citado 6 vezes nas páginas 41, 42, 47, 50, 51.

MCFADDEN, Daniel *et al.* Conditional logit analysis of qualitative choice behavior. Institute of Urban e Regional Development, University of California ..., 1973. Citado 1 vez na página 82.

MCINNES, Leland; HEALY, John; MELVILLE, James. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. en-US. 17 set. 2020. DOI: 10.48550/arXiv.1802.03426. arXiv: 1802.03426 [cs, stat]. Disponível em: http://arxiv.org/abs/1802.03426. Acesso em: 10 dez. 2023. preprint. Citado 2 vezes nas páginas 75, 76.

MEDVEDEVA, Masha; VOLS, Michel; WIELING, Martijn. Using Machine Learning to Predict Decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, v. 28, n. 2, p. 237–266, jun. 2020. ISSN 0924-8463, 1572-8382. DOI: 10.1007/s10506-019-09255-

y. Disponível em: https://doi.org/10.1007/s10506-019-09255-y. Acesso em: 14 dez. 2022. Citado 2 vezes na página 110.

MIKOLOV, Tomas; CHEN, Kai; CHEN, Kai; CORRADO, Greg S.; DEAN, Jeffrey. Efficient Estimation of Word Representations in Vector Space. *arXiv: Computation and Language*, 2013. DOI: 10.48550/arxiv.1301.3781. pmid: null. Citado 3 vezes nas páginas 23, 30.

MIKOLOV, Tomas; SUTSKEVER, Ilya; CHEN, Kai *et al.* Distributed Representations of Words and Phrases and Their Compositionality. *arXiv: Computation and Language*, 2013. pmid: null. Citado 1 vez na página 32.

MIMNO, David; WALLACH, Hanna; TALLEY, Edmund; LEENDERS, Miriam; MCCALLUM, Andrew. Optimizing Semantic Coherence in Topic Models. en-US. *In:* CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK: Association for Computational Linguistics, 27 jul. 2011. p. 262–272. Citado 1 vez na página 79.

MONTEIRO, Carmem Diva. *Sobre a Política Antitruste No Brasil e Seus Aspectos Críticos*. en-US. 2002. Disponível em: https://web.archive.org/web/20120205230339/http://www.seae.fazenda.gov.br/central_documentos/documento_trabalho/2002-1/doctrab27.pdf. Acesso em: 30 nov. 2021. Citado 2 vezes na página 56.

MONTEIRO, Gabriela Reis Paiva. *Big Data e Concorrência : Uma Avaliação Dos Impactos Da Exploração de Big Data Para o Método Antitruste Tradicional de Análise de Concentrações Econômicas /*. 2017. 152 f. Dissertação (Mestrado em Direito) – Fundação Getúlio Vargas, Rio de Janeiro. Citado 3 vezes nas páginas 10, 11.

MONTELLA, Maura. *Micro e Macroeconomia: Uma abordagem conceitual e prática*. Atlas, 2012. Citado 3 vezes nas páginas 43–45.

NEWMAN, David; LAU, Jey Han; GRIESER, Karl; BALDWIN, Timothy. Automatic Evaluation of Topic Coherence. en-US. *In:* NAACL-HLT 2010. Edição: Ron Kaplan, Jill Burstein, Mary Harper e Gerald Penn, p. 100–108. Disponível em: https://aclanthology.org/N10-1012. Acesso em: 28 nov. 2023. Citado 1 vez na página 79.

NITTA, Katsumi; SATOH, Ken. AI Applications to the Law Domain in Japan. *Asian Journal of Law and Society*, Cambridge University Press, v. 7, n. 3, p. 471–494, out. 2020. ISSN 2052-9015, 2052-9023. DOI: 10.1017/als.2020.35. Disponível em: https://www.cambridge.org/core/journals/asian-journal-of-law-and-society/article/ai-applications-to-the-law-domain-in-japan/B0E405E2EBC9BED36D9D38E736B4A099. Acesso em: 8 dez. 2022. Citado 1 vez na página 11.

O'CALLAGHAN, Derek; GREENE, Derek; CARTHY, Joe; CUNNINGHAM, Pádraig. An Analysis of the Coherence of Descriptors in Topic Modeling. *Expert Systems with Applications*, v. 42, n. 13, p. 5645–5657, 2015. ISSN 09574174. DOI: 10.1016/j.eswa.2015.02.055.

pmid: null. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0957417415001633. Citado 1 vez na página 26.

OCDE. *Report on the Nature and Impact of Hard Core Cartels and Sanctions against Cartels under National Competition Laws*. Organization for Economic Co-operation e Development Paris, 2002. Disponível em: https://www.oecd.org/competition/cartels/1841891.pdf. Acesso em: 6 ago. 2021. Citado 2 vezes nas páginas 48, 51.

OECD. *Combate a Cartéis Em Licitações No Brasil: Uma Revisão Das Compras Públicas Federais*. Paris, France, 2021. p. 120. Disponível em: https://www.oecd.org/competition/fighting-bid-rigging-in-brazil-a-review-of-federal-public-procurement-pt.htm. Acesso em: 5 out. 2022. Citado 7 vezes nas páginas 50, 52.

PAPADIA, Gabriele; PACELLA, Massimo; PERRONE, Massimiliano; GILIBERTI, Vincenzo. A Comparison of Different Topic Modeling Methods through a Real Case Study of Italian Customer Care. *Algorithms*, Multidisciplinary Digital Publishing Institute, v. 16, n. 2, p. 94, 2 8 fev. 2023. ISSN 1999-4893. DOI: 10.3390/a16020094. Disponível em: https://www.mdpi.com/1999-4893/16/2/94. Acesso em: 25 maio 2023. Citado 1 vez na página 75.

PARK, Sangchul; KO, Haksoo. Machine Learning and Law and Economics: A Preliminary Overview. *Asian Journal of Law and Economics*, De Gruyter, v. 11, n. 2, p. 20200034, 18 set. 2020. ISSN 2154-4611, 2194-6086. DOI: 10.1515/ajle-2020-0034. Disponível em: https://www.degruyter.com/document/doi/10.1515/ajle-2020-0034/html. Acesso em: 14 dez. 2022. Citado 5 vezes nas páginas 11, 12.

PEINELT, Nicole; NGUYEN, Dong; LIAKATA, Maria. tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection. en-US. *In:* ACL 2020. Edição: Dan Jurafsky, Joyce Chai, Natalie Schluter e Joel Tetreault, p. 7047–7055. DOI: 10.18653/v1/2020.acl-main.630. Disponível em: https://aclanthology.org/2020.acl-main.630. Acesso em: 28 nov. 2023. Citado 1 vez na página 80.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. Glove: Global Vectors for Word Representation. *null*, 2014. DOI: 10.3115/v1/d14-1162. pmid: null. Citado 3 vezes na página 31.

PEREIRA NETO, Caio Mário Da Silva; CASAGRANDE, Paulo Leonardo. *Direito Concorrencial: Doutrina, Jurisprudência e Legislação*. São Paulo: Saraiva, 2016. Citado 1 vez na página 54.

POLO, Felipe Maia; CIOCHETTI, Itamar; BERTOLO, Emerson. Predicting Legal Proceedings Status: Approaches Based on Sequential Text Data. *In:* (ICAIL '21), p. 264–265. ISBN 978-1-4503-8526-8. DOI: 10.1145/3462757.3466138. Disponível em: https://doi.org/10.1145/3462757.3466138. Acesso em: 2 out. 2022. Citado 1 vez na página 8.

POWERS, David MW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020. Citado 1 vez na página 82.

ROBERTS, Margaret E; TINGLEY, Dustin; STEWART, Brandon M; AIROLDI, Edoardo M. The Structural Topic Model and Applied Social Science, 2013. Citado 3 vezes na página 29.

ROBERTS, Margaret E.; STEWART, Brandon M.; TINGLEY, Dustin. Navigating the Local Modes of Big Data: The Case of Topic Models. *In:* ALVAREZ, R. Michael (ed.). 1. ed.: Cambridge University Press, 31 jan. 2016. p. 51–97. ISBN 978-1-107-10788-5. DOI: 10.1017/CBO9781316257340.004. Disponível em: https://www.cambridge.org/core/product/identifier/CBO9781316257340A009/type/book_part. Acesso em: 8 jan. 2023. Citado 1 vez na página 29.

ROBERTS, Margaret E.; STEWART, Brandon M.; TINGLEY, Dustin. STM: An *R* Package for Structural Topic Models. *Journal of Statistical Software*, v. 91, n. 2, 2019. ISSN 1548-7660. DOI: 10.18637/jss.v091.i02. Disponível em: http://www.jstatsoft.org/v91/i02/. Acesso em: 5 fev. 2023. Citado 2 vezes na página 29.

ROBERTS, Margaret E.; STEWART, Brandon M.; TINGLEY, Dustin *et al.* Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, v. 58, n. 4, p. 1064–1082, 2014. ISSN 1540-5907. DOI: 10.1111/ajps.12103. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12103. Acesso em: 4 nov. 2023. Citado 4 vezes na página 29.

ROSA, Rodrigo Augusto. Leis Empíricas e as Máximas Da Razão Em Kant. *Intuitio*, v. 3, n. 1, p. 139–156, 2010. ISSN 1983-4012. Disponível em: https://revistaseletronicas.pucrs.br/index.php/iberoamericana/N%C3%83%C6%92O%20https:/www.scimagojr.com/index.php/intuitio/article/view/6932. Acesso em: 29 set. 2022. Citado 1 vez na página 14.

ROSNER, Frank; HINNEBURG, Alexander; RÖDER, Michael; NETTLING, Martin; BOTH, Andreas. *Evaluating Topic Coherence Measures*. en-US. 25 mar. 2014. DOI: 10.48550/arXiv.1403.6397. arXiv: 1403.6397 [cs]. Disponível em: http://arxiv.org/abs/1403.6397. Acesso em: 28 nov. 2023. preprint. Citado 2 vezes na página 78.

ROSSETTI, José Paschoal. *Introdução à economia*. 17. ed. São Paulo: Atlas, 1997. Citado 1 vez nas páginas 41, 46.

RUSSELL, Stuart J.; NORVIG, Peter. *Artificial Intelligence: A Modern Approach*. Fourth edition. Hoboken: Pearson, 2021. 2145 p. (Pearson Series in Artificial Intelligence). ISBN 978-0-13-461099-3. Disponível em: https://lccn.loc.gov/2019047498. Citado 10 vezes nas páginas 1, 6–8.

SALOMÃO FILHO, Calixto. *Direito Concorrencial: As Estruturas*. São Paulo: Malheiros Editores, 2002. Citado 3 vezes nas páginas 39, 40, 56.

SAMUELSON, William F.; MARKS, Stephen Gary. *Managerial Economics*. Wiley, 2012. v. 7. 765 p. ISBN 978-1-118-04158-1. pmid: 25246403. Citado 10 vezes nas páginas 43–45.

SÁNCHEZ-FRANCO, Manuel J.; REY-MORENO, Manuel. Do Travelers' Reviews Depend on the Destination? An Analysis in Coastal and Urban Peer-to-peer Lodgings. en-US. *Psychology & Marketing*, v. 39, n. 2, p. 441–459, fev. 2022. ISSN 0742-6046, 1520-6793. DOI: 10.1002/mar.21608. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/mar.21608. Acesso em: 17 dez. 2023. Citado 1 vez na página 76.

SANTOS, Aguinaldo dos. *Seleção do método de pesquisa: guia para pós-graduando em design e áreas afins*. 22. ed. Curitiba, Paraná: Insight Editora, 11 maio 2018. 230 p. ISBN 978-85-62241-46-8. Citado 1 vez na página 8.

SANTOS, Keila Barbosa Costa dos. *Categorização de textos por aprendizagem de máquina*. 2019. 85 f. Dissertação (Mestrado em Modelagem Computacional de Conhecimento) – Universidade Federal de Alagoas, Maceió. Disponível em: http://www.repositorio.ufal.br/jspui/handle/riufal/6174. Acesso em: 5 out. 2022. Citado 3 vezes nas páginas 15, 19.

SCHREPEL, Thibault. *Computational Antitrust: An Introduction and Research Agenda*. 15 jan. 2021. Disponível em: https://papers.ssrn.com/abstract=3766960. Acesso em: 14 dez. 2022. preprint. Citado 2 vezes na página 11.

SCHWARZ, Gideon. Estimating the dimension of a model. *The annals of statistics*, JSTOR, p. 461–464, 1978. Citado 1 vez na página 82.

SILVA, Fernando Rodrigues da. *Análises Ecológicas No R*. Em colaboração com Hiago Gonçalves Souza, Gustavo Brant Paterno, Diogo Borges Provete, Maurício Humberto Vancine e Ulysses Paulino de Albuquerque. Bauru, SP: Canal 6 Editora, 30 mar. 2022. 623 p. ISBN 978-85-7917-564-0. Citado 3 vezes nas páginas 13, 14, 108.

SILVA FILHO, Jadir Rafael da; HARO, Guilherme Prado Bohac de. A evolução do direito concorrencial e o papel do Conselho Administrativo de Defesa Econômica. *Encontro de Iniciação Científica*, v. 9, n. 9, 2013. Citado 5 vezes nas páginas 39, 40.

SILVA FILHO, Jadir Rafael da; HARO, Guilherme Prado Bohac de de. A evolução do direito concorrencial e o papel do conselho administrativo de defesa econômica. pt-BR. *Encontro de Iniciação Científica (ETIC)*, v. 9, n. 9, 2013. ISSN 21-76-8498. Citado 10 vezes nas páginas 46, 47, 53, 55, 56.

STUCKE, Maurice E; GRUNES, Allen P. Debunking the Myths over Big Data and Antitrust. *CPI Anitrust Chronicle*, p. 10, May 2015. Disponível em: http://ssrn.com/abstract=2612562. Citado 1 vez na página 11.

TAUFICK, Roberto Domingos. *Nova Lei Antitruste Brasileira - a Lei 12.529/2011 Comentada e a Análise Prévia No Direito Da Concorrência*. São Paulo: Forense, 2012. Citado 4 vezes nas páginas 54, 56, 57.

TEH, Yee Whye; JORDAN, Michael I.; BEAL, Matthew J.; WORK(S): David M. Blei Reviewed. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, v. 101, n. 476, p. 1566–1581, 2006. Disponível em: http://www.jstor.org/stable/27639773. Citado 1 vez na página 28.

TERRAGNI, Silvia; FERSINI, Elisabetta. An Empirical Analysis of Topic Models: Uncovering the Relationships between Hyperparameters, Document Length and Performance Measures. en-US. *In:* RANLP 2021. Edição: Ruslan Mitkov e Galia Angelova, p. 1408–1416. Disponível em: https://aclanthology.org/2021.ranlp-1.157. Acesso em: 7 nov. 2023. Citado 1 vez na página 77.

TERRAGNI, Silvia; FERSINI, Elisabetta; GALUZZI, Bruno Giovanni; TROPEANO, Pietro; CANDELIERI, Antonio. OCTIS: Comparing and Optimizing Topic Models Is Simple! *In:* p. 263–270. DOI: 10.18653/v1/2021.eacl-demos.31. Disponível em: https://aclanthology.org/2021.eacl-demos.31. Acesso em: 12 maio 2023. Citado 1 vez na página 77.

TERRAGNI, Silvia; HARRANDO, Ismail; LISENA, Pasquale; TRONCY, Raphael; FERSINI, Elisabetta. *One Configuration to Rule Them All? Towards Hyperparameter Transfer in Topic Models Using Multi-Objective Bayesian Optimization*. en-US. 15 fev. 2022. arXiv: 2202.07631 [cs]. Disponível em: http://arxiv.org/abs/2202.07631. Acesso em: 7 nov. 2023. preprint. Citado 1 vez na página 77.

TIROLE, Jean. *Economics for the Common Good*. Tradução: Steven Rendall. First paperback printing. Princeton Oxford: Princeton University Press, 2019. 563 p. ISBN 978-0-691-17516-4. Citado 17 vezes nas páginas 36, 37, 39, 57, 58.

TIROLE, Jean. *The theory of industrial organization*. MIT press, 1988. Citado 1 vez na página 43.

ULENAERS, Jasper. The Impact of Artificial Intelligence on the Right to a Fair Trial: Towards a Robot Judge? *Asian Journal of Law and Economics*, De Gruyter, v. 11, n. 2, p. 20200008, 18 set. 2020. ISSN 2154-4611, 2194-6086. DOI: 10.1515/ajle-2020-0008. Disponível em: https://www.degruyter.com/document/doi/10.1515/ajle-2020-0008/html. Acesso em: 14 dez. 2022. Citado 2 vezes na página 12.

VAN RIJSBERGEN, C. Information retrieval: theory and practice. *In:* PROCEEDINGS of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems. 1979. v. 79. Citado 1 vez na página 82.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki *et al.* Attention Is All You Need. *null*, 2017. DOI: null. pmid: null. Citado 3 vezes nas páginas 34, 69.

VEIGA, Fábio Da Silva; ZAŁUCKI, Mariusz (ed.). *Legaltech, Artificial Inteligence and the Future of Legal Practice*. 2022. ISBN 978-989-53-2813-0. Disponível em: https://eprints.ucm.es/id/eprint/73947/1/Ebook%20I%20JURISTECH%20-%20press.pdf#page=12. Acesso em: 29 mar. 2023. Citado 1 vez na página 12.

VERAS, Karina De Oliveira; BARRETO, Gabriela. A Inteligência Artificial No Setor Público: Uma Análise Do Projeto VICTOR No Poder Judiciário, 2022. Citado 6 vezes nas páginas 8, 11, 59.

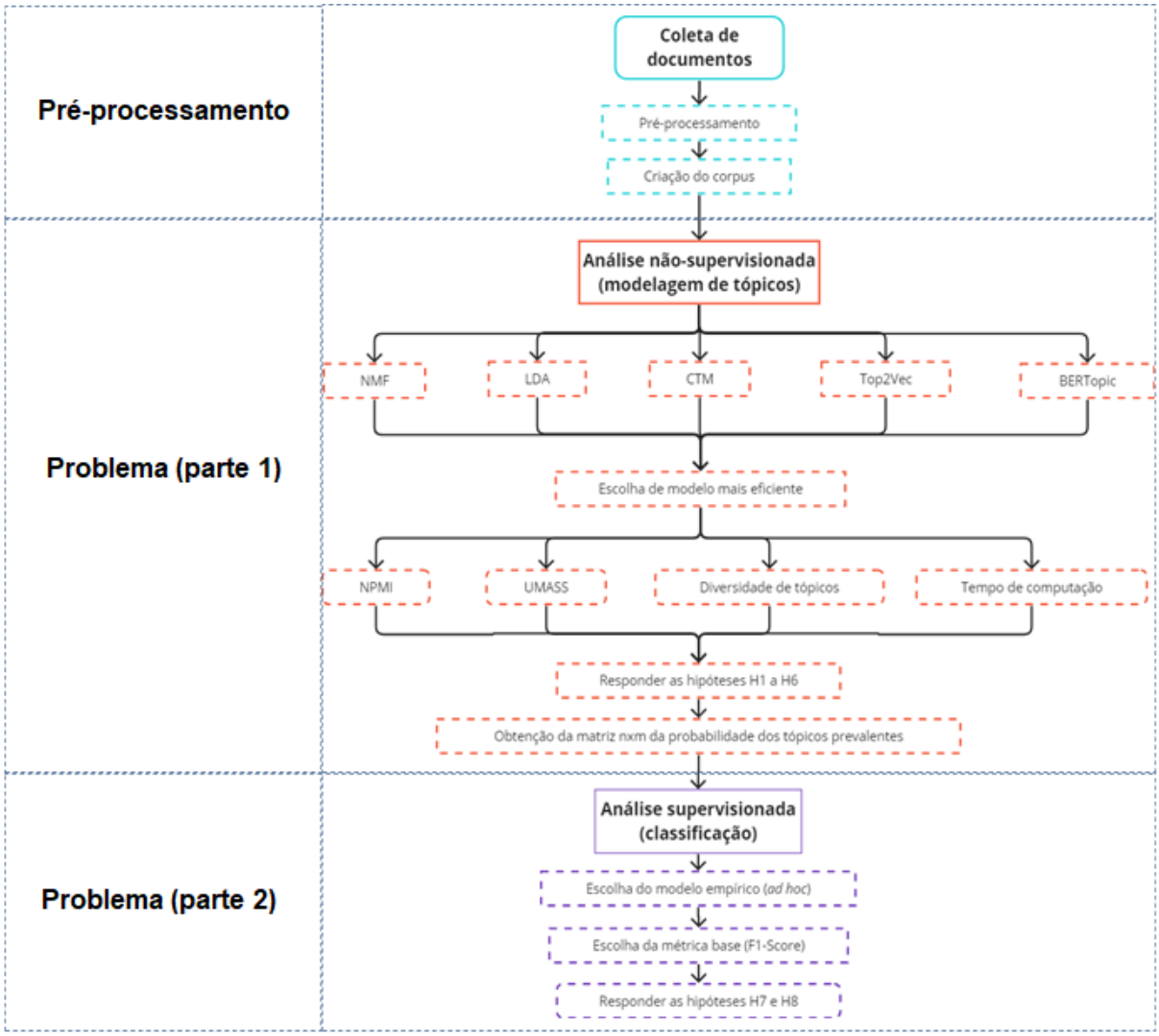
WANG, Yongming. Using Machine Learning and Natural Language Processing to Analyze Library Chat Reference Transcripts. *Information Technology and Libraries*, v. 41, n. 3, 19 set. 2022. ISSN 2163-5226, 0730-9295. DOI: 10.6017/ital.v41i3.14967. Disponível em: https://ital.corejournals.org/index.php/ital/article/view/14967. Acesso em: 16 out. 2023. Citado 1 vez na página 11.

WICKHAM, Hadley; GROLEMUND, Garret. *R para data science: importe, arrume, transforme, visualize e modele dados*. Tradução: Samantha Batista. Rio de Janeiro, RJ, Brasil: Alta Books, 14 out. 2021. 528 p. ISBN 978-85-508-0324-1. Citado 3 vezes nas páginas 13, 14, 108.

WITTEN, Ian. Text Mining. *In:* SINGH, Munindar (ed.). Chapman and Hall/CRC, 29 set. 2004. v. 20042960. ISBN 978-1-58488-381-4. DOI: 10.1201/9780203507223.ch14. Disponível em: http://www.crcnetbase.com/doi/abs/10.1201/9780203507223.ch14. Acesso em: 14 nov. 2023. Citado 2 vezes na página 19.

APÊNDICE A – DESENHO DA PESQUISA

Quadro 3 – Desenho da pesquisa segundo o problema



Fonte: elaborado pelo autor.

APÊNDICE B – ESTADO DA ARTE SOBRE MODELAGEM DE TÓPICOS

Quadro 4 – Modelos de Tópicos abordados na Revisão de Literatura

Ano	Modelo	Pesquisa	Tipologia	Principal Referência	Outros Autores	Tipos de Pesquisa	Será avaliado?
1990	LSA	4	Alocação Latente	DEERWESTER et al., 1990	HOFMANN, 2001; KHERWA; BANSAL, 2018; ZENGUL et al., 2023	04 artigos	Não
1999	NMF	10	Híbrido	LEE et al., 1999	ABUHAY; NIGATIE; KOVALCHUK, 2018; ADHIKARY; MURTY, 2012; ALBALAWI; YEAP; BENYOUCEF, 2020; EGGER; YU, 2022; FU et al., 2021; JANSEN et al., 2021; KRISHNAN, 2023; O'CALLAGHAN et al., 2015; PAPADIA et al., 2023	10 artigos	Sim
2001	PLSA	2	Alocação Latente	HOFMANN, 2001	ZENGUL et al., 2023	02 artigos	Não
2003	LDA	25	Alocação Latente	BLEI; NG; JORDAN, 2003	AGUIRRE; VALLEJO, 2022; AYUSO, 2022; BASILIO; PEREIRA, 2020; BLEI; NG; JORDAN, 2003; CHEN et al., 2023; EGGER; YU, 2022; FARACO, 2020; FERREIRA et al., 2023; JO; OH, 2011; KARAS et al., 2022; KHERWA; BANSAL, 2018; KRISHNAN, 2023; MA; ZENG-TREITLER; NELSON, 2021; MIRANDA et al., 2019; MOREIRA; CÉSAR, 2019; NOWLIN, 2016; O'CALLAGHAN et al., 2015; OZAKI; KOBAYASHIE, 2022; PAPADIA et al., 2023; PARK; HASSAIRI, 2021; RODRIGUES, 2019; TUFTS, [s.d.]; VALVERDE; ARIAS, 2020; ZENGUL et al., 2023	01 livro; 03 teses; 19 artigos; 02 conferência	Sim
2006	DTM	2	Tópico Temporal	BLEI; LAFFERTY, 2006	ZENGUL et al., 2023	02 artigos	Não
2013	Word2Vec	8	Embeddings	MIKOLOV et al., 2013b, 2013a	HARTMANN et al., 2017; KOWSARI et al., 2019; MIKOLOV et al., 2017; O'CALLAGHAN et al., 2015; POLO et al., 2021; ZENGUL et al., 2023	07 artigos; 01 pré-impressão	Não, pois será utilizado o modelo Top2Vec que é considerado um aprimoramento do Word2Vec.
2013	STM	4	Alocação Latente	ROBERTS et al., 2013	ROBERTS et al., 2014; ROBERTS; STEWART; TINGLEY, 2016, 2019	03 artigos; 01 seção de livro	Não
2014	Doc2Vec	2	Embeddings	LE; MIKOLOV, 2014	POLO et al., 2021	01 artigo; 01 conferência	Não
2014	GloVe	1	Embeddings	PENNINGTON; SOCHER; MANNING, 2014		01 artigo	Não
2016	FastText	5	Embeddigns	JOULIN et al., 2016b, 2016a	BOJANOWSKI et al., 2017; HARTMANN et al., 2017;; MIKOLOV et al., 2017	02 artigos; 03 pré-impressão	Não
2017	NTM	1	Rede Neural	DIENG et al., 2017		01 pré-impressão	Não
2019	BERT	17	Embeddings	DEVLIN et al., 2019	ALHASSAN; ZHANG; SCHLEGEL, 2022; AYUSO, 2022; CAI et al., 2022; CAPELLARO; CASELI, 2021; CHEN et al., 2023; EGGER; YU, 2022; GROOTENDORST, 2022; KIM; YOON, 2021; KOZBAGAROV; MUSSABAYEV; MLADENOVIC, 2021; KRISHNAN, 2023; LI, 2022; PENG et al., 2021; POLO et al., 2021; SOUZA; NOGUEIRA; LOTUFO, 2020; VIEGAS; COSTA; ISHII, 2022; ZHANG; MILIOS, 2023	02 teses; 10 artigos; 04 conferência; 01 pré-impressão	Não, pois considerando o BERTopic como aprimoramento do BERT, aquele será escolhido em detrimento deste.
2020	Top2Vec	18	Híbrido	ANGELOV, 2020	AKBAY, 2022; AUSTIN; ZAÏANE; LARGERON, 2022; CAI et al., 2022; CHEN et al., 2023; DIAF; FRITSCHE, 2022; EGGER; YU, 2022; GHASIYA et al., 2021; GHASIYA; OKAMURA, 2021; KARAS et al., 2022; KRISHNAN, 2023; LASSCHE; KOSTKAN; NIELBO, 2022; MA; ZENG-TREITLER; NELSON, 2021; MEMON; SYED; MEMON, 2023; SCHÜTZE; DIAF, 2023; ZENGUL et al., 2023; ZHANG; MILIOS, 2023; ZUNDERT et al., 2022	17 artigos; 01 conferência;	Sim
2022	BERTopic	6	Embeddings	GROOTENDORST, 2022	AYUSO, 2022; CHEN et al., 2023; EGGER; YU, 2022; KRISHNAN, 2023; ZHANG; MILIOS, 2023	01 tese; 04 artigos; 01 conferência	Sim

Fonte: elaborado pelo autor.

APÊNDICE C – *SCRIPTS*

Todos os scripts da modelagem supervisionada e não supervisionada estão contidos no seguinte repositório do Google Colab:

Figura 29 – Acesso ao repositório do Google Colab

https://colab.research.google.com/drive/1EX6lwfjZaTH1soO9t01ROmOCOGE4T57a?usp=sharing

Fonte: Elaborado pelo autor.

APÊNDICE D – RESULTADO DA ANÁLISE NÃO SUPERVISIONADA

Figura 30 – Comparação das métricas de todos os modelos

Modelo	Tópicos	Unigramas	NPMI (médio)	Umass (médio)	Diversidade de tópicos (media)	Tempo de computação (s)(médio)
BERTopic-BERTimbau	20	15	-0,03410057160	0,00000000000	0,55000000000	4,03039360046
BERTopic-BERTimbau	50	15	-0,03410057160	0,00000000000	0,55000000000	4,03209972382
BERTopic-BERTimbau	20	5	-0,02179071713	-0,19219876836	0,50000000000	4,04976367950
BERTopic-BERTimbau	40	15	-0,03005230688	-0,00097561749	0,60000000000	4,08739256859
BERTopic-BERTimbau	50	5	-0,02988474673	-0,17526537238	0,40000000000	4,25029253960
BERTopic-BERTimbau	30	5	-0,03052453394	-0,14170777432	0,33333333333	4,28278994560
BERTopic-BERTimbau	40	5	-0,02892728778	-0,16978300969	0,40000000000	4,33255386353
BERTopic-BERTimbau	10	15	-0,03005230688	-0,00097561749	0,60000000000	4,33474159241
BERTopic-BERTimbau	10	5	-0,03149178016	-0,14148548540	0,33333333333	4,66564631462
BERTopic-BERTimbau	30	15	-0,03410057160	0,00000000000	0,55000000000	5,24143695831
BERTopic-DistilUSE	30	15	-0,02717878374	-0,00151762721	1,00000000000	3,23609399796
BERTopic-DistilUSE	30	5	-0,03441329877	0,00000000000	1,00000000000	3,29782247543
BERTopic-DistilUSE	40	5	-0,03441329877	0,00000000000	1,00000000000	3,29998564720
BERTopic-DistilUSE	10	5	-0,03441329877	0,00000000000	1,00000000000	3,30263185501
BERTopic-DistilUSE	20	15	-0,02717878374	-0,00151762721	1,00000000000	3,33237671852
BERTopic-DistilUSE	50	15	-0,02717878374	-0,00151762721	1,00000000000	3,33322191238
BERTopic-DistilUSE	20	5	-0,02717878374	-0,00151762721	1,00000000000	3,43432521820
BERTopic-DistilUSE	10	15	-0,03441329877	0,00000000000	1,00000000000	3,76840829849
BERTopic-DistilUSE	50	5	-0,02717878374	-0,00151762721	1,00000000000	3,81814908981
BERTopic-DistilUSE	40	15	-0,03441329877	0,00000000000	1,00000000000	3,93821692467
BERTopic-Legal BERT	40	15	-0,03435621805	-0,00119242138	0,30000000000	3,59875988960
BERTopic-Legal BERT	50	15	-0,02831076658	-0,00075881361	0,35000000000	3,61523771286
BERTopic-Legal BERT	30	5	-0,02889397018	-0,04365795619	0,30000000000	3,61639857292
BERTopic-Legal BERT	20	15	-0,02902959389	-0,00081301458	0,35000000000	3,61924314499
BERTopic-Legal BERT	50	5	-0,02792951672	-0,04340501832	0,31666666667	3,63250780106
BERTopic-Legal BERT	40	5	-0,03312640308	-0,04605805623	0,28333333333	3,70013308525
BERTopic-Legal BERT	20	5	-0,03415949767	-0,00494033602	0,28333333333	3,82734465599
BERTopic-Legal BERT	10	15	-0,02902959389	-0,00081301458	0,35000000000	3,95727419853
BERTopic-Legal BERT	30	15	-0,02856811028	-0,00075881361	0,35000000000	4,04739856720
BERTopic-Legal BERT	10	5	-0,02792951672	-0,04340501832	0,31666666667	4,18514347076
BERTopic-Legal BERTimbau	40	15	-0,03441329877	0,00000000000	1,00000000000	3,31642651558
BERTopic-Legal BERTimbau	10	15	-0,02717878374	-0,00151762721	1,00000000000	3,31648325920
BERTopic-Legal BERTimbau	30	15	-0,03441329877	0,00000000000	1,00000000000	3,35509276390
BERTopic-Legal BERTimbau	20	5	-0,03237586330	-0,00108401944	0,46666666667	3,37985706329
BERTopic-Legal BERTimbau	40	5	-0,03226743714	-0,00050587574	0,50000000000	3,39999461174
BERTopic-Legal BERTimbau	50	5	-0,03226743714	-0,00050587574	0,50000000000	3,43594503403
BERTopic-Legal BERTimbau	20	15	-0,03441329877	0,00000000000	1,00000000000	3,53421258926
BERTopic-Legal BERTimbau	10	5	-0,03226743714	-0,00050587574	0,50000000000	3,54966688156
BERTopic-Legal BERTimbau	50	15	-0,03457593801	-0,00086721555	0,60000000000	3,78448724747
BERTopic-Legal BERTimbau	30	5	-0,03226743714	-0,00050587574	0,50000000000	3,83483481407
BERTopic-LegalNLP	10	15	-0,02388378384	0,00000000000	0,60000000000	4,03863096237
BERTopic-LegalNLP	20	15	-0,01983551912	-0,00097561749	0,65000000000	4,05027770996
BERTopic-LegalNLP	40	15	-0,02388378384	0,00000000000	0,60000000000	4,05761146545
BERTopic-LegalNLP	30	15	-0,02388378384	0,00000000000	0,60000000000	4,06655764580
BERTopic-LegalNLP	50	5	-0,02246671157	-0,19284918003	0,53333333333	4,09345245361
BERTopic-LegalNLP	40	5	-0,02211417260	-0,10306750589	0,32000000000	4,18324708939
BERTopic-LegalNLP	20	5	-0,01997309730	-0,12873420797	0,40000000000	4,21700024605
BERTopic-LegalNLP	10	5	-0,01997309730	-0,12873420797	0,40000000000	4,32492804527
BERTopic-LegalNLP	50	15	-0,02388378384	0,00000000000	0,60000000000	4,42459535599
BERTopic-LegalNLP	30	5	-0,02298470949	-0,10306750589	0,32000000000	4,73355507851
BERTopic-MiniLM	40	15	-0,03441329877	0,00000000000	1,00000000000	3,24785041809
BERTopic-MiniLM	30	5	-0,03079604126	-0,00075881361	0,60000000000	3,33185005188
BERTopic-MiniLM	10	5	-0,03034470733	-0,00173443110	0,60000000000	3,34026312828
BERTopic-MiniLM	10	15	-0,03441329877	0,00000000000	1,00000000000	3,36170291901
BERTopic-MiniLM	30	15	-0,03441329877	0,00000000000	1,00000000000	3,37913584709
BERTopic-MiniLM	20	15	-0,03441329877	0,00000000000	1,00000000000	3,40258121490
BERTopic-MiniLM	50	5	-0,03034470733	-0,00173443110	0,60000000000	3,42263770103
BERTopic-MiniLM	40	5	-0,03034470733	-0,00173443110	0,60000000000	3,42683839798
BERTopic-MiniLM	50	15	-0,03441329877	0,00000000000	1,00000000000	3,75735545158
BERTopic-MiniLM	20	5	-0,03079604126	-0,00075881361	0,60000000000	3,92497825623
BERTopic-OpenAI	40	5	-0,00079883221	-0,01066986839	1,00000000000	14,22042155266

BERTopic-OpenAI	40	15	-0,03410057160	0,00000000000	0,55000000000	14,30513834953
BERTopic-OpenAI	30	5	-0,00079883221	-0,01066986839	1,00000000000	14,37499713898
BERTopic-OpenAI	30	15	-0,03410057160	0,00000000000	0,55000000000	14,47093176842
BERTopic-OpenAI	20	5	-0,00079883221	-0,01066986839	1,00000000000	14,59124517441
BERTopic-OpenAI	10	15	-0,03410057160	0,00000000000	0,55000000000	14,75024199486
BERTopic-OpenAI	10	5	-0,00079883221	-0,01066986839	1,00000000000	14,85164046288
BERTopic-OpenAI	20	15	-0,03410057160	0,00000000000	0,55000000000	14,92786192894
BERTopic-OpenAI	50	15	-0,03410057160	0,00000000000	0,55000000000	15,16747903824
BERTopic-OpenAI	50	5	-0,00079883221	-0,01066986839	1,00000000000	19,17563199997
BERTopic-STJIris	50	15	-0,02361541587	-0,00086721555	0,60000000000	3,54382777214
BERTopic-STJIris	40	15	-0,03323181836	-0,00162602916	0,55000000000	3,56023120880
BERTopic-STJIris	10	5	-0,00079883221	-0,01066986839	1,00000000000	3,56994080544
BERTopic-STJIris	30	15	-0,03015722762	-0,00184283304	0,60000000000	3,62713837624
BERTopic-STJIris	20	5	-0,00079883221	-0,01066986839	1,00000000000	3,63616228104
BERTopic-STJIris	40	5	-0,00079883221	-0,01066986839	1,00000000000	3,67371678352
BERTopic-STJIris	20	15	-0,02945192161	-0,00184283304	0,60000000000	3,70599961281
BERTopic-STJIris	30	5	-0,00079883221	-0,01066986839	1,00000000000	4,07072734833
BERTopic-STJIris	50	5	-0,00079883221	-0,01066986839	1,00000000000	4,21503901482
BERTopic-STJIris	10	15	-0,03211200599	-0,00184283304	0,55000000000	4,28187680244
BERTopic-USE Multi	40	15	-0,01762767752	-0,00057814370	0,46666666667	3,43752026558
BERTopic-USE Multi	10	15	-0,02486409372	-0,00115628740	0,50000000000	3,44602966309
BERTopic-USE Multi	50	15	-0,01897539561	-0,00086721555	0,65000000000	3,48635101318
BERTopic-USE Multi	30	15	-0,01949008302	-0,00097561749	0,70000000000	3,49594330788
BERTopic-USE Multi	20	15	-0,02420948583	-0,00086721555	0,65000000000	3,52342033386
BERTopic-USE Multi	10	5	-0,02137210856	-0,11394773776	0,37500000000	3,65302848816
BERTopic-USE Multi	30	5	-0,02551157985	-0,06862987238	0,35000000000	3,80309081078
BERTopic-USE Multi	20	5	-0,02537645120	-0,05454617149	0,35000000000	4,01650547981
BERTopic-USE Multi	50	5	-0,02150723721	-0,11443554651	0,37500000000	4,03059530258
BERTopic-USE Multi	40	5	-0,02551157985	-0,06862987238	0,35000000000	4,85288190842
CTM	20	0	-0,21901616465	-0,92879848125	0,64166666667	1885,52557047208
CTM	30	0	-0,25737579706	-1,37722073184	0,57222222222	1888,51473665237
CTM	40	0	-0,25328778125	-1,35417118556	0,54583333333	1906,58694529533
CTM	10	0	-0,25436628625	-1,25332717872	0,77666666667	2018,57296395302
CTM	50	0	-0,26631826785	-1,81655429645	0,54533333333	2077,39458680152
LDA	10	0	-0,03235529458	0,00000000000	0,13333333333	2,48639496168
LDA	20	0	-0,03237191562	-0,00060956137	0,06833333333	3,04221820831
LDA	30	0	-0,03223656477	-0,00071796332	0,05333333333	3,54726521174
LDA	40	0	-0,03230565009	-0,00079380033	0,04083333333	3,93237463633
LDA	50	0	-0,03238556795	-0,00125562694	0,03266666667	4,18480690320
NMF	10	0	-0,08098048136	-3,35770176608	0,32333333333	1,02264062564
NMF	20	0	-0,08903850816	-3,90329921987	0,32333333333	1,20943117142
NMF	30	0	-0,12982863426	-6,76196319273	0,47666666667	1,69451340040
NMF	40	0	-0,16586623462	-9,47243108074	0,61166666667	2,24549873670
NMF	50	0	-0,16781009047	-10,11283411302	0,64466666667	2,71879410744
Top2Vec-DistilUSE	2	0	-0,11624364916	-1,99171767904	0,75000000000	23,42677984238
Top2Vec-Legal BERT	7	0	-0,30042465508	-2,87586809545	0,66000000000	30,16225013733
Top2Vec-MiniLM	3	0	-0,01032814912	-0,52522206148	0,30000000000	19,69827884436
Top2Vec-MiniLM	2	0	-0,07182897418	-0,66675685957	0,45000000000	19,76377892494
Top2Vec-STJIris	1	0	-0,26385543666	-4,43663846240	1,00000000000	371,61058940887
Top2Vec-USE Multi	4	0	0,18373955258	-0,18380001154	0,32500000000	69,20870113373
Top2Vec-USE Multi	3	0	0,17692296883	-0,22278346424	0,40000000000	73,75959769885
Top2Vec-USE Multi	2	0	0,12084016878	-0,27840790390	0,55000000000	75,83651685715

Fonte: elaborado pelo autor. Destacado em verde são os melhores métricas médias intramodelos. Os scores médios foram computados em um range de 10 a 50 tópicos com passos de 10. A coerência do tópico e a diversidade do tópico foram calculadas em cada passo para cada modelo de tópico. Todos os resultados foram calculados em média para 3 execuções de cada etapa. Assim, cada pontuação é a média de 15 execuções separadas.

APÊNDICE E – SUMÁRIO ESTATÍSTICO DO MODELO LOGÍSTICO

Tabela 8 – Resumo Estatístico do Modelo Logit

Dep. Variable:	merito	**No. Observations:**	7633
Model:	Logit	**Df Residuals:**	7590
Method:	MLE	**Df Model:**	42
Date:	Sun, 28 Jan 2024	**Pseudo R-squ.:**	0.1199
Time:	18:29:22	**Log-Likelihood:**	-4656.3
converged:	False	**LL-Null:**	-5290.7
Covariance Type:	nonrobust	**LLR p-value:**	1.386e-238

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-1.4233	0.181	-7.854	0.000	-1.779	-1.068
mpf_topico_0	0.0341	0.075	0.458	0.647	-0.112	0.180
mpf_topico_1	0.2355	0.090	2.607	0.009	0.058	0.413
mpf_topico_2	-0.8340	0.144	-5.790	0.000	-1.116	-0.552
mpf_topico_3	0.7227	0.147	4.921	0.000	0.435	1.010
mpf_topico_4	0.7712	0.210	3.668	0.000	0.359	1.183
mpf_topico_5	-0.8069	0.169	-4.768	0.000	-1.139	-0.475
mpf_topico_6	-0.8813	0.174	-5.063	0.000	-1.223	-0.540
mpf_topico_7	-0.5219	0.193	-2.709	0.007	-0.900	-0.144
mpf_topico_8	0.5229	0.264	1.980	0.048	0.005	1.041
mpf_topico_9	-0.3761	0.226	-1.667	0.096	-0.818	0.066
mpf_topico_10	-0.3156	0.242	-1.305	0.192	-0.790	0.158
mpf_topico_11	-2.3754	0.394	-6.022	0.000	-3.148	-1.602
mpf_topico_12	-0.9798	0.377	-2.597	0.009	-1.719	-0.240
mpf_topico_13	-0.8467	0.349	-2.427	0.015	-1.531	-0.163
mpf_topico_14	0.0597	0.379	0.158	0.875	-0.683	0.802
mpf_topico_15	-0.0836	0.537	-0.156	0.876	-1.136	0.969
mpf_topico_16	20.2505	9002.082	0.002	0.998	-1.76e+04	1.77e+04
mpf_topico_17	-17.9731	1920.298	-0.009	0.993	-3781.688	3745.742
mpf_topico_18	0.5218	1.012	0.515	0.606	-1.462	2.506
rel_topico_0	-0.0192	0.076	-0.252	0.801	-0.169	0.130
rel_topico_1	-0.2531	0.127	-1.987	0.047	-0.503	-0.004

rel_topico_2	-0.3030	0.129	-2.340	0.019	-0.557	-0.049
rel_topico_3	0.0683	0.152	0.449	0.654	-0.230	0.366
rel_topico_4	-0.0068	0.267	-0.026	0.980	-0.530	0.517
rel_topico_5	-0.6409	0.212	-3.018	0.003	-1.057	-0.225
rel_topico_6	-0.3654	0.255	-1.431	0.153	-0.866	0.135
rel_topico_7	-0.6902	0.255	-2.704	0.007	-1.190	-0.190
rel_topico_8	-0.6318	0.336	-1.881	0.060	-1.290	0.027
rel_topico_9	-0.6205	0.422	-1.471	0.141	-1.448	0.207
rel_topico_10	-1.0122	0.347	-2.914	0.004	-1.693	-0.331
rel_topico_11	-0.7512	0.360	-2.086	0.037	-1.457	-0.046
rel_topico_12	-0.4208	0.388	-1.084	0.278	-1.182	0.340
rel_topico_13	-0.3108	0.506	-0.615	0.539	-1.302	0.680
rel_topico_14	0.4764	0.530	0.899	0.369	-0.563	1.515
rel_topico_15	-0.2861	0.468	-0.612	0.541	-1.203	0.631
rel_topico_16	-0.5912	0.680	-0.870	0.385	-1.924	0.741
rel_topico_17	-0.7617	0.491	-1.551	0.121	-1.724	0.201
rel_topico_18	1.3136	1.143	1.150	0.250	-0.926	3.553
genero_mas	0.4838	0.071	6.812	0.000	0.345	0.623
formacao_direito	0.9089	0.176	5.179	0.000	0.565	1.253
formacao_economia	2.2340	0.177	12.590	0.000	1.886	2.582
servidor_publico_sim	-0.2077	0.064	-3.261	0.001	-0.333	-0.083

Fonte: Elaborado pelo autor.

APÊNDICE F – SUMÁRIO ESTATÍSTICO - ODDS RATIO

Tabela 10 – Odds Ratios e Inverse Odds Ratios do Modelo de Regressão Logística

Variável	Odds Ratio	Inverse Odds Ratio
Intercepto	0.2413	4.1446
mpf_topico_0	1.0347	0.9665
mpf_topico_1	1.2655	0.7902
mpf_topico_2	0.4344	2.3019
mpf_topico_3	2.0596	0.4855
mpf_topico_4	2.1616	0.4626
mpf_topico_5	0.4464	2.2403
mpf_topico_6	0.4144	2.4132
mpf_topico_7	0.5937	1.6844
mpf_topico_8	1.6859	0.5932
mpf_topico_9	0.6869	1.4558
mpf_topico_10	0.7297	1.3704
mpf_topico_11	0.0931	10.7377
mpf_topico_12	0.3758	2.6608
mpf_topico_13	0.4294	2.3290
mpf_topico_14	1.0600	0.9434
mpf_topico_15	0.9225	1.0841
mpf_topico_16	768.2732	0.0013
mpf_topico_17	0.0006	1732.5565
mpf_topico_18	1.6670	0.5999
rel_topico_0	0.9811	1.0193
rel_topico_1	0.7766	1.2877
rel_topico_2	0.7388	1.3535
rel_topico_3	1.0705	0.9341
rel_topico_4	0.9939	1.0061
rel_topico_5	0.5271	1.8971
rel_topico_6	0.6944	1.4400
rel_topico_7	0.5018	1.9927
rel_topico_8	0.5323	1.8788
rel_topico_9	0.5386	1.8565
rel_topico_10	0.3639	2.7481
rel_topico_11	0.4725	2.1164

Continua na próxima página

Tabela 10 – *Continuação da página anterior*

Variável	Odds Ratio	Inverse Odds Ratio
rel_topico_12	0.6576	1.5207
rel_topico_13	0.7348	1.3609
rel_topico_14	1.6057	0.6228
rel_topico_15	0.7533	1.3275
rel_topico_16	0.5563	1.7976
rel_topico_17	0.4680	2.1366
rel_topico_18	3.6724	0.2723
genero_mas	1.6219	0.6166
formacao_direito	2.4777	0.4036
formacao_economia	9.3221	0.1073
servidor_publico_sim	0.8125	1.2308

Fonte: Elaborado pelo autor.

APÊNDICE G – SUMÁRIO ESTATÍSTICO - VIF

Tabela 11 – Fator de inflação da variância (VIF)

índice	variável	VIF
0	mpf_topico_0	1.228820
1	mpf_topico_1	1.158550
2	mpf_topico_2	1.097044
3	mpf_topico_3	1.062285
4	mpf_topico_4	1.032685
5	mpf_topico_5	1.048239
6	mpf_topico_6	1.045270
7	mpf_topico_7	1.032198
8	mpf_topico_8	1.020935
9	mpf_topico_9	1.024739
10	mpf_topico_10	1.020937
11	mpf_topico_11	1.048282
12	mpf_topico_12	1.019893
13	mpf_topico_13	1.009936
14	mpf_topico_14	1.007853
15	mpf_topico_15	1.005871
16	mpf_topico_16	1.009855
17	mpf_topico_17	1.013190
18	mpf_topico_18	1.011723
19	rel_topico_0	1.180112
20	rel_topico_1	1.059949
21	rel_topico_2	1.060842
22	rel_topico_3	1.045437
23	rel_topico_4	1.014256
24	rel_topico_5	1.025566
25	rel_topico_6	1.018830
26	rel_topico_7	1.019977
27	rel_topico_8	1.011429
28	rel_topico_9	1.011470
29	rel_topico_10	1.016975
30	rel_topico_11	1.010365

Continua na próxima página

Tabela 11 – *Continuação da página anterior*

índice	variável	VIF
31	rel_topico_12	1.011413
32	rel_topico_13	1.007843
33	rel_topico_14	1.006044
34	rel_topico_15	1.014156
35	rel_topico_16	1.004080
36	rel_topico_17	1.011428
37	rel_topico_18	1.002863
38	genero_mas	5.459706
39	formacao_direito	2.946222
40	formacao_economia	2.499684
41	servidor_publico_sim	3.393398

Fonte: Elaborado pelo autor.

APÊNDICE H – SUMÁRIO ESTATÍSTICO - RESUMO DOS COEFICIENTES

Tabela 12 – Resumo dos coeficientes do modelo

índice	variável	coeficiente	p-valor	magnitude
0	mpf_topico_11	-2.375370	0.000000	2.375370
1	formacao_economia	2.233990	0.000000	2.233990
2	const	-1.423340	0.000000	1.423340
3	rel_topico_10	-1.012220	0.003570	1.012220
4	mpf_topico_12	-0.979830	0.009410	0.979830
5	formacao_direito	0.908920	0.000000	0.908920
6	mpf_topico_6	-0.881350	0.000000	0.881350
7	mpf_topico_13	-0.846740	0.015240	0.846740
8	mpf_topico_2	-0.833970	0.000000	0.833970
9	mpf_topico_5	-0.806870	0.000000	0.806870
10	mpf_topico_4	0.771200	0.000240	0.771200
11	rel_topico_11	-0.751200	0.036950	0.751200
12	mpf_topico_3	0.722650	0.000000	0.722650
13	rel_topico_7	-0.690210	0.006840	0.690210
14	rel_topico_5	-0.640900	0.002550	0.640900
15	mpf_topico_8	0.522940	0.047690	0.522940
16	mpf_topico_7	-0.521900	0.006760	0.521900
17	genero_mas	0.483800	0.000000	0.483800
18	rel_topico_2	-0.302960	0.019260	0.302960
19	rel_topico_1	-0.253080	0.046870	0.253080
20	mpf_topico_1	0.235550	0.009140	0.235550
21	servidor_publico_sim	-0.207740	0.001110	0.207740

Fonte: Elaborado pelo autor.

APÊNDICE I – SUMÁRIO ESTATÍSTICO - TESTES

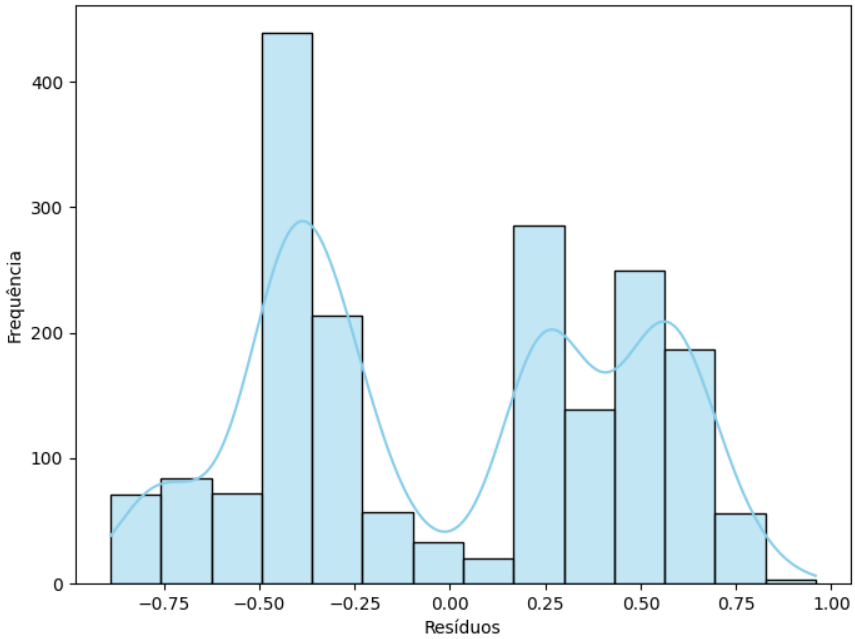
Tabela 13 – Testes de Ajuste do Modelo

Pressuposto	**Teste**	**Estat.**	**df**	**Valor-p**	**Rejeita H0?**
Bondade do ajuste	Hosmer-Lemeshow (χ^2)	9.276	18	0.953	Não
Normalidade dos resíduos	Shapiro-Wilk (W)	0.919	-	<0.001	Sim

Fonte: Elaborado pelo autor.

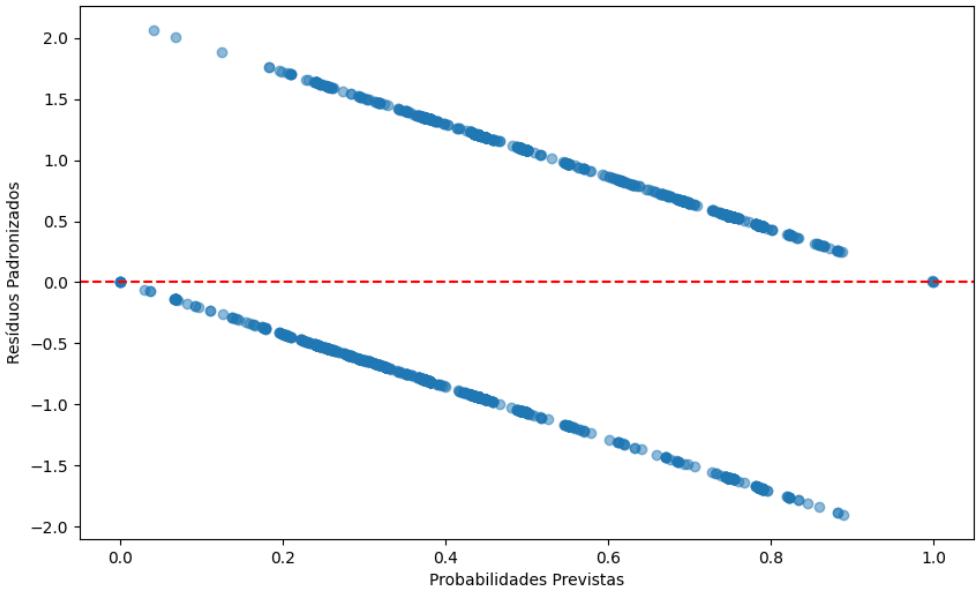
APÊNDICE J – ANÁLISE GRÁFICA DOS RESÍDUOS

Figura 31 – Histograma dos resíduos do modelo logístico



Fonte: elaborado pelo autor.

Figura 32 – Resíduos padronizados vs. probabilidades previstas



Fonte: elaborado pelo autor.