UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MATHEUS ALVES DOS SANTOS

MODELAGEM DE TÓPICOS NA ESTIMATIVA DE PONTOS IDEAIS BASEADOS EM DISCURSOS DE PARLAMENTARES BRASILEIROS

CAMPINA GRANDE - PB
2024

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Modelagem de tópicos na estimativa de pontos ideais baseados em discursos de parlamentares brasileiros

Matheus Alves dos Santos

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Metodologia e Técnicas da Computação

Nazareno Ferreira de Andrade (Orientador)
Fabio Jorge Almeida Morais (Orientador)

Campina Grande, Paraíba, Brasil

S237m Santos, Matheus Alves dos.
Modelagem de tópicos na estimativa de pontos ideais baseados em discursos de parlamentares brasileiros / Matheus Alves dos Santos. – Campina Grande, 2024.
93 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2024.
"Orientação: Prof. Dr. Nazareno Ferreira de Andrade, Prof. Dr. Fabio Jorge Almeida Morais".
Referências.

1. Processamento de Linguagem Natural. 2. Modelagem de Tópicos Latentes. 3. Estimativa de Pontos Ideais. 4. Política – Câmara dos Deputados – Discursos. I. Andrade, Nazareno Ferreira de. II. Morais, Fabio Jorge Almeida. III. Título.

CDU 004.432.45(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM CIENCIA DA COMPUTACAO
Rua Aprígio Veloso, 882, Edifício Telmo Silva de Araújo, Bloco CG1, - Bairro Universitário, Campina Grande/PB, CEP 58429-900
Telefone: 2101-1122 - (83) 2101-1123 - (83) 2101-1124
Site: http://computacao.ufcg.edu.br - E-mail: secretaria-copin@computacao.ufcg.edu.br / copin@copin.ufcg.edu.br

FOLHA DE ASSINATURA PARA TESES E DISSERTAÇÕES

MATHEUS ALVES DOS SANTOS

MODELAGEM DE TÓPICOS NA ESTIMATIVA DE PONTOS IDEAIS BASEADOS EM DISCURSOS DE PARLAMENTARES BRASILEIROS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação como pré-requisito para obtenção do título de Mestre em Ciência da Computação.

Aprovada em: 16/02/2024

Prof. Dr. NAZARENO FERREIRA DE ANDRADE, Orientador, UFCG

Prof. Dr. FÁBIO JORGE ALMEIDA MORAIS, Orientador, UFCG

Prof. Dr. CLÁUDIO ELÍZIO CALAZANS CAMPELO, Examinador Interno, UFCG

Prof. Dr. FLAVIO VINICIUS DINIZ DE FIGUEIREDO, Examinador Externo, UFMG

Documento assinado eletronicamente por **FABIO JORGE ALMEIDA MORAIS**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 21/02/2024, às 08:15, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **CLAUDIO ELIZIO CALAZANS CAMPELO**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 27/02/2024, às 10:32, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **NAZARENO FERREIRA DE ANDRADE**, **PROFESSOR 3 GRAU**, em 27/02/2024, às 15:06, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **4199024** e o código CRC **388BF313**.

Agradecimentos

Agradeço à minha família, sobretudo à minha mãe Suely e ao meu pai Francisco, pela dedicação incansável, pela crença em meu potencial e por todo o incentivo à minha jornada acadêmica. À minha irmã Dannyelle, por sua cumplicidade e por seu apoio incondicional. E ao meu sobrinho Heitor, por me manter esperançoso quanto ao futuro.

Agradeço a Tarso Jabbes por seu carinho, tranquilidade e otimismo, mas também por tantas horas dedicadas à leitura dos meus rascunhos. Você tornou os desafios mais leves e as conquistas mais significativas.

Agradeço aos meus orientadores, Nazareno e Fábio, pelos ensinamentos, pela dedicação e pela paciência durante todo o processo de orientação desta pesquisa. Tê-los como mentores moldou não apenas este trabalho, mas também meu desenvolvimento pessoal, acadêmico e profissional.

Agradeço a Victor Freire por nossas discussões sobre este trabalho, por suas valiosas contribuições e por ouvir meus desabafos acadêmicos. Estendo meus agradecimentos a Daniel Duarte, Deilton Lopes, Juan Barros, Melissa Marques, Rhuan Isllan e Samara Sampaio por cada risada, memória e estresse que compartilhamos. E aos demais amigos e amigas que, de perto ou de longe, no cotidiano ou só de vez em quando, me acompanham e me conduzem por essa vida.

Agradeço à equipe da Odd Data & Design Studio pela colaboração e pela compreensão que me demonstraram nesses últimos anos. Vocês me inspiram profundamente.

Agradeço aos professores, aos servidores e aos demais funcionários que compõem o Departamento de Sistemas e Computação pelo esforço, pela dedicação e pela solicitude.

Por fim, agradeço à Fundação de Apoio à Pesquisa do Estado da Paraíba pelo apoio financeiro à condução desta pesquisa. Que nos mantenhamos sempre em defesa do acesso à educação pública, gratuita e de qualidade.

A todos, muito obrigado!

Resumo

Para a construção de democracias fortes e verdadeiramente representativas é de suma importância que a sociedade civil seja capaz de compreender e monitorar a atuação política de seus representantes. Entretanto, apesar dos notáveis avanços na transparência governamental, a população brasileira tende a não acompanhar as atividades parlamentares. Esse cenário se consolida em decorrência de múltiplas questões socioculturais, mas também da intrínseca complexidade do Poder Legislativo. Assim, são necessários métodos e ferramentas que proporcionem acesso à informação para a sociedade civil e, mais do que isso, que colaborem com seu entendimento e uso dessas informações. Nesse âmbito, as técnicas de Processamento de Linguagem Natural têm se difundido na análise dos volumosos conjuntos de dados textuais que permeiam o contexto político, como os discursos ou as proposições de lei. Neste trabalho, avaliamos o uso individual e conjunto de duas técnicas do estado-da-arte para a modelagem de tópicos latentes e a estimativa de pontos ideais baseados em texto, aplicando-as à caracterização dos discursos e posicionamentos políticos de parlamentares brasileiros. Em específico, utilizamos os modelos BERTopic e *Text-Based Ideal Point* para analisar a 55ª e a 56ª Legislaturas da Câmara dos Deputados, abrangendo o período de 2015 a 2022. Durante esse processo, também construímos e publicamos uma base de dados abertos contendo as transcrições dos discursos em eventos realizados por essa casa legislativa entre 2003 e 2022. A avaliação das técnicas adotadas teve caráter quantitativo e qualitativo, considerando métricas como a coerência e a diversidade de tópicos latentes, mas também a validade aparente e o comparativo com a percepção de especialistas da Ciência Política. O desempenho dos modelos nessa avaliação e as análises baseadas em seus resultados apontam essas técnicas como viáveis, promissoras e capazes de fundamentar novos estudos políticos no cenário brasileiro. Contudo, devido às características inerentes ao Poder Legislativo de nosso país, nossas estimativas divergem da interpretação original dos pontos ideais e, substituindo a tradicional dicotomia esquerda-direita, demonstram o quão "ideológicos" ou "pragmáticos" são os indivíduos analisados.

Palavras-chave: Processamento de Linguagem Natural; Modelagem de Tópicos Latentes; Estimativa de Pontos Ideais; Política; Câmara dos Deputados.

Abstract

Strong and truly representative democracies can be built only when civil society can understand and monitor the political activities of its representatives. However, despite remarkable progress towards government transparency, people in Brazil tend not to be aware of parliamentary activities. This scenario emerges from multiple socio-cultural aspects, as well as from the structural complexity surrounding the Legislative Branch. Therefore, methods and tools that provide access to information for civil society and, more importantly, contribute to its understanding and usage of such information are essential. In this regard, Natural Language Processing techniques have been increasingly employed to analyze huge textual datasets surrounding political contexts, such as speeches and law proposals. In this study, we evaluated both the individual and the combined use of two state-of-the-art techniques for latent topic modeling and text-based ideal point estimation, applying them to characterize the speeches and political views of Brazilian parliamentarians. Specifically, we used the BERTopic and the Text-Based Ideal Point models to analyze the 55th and 56th Legislatures of the Brazilian Chamber of Deputies, spanning the period from 2015 to 2022. In this process, we also built and published an open database containing speech transcriptions from events held by this legislative house between 2003 and 2022. The evaluation of these techniques was quantitative and qualitative, considering metrics such as coherence and diversity of latent topics, but also the face validity and the comparison to Political Science experts' opinions. The performance of these models and our analysis of their results suggest that these techniques are viable, promising, and suitable for new political studies in Brazil. Nevertheless, due to the inherent features of the Brazilian Legislative Branch, our estimations differ from the original interpretations regarding these ideal points by replacing the traditional left-right dichotomy and demonstrating how "ideological" or "pragmatic" the analyzed individuals are.

Keywords: Natural Language Processing; Topic Modeling; Ideal Point Estimation; Politics; Chamber of Deputies.

Sumário

Lista de Abreviaturas e Siglas

ABCP	Associação Brasileira de Ciência Política
BERT	*Bidirectional Encoder Representations from Transformers*
CSV	*Comma-Separated Values*
c-TF-IDF	*Class-based Term Frequency-Inverse Document Frequency*
GPT	*Generative Pre-Trained Transformer*
HDBSCAN	*Hierarchical Density-Based Spatial Clustering of Applications with Noise*
HTML	*HyperText Markup Language*
LDA	*Latent Dirichlet Allocation*
LLMs	*Large Language Models*
PDF	*Portable Document Format*
PLN	Processamento de Linguagem Natural
TBIP	*Text-Based Ideal Point*
TSE	Tribunal Superior Eleitoral
TF-IDF	*Term Frequency-Inverse Document Frequency*
UMAP	*Uniform Manifold Approximation and Projection for Dimension Reduction*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A República Federativa do Brasil é, segundo a Constituição Federal vigente, um Estado Democrático de Direito constituído por três poderes independentes e harmônicos entre si: o Legislativo, o Executivo e o Judiciário [14]. Derivada do princípio da *trias politica*, essa tripartição de poderes define o Legislativo como o poder responsável pela elaboração e revisão do conjunto de leis que rege tanto a vida das pessoas quanto o próprio funcionamento do Estado [47].

O Poder Legislativo brasileiro é composto, em sua esfera federal, por duas casas legislativas: a Câmara dos Deputados e o Senado Federal. Em cada legislatura, 513 deputados federais e 81 senadores — com 2 suplentes cada — se distribuem entre os plenários e as comissões (permanentes ou temporárias) dessas duas casas legislativas. Nesses ambientes, os parlamentares se reúnem para debater, votar e analisar os aspectos técnicos e/ou legais das proposições de lei. Existem, ainda, diversos outros espaços e funções que permeiam o Poder Legislativo brasileiro, como a mesa diretora e as bancadas parlamentares.

Embora sejam escolhidos por voto popular e tenham profundo impacto sobre o cotidiano de milhões de brasileiros e brasileiras, acompanhar os numerosos indivíduos que compõem o Congresso Nacional não é uma tarefa trivial, especialmente no que se refere àqueles que possuem pouca visibilidade midiática ou menor influência no cenário nacional. A Ciência Política é a ciência social que se dedica ao estudo dessas instituições políticas, mas também das figuras, dos acontecimentos e das ideias que permeiam esse contexto [12]. Nos últimos anos, pesquisas e ferramentas cívicas baseadas em dados têm se difundido nessa área, visando facilitar o monitoramento e a compreensão da sociedade civil acerca da atuação dos

parlamentares. O Radar Legislativo[1], o Elas no Congresso[2] e o Parla[3] são exemplos notáveis desse tipo de ferramenta em nosso país.

Contudo, as abordagens baseadas em dados ou o uso de métodos estatísticos e computacionais não são fenômenos recentes na Ciência Política. A princípio, esses métodos eram utilizados quase exclusivamente em eleições, nas quais se tornaram fundamentais para as estratégias de campanha e de divulgação com base em características sociodemográficas do eleitorado. Essas abordagens foram disseminadas, aprimoradas e se estenderam para várias outras tarefas, como a comparação de alinhamento político e a checagem de fatos [33]. Nomeada como Política Computacional, consolidou-se uma área de estudo dedicada à aplicação de métodos computacionais sobre conjuntos de dados para divulgação, persuasão e mobilização com o objetivo de eleger, promover ou se opor a candidatos, políticas ou legislações [72]. Assim como nas ferramentas cívicas supracitadas, são as análises quantitativas que se destacam nos estudos da Política Computacional e viabilizam o melhor entendimento do contexto político vigente.

No cenário brasileiro, é perceptível como as ferramentas e pesquisas relacionadas à atuação política de parlamentares se concentram majoritariamente nas votações do Congresso Nacional. No entanto, ainda que essas votações explicitem o posicionamento de cada parlamentar quanto às propostas em pauta, elas não são seu único meio de incidir sobre o Poder Legislativo. Nos plenários e comissões, os parlamentares discursam e debatem com seus pares não só acerca dessas propostas, mas sobre tantos outros temas pertinentes à época. Os discursos constituem dados textuais que conseguem, portanto, evidenciar aspectos e ênfases que informações binárias, como os votos favoráveis ou contrários, não são capazes de representar.

Porém, mesmo com o notório desempenho do Brasil em termos de acesso à informação e disponibilização de dados abertos, ainda persistem diversos desafios para o acesso e o uso desses dados textuais. Além das transcrições não serem necessariamente produzidas para todos os eventos da Câmara dos Deputados e do Senado Federal, os documentos não são publicados em formato aberto, podem apresentar ruídos (como erros de digitação) e estar mal estruturados ou, até mesmo, incompletos. Ademais, considerando o volume de dados gerado,

[1] https://www.radarlegislativo.org

[2] https://www.elasnocongresso.com.br

[3] https://parla.camara.leg.br

a análise por métodos não computacionais torna-se inviável mesmo para intervalos de tempo relativamente curtos.

Nesse sentido, além da crescente demanda por acesso a essas informações, há a necessidade de métodos que automatizem o processamento, a análise e o reconhecimento de padrões nesses discursos. Atualmente, as principais abordagens com esse propósito utilizam modelos de Aprendizagem de Máquina e Processamento de Linguagem Natural. Tanto no cenário brasileiro quanto no internacional, esses algoritmos têm se consolidado enquanto ferramenta para, por exemplo, identificar ênfases temáticas em discursos de parlamentares, explorar a agenda política de parlamentos, associar discursos a filiações partidárias e, inclusive, identificar e analisar preferências políticas de legisladores, juízes e eleitores [19] [32] [76] [78].

Dentre os métodos adotados para realizar atividades dessa natureza, estão a modelagem de tópicos latentes e a estimativa de pontos ideais. A modelagem de tópicos é uma tarefa de Processamento de Linguagem Natural que visa identificar e extrair automaticamente os temas abstratos que são abordados num conjunto de documentos textuais específico. Tipicamente, essas técnicas se baseiam em distribuições probabilísticas e diferenças de vocabulário, mas as abordagens modernas têm extrapolado as características exclusivamente sintáticas e, em certa medida, incorporado a semântica das palavras em seu aprendizado [51] [57].

A estimativa de pontos ideais, por sua vez, é uma abordagem analítica da Ciência Política que visa representar quantitativamente o posicionamento político de indivíduos ou grupos. Desse modo, os valores estimados podem ser distribuídos num espaço teórico (uni ou multidimensional) tal que as distâncias entre os valores — denominados pontos ideais — correspondam ao quão similares (ou dissimilares) são os posicionamentos dos atores políticos, permitindo a identificação de alinhamentos partidários, coalizões, polarizações etc. Enquanto representações numéricas para as preferências político-ideológicas, os pontos ideais simplificam análises comparativas entre atores políticos e tornam intuitivas as respostas para questionamentos como “A atuação política de X se assemelha mais à de Y ou à de Z?” ou “O parlamentar X está mais à esquerda ou à direita do espectro político?”. Essas estimativas costumam ser baseadas em votações nominais do Poder Legislativo, o que traz inerentes limitações quanto ao período de tempo e às funções desempenhadas pelos atores analisados. Buscando solucionar esses entraves, estudos recentes têm explorado novas técnicas e fontes de dados para estimar os pontos ideais, particularmente aplicando as técnicas de Processamento

de Linguagem Natural aos discursos legislativos.

Tanto a modelagem de tópicos latentes quanto a estimativa de pontos ideais têm consistentemente apresentado bons desempenhos nos cenários nacional e internacional, consolidando-se como ferramentas importantes para a Ciência Política e a Política Computacional no decorrer dos últimos anos. Entretanto, há forte predominância das técnicas mais tradicionais no que se refere ao uso dessas tarefas de Processamento de Linguagem Natural na esfera política brasileira. Embora produzam resultados satisfatórios, essas abordagens convencionais não são capazes de explorar a semântica dos dados textuais ou os aspectos que fogem às votações legislativas, por exemplo. Não obstante, os principais trabalhos nesse contexto se dedicam à análise de períodos encerrados pela 54ª Legislatura do Congresso Nacional (que se estendeu de 2011 a 2014), sugerindo certa defasagem temporal.

Além disso, não foram identificados estudos que vinculem outras tarefas de Processamento de Linguagem Natural, como a modelagem de tópicos latentes, à estimativa de pontos ideais baseados em dados textuais. Essa associação poderia contribuir com seu desempenho ou, ao menos, para a análise de seus resultados. No Brasil, em que centenas de pessoas e dezenas de espaços compõem cada legislatura, o uso individual e/ou conjunto dessas técnicas modernas se mostra bastante promissor e, potencialmente, capaz de enriquecer o entendimento de especialistas e da própria sociedade civil acerca de nossos representantes no Poder Legislativo.

Diante do exposto, adotando técnicas do estado-da-arte para a modelagem de tópicos latentes e para a estimativa de pontos ideais baseados em texto, neste trabalho analisamos os discursos de deputados(as) federais brasileiros(as) para caracterizar seus respectivos posicionamentos e atuações políticas, visando produzir novos *insights* nesse âmbito. Em específico, recorremos aos modelos BERTopic e *Text-Based Ideal Point* (respectivamente). Além do uso individual dessas técnicas, validamos a modelagem de tópicos latentes enquanto ferramenta para a seleção de documentos textuais mais relevantes no contexto político. Através destes, foi possível aprimorar os resultados produzidos pela estimativa de pontos ideais e facilitar as análises sobre o Poder Legislativo brasileiro. Por sua vez, os pontos ideais estimados nos permitiram caracterizar o quão "ideológicos" ou "pragmáticos" são os deputados(as) federais e os partidos políticos que compõem a 55ª e a 56ª Legislaturas da Câmara dos Deputados.

1.1 Objetivos e escopo

O objetivo principal deste trabalho é avaliar o uso de técnicas modernas de Processamento de Linguagem Natural na caracterização de discursos e posicionamentos políticos de parlamentares brasileiros. Mais precisamente, a modelagem de tópicos latentes com o BERTopic e a estimativa de pontos ideais baseada em dados textuais com o *Text-Based Ideal Point* foram as técnicas avaliadas neste estudo. Dedicamo-nos a analisar as transcrições dos discursos de deputados e deputadas federais em eventos realizados pela Câmara dos Deputados durante a 55ª e a 56ª Legislaturas — que se estendem de 2015 a 2018 e de 2019 a 2022, respectivamente. As transcrições utilizadas foram extraídas dos documentos disponibilizados pela Câmara dos Deputados através de páginas web.

Considerando o objetivo principal supracitado, visamos especificamente construir e disponibilizar uma base de dados abertos contendo as transcrições de eventos realizados pela Câmara dos Deputados. É importante destacar que essa base de dados abrange também os discursos de indivíduos não parlamentares e estende o período analisado neste estudo, agregando informações sobre a 52ª, a 53ª e a 54ª Legislaturas da Câmara dos Deputados. Isto é, englobando duas décadas inteiras (2003 a 2022). Além deste, nossos objetivos específicos também incluem a produção de informações inéditas acerca dos discursos de parlamentares brasileiros, analisando-os não apenas por indivíduo, mas também agregando-os conforme filiações partidárias e características sociodemográficas.

Por fim, indo além do uso individual das técnicas de Processamento de Linguagem Natural previamente mencionadas, objetivamos validar um método para aplicá-las conjuntamente: adotando os resultados da modelagem de tópicos latentes na filtragem dos documentos textuais com maior relevância para o contexto político que, posteriormente, sejam utilizados e produzam melhorias na estimativa de pontos ideais para os parlamentares analisados.

1.2 Contribuições

As contribuições deste trabalho podem ser descritas sob três perspectivas distintas. No âmbito social, o ferramental de extração e a base de dados abertos que foram construídos e disponibilizados colaboram com a publicização de informações governamentais de interesse

da sociedade. Junto a esses dados, as análises apresentadas neste documento contribuem com a compreensão da sociedade civil brasileira quanto às atividades de seus representantes na esfera federal do Poder Legislativo, uma vez que nossos resultados simplificam a representação das características políticas inerentes aos discursos de deputados(as) e, ainda, proporcionam novos *insights* nesse contexto.

Já no âmbito da Ciência Política e da Política Computacional, esses dados e ferramentas podem auxiliar as pessoas pesquisadoras (e diferentes profissionais dessas áreas) na condução de seus próprios estudos. Ademais, as técnicas e os métodos apresentados neste documento simplificam a análise dos volumosos conjuntos de dados textuais oriundos do cenário político, tornando mais eficiente o acompanhamento da atuação política dos parlamentares.

Por fim, no âmbito do Processamento de Linguagem Natural, demonstramos e avaliamos duas técnicas do estado-da-arte no domínio dos discursos de deputados e deputadas federais brasileiros, bem como validamos a modelagem de tópicos enquanto instrumento para a seleção dos documentos com maior relevância para a caracterização de atuações políticas. Nesse sentido, também propomos a filtragem do *corpus* usando os tópicos latentes identificados pelo BERTopic como método para a melhoria dos pontos ideais estimados pelo *Text-Based Ideal Point*.

Destacamos ainda que, durante as etapas iniciais desta pesquisa, também conduzimos experimentos com o modelo *Latent Dirichlet Allocation* [10]. Nestes, aplicamos a modelagem de tópicos a discursos em eventos de comissões permanentes da Câmara dos Deputados. Em razão do menor volume de dados textuais, esses experimentos puderam expandir o período analisado, abrangendo os anos de 2008 a 2020. Apesar de não terem sido incluídos neste documento, esses resultados iniciais estão apresentados (e integralmente discutidos) em duas publicações prévias. São elas:

- SANTOS, Matheus A.; ANDRADE, Nazareno; MORAIS, Fábio. Topic Modeling of Discussions in the Standing Committees of the Brazilian Chamber of Deputies. **Journal of Information and Data Management**, *[S. l.]*, v. 13, n. 6, 2023. DOI: https://doi.org/10.5753/jidm.2022.2705.
- SANTOS, Matheus A.; ANDRADE, Nazareno; MORAIS, Fábio. Topic Modeling of Committee Discussions in the Brazilian Chamber of Deputies. *In:* Symposium on Knowledge Discovery, Mining and Learning (KDMILE), 9., 2021, Rio de Janeiro.

Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021 . p. 49-56. DOI: https://doi.org/10.5753/kdmile.2021.17460.

1.3 Estrutura do documento

O conteúdo remanescente deste documento está estruturado conforme descrito a seguir. No Capítulo 2, introduzimos os conceitos fundamentais e o arcabouço teórico que permeiam e sustentam esta pesquisa, bem como discutimos trabalhos correlatos que se dedicam à análise de atuações políticas através de dados textuais. Em seguida, no Capítulo 3, descrevemos o método de extração e as características dos conjuntos de dados que foram produzidos e adotados no decorrer deste estudo. Nos Capítulos 4 e 5 são apresentadas, respectivamente, as etapas de modelagem de tópicos e de estimativa de pontos ideais que compõem nossa pesquisa. Nestes, detalhamos as técnicas adotadas, descrevemos o treinamento dos modelos e avaliamos seus desempenhos quanto às métricas de avaliação e/ou em contraste à percepção de especialistas da Ciência Política. Posteriormente, aplicamos os resultados obtidos na análise da atuação política de parlamentares e de partidos políticos representados em nossos conjuntos de dados. Por fim, no Capítulo 6, discorremos sobre as principais conclusões desta dissertação, suas limitações e as perspectivas para trabalhos futuros.

Capítulo 2

Fundamentação Teórica e Trabalhos Relacionados

Neste capítulo, discorremos sobre os principais conceitos que fundamentam nossa pesquisa, buscando eliminar eventuais ambiguidades e contribuir com a adequada compreensão deste estudo. Também discutimos trabalhos correlatos que fazem uso de técnicas de Processamento de Linguagem Natural para analisar a atuação política de indivíduos parlamentares e/ou não parlamentares. Esses trabalhos foram identificados através de uma revisão bibliográfica de caráter narrativo e abrangem os cenários brasileiro e internacional.

2.1 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) — em inglês, *Natural Language Processing* — é a área de conhecimento dedicada ao estudo e aplicação de técnicas computacionais com o objetivo de compreender e manipular linguagens naturais, isto é, as linguagens que surgiram e evoluíram naturalmente entre os seres humanos [16]. Tipicamente, se entende PLN como uma intersecção entre as áreas da Ciência da Computação, Inteligência Artificial e Linguística.

As primeiras pesquisas dessa natureza tinham foco na tradução automática de texto e, à despeito de seus resultados pouco satisfatórios, foram capazes de fomentar a criação de novas abordagens conexionistas, estatísticas e simbolistas [42]. No decorrer do tempo, tarefas de PLN cada vez mais complexas foram concebidas, aprimoradas e adotadas em diferentes

domínios. São exemplos o reconhecimento de voz [28] e a síntese de fala [38]. Esses avanços foram possíveis, principalmente, devido à redução do custo de hardware e ao aumento da capacidade de processamento computacional.

Nos últimos anos, as técnicas de PLN também se difundiram nas áreas da Ciência Política e da Política Computacional. Essa disseminação é bastante compreensível, uma vez que os discursos são uma importante ferramenta de atuação política e, no âmbito do Poder Legislativo, se somam a diversos outros tipos de documentos (oficiais ou não), como as proposições de lei e as notas à imprensa. No entanto, é notável que essas pesquisas se concentram majoritariamente no Norte Global e, mesmo quando fogem a esse padrão, costumam se originar em países cujos regimes democráticos são particularmente fortes [33].

2.2 Pré-processamento e representação de dados textuais

No contexto de dados textuais, os registros (ou observações) costumam ser denominados "documentos", enquanto uma coleção de documentos é denominada "*corpus*". Assim, podemos definir o pré-processamento de dados textuais como o conjunto de operações aplicadas aos documentos para adequá-los a análises e/ou tarefas de PLN. Geralmente, essas operações priorizam padronizar o conjunto de dados, reduzir sua dimensionalidade e remover informações desnecessárias [3]. Abaixo, elencamos e descrevemos resumidamente as operações de pré-processamento de dados textuais que serão mencionadas no decorrer deste documento.

- **Correção de espaçamentos:** Substituição de caracteres especiais para espaçamento de texto (como `\t` ou `\r`) por espaçamentos simples.
- **Padronização de capitalização:** Conversão de caracteres alfabéticos dos documentos para uma mesma caixa (alta ou baixa).
- **Padronização de datas:** Adequação de valores que representem datas para um mesmo formato. Em nossos dados, adotamos o formato `YYYY-MM-DD`.
- **Remoção de duplicatas:** Remoção dos documentos que se repetem no *corpus*, garantindo que cada documento ocorra uma única vez.
- **Remoção de pontuação:** Remoção de caracteres não alfanuméricos dos documentos.

- **Remoção de *stopwords*:** Remoção de palavras com pouca ou nenhuma contribuição semântica para os documentos em que estão inseridas [37]. Essas palavras, denominadas *stopwords*, podem ser definidas a partir do *corpus* ou de classes gramaticais.
- **Tokenização:** Decomposição dos documentos em unidades menores. Essas unidades são chamadas de *tokens* e podem ser formadas por conjuntos de palavras, denominados n-gramas, ou por segmentos dessas palavras, denominados subpalavras.

Mesmo após a execução das operações de pré-processamento, os dados textuais ainda preservam características inerentes (como o vocabulário potencialmente infinito) que podem tornar seu uso bastante desafiador. Por essa razão, é comum convertê-los para um formato numérico. Esse processo, denominado vetorização, transforma cada documento num vetor de valores numéricos (discretos ou contínuos). As representações numéricas mais tradicionais se baseiam na frequência das palavras, mas tendem a produzir dados esparsos, de alta dimensionalidade e capturar pouca ou nenhuma semântica dos textos.

O *Term Frequency-Inverse Document Frequency* (TF-IDF) é uma dessas abordagens tradicionais e consiste em produzir uma matriz M de frequências relativas, com formato $D \times T$, tal que D é o número de documentos e T é o número de *tokens* distintos no *corpus* [61]. Desse modo, o valor M_{ij} será definido pelo quociente entre os números de ocorrências do j-ésimo *token* em dois contextos diferentes: no i-ésimo documento e no *corpus* inteiro. Assim, se produz uma representação dos documentos que atribuirá maior importância aos termos mais raros e menor importância àqueles que são mais frequentes.

Por outro lado, com o surgimento dos *Neural Network Language Models* [7] e do modelo Word2Vec [45], a vetorização através de *embeddings* se tornou proeminente e cada vez mais adotada nas tarefas de PLN. Os *embeddings* são vetores densos, distribuídos, de comprimento fixo e construídos a partir de estatísticas de coocorrência dos termos [2]. Esses vetores podem ser adaptados para representar palavras, sentenças ou, até mesmo, documentos inteiros. Além de reduzir drasticamente a dimensionalidade dos dados textuais sem torná-los esparsos, os *embeddings* são capazes de capturar as características semânticas dos documentos. Não obstante, as limitações identificadas nesses modelos iniciais têm sido mitigadas (ou solucionadas) por técnicas mais recentes. O fastText [36], por exemplo, utiliza subpalavras como seus *tokens* e, assim, evita que as palavras ausentes no vocabulário de seu *corpus* de treinamento se tornem um entrave.

Nesse cenário de melhoria contínua, surge uma nova arquitetura de redes neurais baseada em mecanismos de autoatenção (ou *self-attention*): a arquitetura Transformer [74]. Esses mecanismos permitem que o modelo se concentre seletivamente nas diferentes partes da sequência de entrada e que, assim, estabeleça dependências globais entre as sequências de entrada e de saída. Dispensando as recorrências e convoluções que são típicas em arquiteturas anteriores para redes neurais, os *transformers* se mostraram eficientes e altamente paralelizáveis. Essa nova arquitetura se disseminou rapidamente na área de PLN e foi, inclusive, adotada durante a criação dos modelos mais relevantes do estado-da-arte: o *Generative Pre-Trained Transformer* (GPT) [58] e o *Bidirectional Encoder Representations from Transformers* (BERT) [20].

Em específico, o BERT é um modelo de aprendizagem não supervisionada, fundamentado em redes neurais profundas e que se destaca por seu método bidirecional para criação de *embeddings*. Isso significa dizer que, ao avaliar um *token* específico, o modelo considera tanto os *tokens* que lhe antecedem quanto os que lhe sucedem no documento. Ainda, cerca de 15% das palavras são mascaradas durante o treinamento, de modo que o próprio modelo precise prevê-las. Essas técnicas contribuem com a captura de nuances semânticas e com a compreensão do contexto das palavras, colaborando com o proeminente desempenho do BERT (e de suas variantes) nas diversas tarefas de PLN — como a análise de sentimentos [15] ou o reconhecimento de entidades nomeadas [54]. Além disso, uma vez que o BERT tenha sido pré-treinado em grandes volumes de dados textuais, não se faz necessário produzir um novo modelo para cada tipo de tarefa. Essa adaptação requere apenas uma etapa de ajuste fino — que treinará a camada final desse modelo — e pode ser feita com um *corpus* substancialmente menor.

2.3 Modelagem de tópicos latentes

A modelagem de tópicos é uma tarefa de PLN que, através de aprendizagem semi-supervisionada ou não supervisionada, busca identificar estruturas semânticas implícitas numa coleção de documentos de interesse [9]. Tipicamente, esses modelos fazem uso da estrutura sintática dos documentos para extrair conhecimento acerca de sua semântica e, assim, distinguir os tópicos latentes associados ao *corpus* analisado. Entretanto, os tópicos

identificados são abstratos e se faz necessário submetê-los à avaliação humana, para que sejam validados e rotulados de maneira objetiva e interpretável.

Entre os algoritmos tradicionais para a modelagem de tópicos, o *Latent Dirichlet Allocation* (LDA) é aquele que possui maior destaque. Esse modelo probabilístico utiliza uma hierarquia de três níveis para descrever os documentos de um *corpus* como combinações finitas de tópicos implícitos [10]. Isso significa dizer que, através de distribuições probabilísticas de Dirichlet, cada documento será representado por um vetor numérico de K associações tópico-documento e, por sua vez, cada valor nesse vetor representará a proporção do documento que está associada ao K-ésimo tópico latente. Apesar de suas inerentes limitações (como o vocabulário fixo e o número de tópicos pré-estabelecido), o LDA apresenta bom desempenho nos mais diversos domínios, abrangendo desde matérias jornalísticas [41] até artigos científicos [77]. Assim, esse modelo se consolidou como importante ferramenta para a classificação automática de documentos em *corpora* muito volumosos e/ou em constante crescimento.

Porém, nos últimos anos, as tradicionais abordagens probabilísticas têm sido gradativamente substituídas. Essas mudanças são motivadas, particularmente, pelo surgimento de novas técnicas para modelagem de tópicos que se baseiam em redes neurais profundas, fazendo uso de *embeddings* e de *transformers*. Hoje, o estado-da-arte dessa tarefa de PLN é composto por modelos como o LDA2Vec [48], o Top2Vec [4] e o BERTopic [31]. Em razão de suas capacidades aprimoradas na captura de nuances semânticas e na compreensão de contexto nos documentos, essas técnicas modernas tendem a produzir representações mais precisas e interpretáveis dos tópicos latentes no *corpus* analisado e, por isso, têm produzido avanços notáveis na modelagem de tópicos [1].

Em face desse contexto, não é surpreendente que essa tarefa de PLN seja amplamente adotada nas áreas da Ciência Política e da Política Computacional. Greene e Cross [29], por exemplo, aplicaram a modelagem de tópicos por Fatoração de Matrizes Não Negativas a discursos em plenário, não só para acompanhar a evolução da agenda política do Parlamento Europeu de 1999 a 2014, como também para identificar o impacto de eventos internos e externos sobre esse parlamento. Já no contexto estadunidense, Grimmer [30] propôs um modelo de tópicos, denominado *Expressed Agenda Model*, para identificar e comparar as prioridades políticas dos senadores desse país a partir de seus comunicados à imprensa.

À parte das nações do Norte Global, Oliveira *et al.* [53] demonstrou como a modelagem de tópicos pode ser adotada para avaliar a presença (ou ausência) de conteúdo político em documentos, mesmo que não estejam diretamente associados ao Poder Legislativo. Nesse estudo, o *Biterm Topic Model* foi utilizado para identificar tópicos latentes num *corpus* constituído por *tweets* dos deputados brasileiros em exercício entre 2013 e 2017. Embora tenha sido observada uma forte predominância de assuntos que não estão relacionados ao contexto político, também foram identificados tópicos latentes que descrevem a atuação parlamentar desses indivíduos. Ainda no cenário brasileiro, Batista [6] e Moreira [49] demonstraram que os modelos de tópicos — particularmente o LDA e suas variantes — permitem a identificação de ênfases temáticas em discursos dos parlamentares brasileiros e em suas respectivas proposições de lei. Através das ênfases identificadas, contrastaram e mensuraram as diferenças entre as atuações políticas do conjunto de parlamentares analisado. Seus resultados corroboram a percepção de que os discursos de deputados e deputadas federais não se limitam aos temas discutidos durante as votações ou em proposições de lei. Além disso, também demonstram como as ênfases nas agendas social ou econômica são capazes de contrastar a atuação política desses parlamentares.

Esses estudos não apenas evidenciam a relevância da modelagem de tópicos para a Ciência Política e a Política Computacional, mas também destacam sua capacidade de fundamentar uma ampla gama de análises nessas áreas, ultrapassando a classificação de documentos nos volumosos *corpora* comuns ao contexto político. No cenário brasileiro, marcado por numerosos atores políticos e pela diversidade de documentos que tangenciam suas atividades parlamentares (como discursos, proposições de lei ou interações em redes sociais), essas técnicas são capazes de simplificar e aprimorar a análise desses dados textuais, contribuindo com o entendimento acerca da atuação política dos nossos representantes no Poder Legislativo.

2.4 Estimativa de pontos ideais

Na Ciência Política, a estimativa de pontos ideais é um método analítico que busca produzir representações espaciais para as preferências político-ideológicas de indivíduos ou grupos [5]. Denominadas pontos ideais, essas representações estão fundamentadas em espaços políticos abstratos, que podem ser unidimensionais ou multidimensionais. Não é incomum, inclusive,

que essas dimensões sejam descritas através das dicotomias que convencionalmente permeiam o vocabulário desse contexto, como esquerda-direita ou libertário-autoritário. Quando apresentados numericamente ou visualmente, os pontos ideais contribuem e simplificam a compreensão acerca da similaridade (ou dissimilaridade) entre atuações ou posicionamentos de diferentes atores políticos.

Inicialmente, a estimativa de pontos ideais se baseava apenas em votações no âmbito do Poder Legislativo. Essas técnicas tradicionais, conhecidas como modelos espaciais de votação, requerem que uma ou mais votações nominais tenham sido compartilhadas pelo conjunto de legisladores a ser analisado [68]. Assim, cada indivíduo pode ser representado como um ponto em um espaço euclidiano n-dimensional, tal que as n votações definam dimensões binárias equivalentes aos votos favoráveis ou contrários. São essas dimensões que determinam as coordenadas dos pontos ideais no espaço euclidiano. Ademais, quando o número de votações é elevado, é comum que sejam utilizadas técnicas de redução de dimensionalidade para viabilizar a representação gráfica dos pontos ideais. O Ideal [18] e o Nominate [55] são os modelos espaciais de votação mais relevantes e amplamente adotados.

Nas últimas duas décadas, numerosos estudos se propuseram a estimar, quantificar e analisar as posições político-ideológicas de legisladores a partir de suas votações nominais, colocando a estimativa de pontos ideais no cerne de diversas subáreas da Ciência Política [35]. Todavia, o processo político não é constituído somente por legisladores e, ainda que o fosse, as preferências políticas desses indivíduos não são expressas apenas por meio de seus votos. Nesse contexto, novos métodos e algoritmos foram desenvolvidos para viabilizar a estimativa de pontos ideais para atores políticos não parlamentares e/ou que não atuaram no mesmo período [17]. Com a crescente disponibilidade de dados e a difusão das redes sociais enquanto instrumento político, também surgiram técnicas para estimativa de pontos ideais baseados em textos como, por exemplo, o Wordfish [67] e o *Text-Based Ideal Point* (TBIP) [73].

Nesse nicho, os modelos tiram proveito do *political framing* [79]. Esse fenômeno descreve como atores políticos adaptam seus discursos com o objetivo de influenciar a opinião pública a respeito de questões políticas, destacando (ou ocultando) os aspectos de interesse através de ajustes no vocabulário adotado. Por exemplo, num debate sobre a criminalização das *fake news* no Brasil, parlamentares contrários tendem a utilizar a palavra “censura”, enquanto os favoráveis estão mais propensos ao uso do termo “desinformação”. Convém destacar

que, mesmo quando baseada em dados textuais, a estimativa de pontos ideais ainda difere substancialmente de outras tarefas de PLN. A análise de sentimentos e a classificação de intenções, por exemplo, empregam categorias predeterminadas para identificar quais são as emoções ou os objetivos subjacentes nos documentos. Em contraste, os pontos ideais são valores numéricos contínuos, cuja distribuição visa representar exclusivamente as características abstratas que definem as preferências político-ideológicas de um indivíduo ou grupo.

Predominantemente, as técnicas para estimativa de pontos ideais têm sido fundamentadas, implementadas e validadas em contextos políticos do Norte Global. Por exemplo, Proksch e Slapin [56] produziram representações para as preferências político-ideológicas dos membros e dos partidos políticos que constituíam o Parlamento Europeu à época. Essas representações foram baseadas em frequências de palavras nos discursos legislativos e, embora não descrevam o convencional espectro esquerda-direita, permitiram identificar os posicionamentos partidários quanto à integração europeia. Lauderdale e Herzog [40], por sua vez, usaram os discursos legislativos para estimar pontos ideais de senadores estadunidenses e dos membros da *Dáil Éireann* (equivalente irlandês da Câmara dos Deputados). Esses pontos ideais foram capazes de contrastar os indivíduos através da dinâmica de situação-oposição, que se manteve no decorrer das diversas legislaturas analisadas.

No contexto brasileiro, Souza, Graça e Silva [69] conseguiram estimar pontos ideais para parlamentares e partidos políticos através das conexões em uma rede social. Essas estimativas se mostraram equiparáveis às baseadas em votações nominais do Poder Legislativo e, ainda, foram estendidas para outros tipos de atores políticos, como jornalistas e comentaristas. Os autores também destacaram a interpretação dos valores estimados como um desafio intrínseco aos pontos ideais em espaços latentes, isto é, cujas dimensões não são definidas por variáveis explícitas. Silva [63], por outro lado, estimou pontos ideais para ministros e ministras do Supremo Tribunal Federal brasileiro. Para isso, aplicou a Análise de Componentes Principais aos votos em julgamentos de ações de inconstitucionalidade ocorridos entre 2012 e 2017. Essas representações ideológicas indicaram que as principais divergências entre esses juristas estão associadas à organização federativa e aos poderes em nível estadual. Oriundos de diferentes domínios e adotando técnicas diversas, esses estudos corroboram a relevância e abrangência da estimativa de pontos ideais na área da Ciência Política e, naturalmente, na compreensão dos atores políticos.

2.5 Métricas de avaliação

Neste estudo, buscamos demonstrar a validade aparente de nossos resultados. A validade aparente consiste na percepção superficial, baseada na lógica e no próprio instrumento, de que um teste ou medida aparenta avaliar adequadamente o construto que pretende mensurar [50]. Nesse sentido, empregamos métodos quantitativos e qualitativos para avaliar os tópicos latentes e os pontos ideais produzidos por nossos modelos. Em razão da especificidade de algumas métricas adotadas, nesta seção dedicamo-nos a descrevê-las mais detalhadamente.

A coerência é uma métrica que busca quantificar a interpretabilidade e a coesão semântica de tópicos latentes, reduzindo o esforço humano necessário para aferir essas características subjetivas [46]. Considerando os K tópicos latentes identificados num *corpus* qualquer, cada tópico k pode ser representado pelo conjunto $\{w_1^{(k)}, \ldots, w_N^{(k)}\}$ que contém as N palavras distintas a que está mais associado. Baseada no conceito probabilístico de *Normalized Pointwise Mutual Information* [13], a coerência avalia o quão dependentes entre si são as palavras que descrevem o mesmo tópico latente, conforme apresentado na Equação 2.1.

$$Coerência = \frac{1}{K} \sum_{k=1}^{K} \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{log \frac{P\left(w_i^{(k)}, w_j^{(k)}\right)}{P\left(w_i^{(k)}\right) P\left(w_j^{(k)}\right)}}{-\log P\left(w_i^{(k)}, w_j^{(k)}\right)} \tag{2.1}$$

Nessa equação, $P\left(w_i^{(k)}\right)$ e $P\left(w_j^{(k)}\right)$ denotam a probabilidade de ocorrência das palavras $w_i^{(k)}$ e $w_j^{(k)}$ em documentos atribuídos ao tópico k. Por sua vez, $P\left(w_i^{(k)}, w_j^{(k)}\right)$ denota a probabilidade de que ambas as palavras ocorram num mesmo documento desse subconjunto. Essas probabilidades são aproximadas através das frequências das palavras no *corpus* e, no caso ideal, a coocorrência deverá ser tão provável quanto as ocorrências individuais. Os valores da coerência de tópicos latentes estão limitados ao intervalo $[-1, 1]$, tal que a coesão semântica e a interpretabilidade por humanos são indicadas pelos valores positivos e mais altos.

Já a diversidade é uma métrica que se propõe a mensurar o quão distintos e disjuntos são os tópicos latentes [21]. Descrita na Equação 2.2, essa métrica avalia o percentual de palavras únicas dentre todos os conjuntos de N palavras mais associadas aos K tópicos latentes identificados no *corpus*. Sendo um percentual, os valores da diversidade de tópicos latentes

estão limitados ao intervalo $[0, 1]$. Percentuais elevados demonstram pouca ou nenhuma sobreposição entre os tópicos latentes, enquanto percentuais mais baixos demonstram a existência de redundância.

$$Diversidade \;=\; \frac{1}{K \cdot N} \left| \bigcup_{k=1}^{K} \{w_1^{(k)}, \; \ldots, \; w_N^{(k)}\} \right| \tag{2.2}$$

Derivada das métricas supracitadas, a qualidade de tópicos latentes é mensurada através do produto entre a coerência e a diversidade [22], conforme apresentado na Equação 2.3. Novamente limitados ao intervalo $[-1, 1]$, os valores mais altos e positivos indicam que a modelagem de tópicos produziu resultados com as três principais características desejadas: sem redundância, interpretáveis por humanos e semanticamente coesos.

$$Qualidade \;=\; Coerência \;\cdot\; Diversidade \tag{2.3}$$

Capítulo 3

Conjuntos de Dados

Neste capítulo, descrevemos os conjuntos de dados adotados no decorrer desta pesquisa. Serão apresentados os métodos utilizados para a extração desses dados, as principais características identificadas e seu uso na construção do *corpus* para treinamento de nossos modelos de PLN.

3.1 Extração e pré-processamento de dados

Neste estudo, propusemo-nos a analisar os discursos de parlamentares por meio de técnicas modernas de PLN. Esses dados textuais são, portanto, a matéria-prima para a extração de informações acerca dos posicionamentos políticos de nossos representantes na esfera federal do Poder Legislativo. Todavia, não conseguimos identificar nenhuma base de dados que disponibilizasse esses discursos, sobretudo considerando a completude e o alcance temporal almejados.

Esse cenário desfavorável nos fez optar pela criação de uma base de dados própria, aberta, potencialmente inédita, focada em deputados(as) federais e contendo os discursos em eventos da Câmara dos Deputados realizados entre 2003 e 2022. Aqui, adotamos a definição do *Open Data Handbook*[1] que estabelece dados abertos como aqueles que "podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa". Além disso, buscamos seguir os princípios do *Open Government Data*[2], que incluem a gratuidade de acesso, a processabilidade

[1] https://opendatahandbook.org

[2] https://opengovdata.org

por máquina e o uso de formatos não proprietários, por exemplo. Para construir essa nova base de dados, recorremos a duas fontes governamentais distintas: o Portal de Dados Abertos do Tribunal Superior Eleitoral[3] e o Portal da Câmara dos Deputados[4].

No acervo do Tribunal Superior Eleitoral (TSE), estão disponíveis informações sobre todas as eleições desde 1994, incluindo a lista de candidaturas de cada uma delas. Esse conjunto de dados pode ser acessado em arquivos de formato CSV (*Comma-Separated Values*), estratificados conforme o ano da eleição e a representação legislativa dos cargos (unidade federativa ou nível nacional). Além das diversas informações sociodemográficas e eleitorais de cada candidato(a) — como unidade federativa de origem e filiação partidária — também é possível identificar as pessoas que foram eleitas ou se tornaram suplentes. Dessa maneira, um conjunto de dados descrevendo parlamentares brasileiros em exercício entre 2003 e 2022 pôde ser construído com operações de filtragem e concatenação dessas tabelas. Ademais, poucas operações de pré-processamento foram necessárias, limitando-se a pequenas modificações em valores categóricos e à padronização da capitalização dos textos. A estrutura desse conjunto de dados está apresentada na Tabela 3.1.

Tabela 3.1: Estrutura do conjunto de dados sobre deputados e deputadas federais em exercício (2003 a 2022).

Atributo	Descrição
`id_parlamentar`	Identificador único do(a) parlamentar.
`ano`	Ano de eleição do(a) parlamentar.
`uf`	Sigla da unidade federativa que elegeu o(a) parlamentar.
`cargo`	Cargo legislativo ocupado pelo(a) parlamentar.
`nome`	Nome completo do(a) parlamentar.
`nome_urna`	Nome de urna utilizado pelo(a) parlamentar quando eleito(a).
`partido`	Filiação partidária do(a) parlamentar quando eleito(a).
`sigla_partido`	Sigla do partido político ao qual o(a) parlamentar estava filiado(a) quando eleito(a).

[3] https://dadosabertos.tse.jus.br

[4] https://www.camara.leg.br

No Portal da Câmara dos Deputados, por outro lado, o processo de extração revelou-se bem mais complexo e demandou bastante esforço. Isso porque, ainda que essa casa legislativa possua um Portal de Dados Abertos próprio (e com vasto conteúdo), nem todos os seus documentos e informações estão acessíveis através de tal repositório. Por isso, recorremos à raspagem de dados em páginas web para viabilizar a construção dos conjuntos de dados remanescentes. Vale ressaltar, no entanto, que ter esse tipo de conhecimento técnico como um requisito para o uso e acesso desses dados pode se tornar um obstáculo considerável para estudos similares a este.

A priori, extraímos os dados sobre os eventos realizados pela Câmara dos Deputados no período que se estende de 2003 a 2022 e, mais especificamente, para os quais foram produzidas notas taquigráficas. Uma nota taquigráfica é um documento oficial definido, no próprio Portal da Câmara dos Deputados, como "o conjunto de discursos que compõe tudo o que é registrado em sessões plenárias ou em reuniões de comissões". Apesar de seus discursos não serem disponibilizados no Portal de Dados Abertos dessa casa legislativa, esses documentos são publicados e podem ser acessados em formato HTML (*HyperText Markup Language*) e, na maioria dos casos, também em formato PDF (*Portable Document Format*). Entretanto, visto que o Regimento Interno da Câmara dos Deputados só estabelece produção obrigatória das notas taquigráficas para as sessões plenárias, é bastante provável que nem todos os eventos das comissões estejam listados. Quanto às operações de pré-processamento, além de padronizar a capitalização dos textos e de realizar pequenos ajustes nos valores categóricos, também removemos as duplicatas, padronizamos o formato de datas e corrigimos o espaçamento em valores textuais. A estrutura do conjunto de dados produzido está apresentada na Tabela 3.2.

Tabela 3.2: Estrutura do conjunto de dados sobre eventos realizados pela Câmara dos Deputados (2003 a 2022).

Atributo	Descrição
`id_evento`	Identificador único do evento.
`categoria_evento`	Categoria do evento. Permite identificar, por exemplo, as audiências públicas ou as sessões ordinárias.

Tabela 3.2 (Continuação)

Atributo	Descrição
`ambiente_legislativo`	Nome do plenário ou comissão a que o evento está associado.
`categoria_ambiente`	Categoria do ambiente legislativo. Permite identificar se o evento foi realizado em plenário ou por uma comissão.
`casa_legislativa`	Nome da casa legislativa à qual o evento está associado.
`data`	A data em que o evento foi realizado.
`discursos`	Link para a página em que está disponível uma nota taquigráfica associada ao evento.

Em posse desse conjunto de dados sobre os eventos, foi possível identificar as páginas web dedicadas às notas taquigráficas e, assim, iniciar a extração dos discursos em eventos realizados pela Câmara dos Deputados. Essa etapa exigiu abordagens bem específicas como, por exemplo, adaptar o ferramental de extração para que pudesse ser executado repetidamente sem gerar duplicatas nos conjuntos de dados. Isso porque essas páginas podem apresentar tempo de resposta elevado e, após ultrapassar certo limiar de tempo, será exibida uma página de erro mesmo que a nota taquigráfica exista e esteja publicada. A Figura 3.1 exemplifica esse comportamento através de acessos sequenciais à nota taquigráfica do evento 0001/03.

Assim, com o objetivo de aumentar o número de documentos obtidos, reexecutamos a extração diversas vezes e priorizamos fazê-lo em horários com menor fluxo provável de usuários. Também construímos expressões regulares para viabilizar a raspagem desse conjunto de dados, uma vez que as páginas das notas taquigráficas não fazem uso adequado da semântica HTML e que a organização de seu conteúdo em *tags* não se mantém consistente entre os documentos. Em particular, as expressões regulares foram utilizadas para delimitar os discursos individualmente e, em seguida, para distinguir as transcrições dos nomes de seus respectivos oradores. Em contraponto, o pré-processamento dos dados foi pouco complexo e manteve-se restrito às correções de espaçamento e de capitalização dos textos. A Tabela 3.3 apresenta a estrutura do conjunto de dados produzido.

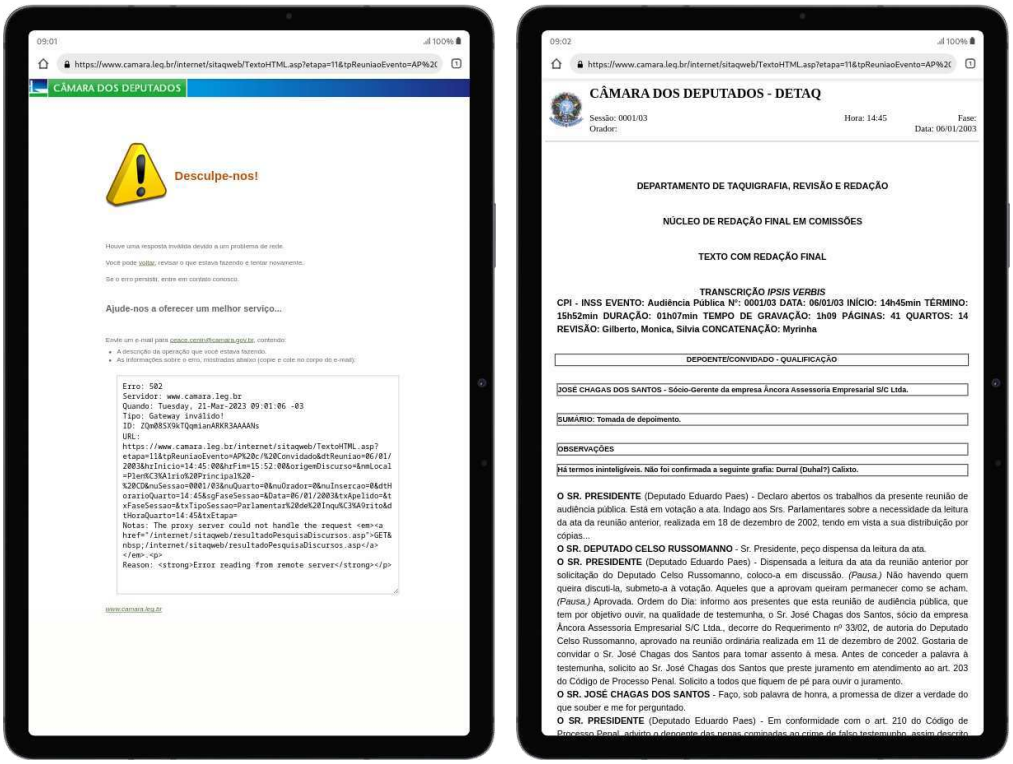

Figura 3.1: Erro no acesso às notas taquigráficas causado por tempo de resposta elevado.

Tabela 3.3: Estrutura do conjunto de dados sobre discursos em eventos realizados pela Câmara dos Deputados (2003 a 2022).

Atributo	Descrição
`id_evento`	Identificador único do evento a que o discurso está associado.
`ordem_discurso`	Atributo para a ordenação dos discursos de um evento.
`orador`	Nome do(a) orador(a) associado(a) ao discurso.
`texto`	Texto da transcrição do discurso.

Todo o ferramental para extração e pré-processamento de dados foi implementado na linguagem *Python* — mais especificamente, com uso das bibliotecas *Pandas*[5] e *Scrapy*[6] — e

[5]https://pypi.org/project/pandas

[6]https://pypi.org/project/scrapy

está disponível em repositório público do *GitHub*[7]. Em paralelo, a base de dados completa foi disponibilizada publicamente para download através do *Google Drive*[8]. Mais detalhes sobre os recursos computacionais utilizados no decorrer dessa etapa estão descritos na Tabela A.1 do Apêndice A.

3.2 Análise exploratória e descritiva

Com o processo de extração finalizado, demos início à análise exploratória e descritiva de nossa base de dados, composta por três conjuntos distintos: um sobre parlamentares, um sobre os eventos realizados pela Câmara dos Deputados e um contendo os discursos proferidos durante esses eventos. Nessa etapa, buscamos não somente descrever e quantificar a base de dados, mas também identificar as características que seriam importantes para a seleção de amostras adequadas ao treinamento dos modelos de PLN.

O conjunto de dados sobre parlamentares traz informações acerca de 1.453 pessoas distintas que foram eleitas para a Câmara dos Deputados entre 2003 e 2022. Dentre esses indivíduos, 651 exerceram mais de um mandato enquanto deputado(a) federal, totalizando 68,73% dos 2.565 cargos disponíveis nesse período. Além desse indicativo de pouca renovação dentre os membros da Câmara dos Deputados, apenas três partidos políticos agregam quase 40% das candidaturas eleitas para essa casa legislativa: o Partido dos Trabalhadores (PT), o Movimento Democrático Brasileiro (MDB) e o Partido da Social Democracia Brasileira (PSDB). Mesmo com drásticas mudanças no cenário político nacional, a distribuição desse número de deputados(as) eleitos(as) por partido não se tornou mais uniforme no decorrer dos últimos anos.

Já no conjunto de dados sobre eventos, temos informações acerca dos 18.686 eventos realizados pela Câmara dos Deputados ou, mais especificamente, sobre 6.706 eventos realizados pelo plenário e 11.980 eventos realizados pelas comissões dessa casa legislativa. A discrepância entre esses valores é considerável, mas está aquém do esperado. Isso porque, enquanto o plenário é o ambiente único que agrega todos os deputados e deputadas, existem 25 comissões permanentes (e tantas outras temporárias) na Câmara dos Deputados. É provável,

[7] https://github.com/beabaparlamentar/congresso-em-texto

[8] https://bit.ly/DiscursosNaCâmaraDosDeputados

portanto, que o número de eventos realizados pelas comissões seja ainda mais elevado, mas que esses eventos não estejam listados devido à ausência de obrigatoriedade para a produção de suas notas taquigráficas. A distribuição anual destes eventos, estratificada entre plenário e comissões, está apresentada na Figura 3.2.

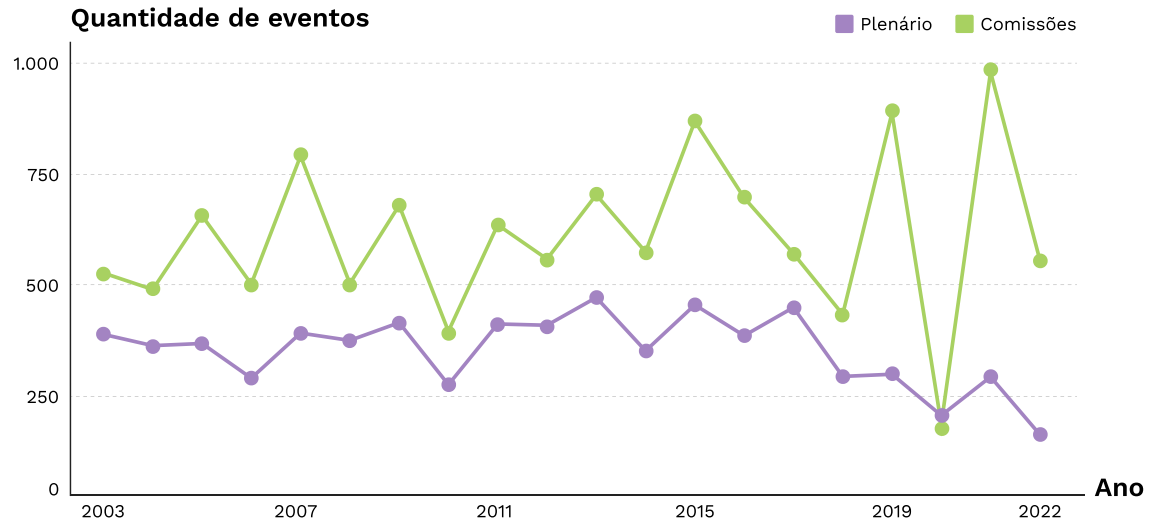

Figura 3.2: Distribuição anual dos eventos realizados pelo plenário e pelas comissões da Câmara dos Deputados (2003 a 2022).

Podemos observar que o plenário costuma realizar entre 250 e 500 eventos por ano. No entanto, essas atividades foram bastante reduzidas após o início da pandemia de COVID-19 (em março de 2020) e, ao menos durante o período avaliado, não retornaram ao seu patamar original. Já nas comissões, a quantidade de eventos por ano apresenta variações maiores e que parecem estar relacionadas à atenção da sociedade civil quanto aos acontecimentos no Poder Legislativo. O ano de 2015, por exemplo, conta com 871 eventos de comissões e foi marcado pelas investigações da Operação Lava Jato contra Eduardo Cunha (presidente da Câmara dos Deputados à época) que, posteriormente, levaram ao acolhimento do pedido de *impeachment* contra a ex-presidenta Dilma Rousseff. Em contrapartida, quando ocorrem eleições gerais ou municipais, o número de eventos realizados pela Câmara dos Deputados tende a sofrer um decréscimo no comparativo com o ano anterior.

Por fim, nosso último conjunto de dados é composto por discursos proferidos em 17.966 eventos da Câmara dos Deputados. Esse número representa apenas 96,15% dos eventos

identificados anteriormente, visto que mesmo com diversas reexecuções do processo de extração, não foi possível acessar integralmente as transcrições. Na Figura 3.3, estão apresentados histogramas da média de palavras por discurso nos eventos da Câmara dos Deputados (divididos conforme o tipo de ambiente legislativo associado). Para melhor visualização dessas distribuições, ocultamos 243 *outliers* cujos valores ultrapassavam o valor médio da distribuição em mais do que três desvios-padrão.

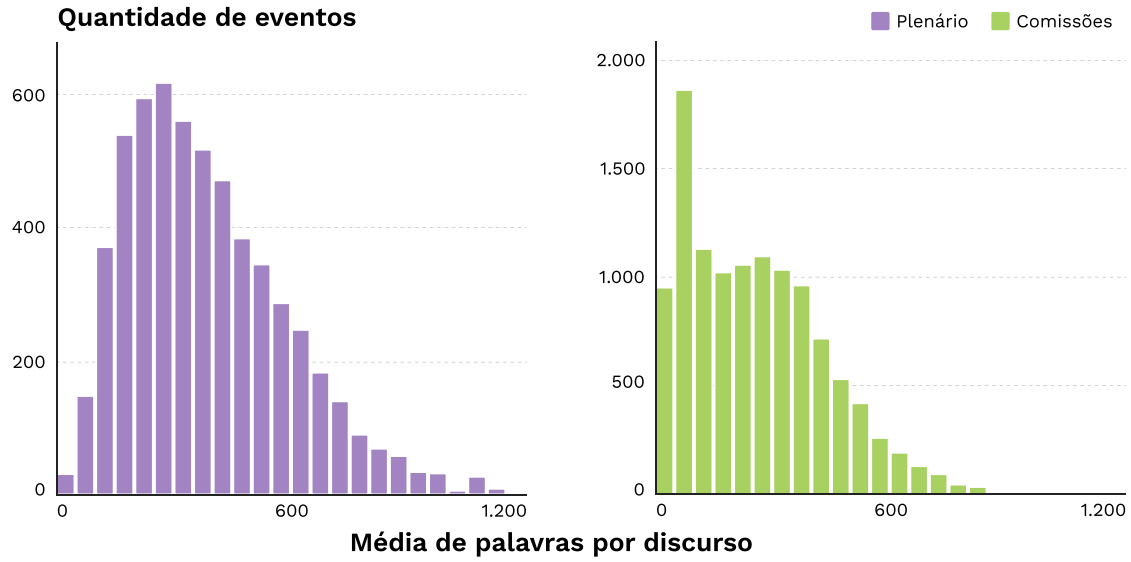

Figura 3.3: Distribuição média de palavras por discurso nos eventos realizados pelo plenário e pelas comissões da Câmara dos Deputados (2003 a 2022).

Os discursos em eventos do plenário tendem a ser consideravelmente mais longos do que aqueles oriundos das comissões. Há uma disparidade de quase 53% no número médio de palavras e essa característica reflete bem o quão diferentes são esses ambientes. As sessões plenárias constituem um cenário relativamente controlado, em que os parlamentares podem discursar mais livremente e no qual estão menos propensos à interferência dos demais deputados e deputadas. Por outro lado, as comissões são espaços mais propícios para debates e discussões entre as poucas dezenas de parlamentares ali reunidas e, portanto, estão mais alinhadas com falas curtas. É notável ainda que, mesmo no comparativo ao subconjunto das sessões plenárias que possuem com mais de 100 oradores distintos, a discrepância nesse número médio de palavras por discurso se mantém relevante.

3.3 Definição do *corpus*

Tendo em vista a dimensão da base de dados construída, se fez necessário que selecionássemos uma amostra menor para, a partir dela, definir o *corpus* de treinamento dos nossos modelos de PLN. Essa amostragem foi motivada não somente por limitações de tempo ou de recursos computacionais, mas também pela viabilidade da análise dos discursos e posicionamentos políticos dos numerosos parlamentares. Nesse sentido, a primeira característica que elencamos para a amostra foi o seu alcance temporal: seriam considerados apenas os eventos realizados nas duas últimas legislaturas da Câmara dos Deputados, isto é, no período que se estende de 2015 a 2022. Nossas análises seriam, portanto, beneficiadas pela contemporaneidade dos acontecimentos discutidos nessa casa legislativa.

Filtramos também os discursos com, no mínimo, 100 caracteres (aproximadamente 20 palavras). Esse limiar nos permite remover falas muito curtas que, em geral, agregariam pouco valor semântico ao conjunto de discursos de cada indivíduo. Essa filtragem é particularmente importante no contexto dos eventos da Câmara dos Deputados, uma vez que os parlamentares costumam interromper uns aos outros durante os discursos e que, em certas situações, se manifestam apenas em razão dos ritos e protocolos legislativos. Em seguida, iniciamos a seleção dos discursos que foram proferidos por deputados e deputadas federais em exercício durante as duas últimas legislaturas. Devido aos diferentes padrões adotados nas notas taquigráficas e à presença de ruído nesses dados textuais, a associação direta entre os nomes de parlamentares e de oradores não era possível. Assim, para viabilizar tal seleção, o mapeamento dessas entidades foi feito iterativamente e com apoio da biblioteca *Difflib*[9], que permite o cálculo de similaridade entre textos baseando-se em *substrings* coincidentes. Visando aumentar a confiabilidade desse mapeamento, os resultados foram submetidos à revisão humana após cada iteração, o que permitiu a correção de erros pontuais.

Além disso, mantivemos apenas os deputados e as deputadas que possuíssem mais de 100 discursos alinhados às características elencadas. Esse filtro nos permitiu evitar parlamentares sobre os quais se tinha uma quantidade muito limitada de informações, um fator que poderia prejudicar o treinamento dos modelos e/ou distorcer análises de maior granularidade. Finalmente, o *corpus* construído a partir da amostra selecionada é composto por 389.562 discursos

[9]https://docs.python.org/3/library/difflib

oriundos de 6.991 eventos da Câmara dos Deputados e associados a 577 parlamentares distintos. Mais detalhes sobre o uso das informações de filiação partidária desses parlamentares estão descritos no Apêndice B.

Capítulo 4

Modelagem de Tópicos Latentes

Neste capítulo, apresentamos a etapa de modelagem de tópicos latentes deste estudo. Após uma breve introdução à técnica adotada, descrevemos o processo de treinamento do modelo, avaliamos qualitativamente os tópicos identificados e analisamos seus resultados quanto a parlamentares e partidos políticos representados em nossos dados.

4.1 BERTopic

O BERTopic é uma técnica do estado-da-arte para modelagem de tópicos latentes, cujo funcionamento costuma ser descrito em três etapas distintas: a criação dos *embeddings*, o agrupamento dos documentos e a representação dos tópicos. Em contraste às tradicionais abordagens probabilísticas (como o LDA), são os grupos de documentos que definem os tópicos latentes para esse modelo. Em sua primeira etapa, utilizando o *framework* Sentence-BERT [59], são gerados *embeddings* que representam os documentos do *corpus*, possibilitando comparações em nível semântico. Na etapa seguinte, esses *embeddings* serão necessários para agrupar os textos semanticamente similares, formando grupos compostos por documentos associados ao mesmo tópico latente. Essa tarefa é realizada através do *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) [43], um algoritmo de agrupamento hierárquico, baseado em densidade e capaz de identificar ruído nos dados. Desse modo, o BERTopic não apenas evita pressupostos sobre as características dos grupos a serem gerados (como o formato ou a distância entre eles), mas também consegue classificar documentos discrepantes como *outliers*.

Visando contribuir com o desempenho do HDBSCAN, os *embeddings* são submetidos à redução de dimensionalidade antes de serem agrupados, isto é, são projetados em um espaço vetorial com menos dimensões. Para isso, utiliza-se o *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP) [44], reconhecido por sua capacidade de preservar as estruturas local e global dos conjuntos de dados. A última etapa do BERTopic consiste em produzir representações para os tópicos latentes identificados, que serão definidas pelo conjunto de palavras mais relevantes para cada grupo de documentos. A relevância dessas palavras será determinada por um esquema de ponderação baseado numa variação do TF-IDF que foi desenvolvida especificamente para esse modelo: o *class-based Term Frequency-Inverse Document Frequency* (c-TF-IDF). Naturalmente, todos os documentos do *corpus* precisarão ser submetidos à tokenização para que seja possível observar as frequências dos termos nesses dados textuais.

Conforme apresentado na Equação 4.1, a relevância de uma palavra p para um grupo g de documentos — e, consequentemente, para o seu tópico latente associado — será denotada por $W_{p,g}$. Nessa equação, M representa o número médio de palavras por grupo, enquanto $tf_{p,g}$ e tf_p denotam a frequência da palavra p nesse grupo g e em todos os grupos gerados, respectivamente. Isso significa dizer que, normalmente, a representação de um tópico latente será composta por palavras com muitas ocorrências em seu grupo de documentos, mas pouco frequentes no restante do *corpus*.

$$W_{p,g} \;=\; tf_{p,g} \;\cdot\; \log\left(1 + \frac{M}{tf_p}\right) \tag{4.1}$$

Os experimentos apresentados na publicação original do BERTopic demonstram, através de métricas de coerência e diversidade, que o desempenho do modelo é superior — ou, no mínimo, equiparável —– às técnicas tradicionais e modernas para a modelagem de tópicos latentes, como o LDA e o Top2Vec, respectivamente. Em razão de sua estrutura de *pipeline*, as técnicas e submodelos que constituem o BERTopic podem ser facilmente substituídas por outras alternativas, a depender do domínio considerado e das limitações existentes. Essa flexibilidade permite uma fácil adaptação do modelo aos recursos computacionais disponíveis e, sobretudo, possibilita o seu aprimoramento contínuo à medida que o estado-da-arte avança na área de PLN. O BERTopic já possui versões que viabilizam, por exemplo, a modelagem

dinâmica de tópicos, a aprendizagem semisupervisionada e, até mesmo, a integração com *Large Language Models* (LLMs) para a representação dos tópicos latentes. Neste trabalho, contudo, restringimo-nos às etapas e técnicas previstas na proposta original do modelo.

4.2 Treinamento do modelo

Para a etapa de treinamento do modelo, recorremos à biblioteca homônima que disponibiliza a implementação do BERTopic na linguagem Python[1]. Visando a criação de *embeddings* que representassem os discursos em nosso *corpus*, adotamos uma versão multilíngue do modelo MiniLM Sentence-BERT[2]. Essa versão do modelo possui 12 camadas de codificação e é dedicada a parafrasear documentos em diferentes idiomas, preservando a semântica textual mesmo entre as diferentes estruturas sintáticas. Ela mapeia os documentos para um espaço vetorial denso com 384 dimensões e foi pré-treinada em *corpora* que abrangem mais de 50 idiomas, inclusive o português brasileiro [60]. Além disso, quando comparada a modelos específicos para nosso idioma, seu desempenho se mostrou equivalente (ou ligeiramente superior) em diversas tarefas de PLN como, por exemplo, a categorização de notícias e a detecção de *fake news* [24]. Os modelos pré-treinados tipicamente tornam desnecessárias — ou até mesmo desaconselháveis — quaisquer operações de pré-processamento de dados. Por essa razão, nossos *embeddings* foram produzidos usando os textos originais dos discursos.

Em contrapartida, a presença de *stopwords* pode ser bastante prejudicial à interpretabilidade dos tópicos latentes. Isso porque, a despeito de seu baixo valor semântico, a alta frequência dessas palavras é capaz de distorcer sua relevância no contexto de cada tópico identificado. O BERTopic possui duas estratégias para mitigar esse problema: a remoção de *stopwords* e a amortização da frequência de palavras. A primeira estratégia consiste em predefinir um conjunto de termos para que sejam desconsiderados durante a tarefa de tokenização do modelo. Nesse sentido, agregamos as listas de *stopwords* disponíveis nas bibliotecas NLTK[3] e Spacy[4] para o português brasileiro, totalizando 500 palavras distintas. Ademais, somente as palavras compostas por dois ou mais caracteres alfabéticos foram consideradas

[1] https://pypi.org/project/bertopic

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[3] https://pypi.org/project/nltk

[4] https://pypi.org/project/spacy

como *tokens* válidos. A segunda estratégia, por sua vez, visa evitar que os termos mais frequentes no *corpus* exerçam influência desproporcional na definição dos tópicos latentes. Para aplicá-la, as frequências das palavras em nossos documentos foram submetidas à operação de radiciação antes do cálculo do c-TF-IDF, diminuindo a amplitude da distribuição desses valores.

Hiperparâmetros relevantes para a nossa modelagem de tópicos também foram identificados nas outras tarefas do BERTopic. Entretanto, visto que essas variáveis específicas costumam estar associadas ao conteúdo semântico dos documentos, optamos por definir seus valores experimentalmente. O *n_components* é o hiperparâmetro que estabelece a quantidade de dimensões do espaço vetorial criado após a redução de dimensionalidade. Ele afeta diretamente o quão bem os dados originais serão preservados, já que mais informação será perdida à medida que se diminui o número de dimensões disponíveis. Em paralelo, o *n_neighbors* determina a quantidade de "vizinhos" que serão analisados durante o mapeamento de cada *embedding* entre esses dois espaços vetoriais. Conforme seu valor se eleva, a estrutura global dos vetores será melhor preservada (em detrimento da estrutura local).

Já na tarefa de agrupamento, o *min_cluster_size* pode ser considerado como hiperparâmetro principal, visto que seu valor define o número mínimo de documentos necessários à criação de um grupo que, depois, representará um tópico latente. Enquanto isso, o *metric* indica a métrica de distância a ser utilizada para avaliar se os pares de documentos estão suficientemente próximos para serem agrupados. Com esses quatro hiperparâmetros selecionados, baseamo-nos na literatura e na documentação da biblioteca para elencar os respectivos valores experimentais (apresentados na Tabela 4.1) e, usando todas as combinações possíveis entre esses valores, produzimos 54 modelos diferentes.

Tabela 4.1: Hiperparâmetros avaliados experimentalmente durante a modelagem de tópicos usando o modelo BERTopic.

Tarefa	**Hiperparâmetro**	**Valores**
Redução de dimensionalidade	*n_components*	[5, 10, 15]
Redução de dimensionalidade	*n_neighbors*	[25, 50, 100]
Agrupamento	*metric*	["euclidean", "manhattan"]

Tabela 4.1 (Continuação)

Tarefa	**Hiperparâmetro**	**Valores**
Agrupamento	*min_cluster_size*	[25, 50, 100]

Adotamos as métricas de coerência e diversidade (descritas na Seção 2.5) para avaliar esses modelos e, assim, determinar a qualidade dos tópicos latentes produzidos por cada um deles. Essa avaliação foi implementada com apoio do *framework* OCTIS [71] ou, mais especificamente, de sua implementação na linguagem Python[5]. Na Tabela 4.2, sumarizamos os resultados obtidos durante essa avaliação e destacamos a performance do BERTopic utilizado neste estudo. O BERTopic selecionado foi aquele com melhor desempenho na métrica de qualidade dos tópicos, mas também apresentou o maior valor de coerência e o sexto maior valor de diversidade. Sua redução de dimensionalidade produziu um espaço vetorial de 5 dimensões, para o qual os *embeddings* foram mapeados com base nos respectivos 25 vetores mais próximos. Já na tarefa de agrupamento, a distância de Manhattan foi utilizada para avaliar a proximidade entre os *embeddings* e, além disso, o tamanho mínimo dos grupos foi fixado em 25 documentos. A priori, esse modelo identificou 117 tópicos latentes em nosso *corpus*, mas 6,13% dos documentos foram considerados *outliers*, isto é, não estavam associados a nenhum desses tópicos.

Tabela 4.2: Desempenho dos modelos BERTopic quanto à coerência, diversidade e qualidade dos tópicos latentes.

	Mínima	**Média**	**Máxima**	**Modelo Final**
Coerência	-0,096	0,046	0,120	0,120
Diversidade	0,467	0,514	0,700	0,481
Qualidade	-0,067	0,023	0,058	0,058

Não é incomum que o BERTopic produza um número tão elevado de tópicos latentes, porém, essa característica amplia bastante o esforço necessário à análise de seus resultados.

[5]https://pypi.org/project/octis

Durante uma investigação preliminar, notamos certa intersecção entre os conjuntos de termos mais relevantes para cada tópico identificado. Essa percepção está alinhada ao desempenho do modelo quanto à métrica de coerência, visto que os valores próximos a zero — sejam positivos ou negativos — costumam indicar sobreposição entre os tópicos. Nesse cenário, implementamos ajustes no treinamento do modelo com o objetivo de reduzir não só o número de tópicos, mas também o percentual de *outliers*. Felizmente, a implementação do BERTopic já prevê soluções pós-treinamento para esses desafios tão recorrentes.

A redução do número de tópicos é feita a partir do esquema de ponderação do modelo. Nela, os vetores dos tópicos (definidos pelo c-TF-IDF) são comparados entre si através da similaridade do cosseno. Em seguida, o tópico menos frequente será mesclado àquele com que possuir maior similaridade. Esse processo será feito iterativamente até que o número de tópicos tenha sido reduzido ao valor desejado. De maneira análoga, a redução de *outliers* representará esse subconjunto de documentos por meio do c-TF-IDF, produzindo vetores que podem ser comparados aos dos tópicos. Assim, cada documento será associado ao tópico com o qual mais se assemelha. Com o uso dessas duas estratégias, a versão final de nosso modelo estratificou os documentos entre 50 tópicos latentes (sem *outliers*). Os recursos computacionais utilizados para o treinamento e a avaliação dos modelos BERTopic estão descritos na Tabela A.2 do Apêndice A.

4.3 Avaliação dos tópicos latentes

Em posse dos resultados do modelo, dedicamo-nos a analisar o conjunto de tópicos latentes produzidos. A priori, substituímos seus identificadores numéricos (gerados automaticamente pelo BERTopic) por rótulos representativos, visando assim construir uma compreensão mais precisa acerca do conteúdo de cada tópico. Devido às limitações de tempo e de escopo, os rótulos foram definidos e revisados exclusivamente pelos autores deste trabalho. Tal processo foi baseado na análise subjetiva dos dez termos mais relevantes para cada tópico latente, apresentados detalhadamente na Tabela C.1 do Apêndice C. Esses conjuntos de termos se mostraram facilmente interpretáveis, tornando a escolha dos rótulos relativamente simples.

Ao passo que encontramos temas como “Agricultura” ou “Esportes”, também constatamos que 14 dos 50 tópicos latentes se dedicam somente ao contexto do Poder Legislativo. São

discursos que se resumem, por exemplo, às orientações partidárias ou aos encerramentos de sessão. A ocorrência desses tópicos já era prevista, mas sua associação a 50,94% do *corpus* nos fez perceber que a maioria dos documentos não seria útil às nossas análises, tendo em vista que não representam os temas defendidos (ou combatidos) por parlamentares durante suas atuações políticas. Denominamos esse conjunto de tópicos como "Processos Legislativos". Em seguida, avaliamos quais tópicos seriam relevantes para o escopo deste estudo e também agrupamo-os em categorias.

A primeira categoria inclui 70.214 documentos e está dividida em nove tópicos latentes. Esses discursos abordam temas associados às obrigações do Estado brasileiro para com seus cidadãos e cidadãs. São exemplos "Saúde", "Segurança" e "Educação". Por isso, nomeamos esse grupo como "Garantias Fundamentais". Já "Desenvolvimento Econômico" é a categoria que inclui temas sobre as atividades produtivas e o desenvolvimento econômico nacional, como "Turismo" ou "Emprego e Renda". Estão inclusos dez tópicos latentes e 38.080 documentos. Por fim, a terceira categoria de tópicos relevantes foi denominada "Equidade e Inclusão" e agrega temas relacionados às lutas sociais e à garantia de direitos para populações subrepresentadas. Esse grupo é composto por 30.755 documentos e nove tópicos latentes, entre os quais figuram "Violência de Gênero", "Igualdade Racial" e "Direitos LGBTQIA+".

Os oito tópicos remanescentes (totalizando 52.060 documentos) não foram considerados relevantes para nossas análises, uma vez que representam temas efêmeros e/ou de escopo muito específico. Nomeado como "Outros", esse grupo inclui tópicos como "Operação Lava Jato" e "Conflitos no Oriente Médio", por exemplo. A divisão dos tópicos latentes entre as cinco categorias está detalhadamente apresentada na Tabela C.2 do Apêndice C. Já na Figura 4.1, apresentamos a quantidade de discursos associados a cada categoria de tópicos.

Em destaque no gráfico acima, as três categorias de tópicos relevantes representam cerca de 35,69% do nosso *corpus*. Apesar desse volume considerável de documentos, é interessante notar como os tópicos temporários geram mais engajamento entre os parlamentares do que as categorias "Desenvolvimento Econômico" e "Equidade e Inclusão". Tal comportamento pode estar relacionado aos diferentes níveis de repercussão desses temas junto à mídia e à sociedade civil. É importante ressaltar, no entanto, que a relevância atribuída às categorias (e aos seus respectivos tópicos) está diretamente associada ao nosso propósito em estabelecer comparação entre todos os deputados e deputadas presentes nesses dados. Logo, pode não ser

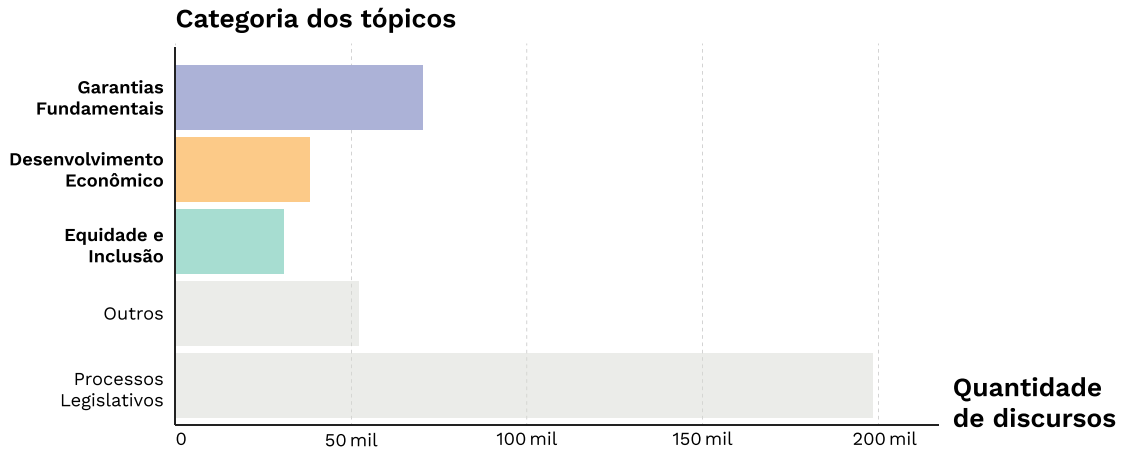

Figura 4.1: Distribuição dos discursos de parlamentares quanto às categorias de tópicos.

aplicável ou pertinente em outros estudos semelhantes.

Para mais detalhes, a Figura 4.2 apresenta a distribuição dos discursos por tópico latente, mas com foco apenas nas três categorias de interesse. É notável que, mesmo dentre esse subconjunto de tópicos, a quantidade de discursos difere bastante. Enquanto o tópico menos comum inclui 729 discursos, o mais comum está associado a 18.124 documentos. Esse último é o tópico "Saúde", cuja frequência foi alavancada durante pandemia de COVID-19. Na categoria "Desenvolvimento Econômico", os tópicos mais frequentes são aqueles que abordam questões governamentais da economia, isto é, "Emprego e Renda" e "Finanças e Tributação". Todos os demais tópicos desse grupo se dedicam a atividades produtivas específicas, como "Agricultura" e "Transporte Aéreo". Essas atividades, porém, são oriundas apenas dos setores primário e terciário, sugerindo menor atenção dos parlamentares quanto à indústria nacional.

Ainda nessa perspectiva, a frequência dos tópicos latentes da categoria "Equidade e Inclusão" apresentou um considerável declínio durante o período analisado (mesmo considerando apenas o pré-pandemia), o que parece se contrapor ao espaço que esses temas ganharam no debate público ao longo dos últimos anos. Apesar dessa redução no número de documentos, dois tópicos latentes apresentaram flutuações marcantes em suas distribuições anuais. Em 2017, ocorreram 1.276 discursos associados à "Previdência Social", o que difere bastante de sua média de 539 documentos anuais. De maneira análoga, o tópico "Direitos Trabalhistas" apresentou picos em 2017 (com 1.181 discursos) e em 2019 (com 1.851 discursos).

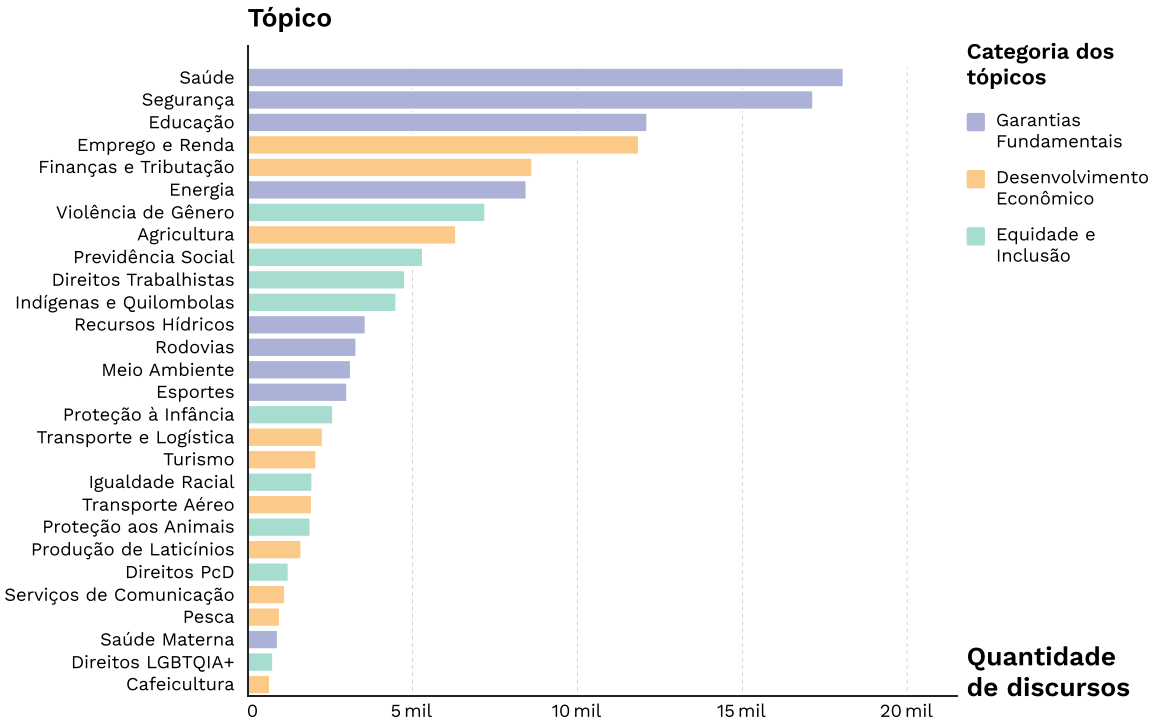

Figura 4.2: Distribuição dos discursos de parlamentares quanto aos tópicos latentes relevantes às nossas análises.

Consultando o *corpus*, constatamos que essas flutuações foram causadas pela tramitação das Reformas Trabalhista e da Previdência na Câmara dos Deputados.

Em resumo, a avaliação qualitativa dos tópicos latentes corrobora os bons resultados do nosso modelo. Foram identificados tópicos facilmente interpretáveis e compatíveis com o contexto do *corpus*, viabilizando uma classificação adequada dos documentos. Por meio desse tópicos, conseguimos selecionar os discursos que abordam temas pertinentes ao escopo deste estudo e, até mesmo, identificar pautas que tramitaram na Câmara dos Deputados em períodos específicos.

4.4 Análise de pautas prioritárias usando tópicos latentes

Os tópicos latentes identificados nos discursos de parlamentares nos permitiram observar quais temas estavam presentes em debates e discussões nos eventos da Câmara dos Deputados durante a 55ª e a 56ª Legislaturas. Todavia, mais do que entender quais assuntos foram

priorizados por esse conjunto de parlamentares (ou como essas prioridades mudaram no decorrer do tempo), os tópicos latentes podem colaborar com nossa compreensão acerca da atuação política individual de deputados e deputadas. Em específico, podemos observar quais desses tópicos foram mais frequentes nos discursos de cada parlamentar e, assim, inferir os temas que lhes são mais importantes no contexto político.

Para exemplificar a viabilidade dessas análises, selecionamos um conjunto de dez parlamentares cujos mandatos na Câmara dos Deputados receberam amplo destaque midiático à época. São eles: Alessandro Molon, Arthur Lira, Benedita da Silva, Carla Zambelli, Eduardo Bolsonaro, Kim Kataguiri, Joice Hasselmann, Marcelo Freixo, Tabata Amaral e Túlio Gadêlha. Tal amostra visa representar a diversidade de atores políticos eleitos para essa casa legislativa, sobretudo quanto ao espectro político e à dinâmica de situação-oposição. O Basômetro do jornal O Estado de São Paulo aponta, por exemplo, que Marcelo Freixo, Alessandro Molon e Benedita da Silva estavam alinhados ao governo Bolsonaro — que se estendeu de 2019 a 2022 — em até 25% das votações na Câmara dos Deputados. Em contrapartida, esse índice de governismo é superior a 90% para Carla Zambelli, Eduardo Bolsonaro, Joice Hasselmann e Kim Kataguiri [23]. Na Figura 4.3, apresentamos detalhadamente a frequência relativa das categorias de tópicos (e os três principais tópicos latentes) nos discursos dos parlamentares selecionados.

A categoria "Garantias Fundamentais" foi aquela que recebeu maior atenção desse subconjunto de deputados e deputadas, o que se assemelha às distribuições identificadas no *corpus* completo. Contudo, o percentual de dedicação divergiu bastante entre os parlamentares e, em alguns casos, essa sequer foi a categoria mais comum nos discursos. Na ausência da pandemia de COVID-19 e de seu profundo impacto na frequência do tópico "Saúde", é possível até que essas variações se tornassem mais evidentes. Além disso, a dinâmica entre as duas categorias restantes parece estar relacionada ao posicionamento desses indivíduos no espectro político. Os parlamentares cuja atuação política está ao centro (como Arthur Lira) tendem a se dedicar mais ao "Desenvolvimento Econômico", enquanto um maior foco em "Equidade e Inclusão" surge daqueles mais à esquerda (como Túlio Gadêlha). Em paralelo, os parlamentares que estão à direita do espectro político aparentam se dividir em dois grupos: um que ascende através das pautas econômicas que figuram na categoria "Desenvolvimento Econômico" (como Kim Kataguiri) e outro que se destaca por seu combate às pautas sociais

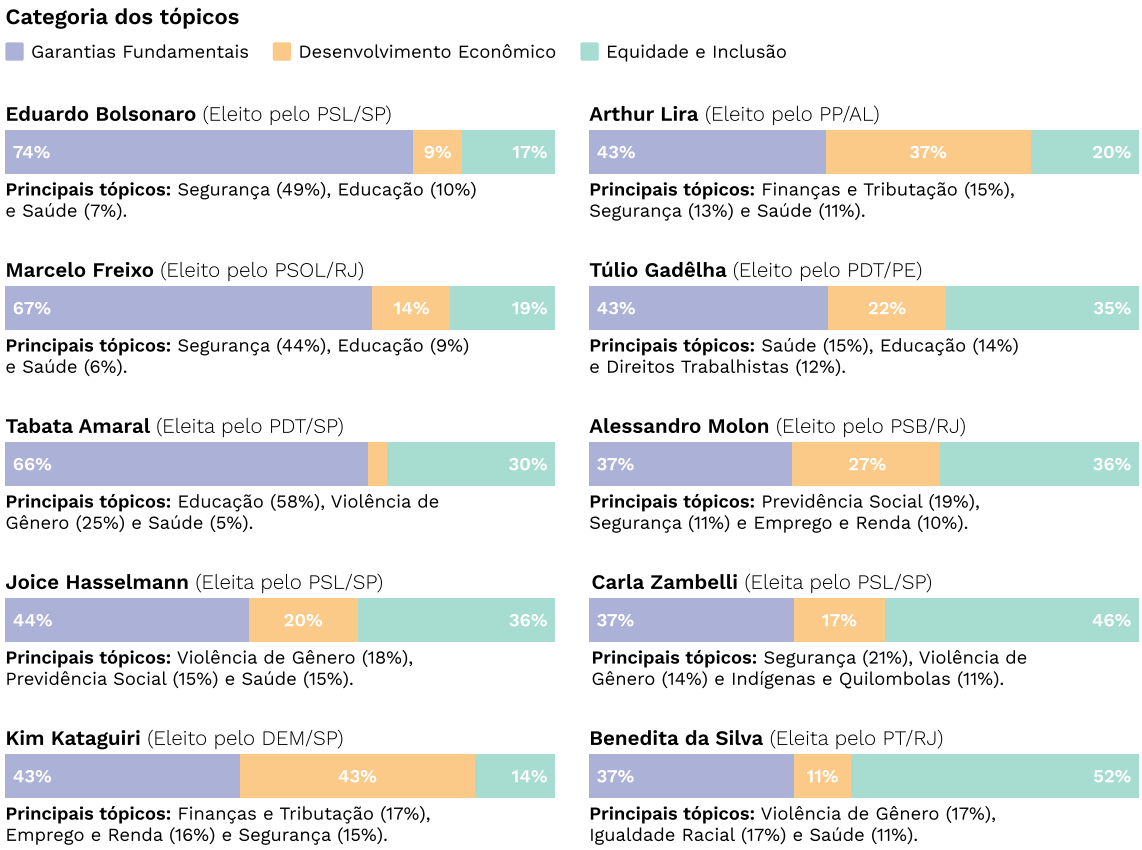

Figura 4.3: Principais tópicos latentes (e categorias de tópicos) nos discursos de uma amostra de dez parlamentares com amplo destaque midiático.

incluídas na categoria "Equidade e Inclusão" (como Carla Zambelli).

Ainda, apesar de suas atuações políticas tão diferentes, as quatro deputadas da amostra possuem "Violência de Gênero" entre os tópicos mais frequentes em seus discursos. Por outro lado, os deputados quase não se dedicam a assuntos dessa natureza, um indício de que as características sociodemográficas podem estar associadas à atuação política de nossos representantes no Poder Legislativo. Vale destacar, entretanto, que tópicos prioritários concomitantes não necessariamente se refletem num mesmo posicionamento sobre esses temas. Por exemplo, os deputados Eduardo Bolsonaro e Marcelo Freixo apresentaram tópicos principais iguais e com percentuais bastante similares. Todavia, quase sempre atuam como antagonistas no cenário político nacional (e estadual no Rio de Janeiro), especialmente no que se refere ao tópico "Segurança" e aos assuntos que o permeiam.

Buscando expandir essas descobertas, também selecionamos seis partidos políticos com

amplo destaque midiático para análise. São eles: Movimento Democrático Brasileiro (MDB), Partido Liberal (PL), Partido da Social Democracia Brasileira (PSDB), Partido Socialismo e Liberdade (PSOL), Partido dos Trabalhadores (PT) e União Brasil (UNIÃO). Na Figura 4.4, a frequência relativa das categorias de tópicos (e os três principais tópicos latentes) em discursos dos parlamentares desses partidos.

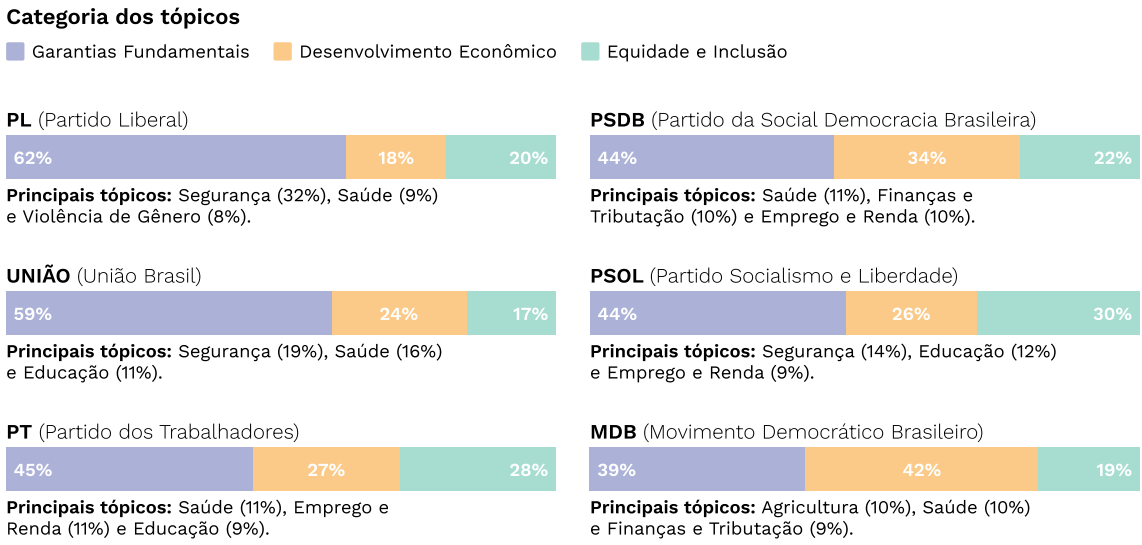

Figura 4.4: Principais tópicos latentes (e categorias de tópicos) nos discursos de uma amostra de seis partidos políticos com amplo destaque midiático.

Sob a perspectiva de partidos políticos, não só a frequência da categoria "Garantias Fundamentais" se torna mais uniforme, como também a dedicação dos parlamentares de centro ao "Desenvolvimento Econômico" se torna mais notória. Por sua vez, discursos associados a "Equidade e Inclusão" são mais frequentes entre os partidos de esquerda. Os principais tópicos de cada partido também apresentaram o comportamento esperado, mas identificamos uma exceção: o único tópico da categoria "Equidade e Inclusão" provém de um partido de direita, o Partido Liberal (PL). Essa ocorrência pode ser um reflexo de como o combate às pautas sociais cresceu entre os parlamentares de direita (especialmente extrema-direita) nos últimos anos, por vezes em detrimento de quaisquer outros temas.

Diante desses resultados, a comparação entre as distribuições dos nossos tópicos latentes com caracterizações prévias, oriundas da mídia e do Basômetro, aponta para a validade aparente da modelagem de tópicos utilizando o BERTopic. Esse modelo permitiu-nos caracterizar

adequadamente os discursos, compreender as pautas prioritárias dos diferentes parlamentares e, através delas, contrastá-los. Em concordância com os resultados de Batista e de Moreira acerca de legislaturas anteriores [6] [49], as ênfases em temas sociais ou econômicos também foram úteis para a caracterização das atuações políticas e apresentaram associação a diferentes características dos parlamentares, como a filiação partidária ou o gênero. Os assuntos de cunho econômico são mais frequentes nos discursos daqueles que se posicionam ao centro ou à direita do espectro político, enquanto os assuntos de cunho social são mais recorrentes entre os indivíduos à esquerda do espectro político. No entanto, identificamos que certos expoentes da extrema-direita brasileira fogem dessa tendência observada nas legislaturas anteriores, apresentando notável dedicação aos tópicos atribuídos à categoria "Equidade e Inclusão".

Esse fenômeno condiz com as profundas transformações no cenário político brasileiro durante os últimos anos, mas sugere que as ênfases temáticas deixaram de ser suficientes para distinguir os indivíduos à esquerda ou à direita no espectro político, o que limita sua capacidade na caracterização de atuações políticas no contexto brasileiro. Nesse sentido, é importante ressaltarmos que as frequências dos nossos tópicos latentes nos discursos apenas demonstram o grau de dedicação dos parlamentares aos diferentes temas, sem necessariamente representar que seus posicionamentos sejam favoráveis — ou, tampouco, contrários — aos assuntos abordados.

Capítulo 5

Estimativa de Pontos Ideais

Neste capítulo, discorremos sobre a etapa de estimativa de pontos ideais deste estudo. A priori, apresentamos a técnica adotada, descrevemos o processo de treinamento do modelo e avaliamos comparativamente as estimativas produzidas. Posteriormente, validamos o uso conjunto da modelagem de tópicos latentes com a estimativa de pontos ideais baseados em texto. Por fim, analisamos parlamentares, partidos políticos e classificações ideológicas através dos pontos ideais estimados.

5.1 *Text-Based Ideal Point*

O *Text-Based Ideal Point* é uma técnica do estado-da-arte para a estimativa de pontos ideais que se propõe a quantificar as preferências político-ideológicas de indivíduos usando exclusivamente dados textuais. Em contraponto às abordagens típicas, que se baseiam em votações no âmbito do Poder Legislativo, esse modelo analisa e contrasta como os diferentes atores políticos discursam (ou escrevem) sobre um conjunto compartilhado de tópicos latentes. Conforme descrito pelo conceito de *political framing*, os atores políticos adaptam seus discursos com o intuito de influenciar a opinião pública quanto aos assuntos discutidos. São essas nuances de vocabulário que caracterizam as preferências político-ideológicas e permitem que o TBIP estime pontos ideais, até mesmo, para indivíduos não parlamentares — como jornalistas ou candidatos(as) à presidência. Convém destacar que o TBIP não faz uso de conjuntos de dados auxiliares ou impõe restrições adicionais ao *corpus* analisado. Em contraponto, outras técnicas baseadas em dados textuais costumam demandar,

por exemplo, rótulos partidários, votos legislativos e/ou que os documentos abordem um único tema [26] [39] [67].

O TBIP adota uma abordagem probabilística, de aprendizagem não supervisionada e produz pontos ideais unidimensionais. Fundamentado na fatoração de Poisson [27], esse modelo considera as frequências das palavras no *corpus* e os atores políticos como variáveis observáveis, enquanto os pontos ideais desses indivíduos e os tópicos discutidos emergem como variáveis latentes nos dados textuais. Na prática, quando um conjunto de atores políticos aborda o mesmo tópico latente em seus discursos, são as palavras que compõem seus respectivos vocabulários que determinam como posicioná-los a respeito desse tema. Esses posicionamentos são expressos numericamente — com valores negativos, neutros ou positivos — para cada tópico latente do *corpus*. Em conjunto, esses valores definem o ponto ideal a ser atribuído a cada indivíduo. É importante notar, contudo, que esses valores são apenas representações espaciais e não equivalem diretamente a preferências político-ideológicas favoráveis, indiferentes ou contrárias aos assuntos debatidos.

Na publicação original do TBIP, a estimativa de pontos ideais foi realizada através dos discursos legislativos e dos *tweets* de senadores estadunidenses. Os pontos ideais estimados apresentaram forte correlação àqueles baseados em votações e, ainda, foram capazes de distinguir os dois principais partidos políticos nesse contexto: os Democratas e os Republicanos. Além disso, também se mostrou possível estimar pontos ideais coerentes para os pré-candidatos e pré-candidatas Democratas à presidência estadunidense durante as eleições de 2020. Esses resultados sugerem que o TBIP pode suprir as limitações mais comuns às técnicas para estimativa de pontos ideais como, por exemplo, a dependência das votações legislativas. Não obstante, o TBIP e suas variações têm contribuído com a expansão dos pontos ideais para novos domínios, como a caracterização das políticas financeiras de bancos centrais europeus [25] ou a análise de variações temporais nos posicionamentos políticos [34].

5.2 Treinamento do modelo

O TBIP não estava incluso em nenhuma biblioteca da linguagem Python à época da condução de nossos experimentos, mas já havia sido disponibilizado em formato de *script*

através de um repositório GitHub[1]. Em específico, adotamos uma implementação baseada em Numpy[2], Scipy[3] e Tensorflow[4]. Para além de uma estrutura distinta no código-fonte, deparamo-nos com uma documentação bastante concisa e com pouquíssimos exemplos de uso em comparação às técnicas mais tradicionais para a estimativa de pontos ideais. Tendo em vista as limitações desse contexto, optamos por reproduzir o método de treinamento adotado para a publicação original do TBIP. As únicas modificações implementadas, portanto, tinham o objetivo de viabilizar o uso de nossos dados e/ou dos recursos computacionais disponíveis.

Nesse sentido, se fez necessário converter o *corpus* em quatro novos arquivos: o mapa de indivíduos, o vocabulário, a matriz de frequências termo-documento e o índice de oradores. No primeiro, estão listados os nomes dos 577 parlamentares associados ao *corpus*, de modo que a posição do nome atue também como o identificador único de cada deputado(a). O vocabulário, por sua vez, é o conjunto de palavras distintas que ocorrem nos 389.562 discursos (ordenadas alfabeticamente). Assim, seja D o número de documentos no *corpus* e P o número de palavras distintas nesses documentos, a matriz de frequência termo-documento pode ser descrita como uma matriz esparsa M com formato $D \times P$, tal que o número de ocorrências da p-ésima palavra no d-ésimo documento está armazenado em M_{dp}. Por fim, o índice de oradores é um vetor D-dimensional que associa os discursos aos respectivos parlamentares ou, mais especificamente, um vetor cujo d-ésimo valor será o identificador único do parlamentar associado ao d-ésimo discurso na coleção de documentos.

Junto à reestruturação dos conjuntos de dados, também buscamos reproduzir as operações de pré-processamento adotadas para publicação original do modelo. Nessa etapa, identificamos que a maioria dessas operações esteve presente na construção de nosso *corpus*, como o descarte de documentos curtos ou a remoção de oradores com poucos discursos. As adições ao pré-processamento dos dados se resumiram, então, a ajustes no vocabulário dos documentos. Foram removidas as palavras mencionadas por menos de dez parlamentares distintos e/ou que pertencessem ao conjunto de *stopwords* descrito na Seção 4.2. Posteriormente, removemos os documentos que se tornaram vazios em decorrência dessas filtragens. Ademais, mantivemos os valores padrão em todos os hiperparâmetros do TBIP durante a execução de nossos ex-

[1] https://github.com/keyonvafa/tbip

[2] https://pypi.org/project/numpy

[3] https://pypi.org/project/scipy

[4] https://pypi.org/project/tensorflow

perimentos. Os principais hiperparâmetros identificados (e seus respectivos valores) estão apresentados na Tabela 5.1.

O *batch_size* é o hiperparâmetro que define a quantidade de discursos que são processados simultaneamente a cada iteração no treinamento do TBIP. Esses subconjuntos de documentos são denominados lotes ou *batches*. Tal variável não só afeta o uso de memória e a velocidade do treinamento, mas também a estabilidade do aprendizado do TBIP. O *learning_rate* estabelece a magnitude dos ajustes que os parâmetros do modelo recebem durante o treinamento (visando minimizar a função de perda). Enquanto valores baixos tendem a desacelerar esse processo, os valores inadequadamente altos podem até impedir a convergência no treinamento. Por sua vez, o *num_epochs* se refere ao número de vezes que o treinamento do modelo percorrerá o *corpus* completo, permitindo-o aprender com as repetições e identificar padrões complexos nos dados.

Tabela 5.1: Principais hiperparâmetros para a estimativa de pontos ideais usando o modelo TBIP.

Hiperparâmetro	**Valor Padrão**
batch_size	512
learning_rate	0,01
num_epochs	10.000
num_topics	50

Esses três hiperparâmetros são recorrentes em diversos algoritmos de Aprendizagem de Máquina e adotam valores tradicionais para experimentos sem ajuste fino. Em contrapartida, o *num_topics* está diretamente associado ao contexto da estimativa de pontos ideais baseados em texto. Essa variável representa o número de tópicos latentes que o modelo assume existir no *corpus* e acerca dos quais os termos do vocabulário serão associados de modo negativo, neutro ou positivo. Ao pressupor 50 tópicos latentes no *corpus*, o TBIP está alinhado aos resultados obtidos durante nossa etapa de modelagem de tópicos, o que poderia contribuir com seu desempenho nesse domínio específico.

Nossos experimentos utilizaram uma amostra dos documentos remanescentes no *corpus*,

abordagem que foi necessária para viabilizar sua execução com os recursos computacionais disponíveis. Para cada parlamentar, selecionamos randomicamente até 200 discursos distintos. Considerando as características selecionadas para a construção do *corpus*, isso significa dizer que cada indivíduo será representado por um conjunto contendo entre 100 e 200 documentos. Mesmo assim, no comparativo com a publicação original do TBIP, mantivemo-nos acima das proporções média e mediana de discursos por parlamentar. Dessa maneira, conseguimos estimar pontos ideais para todos os 577 parlamentares associados ao nosso *corpus* usando, exclusivamente, os seus respectivos discursos nos eventos da Câmara dos Deputados. Os recursos computacionais utilizados durante o treinamento de modelos TBIP estão descritos na Tabela A.3 do Apêndice A.

5.3 Avaliação dos pontos ideais

Com a etapa de treinamento encerrada, deparamo-nos com um desafio subsequente: validar os pontos ideais estimados por nosso modelo. A abordagem típica para essa avaliação consiste em recorrer a pesquisas anteriores e comparar estatisticamente os seus resultados aos novos pontos ideais produzidos. Entretanto, durante a revisão da literatura, não identificamos quaisquer trabalhos que estimassem o posicionamento político individual de deputados e deputadas federais em exercício durante a 55ª e a 56ª Legislaturas. São mais comuns os estudos dessa natureza que se dedicam a analisar agrupamentos desses indivíduos, como as bancadas partidárias [62] ou os partidos políticos [70].

Não é incomum que haja evidente discrepância entre a ideologia de um partido político brasileiro e a atuação de algum (ou alguns) de seus parlamentares eleitos. Um exemplo notório é o Cabo Daciolo que, embora atue (e se autodeclare) à direita no espectro político, foi eleito para a 55ª Legislatura da Câmara dos Deputados enquanto filiado ao Partido Socialismo e Liberdade (PSOL) do Rio de Janeiro. Ainda assim, se agruparmos adequadamente os parlamentares conforme a filiação partidária, é razoável assumir que o efeito desses *outliers* seria pouco perceptível. Os posicionamentos políticos agregados seriam, portanto, aproximadamente compatíveis às ideologias dos respectivos partidos políticos. Nesse sentido, durante a avaliação dos pontos ideais estimados pelo TBIP, baseamo-nos nos dados apresentados por Bolognesi, Ribeiro e Codato em sua publicação "Uma Nova Classificação Ideológica dos

Partidos Políticos Brasileiros" [11].

Para esse artigo, os autores aplicaram uma *survey* junto aos integrantes da Associação Brasileira de Ciência Política (ABCP) em julho de 2018, convidando-os a classificar 35 partidos políticos brasileiros num eixo esquerda-direita. Os respondentes utilizaram uma escala espacial no intervalo de 0 a 10, mas, como estímulo a produzir uma classificação por proximidade, não recebiam acesso aos pontos numéricos (apenas aos espaciais). Ao todo, os autores obtiveram respostas de 519 indivíduos, majoritariamente pessoas com doutorado ou mestrado na área da Ciência Política. Essas respostas os permitiram estimar a posição ideológica média de cada partido político e, posteriormente, propor sua classificação em sete categorias sequenciais: Extrema-esquerda, Esquerda, Centro-esquerda, Centro, Centro-direita, Direita e Extrema-direita.

Após pequenas adaptações (descritas no Apêndice B), associamos a classificação ideológica proposta às informações sobre filiação partidária dos parlamentares que figuram em nosso *corpus*. Assim, apresentamos os 577 pontos ideais estimados por nosso modelo TBIP na Figura 5.1. Em sequência, na Tabela 5.2, estão apresentados os pontos ideais medianos de cada categoria nessa classificação ideológica. Para posterior comparabilidade, os valores estimados foram normalizados para o intervalo de -1,0 a 1,0 (inclusive).

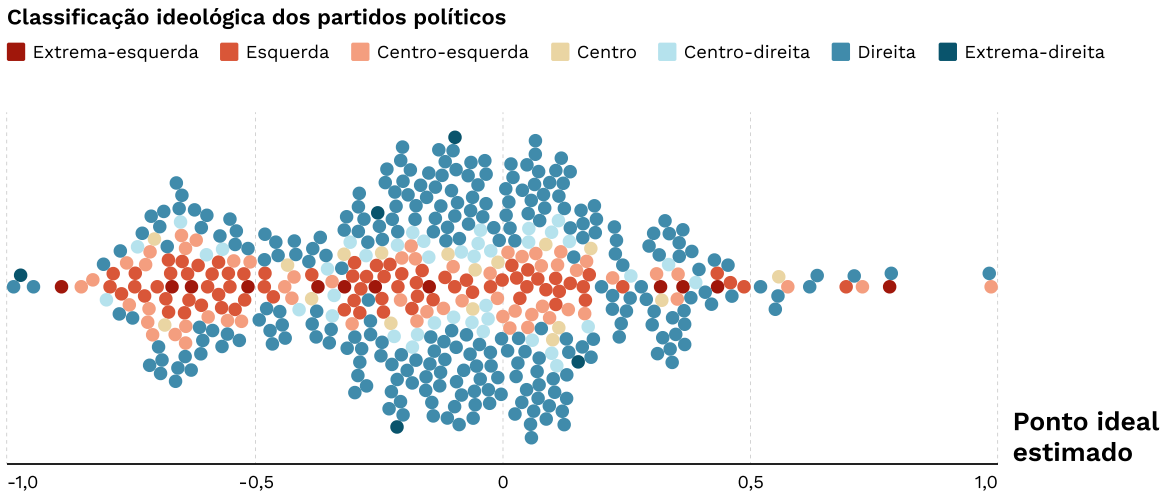

Figura 5.1: Pontos ideais estimados para 577 deputados e deputadas federais com base em seus discursos.

Tabela 5.2: Pontos ideais medianos para as categorias da classificação ideológica.

Classificação Ideológica	**Ponto Ideal Mediano**	**Ranking**
Extrema-esquerda	-0,29	1º
Esquerda	-0,28	2º
Centro-esquerda	-0,16	4º
Centro	-0,06	7º
Centro-direita	-0,08	6º
Direita	-0,09	5º
Extrema-direita	-0,22	3º

A distribuição dos pontos ideais sugere que os parlamentares se concentram em dois ou três grupos principais, tal que os valores estimados estão quase exclusivamente entre -0,25 e 0,50 (abrangendo menos de 40% da amplitude total). Esse comportamento é antagônico à estrutura pluripartidária atual e poderia ser equivocadamente adotado como argumento favorável à redução do número de partidos políticos em nosso país. Entretanto, não há relação aparente das ideologias partidárias com a alocação de seus parlamentares nesses grupos. Além disso, observando os pontos ideais medianos, constatamos que existe discordância entre a classificação ideológica e os valores estimados. Há menor disparidade entre "Extrema-direita" e "Esquerda" do que entre "Extrema-direita" e "Direita", por exemplo. Essas incongruências também foram observadas em análises de menor granularidade, como a diferença de apenas 0,05 entre os pontos ideais atribuídos aos deputados Túlio Gadêlha e Cabo Daciolo.

Mesmo inserido num contexto com pouquíssimos partidos políticos (quase exclusivamente Democratas e Republicanos), o modelo apresentado na publicação original do TBIP foi capaz de estimar pontos ideais que se distribuem mais uniformemente e com evidente associação às ideologias partidárias — e dos próprios parlamentares. Essas características não só permitem uma melhor representação do espectro político, como também contribuem com um dos principais objetivos dos pontos ideais: viabilizar a comparação entre quaisquer pares de parlamentares do conjunto. Nesse sentido, é possível afirmar que os pontos ideais estimados por nosso modelo não representam bem os parlamentares sob análise e que, portanto, seu uso

é inadequado no contexto brasileiro.

Diante do desempenho insatisfatório desse modelo inicial, implementamos algumas modificações em seu processo de treinamento, visando aprimorar os pontos ideais estimados. Essas modificações focaram nos hiperparâmetros do modelo e no pré-processamento dos dados, buscando garantir maior robustez ao aprendizado e mais precisão aos ajustes de parâmetros durante as iterações do treinamento. Em específico, elevamos o *batch_size* de 512 para 1024, reduzimos o *learning_rate* de 0,01 para 0,005 e, ainda, selecionamos apenas os parlamentares cujo número de discursos no *corpus* fosse igual ou superior ao tamanho máximo das amostras (200 documentos). Mesmo assim, os novos pontos ideais produzidos se mostraram bastante semelhantes aos resultados iniciais do TBIP.

5.4 Uso dos tópicos latentes

Posteriormente, observando os tópicos latentes identificados pelo BERTopic e suas respectivas distribuições, hipotetizamos que o desempenho do TBIP poderia ter sido prejudicado pela desproporcionalidade dos temas no *corpus*. Os tópicos da categoria "Processos Legislativos", por exemplo, representam mais da metade dos documentos. No entanto, fornecem pouca informação acerca dos posicionamentos políticos individuais, uma vez que costumam estar atrelados aos ritos e procedimentos mandatórios da Câmara dos Deputados. Já os tópicos da categoria "Equidade e Inclusão" costumam causar polarizações e discordâncias entre os parlamentares, mas agregam menos de 8% dos documentos.

Nesse sentido, retornamos ao treinamento do modelo TBIP com o objetivo de adaptá-lo aos tópicos latentes de nosso *corpus*. Para isso, adotamos o pré-processamento de dados e os valores de hiperparâmetros do primeiro modelo. A principal diferença se deu na amostragem dos discursos por parlamentar, que deixou de considerar todos os documentos disponíveis. Essas amostras foram selecionadas apenas dentre os 139.049 discursos referentes às três categorias de tópicos latentes mais relevantes: "Garantias Fundamentais", "Desenvolvimento Econômico" e "Equidade e Inclusão". Em decorrência dessa modificação, 89 dos 577 parlamentares não mais atingiam o tamanho mínimo das amostras (100 documentos) e, por isso, não tiveram seus pontos ideais estimados. Novamente, todos os valores das estimativas foram normalizados para o intervalo de -1,0 a 1,0 (inclusive) e, por consistência, igualamos

o valor do hiperparâmetro *n_topics* ao número de tópicos inclusos nas três categorias de interesse, reduzindo-o de 50 para 28.

Na Figura 5.2, estão apresentados os 488 pontos ideais estimados pelo novo modelo TBIP. Ademais, apresentamos novos pontos ideais medianos para cada categoria da classificação ideológica na Tabela 5.3. É notável como os valores estimados se distribuem de maneira mais uniforme do que os pontos ideais anteriores, contribuindo com as análises comparativas acerca dos posicionamentos políticos de deputados e deputadas. Não obstante, esses pontos ideais também apresentam evidente associação à classificação ideológica dos partidos políticos. Por exemplo, os parlamentares cuja filiação partidária foi atribuída às categorias "Extrema-esquerda", "Esquerda" e "Centro-esquerda" se concentram em valores inferiores a -0,15. Entre as poucas ocorrências de outras categorias abaixo de tal limiar, identificamos predominância de parlamentares cujas atuações políticas não se alinham às suas respectivas filiações partidárias. É o caso de André Janones do Avante (AVANTE) de Minas Gerais, cujo ponto ideal foi estimado em -0,98 mesmo sendo eleito por um partido político da categoria "Centro-direita", o que está em concordância com seus posicionamentos à esquerda no espectro político.

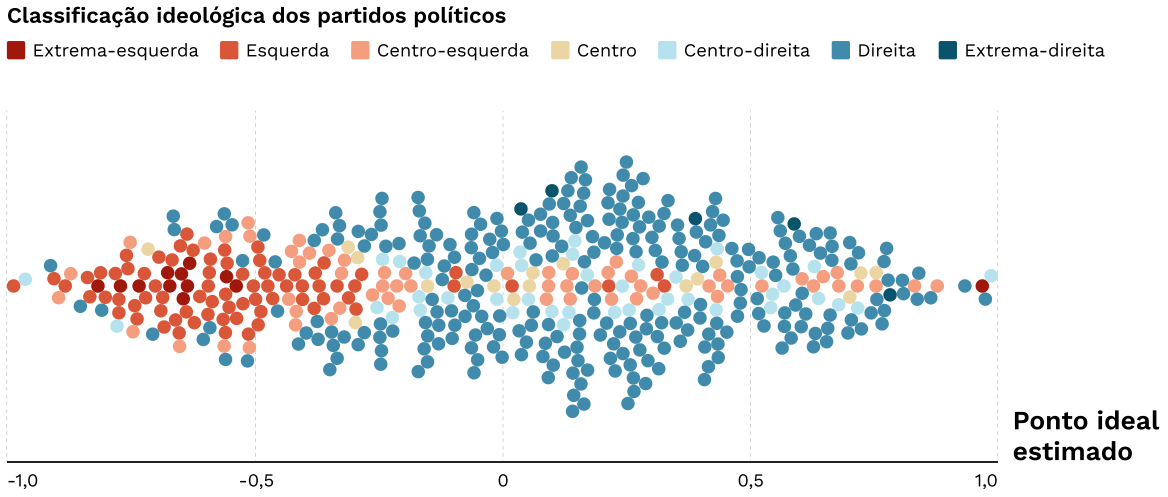

Figura 5.2: Pontos ideais estimados para 488 deputados e deputadas federais após filtragem de seus discursos usando os tópicos latentes do BERTopic.

Tabela 5.3: Pontos ideais medianos para as categorias da classificação ideológica após filtragem dos discursos usando os tópicos latentes do BERTopic.

Classificação Ideológica	**Ponto Ideal Mediano**	**Ranking**
Extrema-esquerda	-0,65	1º
Esquerda	-0,57	2º
Centro-esquerda	-0,19	3º
Centro	0,05	4º
Centro-direita	0,15	5º
Direita	0,17	6º
Extrema-direita	0,39	7º

Com esse novo modelo TBIP, os pontos ideais medianos também se mostraram capazes de ordenar e contrastar adequadamente as categorias da classificação ideológica. Essas estimativas permitem observar, por exemplo, como a categoria "Centro" está mais próxima da "Centro-direita" do que da "Centro-esquerda" ou como as categorias "Direita" e "Centro-direita" se diferenciam em apenas 0,02. Em ambos os casos, os pontos ideais medianos parecem revelar características implícitas do cenário político durante a 55ª e a 56ª Legislaturas da Câmara dos Deputados. Conforme supracitado, a classificação ideológica adotada se baseia nos valores médios atribuídos por especialistas da ABCP às ideologias partidárias. Por simplicidade, denominaremos esses valores como "ideologias médias". Assim, na Figura 5.3, apresentamos as ideologias médias dos partidos políticos brasileiros e comparamo-nas aos pontos ideais medianos que foram estimados por nosso modelo.

O coeficiente de correlação entre as distribuições apresentadas é de 0,79, o que configura uma correlação forte e positiva. Inclusive, considerando a desproporcionalidade das escalas, a diferença média entre os valores atribuídos aos partidos políticos é de apenas 12,7% da amplitude dos eixos. Essa discrepância pequena sugere que nossa estimativa de pontos ideais baseada em discursos dos parlamentares se assemelha à percepção de especialistas acerca do Poder Legislativo brasileiro. Além disso, identificamos padrões que ocorrem tanto nas ideologias médias, quanto nos pontos ideais medianos. Por exemplo, os numerosos

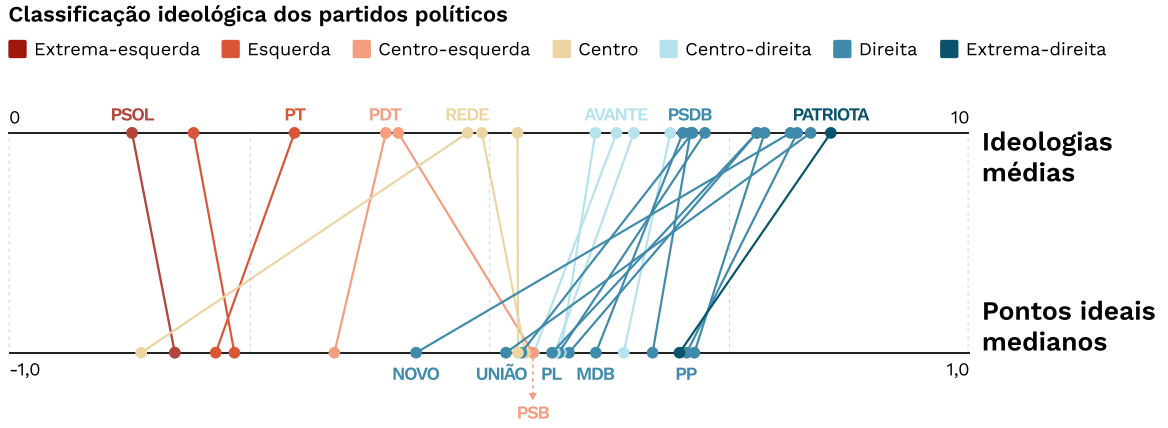

Figura 5.3: Comparativo entre as ideologias médias e os pontos ideais medianos de 23 partidos políticos brasileiros.

partidos políticos das categorias "Centro-direita", "Direita" e "Extrema-direita" receberam valores bastante similares, indicando concordância na atuação política de seus parlamentares. Outrossim, buscamos avaliar detalhadamente os partidos políticos que apresentaram maior divergência entre as distribuições.

A Rede Sustentabilidade (REDE) teve seu ponto ideal mediano estimado em -0,73, posicionando-a junto aos partidos de "Extrema-esquerda" e "Esquerda" apesar de ter sido classificada como "Centro". Contudo, identificamos que tal valor mediano se refere à única parlamentar desse partido político que figura entre os novos pontos ideais: Joenia Wapichana. Eleita pelo estado de Roraima, a atuação dessa deputada federal é reconhecida pela defesa dos povos originários e dos direitos humanos, o que justifica o ponto ideal mediano atribuído à sua filiação partidária. Outros dois partidos políticos apresentaram valores consideravelmente divergentes entre as duas distribuições: o Partido Novo (NOVO) e o União Brasil (UNIÃO), cujos pontos ideais medianos foram estimados em -0,15 e 0,03 (nessa ordem). Eles estão classificados na categoria "Direita" e possuem, respectivamente, 7 e 55 parlamentares com pontos ideais estimados por nosso modelo. Analisando-os, identificamos que ocorre polarização entre os indivíduos eleitos através desses partidos políticos. Por exemplo, enquanto os pontos ideais de quatro parlamentares do Partido Novo (NOVO) estão no intervalo de -0,37 a -0,16, os demais estão no intervalo de 0,26 a 0,44. Esse fenômeno impacta bastante nos valores medianos atribuídos aos partidos políticos e parece esclarecer as discordâncias quanto

à classificação ideológica, visto que ela parece representar apenas o grupo mais numeroso em cada partido político.

Nesse cenário, avaliamos que os tópicos latentes identificados pelo BERTopic contribuíram efetivamente com o desempenho do modelo TBIP, viabilizando seu uso no contexto do Poder Legislativo brasileiro através da filtragem dos documentos mais relevantes em nosso *corpus*. Desse modo, pudemos estimar pontos ideais que, quando agregados por partido político ou classificação ideológica, se mostraram equiparáveis à percepção de especialistas na área de Ciência Política.

5.5 Análise de posicionamento político usando pontos ideais

Quando agregados por mediana, os pontos ideais estimados pelo modelo TBIP (e com apoio dos tópicos latentes identificados pelo BERTopic) se mostraram em consonância com a classificação ideológica de partidos políticos adotada como referência neste estudo. Todavia, essas estimativas também permitem-nos analisar os posicionamentos políticos na menor granularidade, isto é, comparar deputados e deputadas federais individualmente. Para exemplificar, selecionamos os dez parlamentares que compõem a amostra descrita na Seção 4.4 e, na Figura 5.4, destacamos os pontos ideais estimados a partir de seus respectivos discursos. Aqui, é importante ressaltar que as posições verticais dos pontos ideais foram preservadas nesse gráfico exclusivamente para garantir sua comparabilidade à distribuição apresentada na Figura 5.2 e que, portanto, não representam nenhuma outra informação.

Na amostra, cinco parlamentares possuem filiações partidárias classificadas como "Extrema-esquerda", "Esquerda" ou "Centro-esquerda": Alessandro Molon, Benedita da Silva, Marcelo Freixo, Tabata Amaral e Túlio Gadêlha. Excetuando Tabata Amaral (cujo ponto ideal foi estimado em 0,12), os valores associados aos indivíduos desse subconjunto variam de -0,74 a -0,52. A pequena amplitude do intervalo sugere considerável semelhança entre os posicionamentos políticos desses parlamentares, enquanto suas atuações em defesa de pautas sociais e pela redução de desigualdades são traduzidas em pontos ideais dispostos à esquerda no eixo. Até mesmo a divergência do ponto ideal estimado para Tabata Amaral condiz com o contexto político durante seu mandato. Essa parlamentar esteve alinhada ao governo Bolsonaro em 52% das votações na Câmara dos Deputados e, ainda que defenda

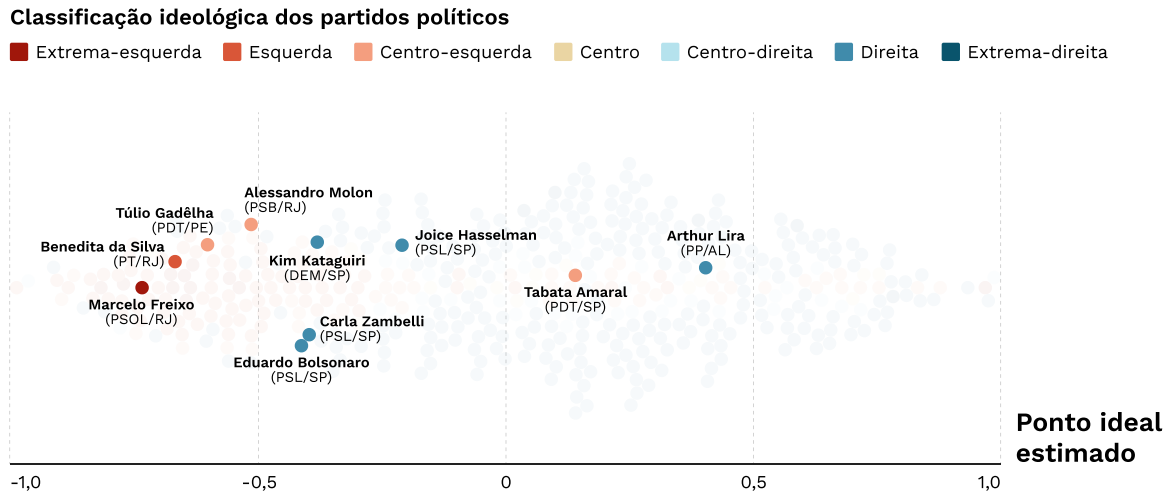

Figura 5.4: Pontos ideais baseados em discursos para uma amostra de dez parlamentares com amplo destaque midiático.

uma agenda política reformista e entenda o Estado como importante agente de transformação social, sua atuação costuma ser descrita como ao centro do espectro político [52]. Logo, é compreensível que seu ponto ideal se distancie da maioria dos deputados e deputadas que possuem filiação partidária similar.

Já Arthur Lira, o atual presidente da Câmara dos Deputados, teve seu ponto ideal estimado em 0,41 e foi posicionado à direita dos demais parlamentares da amostra. O deputado alagoano é uma das principais lideranças do Centrão, um grupo de partidos políticos que costuma receber cargos importantes do Poder Executivo e cujas atuações são frequentemente descritas como fisiológicas e clientelistas [8] [75]. Nesse âmbito, a estimativa associada ao parlamentar parece se adequar ao seu alinhamento com o governo Temer (98% das votações) e, em menor proporção, com o governo Bolsonaro (87% das votações). Seu ponto ideal o distancia profundamente dos indivíduos associados à esquerda, mas sem mesclá-lo ao outro extremo do espectro político. Diversos deputados e deputadas federais do Centrão receberam pontos ideais similares ao de Lira. Por exemplo, os valores atribuídos a Mandetta e André Fufuca foram 0,33 e 0,32 (respectivamente).

Em contrapartida, os pontos ideais atribuídos aos quatro parlamentares remanescentes da amostra divergem das nossas expectativas iniciais. Em específico, temos Eduardo Bolsonaro com -0,40, Kim Kataguiri com -0,38, Carla Zambelli com -0,38 e Joice Hasselmann com

-0,21. Esses indivíduos compunham o núcleo da base parlamentar do governo Bolsonaro durante seus mandatos (ou parte deles), alinhando-se ao Poder Executivo em mais de 90% das votações realizadas entre 2019 e 2022. Nesse sentido, é compreensível que seus pontos ideais estejam tão próximos, sobretudo quando observamos o relativo distanciamento de Joice Hasselmann, que rompeu com o governo Bolsonaro durante seu mandato e tem enfrentado um extenso processo de violência política de gênero desde então [64]. No entanto, inicialmente era esperado que os valores estimados para esses parlamentares fossem próximos a 1,0, posicionando-os na extremidade direita do eixo.

A atenção midiática e a ascensão política desses indivíduos estão intimamente relacionadas às rupturas institucionais que o Brasil vivenciou na última década, particularmente no *impeachment* da ex-presidenta Dilma Rousseff em 2016 e após as eleições gerais de 2018. No decorrer desse período, se observou uma reativação da direita brasileira através de pautas conservadoras, ultraliberais e, até mesmo, reacionárias [65] [66]. Em face desse contexto histórico, investigamos detalhadamente os discursos associados aos dez parlamentares de nossa amostra, buscando aprofundar nossa compreensão acerca de seus pontos ideais e identificar causas para os valores atribuídos aos deputados e às deputadas que compõem essa "nova direita brasileira".

Essa análise revelou um notável contraste entre os discursos de parlamentares cujas atuações são associadas à direita no espectro político: há aqueles que se mostram mais "pragmáticos" e aqueles que tendem a ser mais "ideológicos". No primeiro grupo, os discursos são quase exclusivamente dedicados aos procedimentos legislativos e trâmites burocráticos da Câmara dos Deputados, sem que haja predominância de temas prioritários para cada deputado(a). São discutidas diversas Propostas de Emenda à Constituição, proposições de lei e diretrizes orçamentárias, por exemplo. Já nos discursos do segundo grupo, ocorre uma prevalência de debates e antagonismos quanto às causas sociais e progressistas, frequentemente direcionados a indivíduos e partidos políticos específicos da oposição. O tom de conflito exerce uma influência profunda não apenas no vocabulário desse subconjunto de parlamentares, mas também no de indivíduos alinhados à esquerda. Por exemplo, o termo "Bolsonaro" foi mencionado 128 vezes por Kim Kataguiri, 95 vezes por Marcelo Freixo e 83 vezes por Eduardo Bolsonaro em seus respectivos discursos na amostra. De maneira semelhante, o termo "PT" foi mais citado por Joice Hasselmann (com 57 ocorrências) do que

por Benedita da Silva (com 44 ocorrências).

Essa atuação antagônica às causas sociais e progressistas também foi observada nos discursos de outros parlamentares que, mesmo não pertencendo à nossa amostra de interesse, demonstravam alinhamento com a "ala mais ideológica" da direita brasileira durante seus mandatos. São exemplos Bibo Nunes, Daniel Silveira e Jair Bolsonaro, cujos pontos ideais foram estimados em -0,92, -0,86 e -0,52, respectivamente. Os discursos desses indivíduos incluem desde questionamentos sobre o isolamento social durante a pandemia de COVID-19 até hostilidades direcionadas ao Supremo Tribunal Federal e aos movimentos sociais, como o Movimento dos Trabalhadores Rurais sem Terra. Embora não sejam provenientes de uma análise exaustiva dos discursos, esses resultados sugerem que os parlamentares associados à extrema-direita causam uma "anomalia" na distribuição dos pontos ideais produzidos pelo modelo e, por isso, essas estimativas unidimensionais não descrevem um espectro político esquerda-direita para o cenário político brasileiro.

Logo, em contraponto à publicação original, o TBIP não estimou valores que permitam qualificar os parlamentares como liberais, moderados ou conservadores. Na verdade, os pontos ideais à esquerda parecem representar indivíduos com pautas prioritárias explícitas e ideologias bem definidas, enquanto os pontos ideais à direita representam indivíduos mais pragmáticos e sem ideais políticos evidentes. Essa divergência na interpretação dos pontos ideais pode ter sido causada por diferenças idiomáticas entre os dados textuais analisados, mas está provavelmente relacionada às características intrínsecas ao contexto político brasileiro, como o pluripartidarismo e a diversidade intrapartidária.

Assim, nossas análises evidenciam que as características representadas através dos pontos ideais produzidos pelo TBIP mudam conforme o contexto político sob análise. Para o cenário brasileiro, o modelo aproximou deputados e deputadas cujos discursos incluem assuntos semelhantes, mesmo que haja divergência em seus posicionamentos acerca desses temas. Esse comportamento inesperado mas pode estar relacionado, inclusive, ao elevado número de oradores e de temas discutidos na Câmara dos Deputados. Nossas estimativas ainda são capazes de fundamentar análises políticas importantes, mas substituem as convencionais dicotomias entre esquerda-direita ou liberal-conservador, distinguindo os indivíduos entre os mais ideológicos e os mais pragmáticos. Além disso, os pontos ideais estimados também mostraram-se alinhados às percepções de especialistas da ABCP quanto aos partidos políticos

brasileiros, indicando a validade aparente da estimativa de pontos ideais usando o TBIP.

Nesse cenário, observamos que os parlamentares à esquerda do espectro político tendem a ser mais ideológicos, o que se repete para certos expoentes da extrema-direita brasileira. Por outro lado, os parlamentares associados ao Centrão ou a setores mais tradicionais da direita demonstram maior pragmatismo. Não descartamos, contudo, que exista certa sobreposição entre esses conceitos. Visto que os pontos ideais estimados pelo TBIP são unidimensionais, é possível que estejam representando uma dimensão secundária "esquerda-direita" colapsada sobre a dimensão principal "ideolologia-pragmatismo". Esse fenômeno pode ter sido evidenciado quando, ao agregarmos os pontos ideais conforme a filiação partidária dos indivíduos, encontramos valores fortemente correlacionados à avaliação de especialistas da ABCP, que foi baseada num eixo esquerda-direita. Portanto, supomos que o uso de duas ou três dimensões para a estimativa dos pontos ideais contribuiria com o TBIP em contextos políticos diferentes do estadunidense. Infelizmente, em razão de características desse modelo (como sua baixa explicabilidade), não conseguimos verificar tal hipótese no decorrer deste estudo.

Capítulo 6

Conclusões

Neste capítulo, retomamos os objetivos iniciais deste estudo para discutir suas principais conclusões e contribuições, suas limitações e as eventuais ameaças à validade. Por fim, apresentamos nossas perspectivas quanto às futuras pesquisas na área de PLN aplicado à Ciência Política, particularmente no cenário brasileiro.

6.1 Discussão

Este trabalho se propôs a avaliar o uso de técnicas modernas de PLN na caracterização dos discursos e dos posicionamentos políticos de parlamentares brasileiros. Para isso, adotamos duas abordagens importantes nas áreas da Ciência Política e da Política Computacional: a modelagem de tópicos latentes e a estimativa de pontos ideais. Ainda, o escopo estabelecido neste estudo se refere à análise de transcrições dos discursos de deputados(as) federais em eventos realizados pela Câmara dos Deputados no decorrer da 55ª e 56ª Legislaturas, isto é, abrangendo o período que se estende de 2015 a 2022.

Nesse cenário, desenvolvemos uma base de dados abertos própria, potencialmente inédita e composta por três conjuntos de dados distintos: um sobre parlamentares, um sobre os eventos realizados pela Câmara dos Deputados e um contendo os discursos proferidos nesses eventos. Tendo em vista a importância desses discursos para as pesquisas nas áreas de Ciência Política e Política Computacional, decidimos expandir o intervalo de tempo incorporado em nossa base de dados — em específico, adotando o período de 2003 a 2022 — e, só então, disponibilizá-la publicamente. Com base nos dados textuais extraídos, definimos um *corpus*

composto por 389.562 discursos de 577 parlamentares distintos e dedicado às tarefas de PLN que compõem este estudo. As etapas de modelagem de tópicos latentes e de estimativa de pontos ideais foram conduzidas com duas técnicas do estado-da-arte: o BERTopic e o TBIP, respectivamente.

Nossa avaliação adotou métodos quantitativos e qualitativos, através dos quais se constatou a validade aparente dos resultados obtidos em ambas as etapas. Em outras palavras, demonstramos que os tópicos latentes identificados e os pontos ideais estimados aparentam representar e/ou mensurar adequadamente aquilo a que se propõem. A avaliação quantitativa dos tópicos latentes foi realizada por meio de métricas de coerência, diversidade e qualidade. Quanto à sua avaliação qualitativa, contrastamos as distribuições desses tópicos com caracterizações provenientes da mídia e de ferramentas cívicas sobre parlamentares e partidos políticos brasileiros. De forma análoga, avaliamos o coeficiente de correlação entre os pontos ideais estimados e os valores atribuídos por especialistas da ABCP às ideologias partidárias, além de contrastarmos nossas estimativas com estudos anteriores que descrevessem o posicionamento político dos parlamentares analisados.

Durante a análise exploratória e descritiva do nosso *corpus*, observamos duas nuances importantes acerca da Câmara dos Deputados. A primeira se refere às tendências de concentração da influência política e de baixa renovação nessa casa legislativa. Durante as últimas duas décadas, apenas 1.453 indivíduos ocuparam todos os 2.565 cargos de deputado(a) federal disponíveis e, ainda, três partidos políticos agregaram quase 40% desses cargos. Já a segunda se relaciona às comissões da Câmara dos Deputados e sua reatividade ao interesse público. Diferentemente do plenário e se contrapondo aos princípios de acesso à informação governamental, não há obrigatoriedade na produção (e publicação) de notas taquigráficas dos eventos realizados por comissões, fazendo com que a quantidade de documentos dessa natureza esteja muito aquém do esperado. Todavia, esses números se elevam consideravelmente quando há amplo interesse público por assuntos discutidos nesses ambientes, como ocorreu durante as investigações da Operação Lava Jato ou a tramitação da Reforma da Previdência. Esse fato destaca o impacto da sociedade civil na atuação política dos parlamentares, ressaltando a relevância dos estudos (e ferramentas) que simplificam e contribuem com o monitoramento do Poder Legislativo.

Destacamos também que, embora sejam uma valiosa fonte de informação sobre as ativi-

dades legislativas, nem todos os discursos incluídos em nossa base de dados abertos serão necessariamente úteis para pesquisas futuras nas áreas de Ciência Política ou Política Computacional. Em nossas análises, por exemplo, apenas 28 dos 50 tópicos latentes identificados — correspondendo a 35,70% dos documentos — representavam temas ativamente defendidos (ou combatidos) por deputados e deputadas federais. Enquanto isso, mesmo constituindo maioria em nosso *corpus*, os documentos remanescentes abordavam assuntos efêmeros, de escopo muito específico e/ou intrínsecos ao Poder Legislativo brasileiro, tornando-os pouco relevantes para a caracterização das atuações políticas. Essa dinâmica não somente evidencia o considerável esforço dedicado por parlamentares aos procedimentos mandatórios da Câmara dos Deputados, como também sugere uma crescente demanda por técnicas de classificação automática de textos (como a modelagem de tópicos) em pesquisas nesse domínio. Não à toa, a filtragem dos discursos usando os tópicos latentes se mostrou fundamental para garantir resultados satisfatórios à nossa estimativa de pontos ideais.

As distribuições desses tópicos latentes permitiram-nos compreender e comparar as pautas prioritárias de parlamentares e partidos políticos. Nossos resultados, de maneira geral, alinham-se a estudos que analisaram legislaturas anteriores da Câmara dos Deputados. Por exemplo, também observamos que os temas sociais e econômicos são úteis para distinguir indivíduos que estão, respectivamente, mais à esquerda ou mais à direita no espectro político. Entretanto, identificamos o surgimento de uma notável exceção a essas tendências: os atuais expoentes da extrema-direita brasileira. Apesar de sua oposição às causas sociais e progressistas, as ênfases temáticas desses indivíduos assemelham-se bastante às de parlamentares de esquerda ou extrema-esquerda, apresentando uma dedicação considerável aos tópicos latentes da categoria "Equidade e Inclusão". Além disso, em uma análise mais aprofundada, observamos que essa semelhança também se manifesta no vocabulário desses parlamentares. Esse fenômeno, seja temporário ou não, pode representar novas limitações para o uso de diversas técnicas de PLN no contexto político do nosso país. É possível, por exemplo, que as abordagens baseadas exclusivamente na frequência das palavras (denominadas *bag-of-words*) produzam resultados ambíguos ou que as ênfases temáticas sejam menos capazes de caracterizar as atuações políticas nesse período. Assim, as técnicas baseadas em *embeddings* tendem a se difundir nesse domínio, inclusive abrangendo outros tipos de documentos, como *tweets* ou proposições de lei.

Os pontos ideais estimados também se mostraram adequados à representação de posicionamentos políticos no domínio do Poder Legislativo brasileiro. Nossas estimativas, contudo, destoam da interpretação apresentada na publicação original do modelo e, tampouco, se traduzem nas convencionais dicotomias esquerda-direita ou liberal-conservador. Para o contexto brasileiro, os valores estimados contrastaram o quão "ideológicos" ou "pragmáticos" são os deputados e deputadas federais, contrapondo parlamentares de esquerda e de extrema-direita àqueles que compõem o Centrão e os setores mais tradicionais da direita. A distribuição dos pontos ideais e suas medidas de centralidade apontam que outras características ideológicas podem estar representadas secundariamente na dimensão única desses pontos ideais, mas não foi possível nos aprofundarmos nessa hipótese no decorrer deste trabalho. Nesse sentido, nossos resultados indicam que pontos ideais bidimensionais são mais adequados para cenários políticos amplamente diversos (como o do Brasil) e que, assim como evoluções na explicabilidade do modelo, beneficiariam o TBIP em contextos diferentes do estadunidense. Ademais, o uso conjunto da modelagem de tópicos latentes e da estimativa de pontos ideais pode contribuir profundamente com a análise de preferências político-ideológicas em nosso país, inclusive estratificando-as para temas específicos como educação, segurança pública ou tributação.

Numa perspectiva geral, as descobertas deste estudo permitem-nos avaliar a modelagem de tópicos latentes com o BERTopic e a estimativa de pontos ideais com o TBIP como métodos viáveis e bastante promissores para a caracterização dos discursos e posicionamentos políticos de parlamentares brasileiros. Ao mesmo tempo, as características intrínsecas ao cenário político do Brasil produzem nuances que não são observadas no Norte Global e que aumentam a complexidade da análise e da aplicação dos resultados desses modelos, evidenciando a importância de especialistas da Ciência Política nesse processo. Por fim, em comparação a estudos anteriores, nossos resultados também denotam a ocorrência de profundas transformações no Poder Legislativo nacional no decorrer dos últimos anos. O surgimento e a notoriedade de novos expoentes da extrema-direita na Câmara dos Deputados atenuou a distinção de pautas prioritárias entre os extremos do espectro político. Em particular, a oposição às causas sociais e progressistas ganhou notável espaço nos discursos da "nova direita brasileira", contribuiu com seu alcance e engajamento em mídias sociais e, talvez até, se tornou o cerne de sua atuação política.

6.2 Limitações e trabalhos futuros

Nesta pesquisa, os recursos computacionais disponíveis impuseram restrições às etapas de treinamento da modelagem de tópicos latentes e da estimativa de pontos ideais, tornando necessárias certas adaptações na metodologia adotada. Um avanço natural deste trabalho consiste, portanto, na expansão do conjunto de hiperparâmetros selecionados para os experimentos com o BERTopic e com o TBIP. Novas combinações nos valores desses hiperparâmetros podem ser capazes de aprimorar o aprendizado dos modelos e, por consequência, os resultados obtidos. De maneira análoga, durante a estimativa dos pontos ideais, restringimo-nos ao uso de amostras com até 200 discursos por parlamentar, o que viabilizou a execução dos experimentos. Essa quantidade de documentos parece suficiente à caracterização das preferências político-ideológicas de deputados e deputadas federais, mas, ainda assim, pode ser incrementada com o objetivo de produzir representações mais precisas acerca da atuação política desses indivíduos. Para a rotulagem dos tópicos latentes, consideramos que a inclusão de mais indivíduos no processo ou a automação dessa tarefa através de técnicas baseadas em LLMs podem contribuir com a interpretabilidade e a confiabilidade do rótulos produzidos.

Certas características inerentes ao TBIP também trouxeram obstáculos a este estudo, sobretudo quanto à análise dos pontos ideais estimados. Nesse âmbito, em adição à baixa explicabilidade do modelo, destacamos a unidimensionalidade de suas estimativas. Nossos resultados sugerem que, no contexto brasileiro, o uso de pontos ideais bidimensionais (ou até tridimensionais) seria mais adequado à representação dos posicionamentos políticos de parlamentares. Além disso, devido à abordagem "caixa preta", a análise e a compreensão dos pontos ideais estimados requerem um esforço notável, restringindo sua utilidade junto à sociedade civil. Nesse cenário, estudos correlatos podem contribuir através da implementação de ajustes no TBIP que mitiguem os desafios supracitados, mas também com o desenvolvimento e/ou aplicação de novos modelos para estimativa de pontos ideais baseados em textos.

Sob essa perspectiva, também destacamos o uso de nossa metodologia em outros contextos políticos (como o Senado Federal brasileiro), com outros tipos de documentos (como discursos em debates eleitorais) ou na estimativa de pontos ideais específicos para cada tópico latente identificado. Ademais, as ênfases temáticas podem ser úteis para qualificar e contrastar os diferentes ambientes que compõem o Poder Legislativo brasileiro, como os

plenários, as comissões e as mesas diretoras. Não obstante, sugerimos o desenvolvimento de ferramentas que simplifiquem o acesso e o uso dos dados extraídos durante este estudo, visando fomentar novas pesquisas nas áreas de Ciência Política e de Política Computacional. Por fim, prospectamos a criação de um produto de dados que, através da modelagem de tópicos latentes e da estimativa pontos ideais, facilite o acompanhamento e a compreensão da sociedade civil brasileira acerca de seus atuais representantes no Poder Legislativo.

Referências Bibliográficas

[1] ABDELRAZEK, Aly *et al*. Topic modeling algorithms and applications: A survey. **Information Systems**, v. 112, 2023.

[2] ALMEIDA, Felipe; XEXÉO, Geraldo. Word Embeddings: a survey. **arXiv**, 2019.

[3] ANANDARAJAN, Murugan; HILL, Chelsey; NOLAN, Thomas. Text Preprocessing. In **Practical Text Analytics**. Springer International Publishing, 2018.

[4] ANGELOV, Dimo. Top2Vec: Distributed Representations of Topics. **arXiv**, 2020.

[5] ARAÚJO, Francisco A. S. **Ensaios sobre eleições, financiamento de campanha, ideologia e grupos de interesse**. Tese (Doutorado em Economia Aplicada), Universidade Federal do Ceará, Ceará, 2022.

[6] BATISTA, Mariana. QUAIS POLÍTICAS IMPORTAM? Usando ênfases na agenda legislativa para mensurar saliência. **Revista Brasileira de Ciências Sociais**, v. 35, 2020.

[7] BENGIO, Yoshua; DUCHARME, Réjean; VICENT, Pascal. A Neural Probabilistic Language Model. In **Advances in Neural Information Processing Systems**, volume v. 13. MIT Press, 2000.

[8] BEZERRA, Gabriella M. L.; VIEIRA, Márcia P. C. Interpretações e poderes em disputa: o ressurgimento do Centrão na política brasileira. **Caderno Eletrônico de Ciências Sociais**, v. 10, 2022.

[9] BLEI, David M.; LAFFERTY, John D. Topic Models. In **Text Mining**. Chapman and Hall/CRC, 2009.

[10] BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent Dirichlet Allocation. **Journal of Machine Learning Research**, v. 3, 2003.

[11] BOLOGNESI, Bruno; RIBEIRO, Ednaldo; CODATO, Adriano. Uma Nova Classificação Ideológica dos Partidos Políticos Brasileiros. **Dados**, v. 66, 2023.

[12] BONAVIDES, Paulo. **Ciência política**. Forense, 1976.

[13] BOUMA, Gerlof. Normalized (Pointwise) Mutual Information in Collocation Extraction. In Proceedings of the 2009 Biennial German Society for Computational Linguistics & Language Technology Conference. **GSCL 2009**, Germany, 2009.

[14] BRASIL. [Constituição (1988)]. **Constituição da República Federativa do Brasil**. Presidência da República, Brasília, 2016.

[15] CATELLI, Rosario; PELOSI, Serena; ESPOSITO, Massimo. Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian. **Electronics**, v. 11, n. 3, 2022.

[16] CHOWDHURY, Gobinda G. Natural Language Processing. **Annual Review of Information Science and Technology**, v. 37, 2005.

[17] CLINTON, Joshua D. *et al*. Separated Powers in the United States: The Ideology of Agencies, Presidents, and Congress. **American Journal of Political Science**, v. 56, 2011.

[18] CLINTON, Joshua; JACKMAN, Simon; RIVERS, Douglas. The Statistical Analysis of Roll Call Data. **American Political Science Review**, v. 98, n. 2, 2004.

[19] DALEN, Reinder G.; MELEIN, Léon R.; PLANK, Barbara. Profiling Dutch Authors on Twitter: Discovering Political Preference and Income Level. **Computational Linguistics in the Netherlands Journal**, v. 7, 2017.

[20] DEVLIN, Jacob *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **arXiv**, 2019.

[21] DIENG, Adji B.; RUIZ, Francisco J. R.; BLEI, David M. Blei. The Dynamic Embedded Topic Model. **arXiv**, 2019.

[22] DIENG, Adji B.; RUIZ, Francisco J. R.; BLEI, David M. Blei. Topic Modeling in Embedding Spaces. **Transactions of the Association for Computational Linguistics**, v. 8, 2020.

[23] ESTADÃO DADOS. Basômetro: Quanto apoio o governo tem na Câmara?. **Estadão Dados**, São Paulo, 2022.

[24] FEIJO, Diego V.; MOREIRA, Viviane P. Mono vs Multilingual Transformer-based Models: a Comparison across Several Language Tasks. **arXiv**, 2020.

[25] FELDKIRCHER, Martin; HOFMARCHER, Paul; SIKLOS, Pierre L. One Money, One Voice? Evaluating Ideological Positions of Euro Area Central Banks. **SSRN eLibrary**, 2023.

[26] GENTZKOW, Matthew; SHAPIRO, Jesse M.; TADDY, Matt. Measuring group differences in high-dimensional choices: Method and application to congressional speech. **Econometrica**, v. 87, n. 4, 2019.

[27] GOPALAN, Prem; HOFMAN, Jake M.; BLEI, David M. Scalable Recommendation with Poisson Factorization. **arXiv**, 2013.

[28] GRAVES, Alex; MOHAMED, Abdel R.; HINTON, Geoffrey. Speech Recognition with Deep Recurrent Neural Networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. **ICASSP 2013**, Vancouver, 2013. Institute of Electrical and Electronics Engineers.

[29] GREENE, Derek; CROSS, James P. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. **Political Analysis**, v. 25, 2017.

[30] Justin GRIMMER. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. **Political Analysis**, v. 18, 2010.

[31] GROOTENDORST, Maarten. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. **arXiv**, 2022.

[32] GU, Yupeng *et al*. Topic-factorized Ideal Point Estimation Model for Legislative Voting Network. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **KDD 2014**, New York, 2014. ACM.

[33] HAQ, Ehsan *et al*. A Survey on Computational Politics. **IEEE Access**, v. 8, 2020.

[34] HOFMARCHER, Paul; ADHIKARI, Sourav; GRÜN, Bettina. Gaining Insights on U.S. Senate Speeches Using a Time Varying Text Based Ideal Point Model. **arXiv**, 2022.

[35] IMAI, Kousuke; LO, James; OLMSTED, Jonathan. Fast Estimation of Ideal Points with Massive Data. **American Political Science Review**, v. 110, 2016.

[36] JOULIN, Armand *et al*. Bag of Tricks for Efficient Text Classification. **arXiv**, 2016.

[37] JURAFSKY, Dan; MARGIN, James H. **Speech and Language Processing: an introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. Pearson Prentice Hall, 2009.

[38] KAUR, Navdeep; SINGH, Parminder. Conventional and contemporary approaches used in text to speech synthesis: a review. **Artificial Intelligence Review**, v. 56, n. 7, 2022.

[39] KIM, In S.; LONDREGAN, John; RATKOVIC, Marc. Estimating Spatial Preferences from Votes and Text. **Polytical Analysis**, v. 26, n. 2, 2018.

[40] LAUDERDALE, Benjamin E.; HERZOG, Alexander. Measuring Political Positions from Legislative Speech. **Political Analysis**, v. 24, 2016.

[41] LI, Zhenzhong; SHANG, Wenqian; YAN, Menghan. News text classification model based on topic model. In *15th International Conference on Computer and Information Science. **ICIS 2016***, Japão, 2016. Institute of Electrical and Electronics Engineers.

[42] LIDDY, Elizabeth D. Natural Language Processing for Information Retrieval. In **Encyclopedia of Library and Information Science**. Chapman and Hall/CRC Press, 2017.

[43] MCINNES, Leland; HEALY, John. Accelerated Hierarchical Density Based Clustering. In Proceedings of 2017 IEEE International Conference on Data Mining Workshops. **ICDMW 2017**, New Orleans, 2017.

[44] MCINNES, Leland; HEALY, John; MELVILLE, James. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **arXiv**, 2020.

[45] MIKOLOV, Tomas *et al*. Efficient Estimation of Word Representations in Vector Space. **arXiv**, 2013.

[46] MIMNO, David *et al*. Optimizing Semantic Coherence in Topic Models. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. **EMNLP 2011**, Scotland, 2011.

[47] MONTESQUIEU, Charles-Louis S. **The Spirit of Laws**. Hardpress Publishing, 2020.

[48] MOODY, Christopher E. Mixing Dirichlet Topic Models and Word Embeddings to make LDA2Vec. **arXiv**, 2016.

[49] MOREIRA, Davi. Com a Palavra os Nobres Deputados: Ênfase Temática dos Discursos dos Parlamentares Brasileiros. **Dados**, v. 63, 2020.

[50] NEVO, Baruch. Face Validity Revisited. **Journal of Educational Measurement**, v. 22, n. 4, 1985.

[51] NIU, Liqiang *et al*. Topic2Vec: Learning distributed representations of topics. In International Conference on Asian Language Processing. **IALP 2015**, China, 2015. IEEE.

[52] OLIVEIRA JUNIOR, Robson P.; TELECHI, Acácio V.; FERRÃO, Pedro R. A. Quem é a esquerda brasileira? Uma proposta de classificação empírica. **Entropia**, v. 6, 2022.

[53] OLIVEIRA, Lucas S. *et al*. When Politicians Talk About Politics: Identifying Political Tweets of Brazilian Congressmen. **Proceedings of the International AAAI Conference on Web and Social Media**, v. 12, 2018.

[54] POERNER, Nina; WALTINGER, Ulli; SCHÜTZE, Hinrich. E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. **arXiv**, 2019.

[55] POOLE, Keith T.; ROSENTHAL, Howard. **Congress: A political-economic history of Roll Call Voting**. Oxford University Press, EUA, 2000.

[56] PROKSCH, Sven-Oliver; SLAPIN, Jonathan B. Position Taking in European Parliament Speeches. **British Journal of Political Science**, v. 40, 2009.

[57] QIANG, Jipeng *et al*. Topic Modeling over Short Texts by Incorporating Word Embeddings. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. **PAKDD 2017**, South Korea, 2017. Springer International Publishing.

[58] RADFORD, Alec *et al*. Improving language understanding by generative pre-training. **OpenAI**, 2018.

[59] REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. **arXiv**, 2019.

[60] REIMERS, Nils; GUREVYCH, Iryna. Making monolingual sentence embeddings multilingual using knowledge distillation. **arXiv**, 2020.

[61] ROBERTSON, Stephen. Understanding inverse document frequency: on theoretical arguments for IDF. **Journal of Documentation**, v. 60, n. 5, 2004.

[62] RODRIGUES, Leôncio M. Partidos, ideologia e composição social. **Revista Brasileira de Ciências Sociais**, v. 17, 2002.

[63] SILVA, Jeferson M. Mapeando o Supremo: as posições dos ministros do STF na jurisdição constitucional (2012 – 2017). **Novos Estudos – CEBRAP**, v. 37, 2018.

[64] SILVA, Juliana L. **Violência política contra mulheres: Caso Joice Hasselmann e o bolsonarismo através da misoginia nas redes**. Dissertação (Mestrado em Ciência Política) – Instituto de Ciências Humanas e Filosofia, Universidade Federal Fluminense, Niterói, 2021.

[65] SILVA, Kiane F. **MBL, crise política e conflitos de classe no Brasil**. CRV, Curitiba, 2020.

[66] SINGER, André. A reativação da direita no Brasil. **Opinião Pública**, v. 27, n. 3, 2021.

[67] SLAPIN, Jonathan B.; PROKSCH, Sven-Oliver. A Scaling Model for Estimating Time-Series Party Positions from Texts. **American Journal of Political Science**, v. 52, n. 3, 2008.

[68] SOUZA, Daniela B. M. **Estimação Bayesiana de pontos ideais via dados do Twitter**. Dissertação (Mestrado em Estatística) – Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2017.

[69] SOUZA, Rafael M. S.; GRAÇA, Luís F. G.; SILVA, Ralph S. S. Politics on the web: using twitter to estimate the ideological positions of Brazilian representatives. **Brazilian Political Science Review**, v. 11, 2017.

[70] TAROUCO, Gabriela S.; MADEIRA, Rafael M. Os partidos brasileiros segundo seus estudiosos: análise de um expert survey. **Civitas – Revista de Ciências Sociais**, v. 15, 2015.

[71] TERRAGNI, Silvia *et al*. OCTIS: Comparing and Optimizing Topic Models is Simple!. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. **EACL 2021**, Online, 2021. ACL.

[72] TUFEKCI, Zeynep. Engineering the Public: Big Data, Surveillance and Computational Politics. **First Monday**, 2014.

[73] VAFA, Keyon; NAIDU, Suresh; BLEI, David M. Text-Based Ideal Points. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **ACL 2020**, Online, 2020.

[74] VASWANI, Ashish *et al*. Attention is All you Need. In **Advances in Neural Information Processing Systems**, v. 30. Curran Associates Inc., 2017.

[75] WEISSENBERG, Pedro P. S. **Os evangélicos na Câmara dos Deputados: uma análise quantitativa dos seus discursos (2007 – 2021)**. Dissertação (Mestrado em Ciência Política) – Instituto de Estudos Sociais e Políticos, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2022.

[76] WINDETT, Jason H.; HARDEN, Jeffrey J.; HALL, Matthew E. K. Estimating Dynamic Ideal Points for State Supreme Courts. **Political Analysis**, v. 23, 2015.

[77] YAU, Chyi-Kwei *et al*. Clustering scientific documents with topic modeling. **Scientometrics**, v. 100, n. 3, 2014.

[78] YU, Bei; KAUFMANN, Stefan; DIERMEIER, Daniel. Classifying Party Affiliation from Political Speech. **Journal of Information Technology & Politics**, v. 5, 2008.

[79] YU, QI. Towards a More In-Depth Detection of Political Framing. In Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. **LaTeCHCLfL 2023**, Croatia, 2023.

Apêndice A

Infraestrutura

No decorrer deste estudo, nos deparamos consistentemente com limitações associadas aos recursos computacionais disponíveis, tornando necessário o uso de diferentes ambientes de desenvolvimento. A extração, o pré-processamento e a análise dos dados foram executadas em um computador pessoal, cujas especificações estão detalhadas na Tabela A.1. Já para a modelagem de tópicos latentes, recorremos a uma máquina virtual na *Google Cloud Platform*, descrita na Tabela A.2. Por fim, para a etapa de estimativa de pontos ideais, utilizamos uma máquina cedida pelo Laboratório de Sistemas Distribuídos da Universidade Federal de Campina Grande. As configurações de hardware dessa máquina estão apresentadas na Tabela A.3.

Tabela A.1: Especificação da infraestrutura adotada para etapa de extração e pré-processamento dos dados.

	Especificação
Processador (CPU)	Intel Core i7-12700H
Processador Gráfico (GPU)	-
Memória RAM	16GB DDR4 3200MHz
Armazenamento	128GB SSD
Sistema Operacional	Fedora 37

Tabela A.2: Especificação da infraestrutura adotada para etapa de modelagem de tópicos latentes.

	Especificação
Processador (CPU)	Intel Xeon 2.20GHz
Processador Gráfico (GPU)	-
Memória RAM	64GB DDR4 3600MHz
Armazenamento	50GB HD
Sistema Operacional	Ubuntu 20.04 LTS

Tabela A.3: Especificação da infraestrutura adotada para etapa de estimativa de pontos ideais.

	Especificação
Processador (CPU)	Intel Core i7-8700
Processador Gráfico (GPU)	NVDIA GeForce RTX 2080 Ti
Memória RAM	32GB DDR4 3200MHz
Armazenamento	256GB SSD
Sistema Operacional	Ubuntu 20.04 LTS

Apêndice B

Partidos Políticos

Em outubro de 2023, existiam 30 partidos políticos oficialmente registrados no TSE. Portanto, não é razoável assumir que seus nomes, siglas, parlamentares, ideologias ou atuações sejam amplamente conhecidos pela sociedade civil brasileira. Esse contexto se torna ainda mais complexo ao considerarmos que, no decorrer do período analisado (2015 a 2022), houveram diversas mudanças de nome, fusões e/ou incorporações entre esses partidos. Sabendo disso, na Tabela B.1 apresentamos os nomes, siglas e classificações ideológicas dos partidos políticos brasileiros vigentes à época de nossas análises. A classificação ideológica adotada se origina da publicação "Uma Nova Classificação Ideológica dos Partidos Políticos Brasileiros" de Bolognesi, Ribeiro e Codato [11].

Tabela B.1: Partidos políticos brasileiros registrados no Tribunal Superior Eleitoral em outubro de 2023.

Partido Político	**Classificação Ideológica**
Unidade Popular (UP)	-
Partido Socialista dos Trabalhadores Unificado (PSTU)	Extrema-esquerda
Partido da Causa Operária (PCO)	Extrema-esquerda
Partido Comunista Brasileiro (PCB)	Extrema-esquerda
Partido Socialismo e Liberdade (PSOL)	Extrema-esquerda
Partido Comunista do Brasil (PCdoB)	Esquerda
Partido dos Trabalhadores (PT)	Esquerda

Tabela B.1 (Continuação)

Partido Político	Classificação Ideológica
Partido Democrático Trabalhista (PDT)	Centro-esquerda
Partido Socialista Brasileiro (PSB)	Centro-esquerda
Rede Sustentabilidade (REDE)	Centro
Cidadania (CIDADANIA)	Centro
Partido Verde (PV)	Centro
Partido Trabalhista Brasileiro (PTB)	Centro-direita
Avante (AVANTE)	Centro-direita
Solidariedade (SD)	Centro-direita
Partido da Mobilização Nacional (PMN)	Centro-direita
Partido da Mulher Brasileira (PMB)	Centro-direita
Movimento Democrático Brasileiro (MDB)	Direita
Partido Social Democrático (PSD)	Direita
Partido da Social Democracia Brasileira (PSDB)	Direita
Podemos (PODE)	Direita
Partido Renovador Trabalhista Brasileiro (PRTB)	Direita
Partido Liberal (PL)	Direita
Republicanos (REPUBLICANOS)	Direita
Agir (AGIR)	Direita
Democracia Cristã (DC)	Direita
Partido Novo (NOVO)	Direita
Progressistas (PP)	Direita
União Brasil (UNIÃO)	Direita
Patriota (PATRIOTA)	Extrema-direita

É importante ressaltar que os dados em que essa classificação ideológica se baseia foram coletados em julho de 2018 e, por isso, podem ter ocorrido alterações na percepção dos especialistas em Ciência Política quanto às ideologias partidárias. Ademais, ressaltamos que nem todos os partidos políticos apresentados na Tabela B.1 estão presentes na publicação original. O Unidade Popular (UP), por exemplo, só foi registrado oficialmente em 2019 e não

receberá classificação no decorrer deste estudo. Outros partidos mudaram de nome durante esse período e, nesses casos, mantivemos a classificação ideológica original. Um exemplo é o Republicanos (REPUBLICANOS) que era denominado Partido Republicano Brasileiro (PRB) até 2019. Finalmente, em casos de fusões ou incorporações, calculamos a média das posições ideológicas dos partidos envolvidos e, usando a escala proposta na publicação original, estabelecemos uma classificação ideológica para o partido político criado. Por exemplo, o Partido Social Liberal (PSL) e o Democratas (DEM) possuíam posições ideológicas de 8,11 e 8,57 (respectivamente). Em 2022, esses partidos se fundiram para formar o União Brasil (UNIÃO), ao qual atribuímos a posição ideológica de 8,34 e associamos à categoria "Direita".

Destacamos ainda que as informações sobre partidos políticos em nosso *corpus* se referem exclusivamente às filiações partidárias registradas nas candidaturas dos deputados e deputadas federais, isto é, não há acompanhamento sobre eventuais migrações partidárias desses indivíduos. Ademais, visando garantir simplicidade e melhor adequação à temporalidade dessas informações, adotaremos o seguinte padrão no decorrer deste documento: ao mencionar a filiação partidária de um parlamentar específico, será apresentada a informação mais recente que figure em nossos conjuntos de dados, preservando o nome e a sigla de partidos políticos mesmo que tenham mudado de nome, sido incorporados ou se fundido a outros partidos.

Apêndice C

Tópicos Latentes

Durante a etapa de modelagem de tópicos latentes deste estudo (descrita no Capítulo 4), produzimos um modelo BERTopic que associou os documentos de nosso *corpus* a 50 tópicos distintos. Visando simplificar a compreensão e contribuir com a interpretabilidade desses resultados, nos baseamos nos dez termos mais associados a cada tópico para produzir novos rótulos e, assim, substituir os identificadores numéricos que são gerados automaticamente pelo modelo. Na Tabela C.1 estão apresentados os identificadores numéricos, os rótulos e os termos mais associados a cada tópico latente identificado por nosso modelo.

Tabela C.1: Termos mais relevantes e rótulos atribuídos aos tópicos latentes identificados pelo modelo BERTopic.

Tópico	**Rótulo**	**Termos mais associados**
0	Pausas para Deliberação	“pausa”, “permaneçam”, “item”, “aprovam”, “palavra”, “encontram”, “lula”, “parecer”, “minutos” e “votação”.
1	Saúde	“saúde”, “médicos”, “vacina”, “câncer”, “sus”, “pacientes”, “hospital”, “hospitais”, “vacinas” e “covid”.
2	Segurança	“polícia”, “policiais”, “segurança”, “penal”, “policial”, “militar”, “crime”, “militares”, “crimes” e “pena”.

Tabela C.1 (Continuação)

Tópico	Rótulo	Termos mais associados
3	Educação	“educação”, “ensino”, “professores”, “escola”, “escolas”, “alunos”, “universidades”, “estudantes”, “universidade” e “fundeb”.
4	Lideranças Parlamentares	“exa”, “líder”, “liderança”, “falar”, “deputado”, “vou”, “queria”, “gostaria”, “agradecer” e “ministro”.
5	Impeachment de Dilma	“democracia”, “golpe”, “impeachment”, “dilma”, “eleitoral”, “eleições”, “política”, “corrupção”, “república” e “temer”.
6	Violência de Gênero	“mulheres”, “mulher”, “violência”, “feminina”, “homens”, “gênero”, “feminicídio”, “doméstica”, “penha” e “direitos”.
7	Emprego e Renda	“banco”, “economia”, “econômica”, “desemprego”, “renda”, “juros”, “emprego”, “bancos”, “inflação” e “reais”.
8	Processos de Votação	“votação”, “votar”, “acordo”, “voto”, “texto”, “matéria”, “obstrução”, “partidos”, “destaque” e “pauta”.
9	Governo Bolsonaro	“bolsonaro”, “emancipação”, “cidade”, “município”, “jair”, “aniversário”, “festa”, “prefeito”, “queiroz” e “anos”.
10	Palestras e Exposições	“audiência”, “palavra”, “perguntas”, “convidados”, “minutos”, “palestrantes”, “mesa”, “expositores”, “microfone” e “cultura”.
11	Orientações Partidárias	“psol”, “orienta”, “psb”, “destaque”, “psl”, “vota”, “texto”, “psd”, “obstrução” e “pausa”.

Tabela C.1 (Continuação)

Tópico	Rótulo	Termos mais associados
12	Energia	“energia”, “petrobras”, “petróleo”, “preço”, “eletrobras”, “gás”, “elétrica”, “empresa”, “combustíveis” e “privatização”.
13	Encerramentos de Sessão	“reunião”, “horas”, “sessão”, “feira”, “ordem”, “encerrada”, “min”, “terça”, “havendo” e “amanhã”.
14	Agricultura	“agricultura”, “familiar”, “agricultores”, “rural”, “produtores”, “produção”, “alimentos”, “agrotóxicos”, “agricultor” e “ambiental”.
15	Indígenas e Quilombolas	“indígenas”, “povos”, “indígena”, “terras”, “funai”, “índios”, “terra”, “comunidades”, “quilombolas” e “demarcação”.
16	Homenagens	“falecimento”, “pesar”, “homenagem”, “faleceu”, “homem”, “família”, “amigo”, “amigos”, “pai” e “familiares”.
17	Legislação	“art”, “constitucionalidade”, “juridicidade”, “legislativa”, “técnica”, “lei”, “emendas”, “inciso”, “ii” e “substitutivo”.
18	Investigações	“cpi”, “senhor”, “sa”, “bndes”, “investigação”, “petrobras”, “relatório”, “acho”, “informações” e “carf”.
19	Esportes	“esporte”, “futebol”, “atletas”, “clubes”, “atleta”, “jogos”, “clube”, “olímpico”, “olimpíadas” e “olímpicos”.
20	Meio Ambiente	“amazônia”, “amazonas”, “manaus”, “desmatamento”, “franca”, “zona”, “floresta”, “região”, “amazônica” e “ambiente”.

Tabela C.1 (Continuação)

Tópico	Rótulo	Termos mais associados
21	Gestão Municipal	“prefeito”, “cidade”, “município”, “vereadores”, “vereador”, “prefeitos”, “municípios”, “região”, “governador” e “municipal”.
22	Recursos Hídricos	“água”, “saneamento”, “francisco”, “rio”, “águas”, “transposição”, “chuvas”, “seca”, “região” e “hídrica”.
23	Proteção aos Animais	“animais”, “animal”, “vaquejada”, “tratos”, “maus”, “rodeios”, “rodeio”, “cães”, “izar” e “royal”.
24	Transporte Aéreo	“aeroporto”, “aviação”, “aéreas”, “anac”, “voos”, “aeroportos”, “bagagem”, “passagens”, “aérea” e “aéreo”.
25	Rodovias	“br”, “trecho”, “rodovia”, “dnit”, “barragem”, “obra”, “duplicação”, “vale”, “obras” e “ponte”.
26	Finanças e Tributação	“tributária”, “fiscal”, “bilhões”, “imposto”, “dívida”, “reais”, “estados”, “receita”, “teto” e “pagar”.
27	Direitos Trabalhistas	“trabalhadores”, “trabalhador”, “trabalhista”, “trabalho”, “terceirização”, “clt”, “direitos”, “trabalhistas”, “trabalhadoras” e “sindical”.
28	Religião	“igreja”, “deus”, “pastor”, “jesus”, “cristo”, “fé”, “católica”, “bispo”, “papa” e “dom”.
29	Operação Lava Jato	“jato”, “lava”, “moro”, “operação”, “maranhão”, “dino”, “juiz”, “corrupção”, “flávio” e “sergio”.

Tabela C.1 (Continuação)

Tópico	Rótulo	Termos mais associados
30	Turismo	“turismo”, “turistas”, “cultura”, “embratur”, “turístico”, “jogos”, “setor”, “eventos”, “cassinos” e “turística”.
31	Previdência Social	“previdência”, “aposentadoria”, “salário”, “reforma”, “idade”, “aposentados”, “mínimo”, “aposentar”, “anos” e “contribuição”.
32	Igualdade Racial	“negros”, “racismo”, “negra”, “negro”, “racial”, “negras”, “jovens”, “palmares”, “juventude” e “igualdade”.
33	Transporte e Logística	“transporte”, “caminhoneiros”, “trânsito”, “ferrovia”, “motoristas”, “veículos”, “transportes”, “uber”, “frete” e “cargas”.
34	Intervenções	“palavra”, “deputada”, “erika”, “concedo”, “minutos”, “kokay”, “pausa”, “passo”, “passar” e “deputado”.
35	Agradecimentos	“mandato”, “agradecer”, “partido”, “honra”, “casa”, “brasil”, “orgulho”, “parabéns”, “deus” e “país”.
36	Saúde Materna	“aborto”, “parto”, “grávidas”, “gravidez”, “maternidade”, “bebê”, “mães”, “mãe”, “gestação” e “vida”.
37	Pesca	“pesca”, “pescadores”, “defeso”, “pescador”, “aquicultura”, “seguro”, “artesanais”, “portaria”, “artesanal” e “pescadoras”.
38	Audiências Públicas	“audiência”, “convidados”, “minutos”, “exposição”, “convido”, “informo”, “representante”, “interpelar”, “página” e “palavra”.

Tabela C.1 (Continuação)

Tópico	Rótulo	Termos mais associados
39	Produção de Laticínios	“leite”, “produtores”, “queijo”, “produtor”, “sul”, “produção”, “produto”, “uruguai”, “importação” e “agricultura”.
40	Conflitos no Oriente Médio	“israel”, “palestino”, “embaixador”, “jerusalém”, “palestinos”, “palestina”, “paz”, “judeus”, “refugiados” e “judeu”.
41	Direitos PcD	“deficiência”, “autista”, “autismo”, “autistas”, “espectro”, “síndrome”, “transtorno”, “down”, “pessoa” e “pessoas”.
42	Direitos LGBTQIA+	“lgbt”, “homofobia”, “ódio”, “gay”, “jean”, “homossexual”, “homossexuais”, “wyllys”, “sexual” e “gays”.
43	Análises de Veto	“veto”, “vetos”, “derrubada”, “derrubar”, “congresso”, “senadores”, “pln”, “senado”, “senador” e “vetou”.
44	Proteção à Infância	“crianças”, “criança”, “adolescentes”, “adolescente”, “infância”, “sexual”, “família”, “estatuto”, “pais” e “violência”.
45	Desinformação e Mídia	“fake”, “news”, “imprensa”, “liberdade”, “jornalista”, “jornalistas”, “jornalismo”, “redes”, “expressão” e “censura”.
46	Encaminhamentos Legislativos	“solidariedade”, “encaminha”, “orienta”, “matéria”, “vota”, “urgência”, “texto”, “voto”, “orientamos” e “retirada”.
47	Cafeicultura	“café”, “santo”, “espírito”, “cafeicultura”, “capixaba”, “cafeicultores”, “produtor”, “conilon”, “capixabas” e “incaper”.

Tabela C.1 (Continuação)

Tópico	Rótulo	Termos mais associados
48	Apresentações	“vídeo”, “exibição”, “senhor”, “senhora”, “áudio”, “vídeos”, “vou”, “assistir”, “jean” e “imagem”.
49	Serviços de Comunicação	“whatsapp”, “internet”, “operadoras”, “sites”, “bloqueio”, “aplicativo”, “anatel”, “telefonia”, “pergunta” e “mensagens”.

Durante a avaliação desses resultados, optamos por agrupar os tópicos latentes em cinco categorias. Dentre elas, consideramos apenas “Garantias Fundamentais”, “Desenvolvimento Econômico” e “Equidade e Inclusão” como categorias relevantes às nossas análises. A divisão dos 50 tópicos latentes identificados entre as cinco categorias propostas está apresentada na Tabela C.2.

Tabela C.2: Categorias atribuídas aos tópicos identificados pelo modelo BERTopic nos discursos de parlamentares.

Categoria	Tópicos
Garantias Fundamentais	Educação, Energia, Esportes, Meio Ambiente, Recursos Hídricos, Rodovias, Saúde, Saúde Materna e Segurança.
Desenvolvimento Econômico	Agricultura, Cafeicultura, Emprego e Renda, Finanças e Tributação, Pesca, Produção de Laticínios, Serviços de Comunicação, Transporte Aéreo, Transporte e Logística e Turismo.
Equidade e Inclusão	Direitos LGBTQIA+, Direitos PcD, Direitos Trabalhistas, Igualdade Racial, Indígenas e Quilombolas, Previdência Social, Proteção à Infância, Proteção aos Animais e Violência de Gênero.

Tabela C.2 (Continuação)

Categoria	Tópicos
Outros	Conflitos no Oriente Médio, Desinformação e Mídia, Gestão Municipal, Governo Bolsonaro, Impeachment de Dilma, Investigações, Operação Lava Jato e Religião.
Processos Legislativos	Agradecimentos, Análises de Veto, Apresentações, Audiências Públicas, Encaminhamentos Legislativos, Encerramentos de Sessão, Homenagens, Intervenções, Legislação, Lideranças Parlamentares, Orientações Partidárias, Palestras e Exposições, Pausas para Deliberação e Processos de Votação.