

CENTRO UNIVERSITÁRIO INSTITUTO DE EDUCAÇÃO SUPERIOR DE BRASÍLIA

Bacharelado em
Ciência de Dados e Inteligência Artificial

Trabalho de Conclusão de Curso

Andre Augusto Sak

Mapeando o campo político: Extração de tópicos em discursos de parlamentares governistas e oposicionistas

Brasília
2023

Mapeando o campo político: Extração de tópicos em discursos de parlamentares governistas e oposicionistas

Trabalho de Conclusão de Curso apresentado como pré-requisito para obtenção do título de bacharel em Ciência de Dados e Inteligência Artificial pelo Instituto de Educação Superior de Brasília - IESB.

Orientador: Professor Dr. Sérgio da Costa Côrtes

Mapeando o campo político: Extração de tópicos em discursos de parlamentares governistas e oposicionistas

Andre Augusto Sak

Trabalho de Conclusão de Curso apresentado como pré-requisito para obtenção do título de bacharel em Ciência de Dados e Inteligência Artificial pelo Instituto de Educação Superior de Brasília - IESB.

Brasília – Distrito Federal, em _____ de _____ de _____

Banca Examinadora constituída pelos Professores:

Prof. Dr. Sérgio da Costa Côrtes - Orientador
Professor no IESB

Prof. Dr. Alexandre Magalhães Martins - Membro
Professor no IESB

Prof. Dr. Fernando Antonio Fernandes Anselmo - Membro
Professor no IESB

Prof. Dr. Moises Fabiano Pereira da Silva Júnior - Membro
Professor no IESB

Agradecimentos

À Ana, ao Murilo e à Carolina, meu Universo, sem o qual nenhum dado faria sentido.

Aos gabinetes do Senador Oriovisto Guimarães, da Deputada Adriana Ventura e da Liderança do Novo na Câmara dos Deputados, que inspiraram muitas das análises.

Ao Inutech e, em especial, ao professor Sérgio Medeiros.

Aos colegas de curso e, em especial ao Felipe e ao Marcos.

Aos mestres, por todo esforço e paciência, em especial ao professor Sérgio Côrtes, sempre confiante em nossa capacidade.

Ao IESB, pela estrutura e pela bolsa integral em todo o curso.

"Com muitos cálculos pode-se vencer;
com poucos, não.
E quão menos possibilidades de vitória
tem aquele que não fez nenhum!"
(Sun Tzu)

Resumo

Este estudo apresenta uma análise objetiva dos discursos de parlamentares governistas e oposicionistas na 56ª Legislatura da Câmara dos Deputados. Utilizando dados das APIs do portal de dados abertos do órgão, foram coletados e analisados discursos, votos e orientações de voto. Para além das classificações partidárias tradicionais, a pesquisa adotou um método de identificação de parlamentares governistas e oposicionistas com base na convergência dos seus votos com as orientações do governo. Posteriormente, para a análise dos discursos, foram aplicadas técnicas de processamento de linguagem natural, incluindo limpeza e pré-processamento dos textos, extração de tópicos via *Latent Dirichlet Allocation* (*LDA*) e comparação temporal de tópicos utilizando *word embeddings* e similaridade por cosseno. Os resultados mostraram diferenças nos temas abordados por governistas e oposicionistas. A análise temporal dos tópicos destacou a persistência de certos assuntos, refletindo a dinâmica parlamentar e as divisões ideológicas. Este trabalho contribuiu para o entendimento da comunicação política e dinâmica legislativa, oferecendo percepções sobre a representação de interesses no discurso parlamentar.

PALAVRAS-CHAVE: CIÊNCIA DE DADOS, INTELIGÊNCIA ARTIFICIAL, ANÁLISE DO DISCURSO, PARLAMENTAR, CÂMARA DOS DEPUTADOS, *LATENT DIRICHLET ALLOCATION*, *LDA*, *WORD EMBEDDINGS*.

Abstract

This investigation provides a thorough examination of the speeches in the Brazilian Chamber of Deputies during its 56th Legislature, focusing on both government and opposition congresspersons. Data, including parliamentary speeches, voting records, and government voting orientations, were gathered from the Chamber's APIs. A novel approach was implemented to categorize lawmakers into government or opposition groups based on their voting alignment with governmental directives, moving beyond simplistic party-based classifications. In orther to employ natural language processing techniques, this study embarked on a rigorous process of text cleaning, pre-processing, and topic extraction using Latent Dirichlet Allocation (LDA). This was complemented by an innovative temporal analysis of topics through the application of word embeddings and cosine similarity measures. The analysis uncovered distinct thematic variations in the content articulated by government-aligned and opposition members. A detailed examination of topic evolution over time revealed enduring themes, capturing the intricate dynamics and ideological divides within the parliamentary sphere. This research not only sheds light on the subtleties of political communication within the legislative context but also offers valuable insights into how parliamentary representatives articulate and champion diverse interests.

KEYWORDS: DATA SCIENCE, ARTIFICIAL INTELLIGENCE, DISCOURSE ANALYSIS, PARLIAMENTARY, CHAMBER OF DEPUTIES OF BRAZIL, LATENT DIRICHLET ALLOCATION, LDA, WORD EMBEDDINGS.

Lista de ilustrações

Figura 1 – Representação gráfica do LDA 31
Figura 2 – Total de palavras por tipo de discurso 38
Figura 3 – Percentuais de “Sim”e “Não”na orientação do governo e nos votos . . . 41
Figura 4 – Totais de votos por semestre 41
Figura 5 – Distribuição da quantidade de votos por parlamentar 42
Figura 6 – Frequência por semestre e tipo de discurso 42
Figura 7 – 15 parlamentares que mais falaram (contagem de palavras) 43
Figura 8 – Distribuição da convergência com as orientações do governo com os votos - semestral 44
Figura 9 – Evolução da média semestral de convergência com o governo 45
Figura 10 – Distribuição da convergência com as orientações do governo para toda a legislatura 45
Figura 11 – Tópicos extraídos para o semestre 2019-1 do grupo de oposicionistas . . 47
Figura 12 – Distribuição dos escores de similaridade por cosseno por tópico - oposição 49
Figura 13 – Distribuição dos escores de similaridade por cosseno por tópico - governo 49
Figura 14 – Semelhança temporal dos tópicos extraídos do semestre 2021-1 - oposição 50
Figura 15 – Maiores escores de semelhança com o tópico 46 ao longo do tempo - governo 52
Figura 16 – Maiores escores de semelhança com o tópico 22 ao longo do tempo - governo 54
Figura 17 – Maiores escores de semelhança com o tópico 61 ao longo do tempo - oposição 57
Figura 18 – Maiores escores de semelhança com o tópico 41 ao longo do tempo - oposição 58
Figura 19 – Mapa de calor da diferença da importância da palavra para governistas e oposicionistas 61
Figura 20 – Tópicos extraídos do semestre 2019-1 - governo 90
Figura 21 – Tópicos extraídos do semestre 2019-2 - governo 90
Figura 22 – Tópicos extraídos do semestre 2020-1 - governo 91
Figura 23 – Tópicos extraídos do semestre 2020-2 - governo 91
Figura 24 – Tópicos extraídos do semestre 2021-1 - governo 92
Figura 25 – Tópicos extraídos do semestre 2021-2 - governo 92
Figura 26 – Tópicos extraídos do semestre 2022-1 - governo 93
Figura 27 – Tópicos extraídos do semestre 2022-2 - governo 93
Figura 28 – Semelhança temporal dos tópicos extraídos do semestre 2019-1 - governo 94
Figura 29 – Semelhança temporal dos tópicos extraídos do semestre 2019-2 - governo 95

Figura 30 – Semelhança temporal dos tópicos extraídos do semestre 2020-1 - governo 96
Figura 31 – Semelhança temporal dos tópicos extraídos do semestre 2020-2 - governo 97
Figura 32 – Semelhança temporal dos tópicos extraídos do semestre 2021-1 - governo 98
Figura 33 – Semelhança temporal dos tópicos extraídos do semestre 2021-2 - governo 99
Figura 34 – Semelhança temporal dos tópicos extraídos do semestre 2022-1 - governo 100
Figura 35 – Semelhança temporal dos tópicos extraídos do semestre 2022-2 - governo 101
Figura 36 – Tópicos extraídos do semestre 2019-1 - oposição 102
Figura 37 – Tópicos extraídos do semestre 2019-2 - oposição 102
Figura 38 – Tópicos extraídos do semestre 2020-1 - oposição 103
Figura 39 – Tópicos extraídos do semestre 2020-2 - oposição 103
Figura 40 – Tópicos extraídos do semestre 2021-1 - oposição 104
Figura 41 – Tópicos extraídos do semestre 2021-2 - oposição 104
Figura 42 – Tópicos extraídos do semestre 2022-1 - oposição 105
Figura 43 – Tópicos extraídos do semestre 2022-2 - oposição 105
Figura 44 – Semelhança temporal dos tópicos extraídos do semestre 2019-1 - oposição 106
Figura 45 – Semelhança temporal dos tópicos extraídos do semestre 2019-2 - oposição 107
Figura 46 – Semelhança temporal dos tópicos extraídos do semestre 2020-1 - oposição 108
Figura 47 – Semelhança temporal dos tópicos extraídos do semestre 2020-2 - oposição 109
Figura 48 – Semelhança temporal dos tópicos extraídos do semestre 2021-1 - oposição 110
Figura 49 – Semelhança temporal dos tópicos extraídos do semestre 2021-2 - oposição 111
Figura 50 – Semelhança temporal dos tópicos extraídos do semestre 2022-1 - oposição 112
Figura 51 – Semelhança temporal dos tópicos extraídos do semestre 2022-2 - oposição 113

Lista de tabelas

Tabela 1 – Cronograma do trabalho 27
Tabela 2 – Descrição das variáveis dos dados dos parlamentares 33
Tabela 3 – Descrição das variáveis de orientações de votações 34
Tabela 4 – Descrição das variáveis de votos 34
Tabela 5 – Descrição das variáveis dos discursos 35
Tabela 6 – Possibilidades de fala e dispositivos regimentais 37
Tabela 7 – Tópicos mais semelhantes - governo 51
Tabela 8 – Maiores semelhanças com o tópico 46 - governo 52
Tabela 9 – Maiores semelhanças com o tópico 22 - governo 53
Tabela 10 – Tópicos mais semelhantes - oposição 55
Tabela 11 – Maiores semelhanças com o tópico 61 - oposição 56
Tabela 12 – Maiores semelhanças com o tópico 41 - oposição 58
Tabela 13 – Maiores diferenças das somas dos pesos de palavras em favor de governistas 60
Tabela 14 – Maiores diferenças das somas dos pesos de palavras em favor de oposicionistas 60
Tabela 15 – Frequência e peso das palavras relacionadas à pandemia 63
Tabela 16 – Frequência e peso de palavras associadas à pauta conservadora 63

Lista de abreviaturas e siglas

API	Interface de Programação de Aplicativos (*Application Programming Interface*)
CPI	Comissão Parlamentar de Inquérito (CPI)
GPU	Unidade de Processamento Gráfico (*Graphic Processing Units*)
LDA	Alocação latente de Dirichlet (*Latent Dirichlet Allocation*)
PLN	Processamento de Linguagem Natural

Sumário

1 INTRODUÇÃO 19

2 MOTIVAÇÃO 22

3 OBJETIVOS E ORGANIZAÇÃO DO TRABALHO 25
3.1 Objetivo geral 25
3.2 Objetivos específicos 25
3.3 Organização 25
3.4 Cronograma 27

4 REFERENCIAL TEÓRICO 28

5 DESENVOLVIMENTO 33
5.1 Definição das fontes de dados 33
5.2 Descrição dos dados 33
5.3 Pré-processamento e limpeza dos dados 36
5.3.1 Orientações do governo e votos dos parlamentares 36
5.3.2 Discursos 36
5.4 Análise exploratória 40
5.4.1 Orientações do governo e votos dos parlamentares 40
5.4.2 Discursos 42
5.5 Modelagem 43
5.5.1 Convergência 43
5.5.2 Extração de tópicos com *LDA* 46

6 RESULTADOS 51
6.1 Análise temporal 51
6.2 Tópicos gerais 59

7 CONCLUSÕES 65

REFERÊNCIAS 68

ANEXOS 71

ANEXO A – PRINCIPAIS ALGORITMOS UTILIZADOS 72
A.1 Requisição e tratamento dos votos 72
A.2 Requisição e tratamento dos discursos 74

A.3 Modelagem 80

ANEXO B – TÓPICOS DETALHADOS 90
B.1 Governo 90
B.2 Oposição 102

1 INTRODUÇÃO

A capacidade de analisar dados constitui, antes de tudo, uma vantagem competitiva. Não só atores inseridos no contexto empresarial podem se beneficiar dessa vantagem, mas também a Academia, que compreenderá mais rápido e mais objetivamente uma infinidade de fenômenos sociais. A ciência de dados, sob suas mais diversas denominações e decorrências (inteligência de negócios, mineração de dados, inteligência artificial, aprendizagem de máquina, entre outros) fornece uma miríade de ferramentas e métodos para fornecer, ao que busca, respostas que não teriam sido obtidas há 20 ou 30 anos. Estima-se que existam mais de 137 mil bibliotecas *python*, boa parte delas criadas para atividades diretamente relacionadas a ciência de dados (TEAM, 2023). Some-se a isso, uma capacidade de processamento melhor e mais acessível, especialmente por meio da computação em nuvem. Com poucos dólares, um estudante consegue acessar poderosas *GPUs* (*graphic processing units*) no *Google Collab* para treinar complexas redes neurais.

Além das ferramentas e de uma robustez computacional mais democratizada, também tem-se uma geração frenética de dados, sobre os mais diversos assuntos e sob os mais amplos formatos. “Os usuários são reais, eles estão lá e deixam rastros de dados por onde passam”(DOMO, 2023), relata a conhecida consultoria responsável pelo infográfico *Data Never Sleeps*. As oportunidades são, portanto, inéditas, e os riscos também: nesse oceano de dados, o que é relvante? O que analisar? Quais ferramentas trarão os melhores resultados? Nesse contexto, torna-se inequívoca a importância da formação de profissionais qualificados para responder satisfatoriamente a essas e outras perguntas.

Este trabalho busca sugerir caminhos para responder a importantes questionamentos relativos a uma relevante arena de ocupações e preocupações humanas modernas: o parlamento. Há muito, os cientistas políticos vêm dedicando parte considerável de seus esforços para compreender a dinâmica de funcionamento das assembleias legislativas pelo mundo. Os pesquisadores sempre usufruíram de vastos registros históricos sobre o assunto: é possível facilmente consultar anais e diários de mais de 200 anos (BRASIL, 2023c). Enquanto os documentos mais antigos contêm breves resumos e atas das reuniões, atualmente, é possível obter dados sobre praticamente todos os aspectos relacionados às atividades dos principais colegiados parlamentares. A Câmara dos Deputados, por exemplo, em seu portal de dados abertos (BRASIL, 2023b), disponibiliza aproximadamente 100 *endpoints*[1] por meio dos quais é possível consultar informações atualizadas sobre aquele órgão: de dados sobre votações a tipos de despesas realizadas pelos deputados.

A Ciência de Dados e suas ferramentas e métodos possibilitam uma compreensão

[1] *URL* específica por meio da qual uma *API* pode ser acessada e através da qual ela interage com um sistema, realizando tarefas como ler, modificar ou adicionar dados.

mais rápida, mais objetiva e mais acurada de determinados fenômenos que ocorrem nos parlamentos. Como ser capaz de extrair sentido e agrupar os mais de 68 mil discursos proferidos na Câmara dos Deputados durante a 56ª Legislatura (período compreendido entre 2 de fevereiro de 2019 e 1º de fevereiro de 2023)? Considerando apenas esse período, uma extração manual de tópicos seria inviável em um espaço de tempo razoável. Além do que, o resultado restaria calcado no olhar subjetivo do pesquisador. Assim como ocorre em diversas áreas do conhecimento humano, a ciência política tem muito a lucrar com sua parceria com a ciência dos dados, confirmando ou rechaçando pré-concepções.

Ao explorar e se apropriar dos dados disponibilizados pela Câmara dos Deputados, o presente estudo pretende responder, especialmente, à seguinte pergunta de pesquisa: quais as principais semelhanças e diferenças entre os discursos de parlamentares de apoio ao governo e de oposição, no Plenário da Câmara dos Deputados, durante a 56ª Legislatura? Para que uma resposta pudesse ser construída, foram criadas duas importantes frentes de estudo:

- estabelecimento de um critério objetivo para identificar e diferenciar parlamentares governistas de oposicionistas;
- criação modelos de extração de tópicos que possibilitem a identificação de semelhanças e diferenças entre os assuntos tratados por governistas e oposicionistas.

Quanto ao agrupamento dos parlamentares entre aqueles que apoiavam o governo de então e os que se opunham a ele, optou-se por utilizar o dado que melhor consubstancia essa relação: a voto. Não é incomum que dezenas de parlamentares que pertençam a partidos governistas votem contra projetos de interesse do governo de plantão. Portanto, a métrica mais apurada para a classificação entre governismo e oposicionismo não está relacionada, conforme se verá, à questão partidária, mas sim ao nível de convergência com as orientações dadas pelo governo quanto às matérias que serão apreciadas.

Uma vez identificados os dois grupos, procedeu-se à criação de modelos capazes de extrair, separadamente, os tópicos dos discursos realizados durante o período de análise. Optou-se pelo *Latent Dirichlet Allocation* (*LDA*), uma das ferramentas mais tradicionais para esse fim. O *LDA*, conforme delineado em seu artigo seminal, pretende ser um mecanismo de redução de dimensionalidade, que se propõe a inferir os tópicos subjacentes em determinados documentos (BLEI; NG; JORDAN, 2003, p. 1015). Assume-se que cada documento é uma combinação ou mistura desses tópicos não declarados. Entende-se também que cada tópico possui sua própria distribuição de palavras. Assim, se um documento é principalmente sobre o tópico de economia, provavelmente conterá palavras relacionadas, como “mercado”, “investimento” ou “inflação”.

Fez-se a opção por extrair os tópicos considerando intervalos semestrais, isso porque, conforme estabelece a Constituição Federal de 1988, em seu art. 57, as Casas do Congresso Nacional se reúnem entre 2 de fevereiro e 17 de julho e 1º de agosto e 22 de dezembro (BRASIL, 2023a). Esses períodos compõem etapas bem demarcadas de atividades durante um ciclo de 4 anos: no primeiro semestre, por exemplo, deve ser aprovado o projeto de lei de diretrizes orçamentarias, sem o qual não é permitido haver recesso parlamentar (art. 57, § 2º, da Constituição Federal) (BRASIL, 2023a); no segundo semestre, deve ser aprovado o projeto de lei orçamentária. São, portanto, períodos bem delimitados e que justificam a decisão tomada.

Realizada a extração dos tópicos, eles foram analisados sob dois aspectos:

- a identificação de semelhanças e diferenças com outros tópicos, para o mesmo grupo, ao longo do tempo;
- uma análise mais ampla de semelhanças e diferenças entre governismo e oposição considerando todas as palavras que compunham os tópicos.

Tendo em vista que os tópicos são compostos por uma distribuição de palavras e que não seria esperado que houvessem tópicos idênticos em semestres diversos, foi realizado um cálculo de semelhança entre eles. Assim, tópicos semelhantes foram devidamente identificados, mesmo que, em sua construção, palavras diferentes tenham sido utilizadas. Com essa abordagem, foi possível perceber o quanto determinado tópico era mais ou menos semelhante a outros em períodos distintos.

Ao final, optou-se por considerar apenas as principais palavras de cada tópico identificado para cada grupo e, então, foi possível encontrar as grandes diferenças de escolha de palavras utilizadas por governistas e oposicionistas.

As possibilidades de diferentes tipos de análise de discursos parlamentares foram consideravelmente ampliadas em razão das recentes ferramentas de processamento de linguagem natural. O trabalho objetiva contribuir com o ferramental do analista (seja ele político ou de dados) para criar mecanismos que possibilitem racionalizar a análise e extração de sentido das falas de deputados federais. Os mecanismos criados para o agrupamento de parlamentares, que guiaram a base de criação do modelo, podem ser amplamente aplicados a outros colegiados, de maneira a fomentar uma melhor compreensão de diferenças e semelhanças entre grupos específicos. Assim, espera-se demonstrar como ferramentas disponibilizadas pela ciência de dados podem contribuir para uma melhor e mais rápida compreensão de fenômenos afeitos às ciências sociais, em especial à política.

2 MOTIVAÇÃO

Os debates parlamentares têm sido objeto de crescente interesse da população. Algumas oitivas da chamada Comissão Parlamentar de Inquérito (CPI) da Covid-19 registraram mais de 1,3 milhão de visualizações no canal da TV Senado, no *YouTube* (SAID, 2023). Em outro colegiado, destinado a investigar os atos relativos ao 8 de janeiro, o mesmo canal contabilizou, em um único dia, mais de 400 mil visualizações para uma transmissão de mais de 11 horas em que o depoente se manteve em silêncio (SENADO, 2023).

Com mais de 13 anos de atuação em diversos setores do Congresso Nacional, tive a oportunidade de acompanhar, de perto, debates como os acima descritos. Restou evidente, sob essa ótica, a centralidade do debate entre as atividades que compõem o exercício de um mandato parlamentar. Isso porque, embora não seja garantido a um mandatário que suas propostas (por mais que apresente centenas delas) sejam aprovadas e tornem-se normas jurídicas, a ele(a) será sempre assegurado, segundo os ditames regimentais, o direito à fala.

Atualmente, grande parte das tarefas como analista de dados no Poder Legislativo está focada na tentativa de compreender o comportamento dos parlamentares no que se refere à atuação legislativa. É possível, com razoável facilidade, criar mecanismos de monitoramento, por mandatário, da quantidade de proposições apresentadas, da soma de passos que cada uma delas caminhou em sua tramitação, da quantidade de dias em que houve tramitação, entre outras métricas quantitativas. Contudo, a análise do discurso, apesar de sua importância para o mandato parlamentar, representa um desafio, tanto em termos de fundamentação teórica quanto de ferramentas e métodos para a sua execução.

Entre os métodos mais utilizados para a análise do "discurso como conteúdo" – *speech-as-content* (GOPLERUD, 2021, p. 72), está a extração de tópicos, isto é, a identificação de temas a partir das falas. A despeito de sua aparente trivialidade, essa extração representa um desafio quando se levam em conta os milhares de discursos realizados todos os anos nos principais parlamentos no Brasil.

A utilização de métodos de processamento de linguagem natural e inteligência artificial na identificação de temas e extração de sentido em discursos parlamentares pretende-se relevante, como pesquisa acadêmica, em especial quanto ao seguinte:

- será possível analisar e compreender de forma mais precisa e eficiente os discursos parlamentares. Isso contribuirá para uma melhor gestão da burocracia do Poder Legislativo em várias dimensões, como estratégia, processos e tecnologias. Os modelos propostos podem ajudar a identificar os temas mais discutidos, monitorar as preocupações e necessidades dos legisladores, além de fornecer informações valiosas

para a tomada de decisões e aprimoramento das práticas legislativas;

- com a capacidade aprimorada de análise dos discursos parlamentares, os modelos de processamento de linguagem natural podem contribuir para o monitoramento mais eficiente da avaliação feita pelos parlamentares governistas e oposicionistas a respeito das políticas públicas conduzidas pelo Executivo. Ao se identificar e extrair informações relevantes dos discursos, será possível obter uma compreensão mais profunda dos debates e posicionamentos em relação às políticas em discussão;
- a pesquisa proposta pode ter impacto positivo na promoção da transparência no Poder Legislativo. Ao analisar os discursos parlamentares de forma mais precisa e extrair informações relevantes, os resultados podem ser disponibilizados ao público, permitindo uma maior compreensão das ações, dos debates e das efetivas preocupações dos mandatários. Isso contribuirá para fortalecer a participação cidadã, oferecendo uma visão mais clara, concisa e acessível do trabalho legislativo.

Os resultados esperados também serão relevantes e terão impacto direto nas análises e monitoramentos de atuação parlamentar. Isso porque, para além da mera análise quantitativa, será possível incrementar uma visão de contornos qualitativos sobre a atuação parlamentar, especialmente quanto aos temas de maior predileção dos parlamentares. Essa visão afetará diretamente, por exemplo, a sugestão de relatores para projetos de interesse, o que poderá ser feito a partir da relação entre a pertinência temática da proposta e dos temas a que os parlamentares estão mais afeitos.

A motivação quanto à escolha do período de análise (56ª Legislatura) está relacionado ao fato de ter sido um momento de acontecimentos historicamente importantes que impactaram diretamente o exercício das atividades parlamentares:

- em 2 de fevereiro de 2019, a Legislatura é iniciada;
- em 26 de fevereiro de 2019, é confirmado o primeiro caso de covid-19 no Brasil (SE/UNA-SUS, 2020);
- em 8 de agosto de 2020, o Brasil registra a marca de 100 mil mortes pela covid-19 (BRONZE, 2020);
- em 10 de novembro de 2020, é realizado o primeiro turno das eleições municipais (SENADO, 2020);
- em 17 de janeiro de 2021, a primeira pessoa recebe a vacina contra covid-19 no Brasil (BADDINI; FERNANDES, 2021);
- em 19 de junho de 2021, o Brasil registra a marca de 500 mil mortes pela covid-19 (TITO, 2021);

- em 30 de julho de 2021, o Brasil chega a 100 milhões de aplicações da 1ª dose da vacina contra covid-19 (GARCIA, 2021);
- em 1º de fevereiro de 2023, a Legislatura é encerrada.

Trata-se, portanto, de um período singular na história política e social do país. Este intervalo foi marcado por eventos extraordinários que redefiniram o cenário nacional e impactaram profundamente a atividade parlamentar. O surgimento e a evolução da pandemia de covid-19 trouxeram desafios sem precedentes para a saúde pública e a economia. A resposta legislativa a essa crise, juntamente com as complexidades do manejo de políticas públicas em um contexto de emergência sanitária, ressalta a relevância deste período para análises políticas e sociais. Portanto, a escolha dessa Legislatura como foco de estudo não apenas captura um capítulo crucial da história recente do Brasil, mas também oferece um rico terreno para investigar como eventos de grande magnitude se refletem na esfera legislativa.

Nesse contexto único, espera-se identificar quais foram os tópicos e os temas considerados mais relevantes para os parlamentares governistas e para os oposicionistas, explorando como eles refletiram os desafios e as prioridades de um período marcado por crises e transformações significativas.

3 Objetivos e Organização do Trabalho

Em consonância com o contexto singular da 56ª Legislatura do Congresso Nacional brasileiro e suas implicações políticas e sociais, este trabalho visa explorar e compreender a dinâmica discursiva dentro do parlamento brasileiro. Conforme já foi exposto, a motivação para este estudo surge da necessidade de decifrar como os parlamentares, tanto do governo quanto da oposição, articularam seus discursos em face de um cenário político e social em constante evolução.

3.1 Objetivo geral

O objetivo geral deste estudo é extrair, analisar e comparar tópicos dos discursos de parlamentares governistas e oposicionistas na Câmara dos Deputados durante a 56ª Legislatura, identificando as principais temáticas e padrões.

3.2 Objetivos específicos

- Desenvolver um critério objetivo para classificar parlamentares como governistas e oposicionistas;
- Criar um modelo de extração de tópicos de discursos parlamentares utilizando o *Latent Dirichlet Allocation* (LDA);
- Comparar, para cada grupo, os tópicos extraídos, identificando semelhanças e diferenças ao longo do tempo;
- Identificar diferenças e semelhanças gerais, entre governistas e oposicionistas, de tópicos e palavras utilizadas durante a legislatura.

3.3 Organização

A pesquisa proposta pode ser classificada como uma pesquisa empírica, que busca coletar e analisar dados concretos e observáveis para responder ao problema de pesquisa.

Inicialmente, para a composição do referencial teórico, foi realizada uma busca por trabalhos com objetivos de pesquisa semelhantes, para identificar avanços na área e também desafios a serem esperados. Posteriormente, foram localizados os artigos que tratam do modelo *LDA*, inclusive o texto seminal, datado de 2003 (BLEI; NG; JORDAN, 2003).

A partir da configuração teórica estabelecida, passou-se a extrair os dados necessários para a realização da pesquisa. Essa coleta foi realizada a partir das *APIs* disponibilizadas pelo portal de dados abertos da Câmara dos Deputados.

Com vistas à realização da primeira parte da pesquisa empírica, que era a classificação de parlamentares como governistas ou de oposição, foi realizada a requisição para extração dos dados referentes às orientações de votos feitas pela liderança do governo na Câmara dos Deputados. Posteriormente, foram requisitados os dados relativos a cada votação realizada durante o período, de forma a identificar o exato posicionamento de cada parlamentar que votou. Assim, foi possível comparar os votos dos parlamentares com o posicionamento da liderança do governo, de forma a se gerar um índice de convergência.

Posteriormente, foi realizada a extração dos registros de falas dos parlamentares que atuaram durante o período selecionado para a pesquisa. Esse conteúdo foi devidamente explorado e foram definidos os discursos utilizados para a modelagem de tópicos. Os discursos selecionados foram então processados para a criação do modelo de extração de tópicos com *LDA*.

Foram, então, extraídos os tópicos, semestralmente, para cada grupo. Optou-se por pré-definir a identificação de 10 tópicos, por semestre, e por destacar, em cada tópico, as 10 palavras com os maiores pesos, ou seja, mais relevantes. Com isso, foram obtidos, para cada grupo, 80 tópicos que eram descritos por 800 palavras (não necessariamente únicas).

Para a análise da semelhança entre os discursos de diferentes semestres, realizou-se inicialmente a vetorização das palavras de cada tópico usando *word embeddings*, uma técnica que transforma palavras em vetores em um espaço multidimensional (MIKOLOV et al., 2013). Utilizando um modelo pré-treinado de *word embeddings* do Núcleo Interinstitucional de Linguística Computacional da Universidade de São Paulo (HARTMANN, 2017), cada palavra do tópico foi convertida em um vetor, e esses vetores foram ponderados com base em sua relevância no tópico. A média desses vetores ponderados foi calculada para obter um vetor representativo para cada tópico. Em seguida, utilizou-se a similaridade por cosseno para calcular a semelhança semântica entre os vetores de tópicos, permitindo identificar o grau de semelhança ou de diversidade entre os tópicos ao longo do tempo.

Por fim, foram consolidadas as informações relativas às palavras únicas presentes em cada tópico, tanto para a oposição quanto para o governo, de maneira que se obteve um levantamento que identificou a frequência em que apareceram nos tópicos e também a soma dos pesos que foram conferidos a elas. Assim, foi possível precisar a diferença de frequência e da soma dos pesos entre governo e oposição para cada palavra única. Essa abordagem permitiu uma análise quantitativa e qualitativa das temáticas abordadas por ambos os grupos, revelando não apenas quais assuntos foram mais discutidos, mas também a intensidade com que certas palavras foram utilizadas. A frequência de ocor-

rência e a soma dos pesos atribuídos a cada palavra forneceram percepções importantes sobre as prioridades e os focos de atenção de cada grupo. Por exemplo, palavras com alta frequência e pesos significativos em um grupo, mas não no outro, indicam tópicos de particular interesse ou preocupação para esse grupo. Assim, essa metodologia possibilitou uma compreensão mais profunda e detalhada das diferenças e semelhanças nos discursos e posicionamentos dos parlamentares governistas e oposicionistas ao longo da 56ª Legislatura, refletindo as nuances políticas do período analisado.

Tanto a extração dos dados quanto a sua análise e a criação dos modelos foram realizados em *python*, tendo em vista as robustas bibliotecas disponibilizadas para os fins a que se propôs a pesquisa. Foram especialmente utilizadas as bibliotecas: *pandas*, *seaborn*, *matplotlib*, *requests*, *unicodedata*, *spacy* e *scikit-learn*.

3.4 Cronograma

Tabela 1 – Cronograma do trabalho

Atividade	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**	**Ago**	**Set**	**Out**	**Nov**
Revisão de literatura	X	X							
Coleta de dados		X	X						
Classificação de parlamentares			X	X					
Processamento dos discursos				X	X				
Modelagem de tópicos com LDA					X	X			
Análise de semelhança dos tópicos						X	X		
Consolidação das informações e das análise							X	X	
Redação do relatório final								X	X
Revisão e submissão do trabalho									X

Fonte: Elaboração própria, 2023.

4 Referencial Teórico

O estudo dos discursos parlamentares ocupa um lugar fundamental na compreensão dos processos políticos, refletindo as preocupações, estratégias e posicionamentos dos legisladores. No Brasil, a Câmara dos Deputados, como uma das principais instituições públicas, é o palco em que diversas perspectivas políticas se encontram e se confrontam. A análise dos discursos que lá ocorrem pode oferecer informações valiosas sobre a dinâmica política, os temas prioritários para diferentes grupos e a evolução do discurso político ao longo do tempo. Ao tentar responder à pergunta"porque estudar debates legislativos?", especialistas afirmam que eles constituem fonte de informação mais relevante que o próprio voto no que se refere às posições políticas do mandatário, pois estão menos afeitos às limitações de agenda impostas pelas lideranças partidárias (BACK; DEBUS; FERNANDES, 2021b, p. 7).

Como já se afirmou, a 56ª Legislatura do Congresso Nacional brasileiro (2019-2023) apresenta-se como um período singular na história política recente do Brasil. Marcada por eventos significativos como a pandemia de COVID-19, mudanças políticas substanciais e crises sociais, essa era representa um cenário complexo e multifacetado para a análise discursiva. A maneira como os parlamentares articularam seus discursos durante este período fornece um campo fértil para a investigação, oferecendo uma janela para entender as respostas políticas a desafios sem precedentes.

Com o avanço da ciência de dados e das tecnologias de PLN, surge uma oportunidade ímpar de explorar esses discursos de maneira mais sistemática e profunda. Ferramentas como o *Latent Dirichlet Allocation* (LDA) para modelagem de tópicos e as técnicas de *word embeddings* permitem uma análise qualitativa dos discursos, transcendendo as abordagens quantitativas tradicionais. Estas técnicas fornecem novas perspectivas para desvendar padrões e tendências nos discursos parlamentares, contribuindo até mesmo para uma melhor compreensão do cenário político.

Este estudo busca não apenas descrever os tópicos discutidos, mas também entender as nuances nas abordagens de parlamentares governistas e oposicionistas. Através de uma análise empírica baseada em dados, o trabalho propõe contribuir para o campo de estudos políticos e análise de dados. Este enfoque é particularmente relevante para entender como questões de interesse nacional são moldadas e debatidas no âmbito legislativo, e como isso reflete as dinâmicas políticas mais amplas do país.

A robusta e recente compilação de estudo sobre debates legislativos, *The Politics of Legislative Debates* (BACK; DEBUS; FERNANDES, 2021a), fornece uma ampla possibilidade de compreensão do estado da arte sobre o tema. Além de dezenas de artigos que tratam de diferentes possibilidades de análise a partir dos mais diversos parlamentos no

mundo, a obra conta com importantes contribuições teóricas sobre o assunto. Goplerud (GOPLERUD, 2021, p. 72), por exemplo, descreve os principais métodos para análise dos debates parlamentares: discurso como atividade, discurso como medida e discurso como conteúdo. Cada uma dessas abordagens conta com diferentes métodos para a consecução de seus objetivos.

A análise do discurso como atividade é a mais comum e mais simples de ser realizada. São exemplos a contagem de discurso feitos por parlamentares ao longo de determinado período, a soma das palavras, o momento em que foram realizados, quais partidos mais falaram, entre outros (GOPLERUD, 2021, p. 72).

Na análise do discurso como medida, a fala é utilizada como uma fonte de dados para inferir conceitos relevantes, como a ideologia dos legisladores. Esta abordagem é quantitativa e busca medir aspectos subjacentes do discurso. Pode-se, por exemplo, criar variáveis que representem posicionamentos políticos (GOPLERUD, 2021, p. 76).

Por fim, na análise do discurso como conteúdo, "o objetivo geral é classificar cada discurso como sendo sobre algum tópico político saliente ('economia', 'agricultura') e analisar como os legisladores diferem na utilização desses tópicos..."(GOPLERUD, 2021, p. 81).

O presente trabalho adotará métodos para análise do discurso como conteúdo, objetivando identificar as semelhanças e diferenças entre as escolhas e ênfases dadas por diferentes grupos de parlamentares, no caso governistas e oposicionistas.

Foram ainda, identificados dois importantes trabalhos que também se dedicaram a analisar os discursos parlamentares como conteúdo: "Com a palavra os nobres deputados: frequência e ênfase temática dos discursos dos parlamentares brasileiros"(MOREIRA, 2016) e "Análise do discurso parlamentar por meio da técnica de processamento de linguagem natural: abordagem estatística e aprendizagem de máquina"(SCHWARTZ, 2018).

O estudo conduzido por Davi Moreira representa uma importante contribuição para a compreensão do comportamento parlamentar no Brasil, especialmente no que tange à dinâmica dos discursos no plenário da Câmara dos Deputados. Utilizando o modelo de agenda expressa (*expressed agenda model*), o autor analisou mais de 127 mil pronunciamentos feitos por mais de 2 mil parlamentares ao longo de 15 anos (1999-2014). O objetivo central do estudo foi investigar se os temas proferidos pelos parlamentares em seus discursos eram predominantemente moldados pela dicotomia governo-oposição (MOREIRA, 2016).

No trabalho correlato, Schwartz também objetivou automatizar a extração de sentido dos discursos parlamentares a partir de técnicas de processamento de linguagem natural. Entre as ferramentas utilizadas pelo autor na análise de mais de 340 mil discursos durante um período de 20 anos, também está o *LDA* (SCHWARTZ, 2018).

A contribuição da presente pesquisa com relação aos trabalhos citados reside tanto no método para identificação de parlamentares governistas e oposicionistas, quanto no que possibilitou a determinação da semelhança entre tópicos ao longo do tempo, por meio da utilização de *word embeddings*.

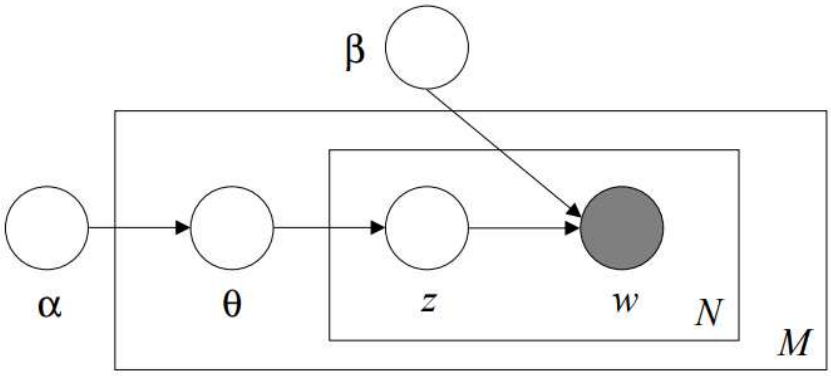
Também foi referência imprescindível o trabalho seminal relacionado ao LDA. O artigo *Latent Dirichlet Allocation*, publicado em 2003, apresenta a ferramenta como um modelo generativo probabilístico para coleções de documentos. A principal inovação do LDA é a sua capacidade de modelar cada documento como uma mistura de tópicos, em que cada tópico é caracterizado por uma distribuição de palavras (BLEI; NG; JORDAN, 2003, p. 996).

Conforme descrição do processo generativo no artigo, para cada documento w em um corpus D tem-se (BLEI; NG; JORDAN, 2003, p. 996):

1. **Escolha de N ($\sim$ Poisson(ξ)):**
   - Cada documento é considerado como composto por um número variável de palavras;
   - Este número, N, é determinado por uma distribuição de Poisson com parâmetro ξ.
2. **Escolha de θ ($\sim$ Dir(α)):**
   - Cada documento é uma mistura de vários tópicos;
   - A mistura de tópicos é representada por θ, uma distribuição de probabilidade sobre os tópicos;
   - θ é gerado a partir de uma distribuição *Dirichlet* com parâmetro α.
3. **Para cada uma das palavras N no documento:**
   a) **Escolha de um tópico z_n ($\sim$ Multinomial(θ)):**
      - Para cada palavra no documento, um tópico z_n é escolhido com base na distribuição multinomial θ;
      - Cada palavra é atribuída a um tópico de acordo com as proporções de tópicos do documento.
   b) **Escolha de uma palavra w_n a partir de $p(w_n|z_n, \beta)$:**
      - Após a escolha do tópico z_n para a palavra, w_n é escolhida de uma distribuição multinomial condicionada ao tópico z_n e ao parâmetro β;
      - β representa a distribuição de palavras para cada tópico, isto é, a probabilidade de uma palavra ser escolhida dado um tópico específico.

O representação gráfica do modelo probabilístico de LDA é a seguinte:

Figura 1 – Representação gráfica do LDA



Fonte: (BLEI; NG; JORDAN, 2003, p. 997)

Estão presentes três níveis: os parâmetros α e β são relativos ao *corpus*, selecionados uma única vez no processo de geração do *corpus*; as variáveis θ_d estão no nível de documento, selecionadas uma vez por documento; as variáveis z_{dn} e w_{dn} estão no nível de palavra, e são selecionadas uma vez para cada palavra em cada documento (BLEI; NG; JORDAN, 2003, p. 997).

Este processo generativo descreve uma maneira hipotética de como um documento poderia ser construído a partir de um conjunto de tópicos. Na prática, o *LDA* inverte esse processo para descobrir os tópicos latentes que melhor explicam os documentos em um *corpus*.

Também foi de grande utilidade o artigo *Probabilistic Topic Models* (BLEI, 2016), que serviu para esclarecer aspectos importantes e complexos do *LDA* ao relacioná-lo como possibilidade de solução para o gerenciamento de grande arquivos digitais de documentos.

Por fim, conforme já mencionado, foi imprescindível para a pesquisa a utilização da vetorização via *word embbeddings*, o que possibilitou a comparação de tópicos ao longo do tempo.

No artigo seminal *Efficient Estimation of Word Representations in Vector Space* (MIKOLOV et al., 2013), os autores introduzem uma abordagem inovadora para modelar representações de palavras em espaços vetoriais. Esta técnica permitiu a captura de semânticas complexas e relações entre palavras ao representá-las como vetores densos em um espaço de alta dimensionalidade. Diferentemente de representações mais antigas baseadas na frequência de palavras individuais, os *word embeddings* são capazes de capturar nuances semânticos, tornando-se uma ferramenta poderosa para diversas aplicações de PLN.

O artigo introduz dois modelos principais para a geração de *word embeddings*:

o *continuous bag-of-words* (C-BOW) e o *skip-gram* (MIKOLOV et al., 2013, p. 4-5). O modelo *C-BOW* foca na previsão de uma palavra-alvo com base no contexto de palavras circundantes, enquanto o *Skip-Gram* faz o inverso, utilizando uma palavra-alvo para prever seu contexto. O modelo *Skip-Gram*, em particular, mostrou-se mais eficaz em capturar relações semânticas (MIKOLOV et al., 2013, p. 7), tornando-o uma escolha ideal para a comparação de tópicos extraídos de discursos parlamentares.

5 Desenvolvimento

5.1 Definição das fontes de dados

Os dados necessários para a realização da pesquisa foram coletados por meio das *APIs* disponibilizadas pelo portal de dados abertos da Câmara dos Deputados (BRASIL, 2023b).

Foram feitas as requisições necessárias e obteve-se as seguintes informações:

- 626 dados sobre os parlamentares;
- 53.267 orientações de voto;
- 800.095 votos;
- 68.224 discursos.

Enquanto os dados relativos às orientações de lideranças e aos votos foram necessários para a implementação de um método para identificação de parlamentares governistas e oposicionistas, os dados sobre parlamentares e discursos foram utilizados para a preparação das falas para a extração de tópicos via *LDA*.

5.2 Descrição dos dados

Os dados disponibilizados para os parlamentares federais continham as seguintes variáveis:

Tabela 2 – Descrição das variáveis dos dados dos parlamentares

Variável	**Descrição**
id	Identificador único do deputado
uri	URL do recurso do deputado, fornecendo acesso direto às informações detalhadas
nome	Nome parlamentar do deputado
siglaPartido	Sigla do partido político ao qual o deputado está filiado
uriPartido	URL do recurso do partido político do deputado
siglaUf	Sigla da Unidade Federativa representada pelo deputado
idLegislatura	Identificador da legislatura durante a qual o deputado está atuando
urlFoto	URL da foto oficial do deputado
email	Endereço de e-mail do deputado

Fonte: Elaboração própria, 2023.

As orientações das votações vieram acompanhadas das seguintes variáveis:

Tabela 3 – Descrição das variáveis de orientações de votações

Variável	**Descrição**
idVotacao	Identificador único para cada votação
uriVotacao	URL do recurso da votação, fornecendo acesso direto às informações detalhadas da votação específica
siglaOrgao	Sigla do órgão legislativo em que a votação ocorreu
descricao	Descrição do resultado da votação
siglaBancada	Sigla representativa da bancada ou do partido político que orientou a votação
uriBancada	URL do recurso da bancada ou do partido, contendo informações adicionai.
orientacao	Orientação da votação dada pela liderança do partido ou da bancada (ex.: sim, não, obstrução, etc.).

Fonte: Elaboração própria, 2023.

Os dados referentes aos votos individuais dos parlamentares eram compostos pelo seguinte:

Tabela 4 – Descrição das variáveis de votos

Variável	**Descrição**
idVotacao	Identificador único para cada votação
uriVotacao	URL do recurso da votação, fornecendo acesso direto às informações detalhadas da votação específica
dataHoraVoto	Data e hora em que o voto foi registrado
voto	Tipo de voto registrado pelo deputado (ex.: sim, não, abstenção, obstrução)
deputado_id	Identificador único do deputado que votou
deputado_uri	URL do recurso do deputado, contendo informações detalhadas
deputado_nome	Nome parlamentar do deputado
deputado_siglaPartido	Sigla do partido político do deputado
deputado_uriPartido	URL do recurso do partido do deputado
deputado_siglaUf	Sigla da Unidade Federativa representada pelo deputado
deputado_idLegislatura	Identificador da legislatura em que o deputado está atuando
deputado_urlFoto	URL da foto oficial do deputado

Fonte: Elaboração própria, 2023.

A coleta de discursos parlamentares foi realizada em duas etapas principais:

- Função de coleta de discursos (`get_speeches`):
  - Esta função teve como objetivo requisitar os discursos de um deputado específico, usando o seu identificador (`deputado_id`);

  - A *API* da Câmara dos Deputados foi consultada com parâmetros de data de início e término, ordenação por data e hora, e número de itens por paginação;
  - Um *loop* foi implementado para coletar os discursos de todas as páginas disponíveis, concatenando os resultados em uma lista de *DataFrames*;
  - Para cada discurso obtido, foi calculado o número total de palavras e adicionado ao *DataFrame* resultante.

- Processamento e coleta de discursos (`process_speeches`)
  - Esta função foi responsável por processar os discursos de uma lista de deputados, provida pelo *DataFrame* com os dados dos deputados;
  - Para cada deputado na lista, a função `get_speeches` foi invocada para coletar seus discursos;
  - O processo incluiu tratamento de erros e tentativas de reconexão, assegurando a coleta completa dos discursos;
  - O progresso da coleta foi monitorado, registrando-se o *status*, a quantidade de discursos coletados e o total de palavras para cada deputado;
  - Os discursos coletados de todos os deputados foram consolidados em um único *DataFrame* (`df_speeches`).

Ao final, os dados relativos aos discurso estavam organizados da seguinte forma:

Tabela 5 – Descrição das variáveis dos discursos

Variável	**Descrição**
dataHoraInicio	Data e hora de início do discurso
dataHoraFim	Data e hora de término do discurso
uriEvento	URL do recurso do evento associado ao discurso
tipoDiscurso	Categoria ou tipo do discurso
urlTexto	URL do texto integral do discurso
urlAudio	URL do arquivo de áudio do discurso
urlVideo	URL do vídeo do discurso
keywords	Palavras-chave associadas ao discurso
sumario	Resumo ou sumário do conteúdo do discurso
transcricao	Transcrição textual completa do discurso
contagemPalavras	Número total de palavras no discurso
faseEvento.titulo	Título da fase do evento em que o discurso ocorreu
faseEvento.dataHoraInicio	Data e hora de início da fase do evento
faseEvento.dataHoraFim	Data e hora de término da fase do evento
deputado_id	Identificador único do deputado que proferiu o discurso
deputado_nome	Nome parlamentar do deputado que proferiu o discurso

Fonte: Elaboração própria, 2023.

5.3 Pré-processamento e limpeza dos dados

5.3.1 Orientações do governo e votos dos parlamentares

Para os dados que foram utilizados para a identificação de parlamentares governistas e oposicionistas não houve grandes necessidades de limpezas ou transformações. Como o interesse era na identificação das orientações feitas pela liderança do governo na Câmara, foi realizado esse filtro, identificando-se o posicionamento esposado em cada votação. Nesse momento, observou-se que havia a identificação, no campo "siglaBancada", dos seguintes valores: "Governo"e "GOV". A solução simples foi a substituição dos valores "GOV"para "Governo".

Foram ainda realizados dois filtros. O primeiro foi a definição apenas das votações realizadas em Plenário, que é a arena em que todos podem votar. As votações nas comissões não foram consideradas justamente porque não seria possível identificar a dicotomia governismo - oposicionismo para aqueles parlamentares que não participaram de votações no colegiado. O segundo filtro ateve-se apenas às orientações "Sim"e "Não", excluindo-se os demais valores como "Abstenção"e "Obstrução", pois pouco dizem sobre o posicionamento da liderança sobre o mérito das votações e mais sobre a dinâmica dos acontecimentos no Plenário. Ao final, restaram 1.778 votações para as quais a liderança do governo na Câmara orientou "Sim"ou "Não".

Quanto aos dados relativos aos votos, foram igualmente filtrados para que restassem apenas os "Sim"e "Não", excluindo-se os demais valores. Isso possibilitou a comparação entre os efetivos posicionamentos da liderança do governo e dos parlamentares, por meio de seus votos. Além disso, foram identificados os votos exclusivamente das votações correspondentes às que haviam sido orientadas pela liderança do governo. Restaram, assim, 698.031 votos que foram comparados com as 1.778 orientações do governo, ou aproximadamente 393 votos por orientação.

5.3.2 Discursos

São diversos os tipos de classificação de discursos de um parlamentar federal. A decisão, portanto, é a de quais tipos devem ser considerados para a criação de um modelo de extração de tópicos.

Quanto isso, vamos considerar a compilação de possibilidades de fala organizada por Davi Moreira e que as relaciona aos dispositivos do Regimento Interno da Câmara dos Deputados (MOREIRA, 2016, p. 45):

Tabela 6 – Possibilidades de fala e dispositivos regimentais

Situação nº	**Possibilidade de fala**	**Dispositivos Regimentais**
1	Apresentar proposição	Art. 132
2	Fazer comunicação ou versar sobre assuntos diversos, a hora das Breves Comunicações, do Pequeno Expediente, do Grande Expediente ou das Comunicações Parlamentares	Art. 66, I, II, e IV; arts. 81, 87 e 90
3	Discutir proposição em pauta	Art. 157, §3º; arts. 171 a 176
4	Formular Questão de Ordem ou contra-argumentar	Art. 95
5	Apresentar reclamação	Art. 96
6	Encaminhar votação	Art. 192
7	Contestar acusação pessoal à própria conduta, feita durante discussão, ou para contradizer o que lhe for atribuído indevidamente como opinião pessoal	—

Fonte: (MOREIRA, 2016, p. 45).

Segundo o autor,

> "Cada um desses momentos de fala possui regras próprias para a manifestação dos Deputados Federais..., mas é no Pequeno Expediente, no Grande Expediente e (*sic*) nas Comunicações Parlamentares e nas Breves Comunicações que os parlamentares têm a previsão regimental para tratar de assuntos diversos à pauta e podem discursar de forma livre na tribuna desde que dentro do tempo e das regras permitidas."(MOREIRA, 2016, p. 44-45)

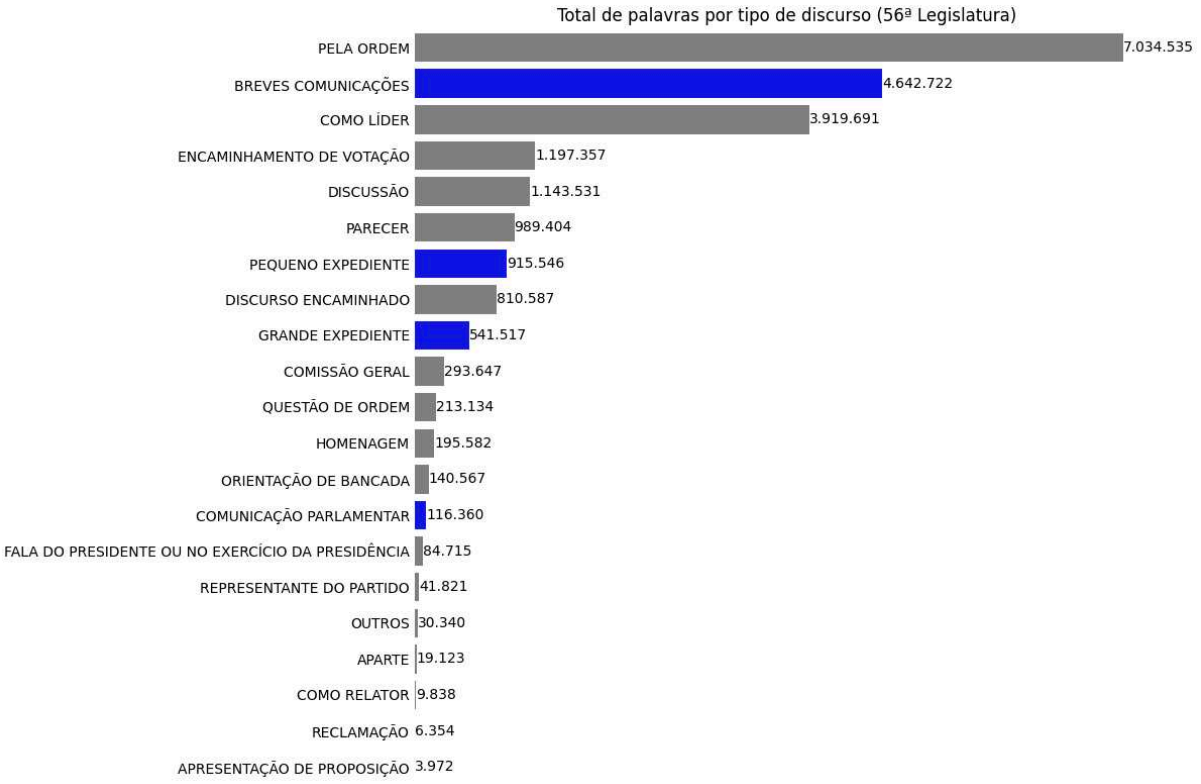
E ainda acrescenta:

> "É justamente pelo fato da fala parlamentar ter a possibilidade de se concretizar em um ambiente de relativa independência em relação ao processo decisório que este trabalho busca investigar como se distinguem os parlamentares em relação à enfase temática dos discursos proferidos em plenário e avançar na análise de outras dimensões da atividade parlamentar."(MOREIRA, 2016, p. 45)

Assim, considerando que diversos outros tipos de discursos estão necessariamente atrelados às questões procedimentais relativas aos andamentos das sessões, optou-se por definir tipos específicos que, por sua natureza e momento, são ideais para a extração de tópicos, quais sejam: Breves Comunicações, Pequeno Expediente, Grande Expediente e Comunicação Parlamentar.

O gráfico a seguir revela a dimensão da importância desses 4 tipos de discursos, considerando a quantidade total de palavras utilizadas:

Figura 2 – Total de palavras por tipo de discurso



Fonte: Elaboração própria, 2023.

O exemplo abaixo, de uma transcrição de discurso exatamente como fornecida pela fonte de dados, ilustra os desafios para o pré-processamento das falas:

```
`O SR. ARLINDO CHINAGLIA (PT - SP. Sem revisão do orador.) - Sr. Presidente, Srs. e
Sras. Parlamentares, hoje teremos reunião da Comissão Especial para tratar da Proposta
de Emenda à Constituição nº 135, aquela que busca estabelecer o chamado voto impresso.
Uma das consequências da movimentação, especialmente do Presidente da República, é que
mais de 200 intelectuais, empresários, economistas, professores, no dia de ontem,
divulgaram uma nota, afirmando que vai haver eleições e que o resultado será
respeitado. Isso é consequência dos ataques que o Presidente da República fez ao
Tribunal Superior Eleitoral, especialmente ao Ministro Barroso. Agora eu quero entrar
no mérito da questão. O Relator afirma no seu parecer que o voto eletrônico foi um
avanço. Mas, nessa toada de levantar suspeitas sem nenhuma prova, que levou o
Presidente da República a um vexame de dizer que iria provar fraude, no meio do seu
pronunciamento, através da live, ele disse que não há provas de fraude. Este é o
fato: nunca houve fraude provada. Nunca houve sequer suspeição fundamentada. Mas
caminha o parecer do Relator para propor que a apuração seja feita na própria mesa
receptora dos votos. Então, serão toneladas de papel com milhares de apuradores, mais
de 400 mil pessoas, quase 500 mil pessoas. Para piorar ainda mais, no seu parecer,
```

```
absolutamente de forma impensada, ele diz que os cidadãos poderão acompanhar a
apuração. Para piorar ainda mais, ele diz que a apuração será feita pela Polícia
Federal e exclui o próprio Tribunal Superior Eleitoral do seu papel que é exatamente
o de coordenar as apurações das eleições no nosso País. Então, imaginem que ele ainda
abre a possibilidade, sem nenhum pretexto, para o partido político ter até 15 dias
para reclamar do resultado das eleições. Ou seja, busca-se criar uma confusão para
abrir a possibilidade... (Desligamento automático do microfone.) O SR. ARLINDO
CHINAGLIA (PT - SP) - Ele diz que, se não forem eleições limpas a critério dele,
exclusivamente dele, contrariando todos os Poderes, inclusive a maioria naquela
Comissão - nós vamos derrotar o parecer do Relator -, vai sair das quatro linhas da
Constituição. Então, nós temos que, desde já, denunciar o objetivo do golpe, mas não
temê-lo porque o povo brasileiro já venceu algumas ditaduras.'
```

Inicialmente, é possível perceber a existência de diversos elementos não orais como, por exemplo:

- "O SR. ARLINDO CHINAGLIA (PT - SP. Sem revisão do orador.) -"
- "(Desligamento automático do microfone.)"
- "SR. ARLINDO CHINAGLIA (PT - SP) -"

A partir dessas e de outras constatações ao longo da pesquisa, foi criado o seguinte processo de limpeza dos dados:

- normalização *Unicode* e remoção de caractéres específicos;
- remoção de textos após frases específicas;
- remoção de identificação de orador e fragmentos subsequentes;
- remoção de conteúdo dentro de parênteses;
- remoção de aspas;
- remoção de texto dentro de colchetes duplos;
- atualização da contagem de palavras.

Na sequência, foram realizados mais dois importantes processos para que os textos dos discursos pudessem ser utilizados para a criação de um modelo de extração de tópicos com *LDA*:

- retirada de *stopwords*: palavras comuns, que não contribuem significativamente para um modelo de extração de tópicos;

- lematização: processo de redução das palavras para o seu *lemma*, ou forma mais básica.

Ao final, os discursos foram reduzidos ao estritamente necessário para que fossem extraídos tópicos relevantes, conforme no exemplo a seguir, que é a fala anteriormente transcrita e agora simplificada:

```
`presidente parlamentar reuniãor comissão especial tratar proposta emendo constituição
busca estabelecer chamar voto impresso consequência movimentação especialmente
presidente repúblicar intelectual empresário economista professor divulgar nota
afirmar haver eleição resultado respeitar consequência ataque presidente repúblicar
Tribunal superior eleitoral especialmente ministro barroso entrar mérito relator
afirmar parecer voto eletrônico avanço toar levantar suspeito prova levar presidente
repúblico vexamer ir provar fraude pronunciamento live dizer prova fraude fato fraude
provar sequer suspeição fundamentar caminhar parecer relator propor apuração fazer
mesa receptor voto tonelar papel milhar apurador quase piorar parecer absolutamente
impensar cidadão poder acompanhar apuração piorar apuração fazer polícia federal
excluir tribunal superior eleitoral papel exatamente coordenar apuração eleição país
imaginem abrir possibilidade nenhum pretexto partido político dia reclamar resultar
eleição buscar criar confusão abrir possibilidade eleição Limpas critério
exclusivamente contrariar poder inclusive comissão derrotar parecer relator sair
linha constituição denunciar objetivo golpe temê lo brasileiro vencer ditadura',
```

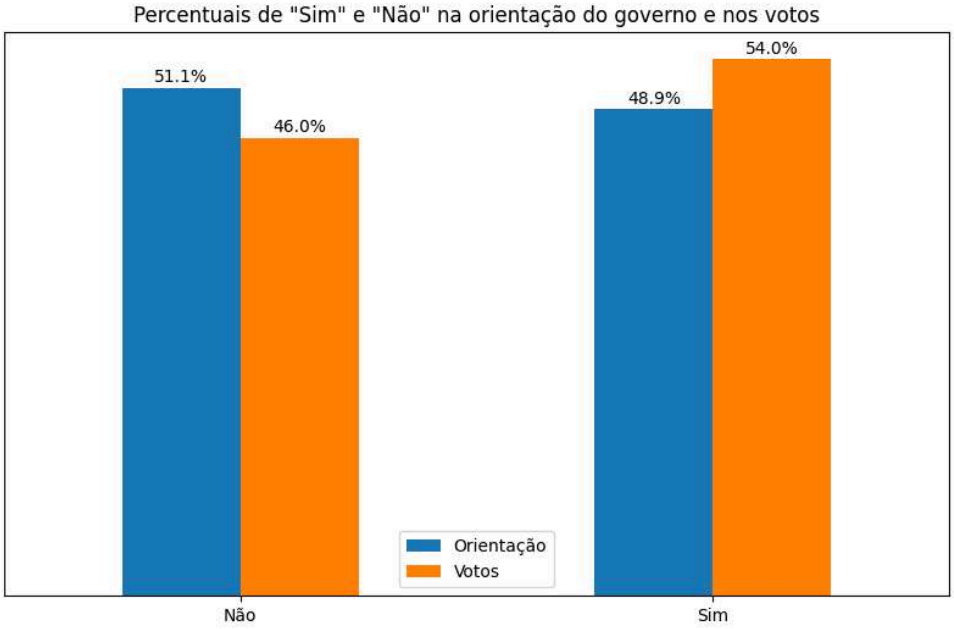
O processo de lematização, especialmente em idiomas complexos como o português, frequentemente apresenta desafios significativos, em razão das nuances linguísticas e da diversidade morfológica do idioma. Bibliotecas como o *spaCy*, que foi utilizada nesse processo, acabam criando palavras que não existem ou são inadequadas no contexto. Por exemplo, no texto processado, palavras como “reuniãor”, “repúblicar”e “vexamer”são claramente produtos de uma lematização imprecisa, demonstrando as limitações da ferramenta em lidar com as peculiaridades do português. Encontrar uma ferramenta de lematização que funcione apropriadamente para o português continua sendo um desafio, refletindo a necessidade de modelos linguísticos mais robustos e adaptados às especificidades do idioma. No entanto, conforme se verá, essas inadequações não afetaram o bom andamento da pesquisa e das análises realizadas.

5.4 Análise exploratória

5.4.1 Orientações do governo e votos dos parlamentares

Considerando as 1.778 orientações do governo, quanto a projetos em pauta, a proporção entre “Sim”e “Não”foi bastante equilibrada, assim como também houve razoável balanceamento para os percentuais de votos “Sim”e “Não”:

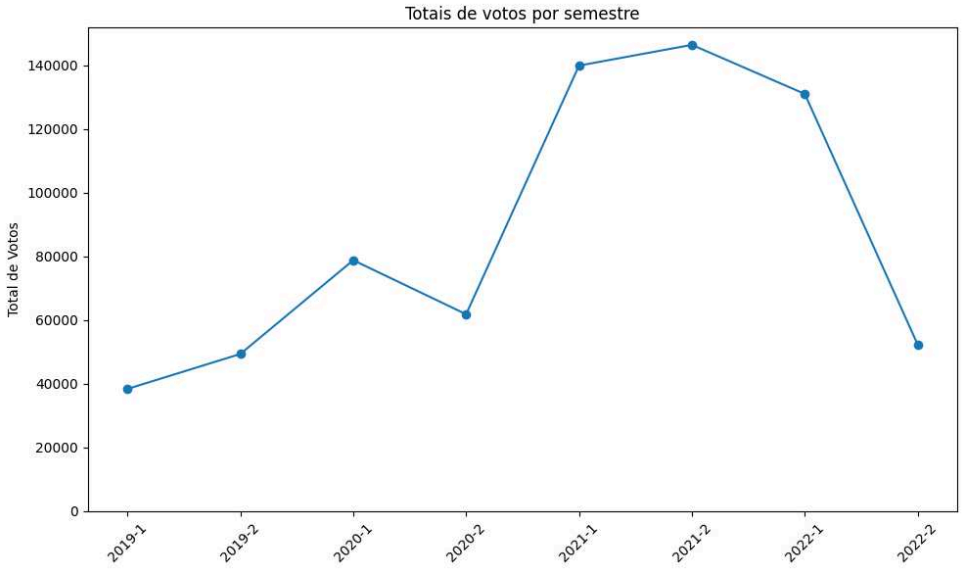
Figura 3 – Percentuais de “Sim”e “Não”na orientação do governo e nos votos



Fonte: Elaboração própria, 2023.

Considerando a evolução semestral da quantidade de votos, é possível perceber uma disparidade bastante considerável entre os primeiros dois anos da legislatura (2019 e 2020) e os dois últimos (2021 e 2022). Enquanto, no primeiro período, a Câmara dos Deputados foi presidida pelo Deputado Rodrigo Maia (DEM/RJ), o segundo foi capitaneado pelo Deputado Arthur Lira (PP/AL):

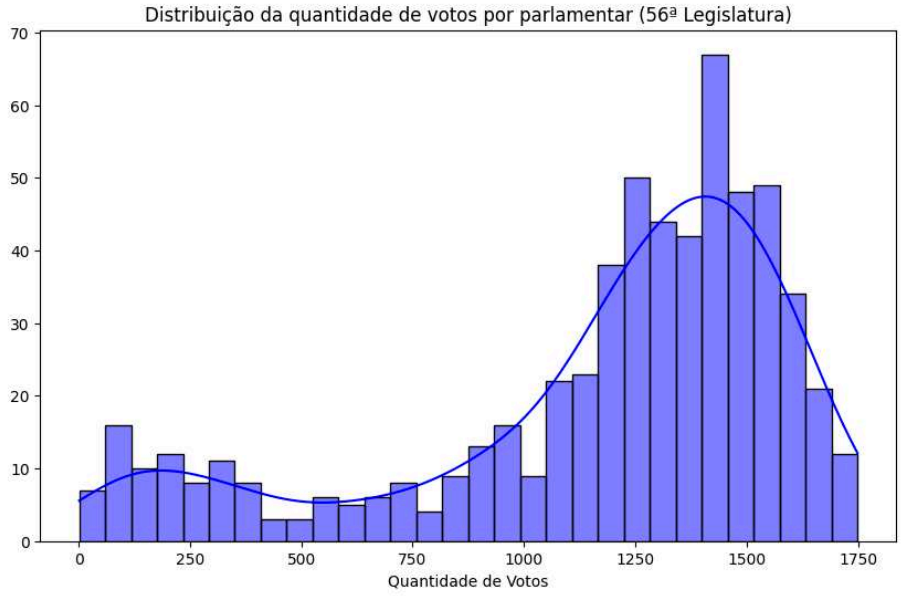
Figura 4 – Totais de votos por semestre



Fonte: Elaboração própria, 2023.

Quando é considerada a distribuição da quantidade de votos por parlamentar durante o período, a sua amplitude vai de 1 a 1750. Isso ocorre porque muitos suplentes acabam assumindo por curtos espaços de tempo quando algum titular precisa se ausentar. Para a realização da análise de convergência, todos os parlamentares foram considerados, mesmo aqueles com poucas votações.

Figura 5 – Distribuição da quantidade de votos por parlamentar

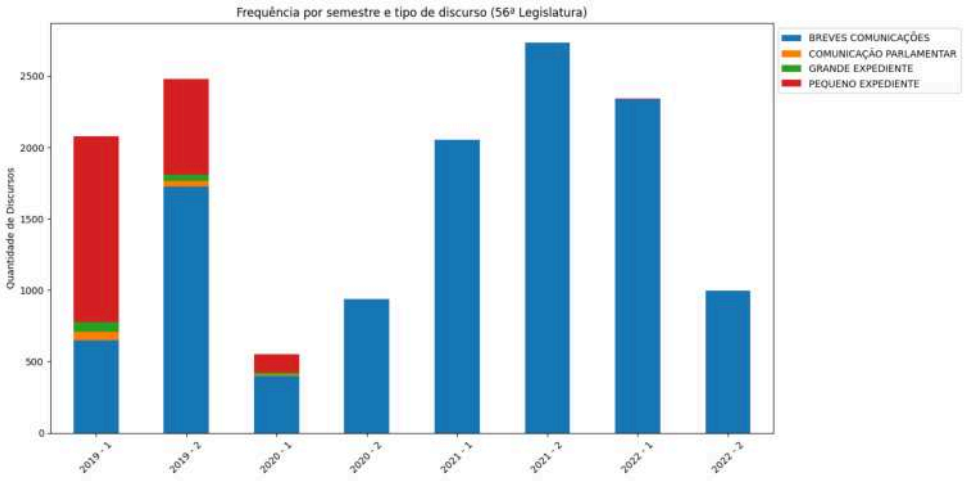


Fonte: Elaboração própria, 2023.

5.4.2 Discursos

A partir da filtragem pelos tipos de discursos selecionados, foram obtidas 14.246 falas, cuja frequência semestral foi a seguinte:

Figura 6 – Frequência por semestre e tipo de discurso

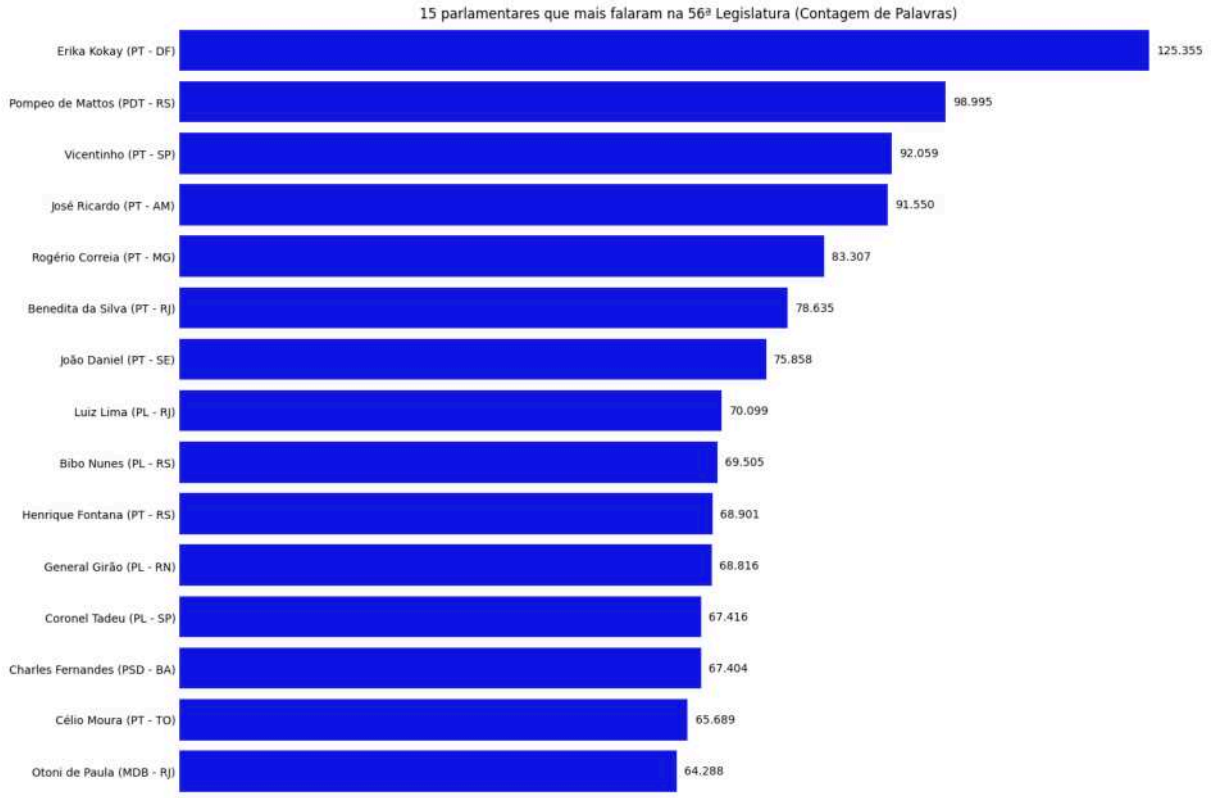


Fonte: Elaboração própria, 2023.

As disparidades podem ser justificadas considerando que os semestres 2020-2 e 2022-2 foram períodos de eleições municipais e gerais, respectivamente. Já o semestre 2020-1 foi justamente o primeiro período da pandemia, que exigiu grande adaptação institucional de meios para as realizações das sessões, que foram dedicadas mais à apreciação de matérias e menos aos discursos parlamentares.

Considerando os 15 parlamentares que mais fizeram uso dos tipos de discursos selecionados, tem-se o seguinte ranqueamento:

Figura 7 – 15 parlamentares que mais falaram (contagem de palavras)



Fonte: Elaboração própria, 2023.

Conforme se verá ao longo to trabalho e está espelhado na figura acima, a tribuna é mais utilizada por parlamentares de oposição, que encontram no púlpito uma ferramenta para expor suas divergências com o governo de plantão.

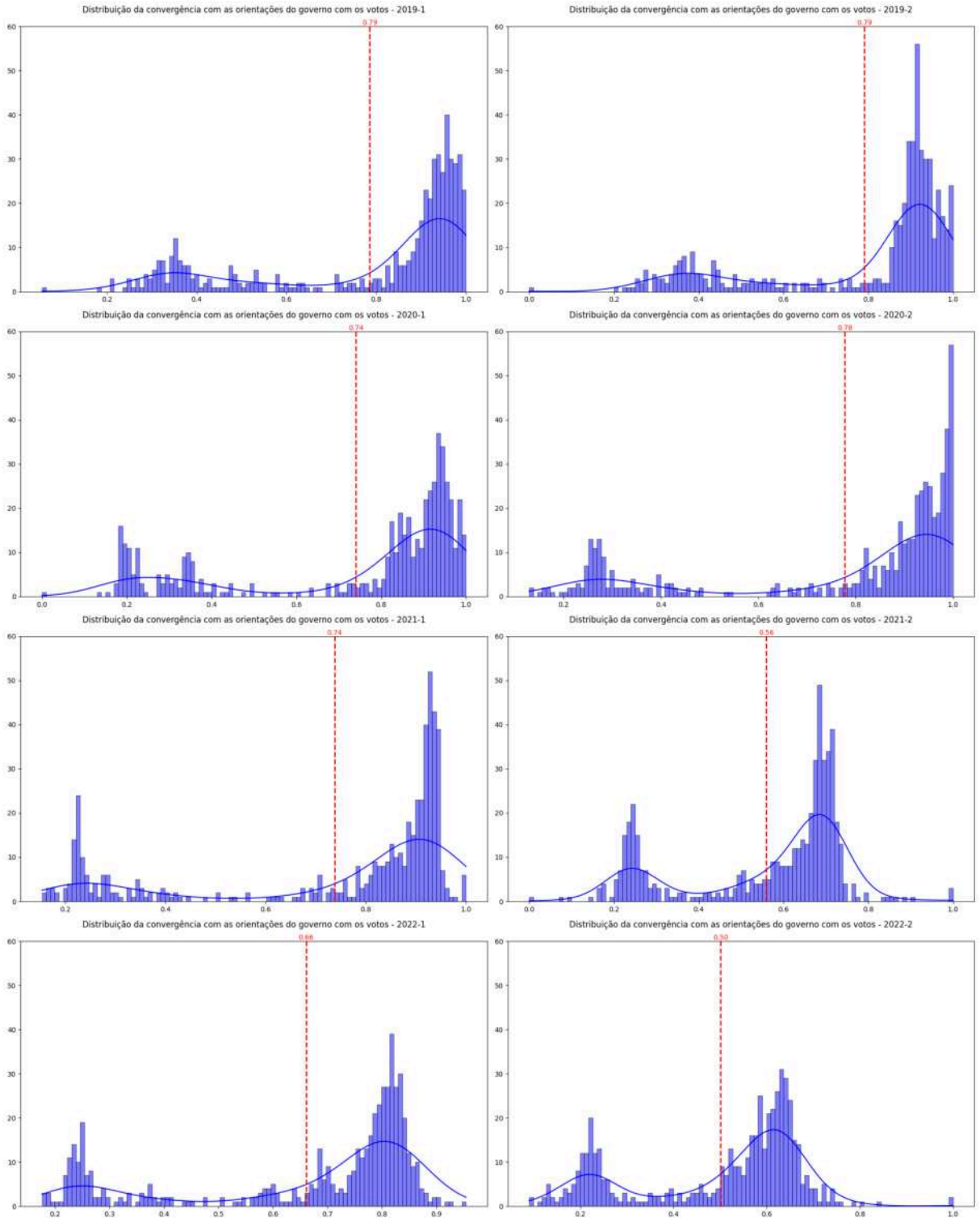
5.5 Modelagem

5.5.1 Convergência

Quando se leva em conta a distribuição da convergência com as orientações do governo (que vai de 0 - nenhuma convergência a 1 - total convergência), percebe-se que

ela varia consideravelmente ao longo do tempo. Quando foi realizado o cálculo para cada semestre da 56ª Legislatura, obteve-se o seguinte:

Figura 8 – Distribuição da convergência com as orientações do governo com os votos - semestral

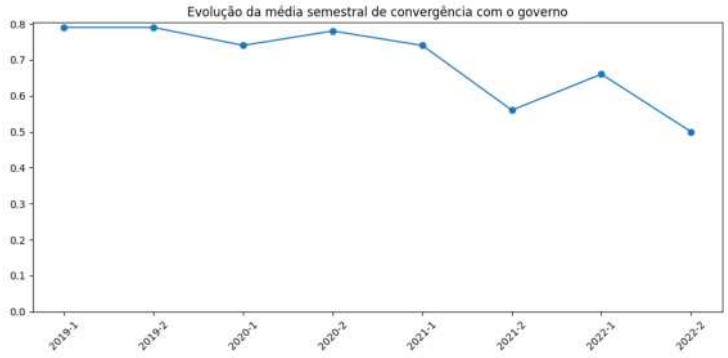


Fonte: Elaboração própria, 2023.

Alguns aspectos dessas distribuições chamam a atenção:

- Em todos os semestres há uma distribuição bimodal, com dois picos distintos de convergência: um mais propenso a seguir as orientações do governo e outro menos alinhados;
- Há uma concentração de índices mais altos de convergência nos primeiro 4 semestres. Nos últimos 4 semestres a curva dos mais convergentes se desloca para a esquerda. Isso fica ainda mais evidente quando são comparadas as médias de convergência com o governo para cada semestre:

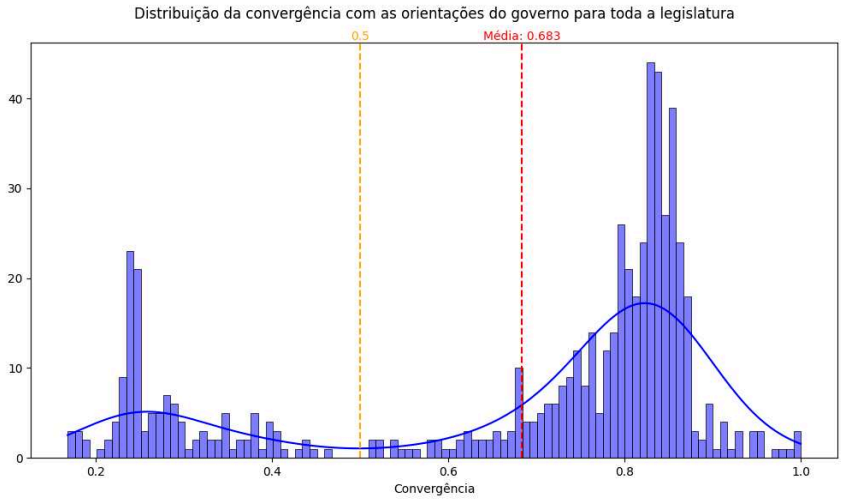
Figura 9 – Evolução da média semestral de convergência com o governo



Fonte: Elaboração própria, 2023.

Quando levamos em conta a frequência para toda a legislatura, obtemos duas curvas bem distintas e claramente separadas:

Figura 10 – Distribuição da convergência com as orientações do governo para toda a legislatura



Fonte: Elaboração própria, 2023.

Dada a clareza da divisão entre dois grupos, optou-se por identificar como governistas aqueles que, ao longo da legislatura, obtiveram um índice de convergência superior ou igual a 0,5 e, como oposicionistas, aqueles abaixo desse índice. Ao final, foram rotulados como governistas 467 parlamentares e como oposicionistas, 137.

5.5.2 Extração de tópicos com *LDA*

Uma vez identificados os discursos com o respectivo rótulo do orador, foram criados dois *datasets*. Os governistas somaram 6.172 discursos e os oposicionistas 8.074. Assim, embora os governistas representassem 77,3% dos parlamentares votantes, foram responsáveis por 43,3% dos discursos. Já os oposicionistas somavam apenas 22,7% dos votantes, mas proferiram 56,7% dos discursos. Enquanto um parlamentar da oposição discursou aproximadamente 60 vezes na legislatura, um governista falou pouco mais de 17 vezes. Ou seja, um oposicionista ocupou a tribuna 3,5 mais vezes que um governista.

Entre diversos parâmetros disponíveis, foram utilizados os seguintes para a construção do modelo:

```
vectorizer = CountVectorizer(max_df=0.90, min_df=0.05, max_features=1000)
lda_model = LatentDirichletAllocation(n_components=10, random_state=42)
```

CountVectorizer é uma técnica de transformação de texto em uma representação de vetor de contagens de termos / palavras(LEARN, 2023b). Vamos examinar os parâmetros utilizados:

- *max_df*: define o valor máximo de frequência de documentos para os termos / palavras a serem considerados. Se uma palavra aparece em mais de 90% dos documentos, ela é excluída. Isso é útil para remover termos muito comuns, que são menos informativos sobre o conteúdo específico do documento;
- *min_df*: define o valor mínimo de frequência de documentos. Se uma palavra aparece em menos de 5% dos documentos, ela é excluída. Isso ajuda a remover termos muito raros, que podem ser específicos demais ou ruídos;
- *max_features*: limita o número de palavras a serem consideradas, com base na sua frequência. Aqui, o número máximo de recursos é definido como 1000. Isso significa que apenas as 1000 palavras mais frequentes serão usadas.

Como já se falou, o *LDA* é um modelo de extração de tópicos utilizado para encontrar palavras temáticas a partir de documentos de texto (LEARN, 2023a). Os parâmetros utilização foram os seguintes:

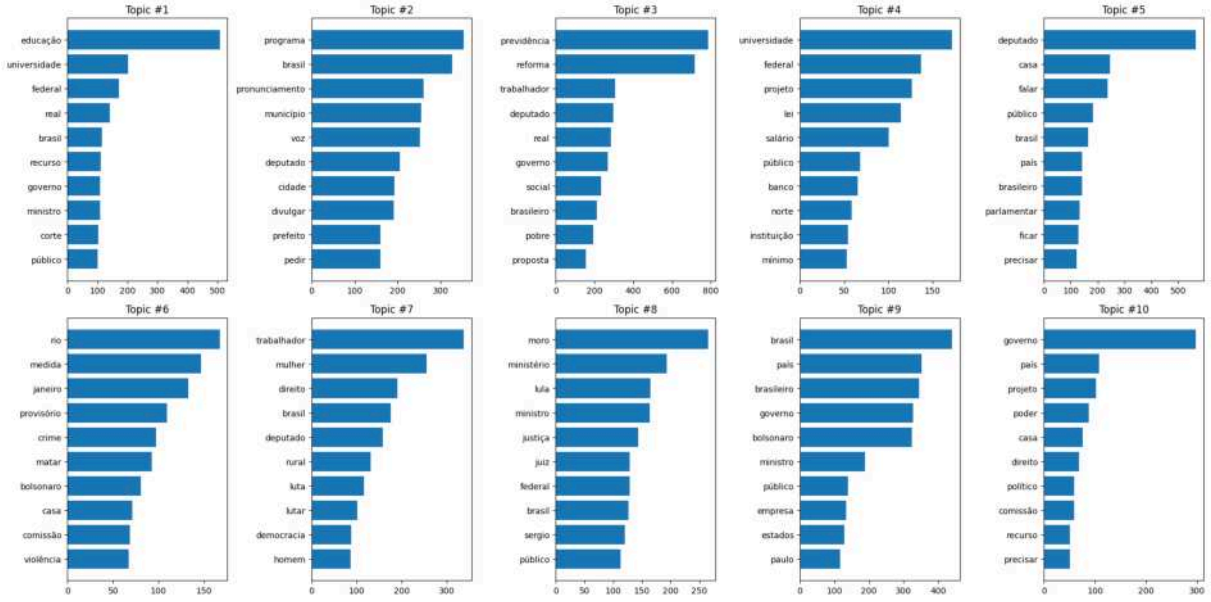
- *n_components*: número de tópicos a serem extraídos do conjunto de documentos. Aqui, está configurado para encontrar 10 tópicos diferentes;
- *random_state*: é a semente usada pelo gerador de números aleatórios. Definir uma semente específica garante que os resultados sejam reproduzíveis.

Esses parâmetros configuram o processo de vetorização do texto para filtrar palavras muito comuns ou muito raras e definem o modelo *LDA* para descobrir 10 tópicos distintos nos documentos, com resultados reproduzíveis. Além disso, foi definido que, para cada tópico, seriam analisados os resultados de até 10 palavras.

Em seguida, foram extraídos os tópicos separadamente, para os discursos dos governistas e oposicionistas, considerando, também separadamente, os discursos para cada semestre. Assim, foram obtidos 10 tópicos, com 10 palavras cada, por semestre, para os dois grupos. Ao final, foram identificados 80 tópicos e 800 palavras para cada grupo.

A seguir estão, por exemplo, os tópicos extraídos para o semestre 2019-1 do grupo de oposicionistas:

Figura 11 – Tópicos extraídos para o semestre 2019-1 do grupo de oposicionistas



Fonte: Elaboração própria, 2023.

Para facilitar as análises subsequentes, foram geradas tabelas contendo a identificação do semestre, o número do tópico, a palavra e o respectivo peso no tópico. Conforme se demonstrará no capítulo dos resultados, isso possibilitou dois tipos de comparação:

- a semelhança de tópicos ao longo do tempo, separadamente para cada grupo;
- as diferenças gerais entre discursos de governistas e de oposicionistas.

Para que a primeira comparação fosse possível, foi necessária a utilização de *word embeddings*, que, conforme se viu no referencial teórico, é um método que possibilita a representação vetorial de palavras, capturando seu significado e contexto em um espaço multidimensional. Cada palavra é mapeada para um vetor em um espaço, o que possibilita que palavras semanticamente similares estejam localizadas próximas umas das outras.

No caso em questão, cada palavra do tópico foi convertida em seu vetor correspondente. Nessa etapa, os pesos das palavras entram em jogo: ao ponderar os vetores de *word embeddings* com os pesos das palavras no tópico, foi possível criar um vetor de tópico que reflete mais precisamente a composição semântica do tópico. O processo envolveu a multiplicação do vetor de *embedding* de cada palavra pelo seu peso no tópico e, em seguida, calculou-se a média desses vetores ponderados. Isso resultou em um único vetor que representa o tópico, que capturou a sua essência matemática em termos de suas palavras-chave.

Cabe destacar que embora a ponderação dos vetores de *word embeddings* com os pesos das palavras obtidos a partir do modelo *LDA* altere a localização espacial dos vetores, ela não distorce arbitrariamente seu significado semântico. Em realidade, ela possibilitou o enriquecendo a representação vetorial dos tópicos. Isso ocorreu porque foram enfatizadas as palavras mais significativas de cada tópico, refletindo assim a relevância contextual dessas palavras e levando a representações mais precisas e contextualizadas. Assim, essa abordagem aprimorou a comparação semântica entre diferentes tópicos, permitindo análises mais profundas, especialmente em contextos temporais ou comparativos.

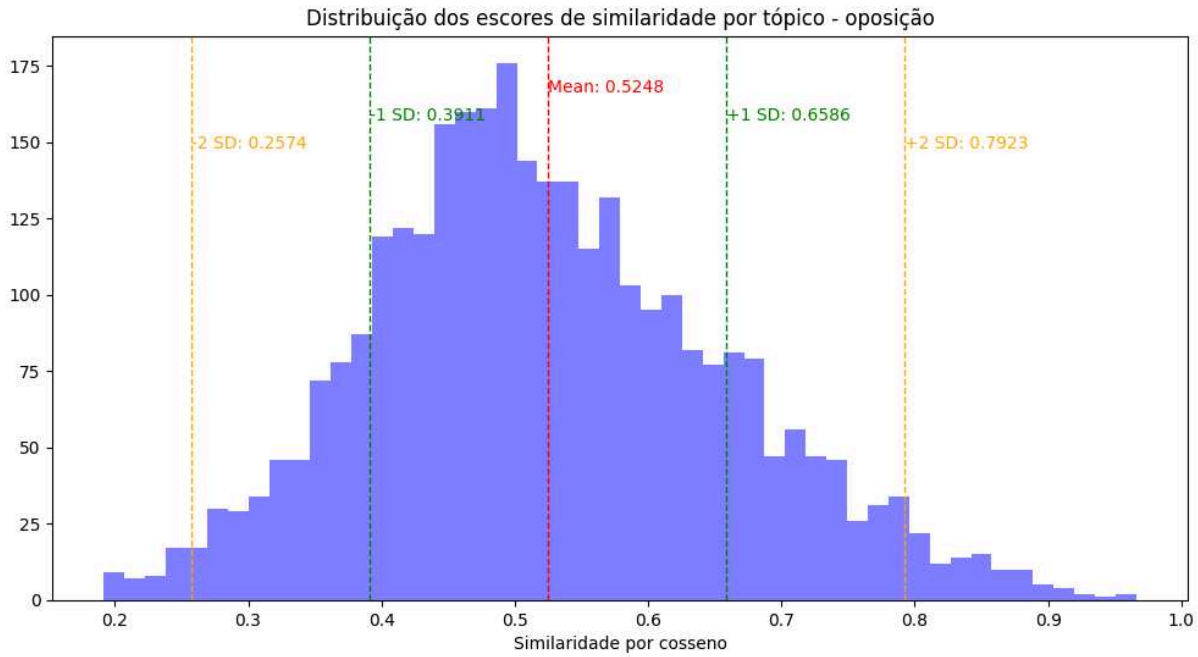
Por fim, a comparação dos tópicos ao longo do tempo foi feita calculando-se a semelhança entre os vetores de tópicos de diferentes semestres, por meio da da similaridade por cosseno, que consegue medir o quão próximos ou distantes estão esses vetores no espaço de *word embeddings*.

Essa abordagem permitiu comparar tópicos mesmo que eles não compartilhem as mesmas palavras. Como os *word embeddings* refletem a semântica das palavras, tópicos com palavras diferentes, mas com significados relacionados, terão vetores de tópicos semelhantes.

A expectativa foi que, ao visualizar essas similaridades ao longo do tempo, seria possível identificar mudanças e padrões na discussão de tópicos, como observar a persistência de certos tópicos, a emergência de novos temas ou a variação na relevância de diferentes assuntos.

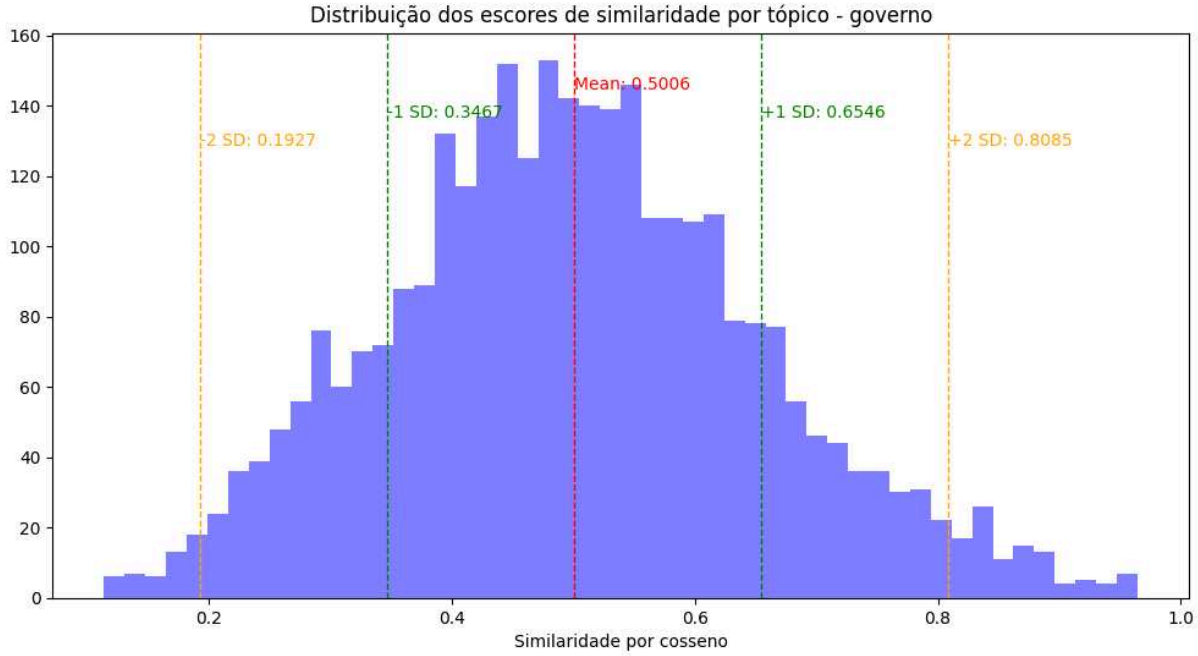
Como critério para definir um alto grau de semelhança entre os tópicos optou-se por definir o valor referente ao segundo desvio-padrão acima da média considerando a distribuição dos valores de similaridade por cosseno entre os tópicos. Essa análise foi realizada separadamente para cada grupo:

Figura 12 – Distribuição dos escores de similaridade por cosseno por tópico - oposição



Fonte: Elaboração própria, 2023.

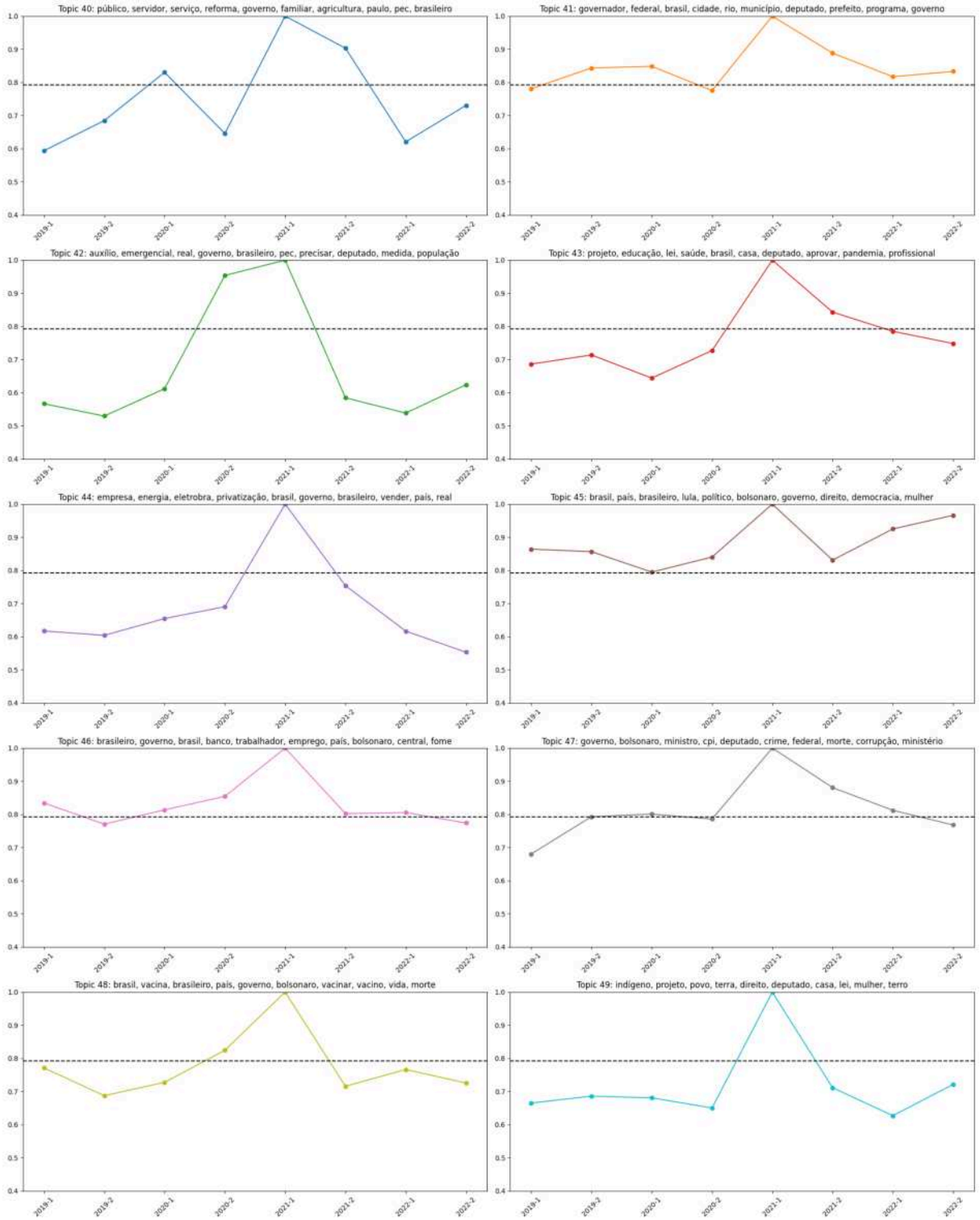
Figura 13 – Distribuição dos escores de similaridade por cosseno por tópico - governo



Fonte: Elaboração própria, 2023.

Com isso, foi possível, objetivamente, estabelecer a persistência ou não de certos tópicos ao longo dos semestres. A seguir, constam os tópicos referentes ao semestre 2021-1 para os discursos oposicionistas:

Figura 14 – Semelhança temporal dos tópicos extraídos do semestre 2021-1 - oposição



Fonte: Elaboração própria, 2023.

Vamos, agora, passar à análise dos resultados referentes às extrações de tópicos realizadas.

6 Resultados

6.1 Análise temporal

Considerando o *cutoff* de similaridade por cosseno estabelecido para os discurso de parlamentares governistas (0,8085), foi possível identificar tópicos com alto nível de semelhança. Com esse método, obteve-se a seguinte tabela que discrimina tópicos para os quais foram encontrados semelhantes em ao menos 6 semestres:

Tabela 7 – Tópicos mais semelhantes - governo

Id	**Semestre**	**Id semestral**	**Principais palavras**	**Qtd semelhantes**	**Qtd semelhantes em diferentes semestres**
46	2020-2	7	brasil, bolsonaro, governo, brasileiro, país, ministro, falar, professor, bom, deputado	8	8
22	2019-2	3	projeto, lei, deputado, casa, comissão, aprovar, votar, federal, público, matéria	8	8
79	2022-1	10	cidade, município, prefeito, deputado, região, vereador, casa, federal, brasil, sul	8	7
66	2021-2	7	município, prefeito, cidade, deputado, vereador, brasil, casa, comunidade, registrar, fazer	6	6
42	2020-2	3	projeto, deputado, lei, votar, casa, precisar, constituição, aprovar, brasil, comissão	7	7
18	2022-2	9	governo, país, brasil, real, bolsonaro, brasileiro, votar, emprego, economia, milhão	5	6
78	2022-1	9	lei, projeto, brasil, comissão, deputado, pec, casa, constituição, aprovar, arthur	5	6
76	2022-1	7	brasil, falar, lula, bolsonaro, dizer, país, deputado, governo, criança, ver	5	6
69	2021-2	10	brasil, bolsonaro, falar, dizer, país, liberdade, brasileiro, cpi, governo, ver	5	6
57	2021-1	8	bolsonaro, brasil, governo, país, brasileiro, falar, deputado, jair, dizer, oposição	5	6
55	2021-1	6	cidade, rio, prefeito, município, região, janeiro, deputado, vereador, fazer, norte	5	6
54	2021-1	5	deputado, projeto, pec, federal, lei, parlamentar, poder, ministro, câmara, votar	6	6
38	2020-1	9	país, brasil, governo, brasileiro, bolsonaro, reforma, precisar, casa, falar, importante	6	6
20	2019-2	1	deputado, falar, país, brasil, família, bolsonaro, dizer, ir, político, vida	5	6
0	2019-1	1	governo, brasil, brasileiro, bolsonaro, país, jair, ministro, banco, milhão, empresa	5	6
10	2022-2	1	prefeito, cidade, município, vereador, população, maranhão, deputado, precisar, recurso, saúde	5	6
4	2019-1	5	projeto, lei, casa, direito, deputado, público, federal, crime, país, trabalhador	6	6

Fonte: Elaboração própria, 2023.

O tópico 46, extraído durante o segundo semestre de 2020, em discursos dos parlamentares governistas, foi considerando altamente semelhante a 8 outros tópicos, presentes em 8 semestres, indicando conter assuntos relevantes para os governistas durante toda a legislatura.

Vejamos quais foram os tópicos mais semelhantes a ele:

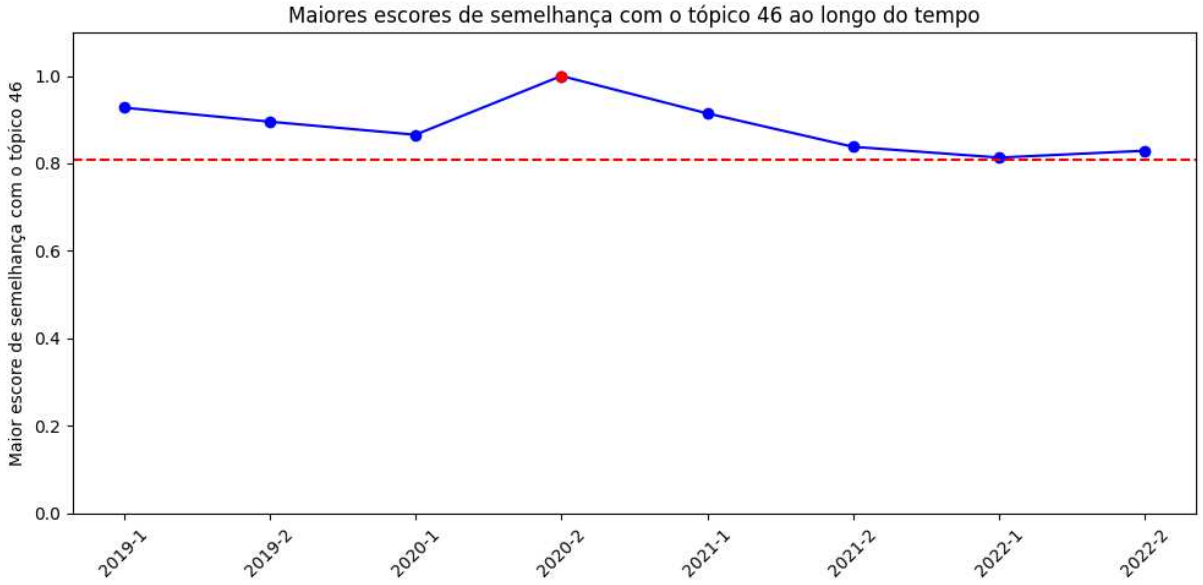
Tabela 8 – Maiores semelhanças com o tópico 46 - governo

Id	**Semestre**	**Id semestral**	**Principais palavras**	**Similaridade por cosseno**
0	2019-1	1	governo, brasil, brasileiro, bolsonaro, país, jair, ministro, banco, milhão, empresa	0.9276
57	2021-1	8	bolsonaro, brasil, governo, país, brasileiro, falar, deputado, jair, dizer, oposição	0.9144
21	2019-2	2	ministro, brasil, brasileiro, governo, parlamentar, país, bolsonaro, federal, frente, justiça	0.8955
38	2020-1	9	país, brasil, governo, brasileiro, bolsonaro, reforma, precisar, casa, falar, importante	0.8655
69	2021-2	10	brasil, bolsonaro, falar, dizer, país, liberdade, brasileiro, cpi, governo, ver	0.8380
2	2019-1	3	brasil, programa, voz, ministro, casa, brasileiro, pronunciamento, divulgar, país, deputado	0.8375
18	2022-2	9	governo, país, brasil, real, bolsonaro, brasileiro, votar, emprego, economia, milhão	0.8290
76	2022-1	7	brasil, falar, lula, bolsonaro, dizer, país, deputado, governo, criança, ver	0.8133

Fonte: Elaboração própria, 2023.

É possível visualizar o quão relevante foi esse tópico ao longo dos semestres:

Figura 15 – Maiores escores de semelhança com o tópico 46 ao longo do tempo - governo



Fonte: Elaboração própria, 2023.

Levando-se em conta a soma dos pesos das palavras presentes nesses 8 tópicos semelhantes ao de nº 46, são mais importantes os termos Brasil (peso 2529,86), Bolsonaro (1624,15), governo (1535,24), país (1477,93) e brasileiro (1320,82). Esses achados revelam uma enfática centralização em termos fortemente associados à identidade nacional e à figura presidencial. O destaque para "Brasil", com o maior peso acumulado, indica um foco predominante em questões ou discursos de âmbito nacional. A significativa presença de "Bolsonaro"reflete uma ênfase particular na figura do presidente, sugerindo que muitas das discussões giram em torno de suas políticas, liderança ou aspectos relacionados à sua administração. Os termos "governo", "país"e "brasileiro"complementam este cenário, indicando um discurso alinhado aos interesses nacionais e à governança de então. Essa tendência nos tópicos de parlamentares governistas sugere uma narrativa focada na promoção ou defesa das ações do governo.

Entre os tópicos mais frequentes, presentes na Tabela 7, e que não constaram entre os mais semelhantes com o tópico 46, temos o tópico 22, com o qual assemelharam-se fortemente os seguintes, também presentes nos 8 semestres:

Tabela 9 – Maiores semelhanças com o tópico 22 - governo

Id	**Semestre**	**Id semestral**	**Principais palavras**	**Similaridade por cosseno**
42	2020-2	3	projeto, deputado, lei, votar, casa, precisar, constituição, aprovar, brasil, comissão	0.9610
19	2022-2	10	projeto, lei, casa, deputado, aprovar, precisar, nacional, comissão, saúde, público	0.9536
78	2022-1	9	lei, projeto, brasil, comissão, deputado, pec, casa, constituição, aprovar, arthur	0.9244
4	2019-1	5	projeto, lei, casa, direito, deputado, público, federal, crime, país, trabalhador	0.8929
54	2021-1	5	deputado, projeto, pec, federal, lei, parlamentar, poder, ministro, câmara, votar	0.8845
70	2022-1	1	lei, projeto, direito, brasil, deputado, médico, pagar, público, muito, precisar	0.8316
33	2020-1	4	projeto, casa, brasil, brasileiro, deputado, lei, governo, federal, importante, apresentar	0.8287
62	2021-2	3	deputado, votar, federal, casa, pec, voto, brasileiro, ministro, justiça, constituição	0.8092

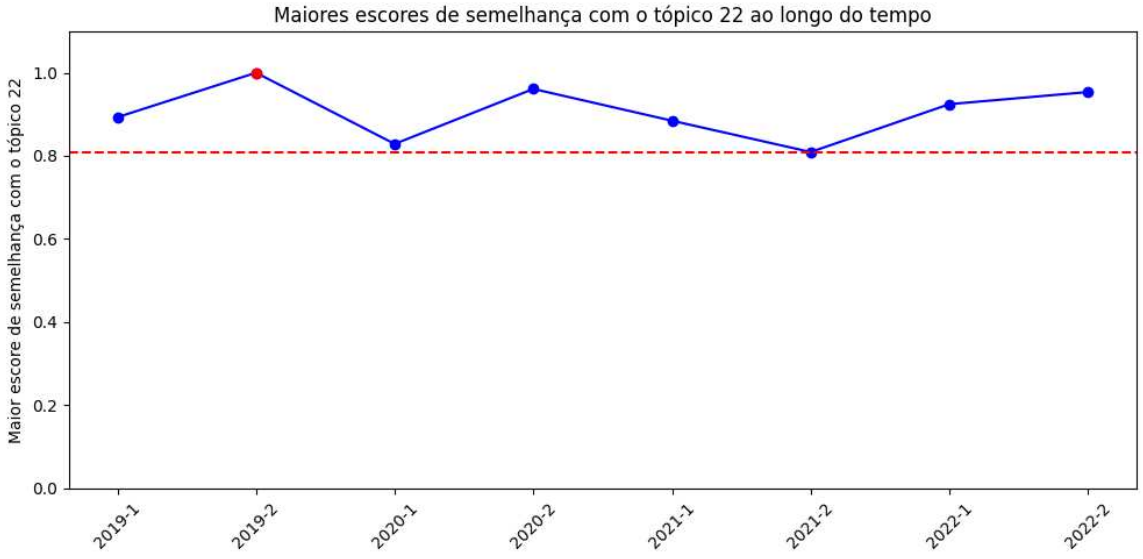
Fonte: Elaboração própria, 2023.

Quando somamos os pesos das palavras presentes nos 8 tópicos semelhantes ao de nº 22, são mais importantes os termos projeto (peso 1946,16), deputado (1886,64), lei (1754,75) e casa (1078,34). É possível perceber aqui uma concentração significativa em temas legislativos e processuais. O termo "projeto", com o maior peso acumulado, ressalta o foco intenso em projetos de lei, refletindo a prioridade dada à formulação e discussão de novas legislações ou emendas. A palavra "deputado"vem logo em seguida em termos de peso, sublinhando o papel ativo dos membros da Câmara dos Deputados nessas

iniciativas legislativas. “Lei”, sendo o terceiro termo mais ponderado, reforça ainda mais esse enfoque legislativo. Interessantemente, “casa”, referindo-se especificamente à Câmara dos Deputados, aparece como um termo chave, indicando que as discussões e deliberações nesse contexto são centrais para os tópicos abordados por esses parlamentares. Esses achados sugerem que, no seio dos discursos governistas, existe uma ênfase considerável na atividade legislativa, no processo de elaboração de leis e nas dinâmicas internas da Câmara dos Deputados.

É possível visualizar o quão relevante foi esse tópico ao longo dos semestres:

Figura 16 – Maiores escores de semelhança com o tópico 22 ao longo do tempo - governo



Fonte: Elaboração própria, 2023.

Agora, tendo em vista o *cutoff* de similaridade por cosseno estabelecido para os discurso de oposicionistas (0,7923), também foi possível identificar tópicos com alto nível de semelhança, considerando a sua presença em ao menos 6 semestres:

Tabela 10 – Tópicos mais semelhantes - oposição

Id	**Semestre**	**Id semestral**	**Principais palavras**	**Qtd semelhantes**	**Qtd semelhantes em diferentes semestres**
61	2022-1	2	brasil, brasileiro, governo, lula, bolsonaro, deputado, país, fome, programa, milhão	13	8
74	2022-2	5	brasil, brasileiro, lula, país, democracia, governo, deputado, eleição, direito, bolsonaro	9	8
65	2022-1	6	brasileiro, país, deputado, brasil, bolsonaro, direito, democracia, político, eleitoral, falar	9	8
39	2020-2	10	bolsonaro, governo, brasil, brasileiro, saúde, país, crime, deputado, população, político	10	8
8	2019-1	9	brasil, país, brasileiro, governo, bolsonaro, ministro, público, empresa, estados, paulo	12	8

Id	**Semestre**	**Id semestral**	**Principais palavras**	**Qtd semelhantes**	**Qtd semelhantes em diferentes semestres**
45	2021-1	6	brasil, país, brasileiro, lula, político, bolsonaro, governo, direito, democracia, mulher	9	8
22	2020-1	3	governo, país, brasil, projeto, brasileiro, real, milhão, bolsonaro, falar, dizer	8	7
19	2019-2	10	bolsonaro, brasil, governo, dizer, brasileiro, ministro, país, paulo, jair, falar	9	7
41	2021-1	2	governador, federal, brasil, cidade, rio, município, deputado, prefeito, programa, governo	6	6
46	2021-1	7	brasileiro, governo, brasil, banco, trabalhador, emprego, país, bolsonaro, central, fome	5	6
14	2019-2	5	município, cidade, deputado, prefeito, região, importante, brasil, bahia, paulo, vereador	5	6
31	2020-2	2	trabalhador, país, governo, brasil, ministro, brasileiro, direito, lula, bolsonaro, ambiente	6	6
56	2021-2	7	governo, bolsonaro, brasil, brasileiro, crime, corrupção, morte, cpi, deputado, saúde	7	6

Fonte: Elaboração própria, 2023.

O tópico 61, extraído durante o primeiro semestre de 2022, em discursos da oposição, foi considerando altamente semelhante a 13 outros tópicos, presentes em todos os 8 semestres, indicando conter assuntos relevantes para a oposição durante toda a legislatura.

Vejamos quais foram os tópicos mais semelhantes a ele:

Tabela 11 – Maiores semelhanças com o tópico 61 - oposição

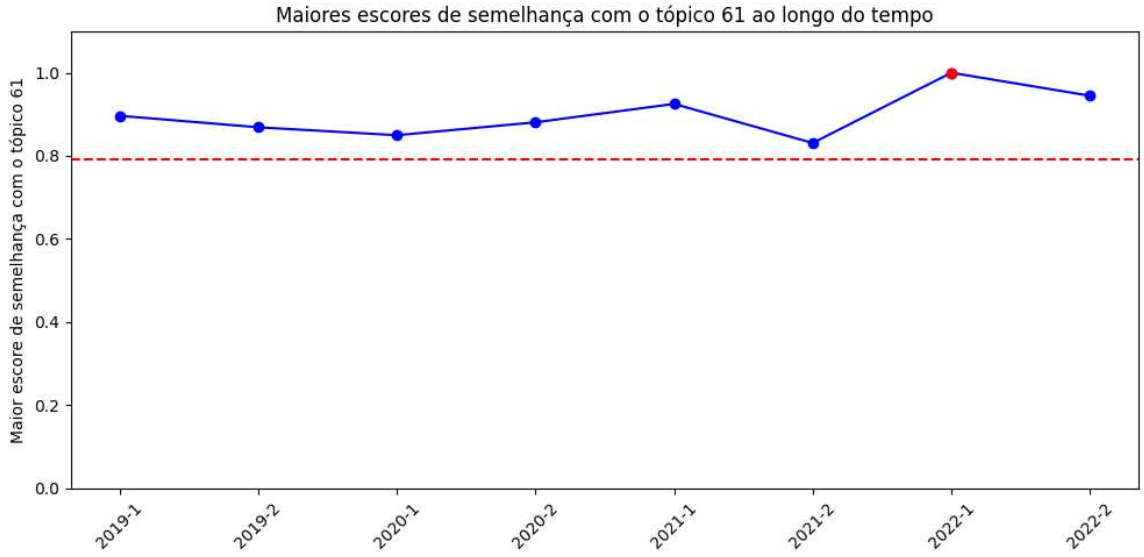
Id	**Semestre**	**Id semestral**	**Principais palavras**	**Similaridade por cosseno**
74	2022-2	5	brasil, brasileiro, lula, país, democracia, governo, deputado, eleição, direito, bolsonaro	0.9446
45	2021-1	6	brasil, país, brasileiro, lula, político, bolsonaro, governo, direito, democracia, mulher	0.9250
8	2019-1	9	brasil, país, brasileiro, governo, bolsonaro, ministro, público, empresa, estados, paulo	0.8965
39	2020-2	10	bolsonaro, governo, brasil, brasileiro, saúde, país, crime, deputado, população, político	0.8806
16	2019-2	7	lula, país, deputado, brasileiro, brasil, polícia, democracia, federal, crime, militar	0.8688
65	2022-1	6	brasileiro, país, deputado, brasil, bolsonaro, direito, democracia, político, eleitoral, falar	0.8628
22	2020-1	3	governo, país, brasil, projeto, brasileiro, real, milhão, bolsonaro, falar, dizer	0.8496
56	2021-2	7	governo, bolsonaro, brasil, brasileiro, crime, corrupção, morte, cpi, deputado, saúde	0.8310
31	2020-2	2	trabalhador, país, governo, brasil, ministro, brasileiro, direito, lula, bolsonaro, ambiente	0.8307
19	2019-2	10	bolsonaro, brasil, governo, dizer, brasileiro, ministro, país, paulo, jair, falar	0.8258

Id	**Semestre**	**Id semestral**	**Principais palavras**	**Similaridade por cosseno**
53	2021-2	4	brasil, governo, brasileiro, bolsonaro, preço, país, real, gás, aumento, combustível	0.8187
54	2021-2	5	família, programa, lula, brasil, governo, auxílio, milhão, brasileiro, fome, bolsonaro	0.8077
46	2021-1	7	brasileiro, governo, brasil, banco, trabalhador, emprego, país, bolsonaro, central, fome	0.8054

Fonte: Elaboração própria, 2023.

É possível visualizar o quão relevante foi esse tópico ao longo dos semestres:

Figura 17 – Maiores escores de semelhança com o tópico 61 ao longo do tempo - oposição



Fonte: Elaboração própria, 2023.

Levando-se em conta a soma dos pesos das palavras presentes nesses 13 tópicos semelhantes ao de nº 61, são mais importantes os termos Brasil (peso 5969,52), brasileiro (5179,13), governo (4920,69), Bolsonaro (4609,06) e país (3643,73). A presença e o peso destas palavras sugere que as discussões da oposição estavam intensamente focadas nas políticas e ações do governo de então, refletindo uma abordagem crítica ou contestatória às decisões e à liderança de Bolsonaro, como normalmente se espera de um papel oposicionista. Este padrão indica que as principais preocupações e debates da oposição giram em torno de questões nacionais, governamentais e da figura do presidente, destacando os desafios, as políticas e as implicações desses aspectos para o Brasil.

Entre os tópicos mais frequentes, presentes na Tabela 10, e que não constaram entre os mais semelhantes com o tópico 61, temos o tópico 41, com o qual assemelharam-se fortemente os seguintes:

Tabela 12 – Maiores semelhanças com o tópico 41 - oposição

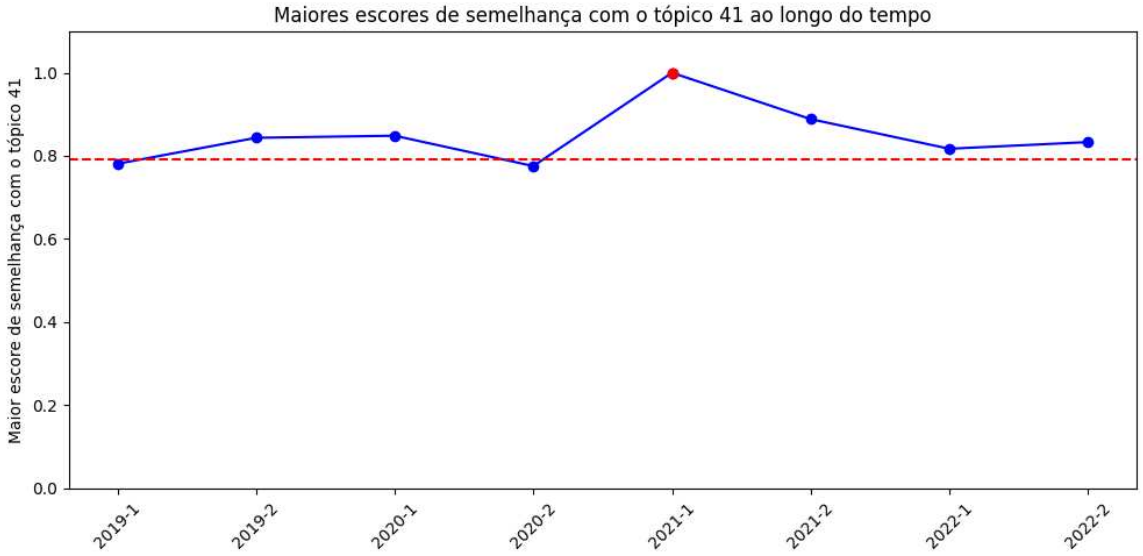
Id	**Semestre**	**Id semestral**	**Principais palavras**	**Similaridade por cosseno**
55	2021-2	6	município, cidade, governo, prefeito, região, rio, agricultura, governador, federal, familiar	0.8884
28	2020-1	9	cidade, paulo, ministro, prefeito, federal, município, companheiro, deputado, guede, ministério	0.8484
14	2019-2	5	município, cidade, deputado, prefeito, região, importante, brasil, bahia, paulo, vereador	0.8434
73	2022-2	4	deputado, federal, brasil, casa, lei, brasileiro, país, governo, esperar, deputar	0.8331
68	2022-1	9	cidade, cultura, município, deputado, lei, governador, paulo, prefeito, região, mina	0.8170
21	2020-1	2	governo, deputado, país, saúde, federal, brasil, brasileiro, público, casa, precisar	0.8079

Fonte: Elaboração própria, 2023.

Quando somamos os pesos das palavras presentes nos 6 tópicos semelhantes ao de nº 41, são mais importantes os termos município (peso 1496,64), cidade (1463,06), deputado (1033,31) e prefeito (978,35). Percebe-se aqui uma clara ênfase em termos relacionados à administração e política local, como evidenciado pela predominância dos termos "município"e "cidade", seguidos de perto por "deputado"e "prefeito". Esses pesos significativos indicam que as discussões nesses tópicos estão fortemente centradas em questões municipais e urbanas, abordando possivelmente aspectos da governança local, desafios enfrentados pelas cidades e municípios, bem como o papel dos deputados e prefeitos nesse contexto.

É possível visualizar o quão relevante foi esse tópico ao longo dos semestres:

Figura 18 – Maiores escores de semelhança com o tópico 41 ao longo do tempo - oposição



Fonte: Elaboração própria, 2023.

A despeito não terem sido identificados discursos acima do *cutoff* para os semestres 2019-1 e 2020-2, pode-se perceber que havia discursos bem próximos ao escore definido.

Cabe justificar aqui que a presença exatamente das mesmas palavras-chave destacadas nos tópicos 46 (governo) e 61 (oposição), ainda que com diferentes pesos, reflete uma dinâmica política intrinsecamente interconectada e uma reação direta um do outro no espectro político. A ênfase em palavras como “Brasil”, “Bolsonaro”, “governo”, “país”e “brasileiro”em ambos os lados indica que tanto governistas quanto oposicionistas estão concentrados em questões nacionais e na liderança do país. Isso é natural em um ambiente político, onde o governo define a agenda e a oposição responde a ela. A figura do presidente e as políticas do governo nacional são, naturalmente, pontos centrais de discussão, seja na defesa ou na crítica.

6.2 Tópicos gerais

Para uma análise mais ampla de temas tratados durante 56ª Legislatura, tanto a partir de discursos de governistas quanto de oposicionistas, optou-se por criar uma tabela que contivesse uma relação de todas as palavras únicas extraídas e, para cada uma delas, as seguintes informações:

- frequência em tópicos de discursos de governistas;
- frequência em tópicos de discursos de oposicionistas;
- diferença das frequências encontradas;
- soma dos pesos em tópicos de discursos de governistas;
- soma dos pesos em tópicos de discursos de oposicionistas;
- diferença das somas dos pesos encontrados.

Essa abordagem, que restou em uma listagem com 249 palavras únicas, permitiu identificar, especialmente, as diferenças entre as escolhas de palavras entre os dois grupos.

Vejamos agora as palavras mais utilizadas pelos governistas e que estão menos presentes entre os oposicionistas, considerando a diferença das somas dos pesos para cada grupo:

Tabela 13 – Maiores diferenças das somas dos pesos de palavras em favor de governistas

Palavra	**Freq. Gov**	**Peso Somado Gov**	**Freq. Opos**	**Peso Somado Opos**	**Diferença de Freq.**	**Diferença de Peso**
falar	19	2576,47	4	533,76	15	2042,71
deputado	51	9757,77	40	7798,96	11	1958,82
policial	6	1265,27	0	0,00	6	1265,27
rio	11	2369,10	7	1192,04	4	1177,06
polícia	7	1376,07	1	215,29	6	1160,77
cidade	14	2920,04	7	1774,57	7	1145,47
janeiro	6	1327,34	3	237,74	3	1089,60
parlamentar	11	1130,74	1	132,50	10	998,24
município	12	2592,95	7	1794,11	5	798,84
militar	6	889,27	1	137,36	5	751,91

Fonte: Elaboração própria, 2023.

Observa-se uma tendência significativa no uso de certas palavras por parte dos governistas em comparação com os oposicionistas. O grande destaque fica por conta da palavra “falar”, que possivelmente se refere a questões de liberdade de expressão pelos parlamentes, assunto amplamente explorado pelos então governistas. Termos como “deputado”, “cidade”e “município”apresentam uma alta frequência e soma de pesos no discurso governista, indicando que temas relacionados à administração local, representação política e aspectos urbanos são importantes nas discussões do governo. Palavras como “policial”, “rio”e “polícia”também surgem com diferenças notáveis, sugerindo uma atenção considerável dos governistas a questões de segurança e ordem pública. A diferença marcante na soma de pesos para a palavra “policial”(1265,27 a favor dos governistas e zero para os oposicionistas) ilustra essa ênfase.

Vejamos agora o outro lado da tabela: as palavras mais utilizadas pelos oposicionistas e que estão menos presentes entre os governistas, considerando, igualmente, a diferença das somas dos pesos entre cada grupo:

Tabela 14 – Maiores diferenças das somas dos pesos de palavras em favor de oposicionistas

Palavra	**Freq. Gov**	**Peso Somado Gov**	**Freq. Opos**	**Peso Somado Opos**	**Diferença de Freq.**	**Diferença de Peso**
governo	21	3842,54	44	13044,08	-23	-9201,54
brasileiro	30	4125,91	44	10391,80	-14	-6265,90
bolsonaro	13	2392,29	26	7881,67	-13	-5489,38
brasil	48	9096,71	51	13364,02	-3	-4267,32
educação	4	847,07	10	3480,56	-6	-2633,49
país	28	4412,79	31	6950,98	-3	-2538,19
lula	2	446,79	10	2869,64	-8	-2422,84
público	18	2043,76	20	4275,66	-2	-2231,90
trabalhador	1	94,94	12	2300,45	-11	-2205,51
real	10	2059,89	14	3907,55	-4	-1847,66

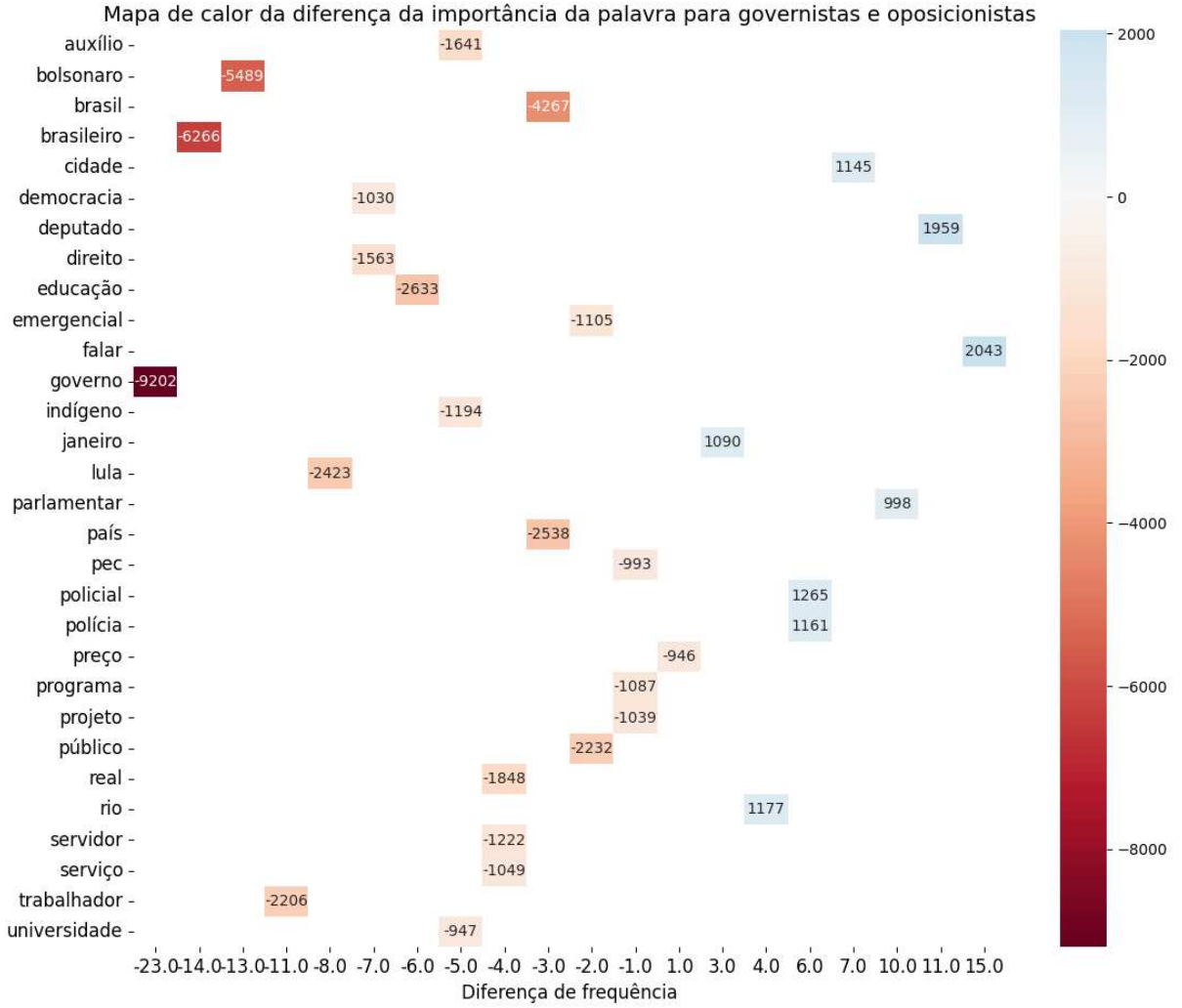
Fonte: Elaboração própria, 2023.

Aqui, observam-se tendências claras nos temas e preocupações dominantes neste

grupo. Palavras como “governo”, “brasileiro”e “Bolsonaro”apresentam diferenças significativas nas somas dos pesos, sugerindo que o foco da oposição está fortemente centrado em questões nacionais, políticas e na figura do presidente. A palavra “governo”, com a maior diferença de peso (-9201,54), indica uma discussão intensa sobre as ações e políticas do governo, enquanto “Bolsonaro”reflete um engajamento profundo com discussões relativas à liderança política do então presidente. Além disso, termos como “educação”e “trabalhador”também se destacam, sugerindo uma ênfase na discussão de questões sociais.

O mapa de calor a seguir permite visualizar as palavras que obtiveram as 30 maiores diferenças de peso (em valores absolutos) entre os grupos:

Figura 19 – Mapa de calor da diferença da importância da palavra para governistas e oposicionistas



Fonte: Elaboração própria, 2023.

A análise dos termos mais significativos na comunicação de governistas e oposicionistas revela padrões interessantes em suas narrativas e prioridades. A palavra “governo”, com uma notável diferença negativa, indica uma discussão mais frequente e crítica por

parte da oposição, o que é esperado, já que a oposição exerce com mais ênfase o papel fiscalizados e crítico com relação às ações do governo. Por outro lado, a presença marcante de "brasileiro"e "Brasil"sugere um discurso voltado à identidade nacional e possivelmente à soberania do país, temas que costumavam ser caros a ambos os lados, mas aqui parecem ser mais enfatizados pela oposição.

O termo "Bolsonaro"segue uma linha semelhante à de "governo", com uma frequência significativamente maior entre os oposicionistas, indicando um foco intenso na figura do chefe do Executivo. Isso pode refletir uma estratégia de personalizar as críticas ao vinculá-las diretamente ao então presidente.

Notavelmente, "educação"e "direito"aparecem com diferenças substanciais, sugerindo que a oposição pode estar mais vocal sobre esses tópicos, talvez como parte de uma agenda de contraposição às políticas governamentais.

Também chama a atenção a utilização da palavra "falar"predominantemente pelos governistas, o que pode indicar uma enfática manifestação de suas perspectivas e políticas. Isso poderia ser associado a uma estratégia de comunicação ativa ou até mesmo a um esforço para manter a narrativa em termos favoráveis. No contexto das discussões sobre liberdade de expressão, isso pode ser visto como um esforço para destacar este valor como um ponto central em sua retórica, reivindicando o direito de comunicar suas políticas e pontos de vista, ou possivelmente como uma resposta às acusações de censura.[1]

A análise dessas palavras-chave possibilitou uma visão mais clara e simplificada dos temas e preocupações que dominaram as discussões políticas na Câmara dos Deputados durante a 56ª Legislatura. Enquanto a oposição se concentrou em questões como educação, direitos e críticas ao governo e ao presidente, os governistas parecem ter focado em afirmar suas posições e destacar questões como a liberdade de expressão e a segurança pública. As diferenças nas palavras usadas refletem estratégias distintas que visam moldar o debate público e influenciar a opinião dos eleitores.

É inevitável abordar, quanto os resultados, a ausência de destaque para palavras diretamente relacionadas à pandemia de COVID-19, como "vacina", "pandemia"e "COVID". Vejamos a frequência e peso dessas e de outras palavras possivelmente relacionadas a essa questão:

[1] Alguns deputados da base do governo chegaram a enfrentar ações no Supremo Tribunal Federal (STF) relacionadas a termos de suas manifestações. O Deputado Daniel Silveira teve o seu mandato cassado em razão do que foi considerado pelo como "estímulo a atos antidemocráticos e ataques a ministros do Supremo e a instituições como o próprio STF." (VIVAS, 2022). Eventos como esse definitivamente tiveram influência nos discursos parlamentares.

Tabela 15 – Frequência e peso das palavras relacionadas à pandemia

Palavra	**Freq. Gov**	**Peso Somado Gov**	**Freq. Opos**	**Peso Somado Opos**	**Diferença de Freq.**	**Diferença de Peso**
saúde	11.0	1930.070742	11.0	2696.667290	0.0	-766.596548
vacina	1.0	125.393496	2.0	664.733761	-1.0	-539.340265
pandemia	1.0	105.988153	3.0	310.929610	-2.0	-204.941457
covid	1.0	115.637633	0.0	0.000000	1.0	115.637633
auxílio	0.0	0.000000	5.0	1641.203069	-5.0	-1641.203069
emergencial	0.0	0.000000	2.0	1105.134338	-2.0	-1105.134338

Fonte: Elaboração própria, 2023.

Enquanto os termos “vacina”, “pandemia”e “covid”não foram tão proeminentes, houve uma discussão significativa sobre o “auxílio emergencial”, especialmente pela oposição.

Esta tendência sugere que os parlamentares, particularmente da oposição, focaram mais nas implicações socioeconômicas da pandemia do que em outros aspectos. O destaque para “auxílio”e “emergencial”reflete uma preocupação com as respostas políticas e as medidas de apoio implementadas para mitigar os impactos da crise sanitária sobre as populações mais vulneráveis.

A diferença na frequência e no peso dessas palavras entre governistas e oposicionistas sugere abordagens distintas na articulação das preocupações relacionadas à pandemia. Enquanto a oposição parece ter usado esses termos para destacar a necessidade de medidas de apoio, o governo pode ter optado por concentrar sua atenção em outras áreas.

Além disso, a aparente ausência de relevância de termos como “vacina”, “pandemia”e “covid”nos discursos pode ser atribuída, em parte, à metodologia de análise de tópicos via *LDA*. Se esses termos foram amplamente usados, mas de maneira mais concentrada por alguns parlamentares e menos distribuída entre outros, o modelo pode não ter capturado esses tópicos como dominantes no conjunto de dados analisado. Isso destacaria um limite do modelo na captura de tópicos que, embora frequentes e relevantes, não são distribuídos uniformemente entre os discursos dos parlamentares.

Foi notável também ausência de destaque para termos frequentemente associados a um discurso mais conservador. Palavras como “família”, “Deus”, “pátria”e “aborto”tiveram pouca ou nenhuma presença nos debates parlamentares:

Tabela 16 – Frequência e peso de palavras associadas à pauta conservadora

Palavra	**Freq. Gov**	**Peso Somado Gov**	**Freq. Opos**	**Peso Somado Opos**	**Diferença de Freq.**	**Diferença de Peso**
família	4.0	484.019872	5.0	1406.784642	-1.0	-922.764771
deus	1.0	138.630266	0.0	0.000000	1.0	138.630266

Fonte: Elaboração própria, 2023.

Estas ausências podem ser interpretadas sob a perspectiva de que, em boa parte do período analisado, as prioridades políticas podem ter sido dominadas por outros temas urgentes, como a resposta à pandemia de COVID-19, questões econômicas e reformas. Em períodos de crise ou mudanças políticas significativas, temas tradicionalmente conservadores podem ter recebido menos atenção em resposta à percepção das necessidades imediatas dos eleitores.

7 Conclusões

O objetivo geral deste trabalho foi identificar, analisar e comparar os tópicos prevalentes nos discursos dos parlamentares governistas e oposicionistas, oferecendo uma visão abrangente da fala parlamentar durante um período historicamente significativo. Através deste estudo, pretendeu-se contribuir para uma compreensão mais abrangente da política brasileira, destacando a interseção entre a linguagem política e ferramentas da ciência de dados e inteligência artificial.

Para que fosse possível atingir esse objetivo, empreendeu-se uma análise detalhada da comunicação parlamentar na 56ª Legislatura, utilizando uma abordagem qualitativa para desvendar os padrões discursivos e as prioridades temáticas de governistas e oposicionistas. Os achados revelam um panorama político em que as narrativas e a retórica são profundamente influenciadas pelas agendas partidárias e pelas dinâmicas do poder legislativo.

A pesquisa envolveu a coleta de dados diretamente das *APIs* da Câmara dos Deputados, reunindo informações detalhadas sobre parlamentares, votações e discursos. Foram obtidos, inicialmente, dados de 626 parlamentares, mais de 53 mil orientações para votações, cerca de 800 mil votos e aproximadamente 68 mil discursos.

Após a comparação dos votos dos parlamentares com as orientações do governo, para se obter um índice de convergência, a análise dos votos revelou uma distribuição bimodal da convergência com o governo, sugerindo uma divisão clara entre governistas e oposicionistas.

Posteriormente, a coleta de discursos focou em tipos específicos de fala que permitiam maior liberdade temática, resultando em um *corpus* relevante para análise. No pré-processamento, os discursos passaram por uma limpeza que incluiu a remoção de elementos não orais e normalização textual, além de retirada de *stopwords* e de um processo de lematização dos termos. Isso refinou os dados para uma melhor modelagem.

Na modelagem, utilizou-se *LDA* para extrair tópicos dos discursos, separadamente para governistas e oposicionistas. Foram observadas as variações semestrais nos tópicos discutidos. A similaridade por cosseno, a partir dos vetores criados via *word embeddings*, permitiu comparar tópicos ao longo do tempo. Os resultados preliminares apontaram para uma utilização mais frequente da tribuna por parte dos oposicionistas e destacaram diferenças nos temas abordados por ambos os grupos.

Do lado dos governistas, percebeu-se uma tendência a enfatizar termos associados ao governo, ao processo legislativo, à segurança pública e à liberdade de expressão. Estes padrões discursivos sugerem uma estratégia de comunicação centrada na afirmação das

políticas e ações do governo, bem como na defesa das práticas e princípios caros ao grupo.

Por outro lado, os oposicionistas concentraram-se em termos que refletem uma postura crítica em relação ao governo e ao presidente, com ênfase também em aspectos como educação, direitos e políticas sociais e também em termos associados à administração e representação local ("prefeito", "município"). A escolha das palavras aponta para uma possível agenda contestatória.

O mapeamento temporal dos tópicos discursivos destacou a persistência de certas questões ao longo do tempo, indicando que tanto governistas quanto oposicionistas mantêm focos consistentes, embora diferenciados. A figura do então presidente Bolsonaro surgiu como um ponto focal comum, catalisando discussões e servindo como eixo para a delimitação das posições políticas. A presença de outras palavras semelhantes em destaque nos tópicos de governo e oposição, ainda que com diferentes pesos, reflete a natureza interconectada da política, onde governistas e oposicionistas se engajam em um diálogo contínuo, muitas vezes focado em torno das mesmas questões-chave, mas abordados de perspectivas distintas.

O estudo evidenciou ainda que, embora a pandemia de COVID-19 tenha sido uma questão central globalmente, nos discursos parlamentares, termos diretamente relacionados à pandemia, como "vacina", "pandemia"e "COVID", não foram tão enfatizados quanto as discussões sobre suas consequências socioeconômicas, especialmente em relação ao "auxílio emergencial". Esta ênfase sugere uma priorização da discussão sobre as respostas políticas e as medidas de apoio, refletindo uma preocupação com os efeitos práticos da pandemia na vida da população.

Outro aspecto interessante é a aparente ausência de temas comumente associados a discursos conservadores nos debates parlamentares. Palavras como "família", "Deus", "pátria"e "aborto"tiveram pouca ou nenhuma relevância destacada nos discursos analisados. Isso pode indicar que, durante o período em questão, os parlamentares focaram mais em responder às questões imediatas e pragmáticas relacionadas à crise da pandemia e às suas implicações econômicas e sociais, relegando para um segundo plano os temas tradicionalmente abordados por conservadores. Essa mudança nos temas de destaque nos discursos parlamentares pode refletir uma adaptação às necessidades e preocupações dos eleitores em um momento de crise, evidenciando a natureza dinâmica da política e do discurso político.

Este trabalho buscou oferecer uma contribuição ao estudo da dinâmica parlamentar, abordando a identificação objetiva de parlamentares governistas e oposicionistas sem se ater a alinhamentos partidários predefinidos. Através de uma análise meticulosa das orientações de votos e dos votos efetivamente manifestados, estabeleceu-se uma metodologia baseada em dados que captura o comportamento real dos deputados no espectro governismo-oposicionismo. Esta abordagem buscou superar limitações dos métodos con-

vencionais que frequentemente se apoiam em suposições ou classificações partidárias que podem não refletir a complexidade das dinâmicas políticas atuais.

Além disso, o estudo avançou na comparação da evolução de tópicos ao longo do tempo com a aplicação de *word embeddings*, uma estratégia que não foi explorada nos trabalhos correlatos consultados. Esta técnica permitiu uma análise mais aprofundada e matizada dos temas discutidos, considerando não apenas a presença de palavras idênticas, mas também a proximidade semântica entre os termos usados nos discursos. Esse método capta a essência dos tópicos discutidos e oferece *insights* importantes sobre as mudanças temáticas, a persistência de certas discussões e a emergência de novas pautas no decorrer da legislatura.

Em suma, esta pesquisa contribuiu para a literatura sobre análise de discursos políticos, demonstrando como as preferências e estratégias de comunicação dos parlamentares refletem e reforçam as linhas de divisão ideológica. Através da mineração e análise dos dados coletados, foram oferecidas percepções importantes sobre como os representantes eleitos articulam e expressam os interesses dos cidadãos no plenário.

No caminho para aprimorar e expandir esta pesquisa, diversos desafios se apresentam, assim como oportunidades promissoras. Uma das direções possíveis envolve expandir a base de dados para incluir outras legislaturas ou esferas legislativas, permitindo uma análise comparativa mais abrangente e possibilitando identificar padrões e tendências a longo prazo na comunicação política. Outro desafio importante é aprimorar a metodologia de análise de tópicos, incorporando modelos mais avançados de processamento de linguagem natural, que possam capturar melhor a complexidade e as sutilezas dos discursos políticos. Estes avanços metodológicos poderiam oferecer percepções mais precisas e detalhadas sobre as estratégias de comunicação e as prioridades dos parlamentares.

Por fim, espera-se que este trabalho possa enriquecer a compreensão da política brasileira, e, ao sugerir abordagens inovadoras para o estudo da linguagem e do comportamento político, também abrir caminho para futuras pesquisas e avanços nesse campo.

Referências

BACK, H.; DEBUS, M.; FERNANDES, J. M. *The politics of legislative debates*. [S.l.]: Oxford University Press, 2021. Citado na página 28.

BACK, H.; DEBUS, M.; FERNANDES, J. M. *The politics of legislative debates: an introduction*. [S.l.]: Oxford University Press, 2021. Citado na página 28.

BADDINI, B.; FERNANDES, D. *Primeira pessoa é vacinada contra Covid-19 no Brasil*. 2021. Acessado em 15 de setembro de 2023. Disponível em: <https://www.cnnbrasil.com.br/nacional/primeira-pessoa-e-vacinada-contra-covid-19-no-brasil/>. Citado na página 23.

BLEI, D. M. *Probabilistic topic models*. 2016. Acessado em 12 de setembro de 2023. Disponível em: <https://www.cs.columbia.edu/~blei/papers/Blei2012.pdf>. Citado na página 31.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. *Latent Dirichlet allocation*. 2003. Acessado em 27 de agosto de 2023. Disponível em: <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>. Citado 4 vezes nas páginas 20, 25, 30 e 31.

BRASIL. *Constituição da República Federativa do Brasil de 1988*. 2023. Acessado em 27 de novembro de 2023. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Citado na página 21.

BRASIL. *Dados abertos*. 2023. Acessado em 27 de novembro de 2023. Disponível em: <https://dadosabertos.camara.leg.br/swagger/api.html>. Citado 2 vezes nas páginas 19 e 33.

BRASIL. *Diários da Câmara dos Deputados*. 2023. Acessado em 27 de novembro de 2023. Disponível em: <https://imagem.camara.leg.br/diarios.asp>. Citado na página 19.

BRONZE, G. *Brasil chega à marca de 100 mil mortes por Covid-19*. 2020. Acessado em 15 de setembro de 2023. Disponível em: <https://www.cnnbrasil.com.br/saude/brasil-registra-100-mil-mortes-por-covid-19-mostra-levantamento-da-cnn/>. Citado na página 23.

DOMO. *Data never sleeps*. 2023. Acessado em 27 de novembro de 2023. Disponível em: <https://web-assets.domo.com/blog/wp-content/uploads/2013/07/DataNeverSleeps.jpeg>. Citado na página 19.

GARCIA, M. *Primeira pessoa é vacinada contra Covid-19 no Brasil*. 2021. Acessado em 15 de setembro de 2023. Disponível em: <https://g1.globo.com/bemestar/vacina/noticia/2021/07/30/brasil-chega-aos-100-milhoes-de-vacinados-com-1a-dose-especialistas-alertam-sobre=a-importancia-das-duas-doses.ghtml>. Citado na página 24.

GOPLERUD, M. *Methods for analyzing parliamentary debates*. [S.l.]: Oxford University Press, 2021. Citado 2 vezes nas páginas 22 e 29.

HARTMANN, N. S. *Repositório de Word Embeddings do NILC*. 2017. Acessado em 27 de maio de 2023. Disponível em: <http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc>. Citado na página 26.

LEARN, S. *sklearn.decomposition.LatentDirichletAllocation*. 2023. Acessado em 12 de outubro de 2023. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html>. Citado na página 46.

LEARN, S. *sklearn.feature_extraction.text.CountVectorizer*. 2023. Acessado em 12 de outubro de 2023. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html>. Citado na página 46.

MIKOLOV, T. et al. *Efficient estimation of word representations in vector space*. 2013. Acessado em 12 de outubro de 2023. Disponível em: <https://arxiv.org/pdf/1301.3781.pdf>. Citado 3 vezes nas páginas 26, 31 e 32.

MOREIRA, D. C. *Com a palavra os nobres deputados: frequência e ênfase temática dos discursos dos parlamentares brasileiros*. 2016. Acessado em 5 de julho de 2023. Disponível em: <https://www.teses.usp.br/teses/disponiveis/8/8131/tde-04112016-124733/publico/2016_DaviCordeiroMoreira_VOrig.pdf>. Citado 3 vezes nas páginas 29, 36 e 37.

SAID, F. *Com CPI da Covid, TV Senado registra recorde nos índices de audiência*. 2023. Acessado em 15 de setembro de 2023. Disponível em: <https://www.metropoles.com/brasil/politica-brasil/com-cpi-da-covid-tv-senado-registra-recorde-nos-indices-de-audiencia>. Citado na página 22.

SCHWARTZ, F. P. *Análise do discurso parlamentar por meio da técnica de processamento de linguagem natural: abordagem estatística e aprendizagem de máquina*. 2018. Acessado em 8 de agosto de 2023. Disponível em: <https://bd.camara.leg.br/bd/handle/bdcamara/35719>. Citado na página 29.

SENADO, A. *Sob pandemia, brasileiros vão às urnas neste domingo*. 2020. Acessado em 15 de setembro de 2023. Disponível em: <https://www12.senado.leg.br/noticias/materias/2020/11/10/sob-pandemia-brasileiros-vao-as-urnas-neste-domingo>. Citado na página 23.

SENADO, T. *CPMI do 8 de janeiro ouve Mauro Cid, ex-ajudante de ordens de Jair Bolsonaro*. 2023. Acessado em 15 de setembro de 2023. Disponível em: <https://www.youtube.com/watch?v=jPWEcRIH2cM&list=PLLLnytnGoqiYN5mRtA6KQQFDfecNrLlJ5&index=2&t=32220s>. Citado na página 22.

SE/UNA-SUS, A. *Coronavírus: Brasil confirma primeiro caso da doença*. 2020. Acessado em 15 de setembro de 2023. Disponível em: <https://www.unasus.gov.br/noticia/coronavirus-brasil-confirma-primeiro-caso-da-doenca>. Citado na página 23.

TEAM, G. L. *Top 30 python libraries to know in 2024*. 2023. Acessado em 27 de novembro de 2023. Disponível em: <https://www.mygreatlearning.com/blog/open-source-python-libraries/>. Citado na página 19.

TITO, F. *Primeira pessoa é vacinada contra Covid-19 no Brasil*. 2021. Acessado em 15 de setembro de 2023. Disponível em: <https://g1.globo.com/bemestar/coronavirus/noticia/2021/06/19/brasil-chega-a-marca-de-500-mil-mortes-por-covid.ghtml>. Citado na página 23.

VIVAS, M. F. e F. *STF condena Daniel Silveira à perda do mandato e dos direitos políticos e a 8 anos e 9 meses de prisão*. 2022. Acessado em 31 de novembro de 2023. Disponível em: <https://g1.globo.com/politica/noticia/2022/04/20/julgamento-stf-daniel-silveira.ghtml>. Citado na página 62.

Anexos

ANEXO A – Principais algoritmos utilizados

A.1 Requisição e tratamento dos votos

```
import pandas as pd

def read_csv_from_url(year):
    url = f"http://dadosabertos.camara.leg.br/arquivos/votacoesOrientacoes/
    csv/votacoesOrientacoes-{year}.csv"
    return pd.read_csv(url, encoding='utf-8', sep=";")

df_list = [read_csv_from_url(year) for year in range(2019, 2023)]
df_orientacoes = pd.concat(df_list, ignore_index=True)

df_plen = df_orientacoes.loc[df_orientacoes['siglaOrgao'] == 'PLEN']

df_plen['siglaBancada'] = df_plen['siglaBancada'].replace('GOV.', 'Governo')
df_plen_gov = df_plen.loc[df_plen['siglaBancada'] == 'Governo']
df_plen_gov = df_plen_gov.loc[df_plen_gov['orientacao'] != 'Liberado']
df_plen_gov = df_plen_gov[["idVotacao", "orientacao"]]

df_votos_filtered = df_votos[df_votos['idVotacao'].isin(df_plen_gov
['idVotacao'])]
df_votos_filtered = df_votos_filtered[["idVotacao", "dataHoraVoto", "voto",
"deputado_id", "deputado_nome", "deputado_siglaPartido", "deputado_siglaUf",
"deputado_urlFoto"]]
df_votos_filtered['dataHoraVoto'] = pd.to_datetime(df_votos_filtered
['dataHoraVoto'])
def get_semester(date):
    if date.month >= 2 and date.month <= 7:
        return 1
    else:
        return 2
df_votos_filtered['Year-Semester'] = df_votos_filtered['dataHoraVoto'].
apply(lambda x: f"{x.year}-{get_semester(x)}")

df_votos_filtered = df_votos_filtered.loc[df_votos_filtered['voto']
```

```
!= 'Obstrução']
df_votos_filtered = df_votos_filtered.loc[df_votos_filtered['voto']
!= 'Artigo 17']
df_votos_filtered = df_votos_filtered.loc[df_votos_filtered['voto']
!= 'Abstenção']

df_merged = pd.merge(df_votos_filtered, df_plen_gov, on='idVotacao',
how='left')
qtd_votos = df_merged.groupby('deputado_id').size().reset_index
(name='qtd_votos')
df_merged['voto_match_gov'] = df_merged['voto'] == df_merged
['orientacao']
qtd_votos_gov = df_merged[df_merged['voto_match_gov']].groupby
('deputado_id').size().reset_index(name='qtd_votos_gov')
final_df = pd.merge(qtd_votos, qtd_votos_gov, on='deputado_id',
how='left')
final_df['qtd_votos_gov'] = final_df['qtd_votos_gov'].fillna(0)
final_df['convergence_gov'] = final_df['qtd_votos_gov'] /
final_df['qtd_votos']
final_df = final_df.merge(df_votos_filtered[['deputado_id',
'deputado_nome']].drop_duplicates(), on='deputado_id')
final_df = final_df[['deputado_id', 'deputado_nome', 'qtd_votos',
'qtd_votos_gov', 'convergence_gov']]
final_df_sorted = final_df.sort_values(by='convergence_gov',
ascending=False)

# Plotando o histograma para 'convergence_gov'
plt.figure(figsize=(10, 6))
sns.histplot(final_df['convergence_gov'], bins=100, color='blue',
kde=True)

# Título com espaço adicional
plt.title('Distribuição da convergência com as orientações do
governo para toda a legislatura', pad=20)

# Definindo os rótulos dos eixos
plt.xlabel('Convergência')
plt.ylabel('')
```

```
# Adicionando a linha vertical vermelha tracejada com a média
mean_value = final_df['convergence_gov'].mean()
plt.axvline(mean_value, color='red', linestyle='--')

# Adicionando o valor da média acima da linha vermelha
plt.text(mean_value, plt.ylim()[1], f'Média: {mean_value:.3f}',
color='red',
         ha='center', va='bottom')

# Adicionando a linha vertical alaranjada tracejada em x = 0.5
plt.axvline(0.5, color='orange', linestyle='--')

# Adicionando o valor '0.5' acima da linha alaranjada
plt.text(0.5, plt.ylim()[1], '0.5', color='orange',
         ha='center', va='bottom')

# Exibindo o gráfico
plt.tight_layout()
plt.show()

gap_threshold = 0.5  # Replace this with the actual value you determined
final_df['status'] = final_df['convergence_gov'].apply(lambda x:
'governo' if x > gap_threshold else 'oposicao')
```

A.2 Requisição e tratamento dos discursos

```
import requests
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import unicodedata
import spacy
import string
import re
import warnings

response = requests.get('https://dadosabertos.camara.leg.br/api/v2/
deputados?idLegislatura=56&ordem=ASC&ordenarPor=nome').json()
```

```
df_deputados = pd.json_normalize(response['dados'])

def get_speeches(deputado_id):
    url = f'https://dadosabertos.camara.leg.br/api/v2/deputados/
    {deputado_id}/discursos'
    params = {
        'dataInicio': '2019-02-01',
        'dataFim': '2023-01-31',
        'ordenarPor': 'dataHoraInicio',
        'ordem': 'ASC',
        'itens': 100,
    }

    df_list = []

    while True:
        response = requests.get(url, params=params)

        if response.status_code == 200:
            data = response.json()
            df_list.append(pd.json_normalize(data["dados"]))

            if len(data["dados"]) < 100:
                break
            else:
                params["pagina"] = params.get("pagina", 1) + 1
        else:
            print(f'Request failed with status code
            {response.status_code}')
            break

    df = pd.concat(df_list, ignore_index=True)

    df.insert(loc=df.columns.get_loc('transcricao')+1,
    column='contagemPalavras', value=df['transcricao'].
    apply(lambda x: len(x.split())))

    return df
```

```
import time

def process_speeches(df_deputados):
    speeches_list = []
    status_list = []
    speeches_count_list = []
    total_words_count_list = []

    for i in range(len(df_deputados)):
        deputado_id = df_deputados['id'][i]
        deputado_nome = df_deputados['nome'][i]

        retries = 0
        status = 'Processed'
        speeches_count = 0
        total_words_count = 0

        while retries < 1:
            try:
                speech_df = get_speeches(deputado_id)
                speech_df['deputado_id'] = deputado_id
                speech_df['deputado_nome'] = deputado_nome
                speeches_list.append(speech_df)
                print(f'Finished processing speeches for {deputado_nome}
                ({deputado_id})')

                speeches_count = len(speech_df)
                total_words_count = speech_df['contagemPalavras'].sum()

                # Add a 1-second delay
                time.sleep(0.5)

                break  # Exit the retry loop if successful

            except Exception as e:
                retries += 1
                status = 'Error'
                print(f'Error processing speeches for {deputado_nome}
```

```
                ({deputado_id}): {str(e)}')
                if retries < 2:
                    print('Retrying...')
                    time.sleep(2)
                    continue
                else:
                    print('Max retries reached. Moving on to the next
                    deputado.')

        status_list.append(status)
        speeches_count_list.append(speeches_count)
        total_words_count_list.append(total_words_count)

    df_speeches = pd.concat(speeches_list, ignore_index=True)
    df_status = pd.DataFrame({
        'deputado_id': df_deputados['id'],
        'deputado_name': df_deputados['nome'],
        'siglaPartido': df_deputados['siglaPartido'],
        'siglaUf': df_deputados['siglaUf'],
        'status': status_list,
        'speeches_count': speeches_count_list,
        'total_words_count': total_words_count_list
    })

    return df_speeches, df_status

df_speeches, df_status = process_speeches(df_deputados)

types_of_speech = ["BREVES COMUNICAÇÕES", "PEQUENO EXPEDIENTE",
"COMUNICAÇÃO PARLAMENTAR", "GRANDE EXPEDIENTE"]
df_speeches_filtered = df_speeches[df_speeches['tipoDiscurso'].
isin(types_of_speech)]
df_speeches_filtered = df_speeches_filtered[['tipoDiscurso',
'dataHoraInicio', 'keywords', 'sumario', 'transcricao',
                                            'deputado_id',
                                            'deputado_nome']]
df_speeches_filtered.dataHoraInicio = pd.to_datetime(df_speeches_filtered.
dataHoraInicio)
```

```
df_speeches_filtered['transcricao_limpa'] = df_speeches_filtered
['transcricao'].apply(lambda x: unicodedata.normalize("NFC", x).
replace("\x97", "").replace("\r\n", " "))

remove_phrases = r'(DISCURSO NA ÍNTEGRA ENCAMINHADO|DISCURSOS NA
ÍNTEGRA ENCAMINHADOS|DOCUMENTOS ENCAMINHADOS PELO|DOCUMENTOS
ENCAMINHADOS PELA|DOCUMENTO ENCAMINHADO PELO|DOCUMENTO ENCAMINHADO
PELA)[\s\S]*'

df_speeches_filtered['transcricao_limpa'] = df_speeches_filtered
['transcricao_limpa'].str.replace(remove_phrases,
'', regex=True)

remove_phrases2 = r'(O SR\.|A SRA\.) .+?\) -\s*'

df_speeches_filtered['transcricao_limpa'] = df_speeches_filtered
['transcricao_limpa'].str.replace(remove_phrases2,
'', regex=True)

remove_phrases3 = r'(?i)(o sr\.|a sra\.) ([^\s]+ ){1,5}-'

df_speeches_filtered['transcricao_limpa'] = df_speeches_filtered
['transcricao_limpa'].str.replace(remove_phrases3,
'', regex=True)

df_speeches_filtered['transcricao_limpa'] = df_speeches_filtered
    ['transcricao_limpa'].str.replace(r'\(.*?\)', '', regex=True).
    str.strip()

df_speeches_filtered['transcricao_limpa'] = df_speeches_filtered
['transcricao_limpa'].str.replace('[\'"]', '',
regex=True)

df_speeches_filtered['transcricao_limpa'] = df_speeches_filtered
['transcricao_limpa'].str.replace(r'\[\[.*?\]\]', '', regex=True)

df_speeches_filtered['contagemPalavras'] = df_speeches_filtered
['transcricao'].apply(lambda x: len(str(x).split()))
```

```
df_speeches_filtered['contagemPalavras_limpa'] = df_speeches_filtered
['transcricao_limpa'].apply(lambda x: len(str(x).split()))

df_speeches_filtered = df_speeches_filtered.merge(df_deputados[['id',
'siglaPartido', 'siglaUf']],
                                                  left_on='deputado_id',
                                                  right_on='id',
                                                  how='left').
                                                  drop(columns='id')

with open('../organizacao/arquivos/stopwords-pt.txt', 'r', encoding=
'utf-8') as file:
    stopwords = file.readlines()
stopwords = [word.strip() for word in stopwords]

nlp = spacy.load('pt_core_news_sm')

def preprocess(text):
    # Replace punctuation with a space
    text = re.sub(f'[{string.punctuation}]', ' ', text)
    # Convert to lowercase and remove stopwords
    text = ' '.join([word for word in text.lower().split() if
    word not in stopwords])
    # Remove numbers
    text = re.sub(r'\d+', '', text)
    # Remove specific special characters
    text = re.sub(r'[ªº]', '', text)
    # Normalize whitespace
    text = ' '.join(text.split())
    # Process the text and lemmatize
    doc = nlp(text)
    text = ' '.join([token.lemma_ for token in doc if token.text
    not in stopwords])
    return text

df_speeches_filtered['transcricao_modelo'] = df_speeches_filtered
['transcricao_limpa'].apply(preprocess)
```

A.3 Modelagem

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib.cm import get_cmap
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
import numpy as np
from sklearn.metrics.pairwise import cosine_similarity

df_speeches = pd.merge(df_speeches, df_convergence[['deputado_id',
'status']], on='deputado_id', how='left')
df_speeches['Year-Semester'] = df_speeches['Year'].astype(str) + '-'
+ df_speeches['Semester'].astype(str)

df_speeches_governo = df_speeches.loc[df_speeches['status'] ==
'governo']
df_speeches_oposicao = df_speeches.loc[df_speeches['status'] ==
'oposicao']

vectorizer = CountVectorizer(max_df=0.90, min_df=0.05,
max_features=1000)

lda_model = LatentDirichletAllocation(n_components=10,
random_state=42)

def plot_top_words(model, feature_names, n_words, title):
    fig, axes = plt.subplots(2, 5, figsize=(15, 10))
    axes = axes.flatten()  # Flattens the 2x5 grid into a
    1D array for easy indexing

    for topic_idx, topic in enumerate(model.components_):
        top_features = topic.argsort()[:-n_words - 1:-1]
        top_words = [feature_names[i] for i in top_features]
        weights = topic[top_features]

        axes[topic_idx].barh(top_words, weights, align='center')
        axes[topic_idx].set_title(f"Topic #{topic_idx + 1}")
```

```
            axes[topic_idx].invert_yaxis()

        plt.suptitle(title)  # Adiciona um título para o conjunto
        de gráficos
        plt.tight_layout()  # Adjusts spacing between plots for
        better layout
        plt.show()

    def process_and_plot_for_year_semester(df_year_semester, title):
        X = vectorizer.fit_transform(df_year_semester
        ['transcricao_modelo'])
        lda_model = LatentDirichletAllocation(n_components=10,
        random_state=42)
        lda_model.fit(X)
        plot_top_words(lda_model, vectorizer.get_feature_names_out(),
        n_words=10, title=title)

    for year_semester in df_speeches_oposicao['Year-Semester'].
    unique():
        title = f"Oposição {year_semester}"
        print(f"Topics for {title}:")
        df_year_semester = df_speeches_oposicao[df_speeches_oposicao
        ['Year-Semester'] == year_semester].copy()
        process_and_plot_for_year_semester(df_year_semester,
        title=title)

    def extract_topic_info(model, feature_names, n_words=10):
        topic_info = []
        for topic_idx, topic in enumerate(model.components_):
            top_features = topic.argsort()[:-n_words - 1:-1]
            top_words = [feature_names[i] for i in top_features]
            weights = topic[top_features]
            for word, weight in zip(top_words, weights):
                topic_info.append({
                    'Year-Semester': year_semester,
                    'Topic': topic_idx + 1,
                    'Word': word,
                    'Weight': weight
                })
```

```
    return topic_info

all_topics_info = []

for year_semester in df_speeches_oposicao['Year-Semester']
unique():
    df_year_semester = df_speeches_oposicao[df_speeches_oposicao
    ['Year-Semester'] == year_semester].copy()

    # Vectorize and fit LDA model
    X = vectorizer.fit_transform(df_year_semester
    ['transcricao_modelo'])
    lda_model.fit(X)

    # Extract topic info
    topics_info = extract_topic_info(lda_model, vectorizer.
    get_feature_names_out())
    all_topics_info.extend(topics_info)

df_topics_oposicao = pd.DataFrame(all_topics_info)

word_frequency = df_topics_oposicao['Word'].value_counts()

weight_sum = df_topics_oposicao.groupby('Word')['Weight'].sum()

df_most_important_topics_oposicao = pd.DataFrame({
    'word': word_frequency.index,
    'frequency': word_frequency.values,
    'weight_sum': weight_sum[word_frequency.index].values
})

df_most_important_topics_oposicao.sort_values(by="weight_sum",
ascending=False).head(30)

def weighted_topic_vector(top_words, weights, model):
    word_vectors = [model[word] * weight for word, weight in
    zip(top_words, weights) if word in model]
    if len(word_vectors) == 0:
        return np.zeros(model.vector_size)  # Return a zero
```

```
        vector if no words are in the model
    return np.mean(word_vectors, axis=0)

weighted_topic_vectors = []
for year_semester in df_topics_oposicao['Year-Semester'].
unique():
    for topic in df_topics_oposicao[df_topics_oposicao
    ['Year-Semester'] == year_semester]['Topic'].unique():
        topic_data = df_topics_oposicao[(df_topics_oposicao
        ['Year-Semester'] == year_semester) & (df_topics_oposicao
        ['Topic'] == topic)]
        top_words = topic_data['Word'].tolist()
        weights = topic_data['Weight'].tolist()
        vec = weighted_topic_vector(top_words, weights,
        model_skip)
        weighted_topic_vectors.append(vec)

weighted_similarity_matrix_oposicao = cosine_similarity
(weighted_topic_vectors)

upper_triangle = weighted_similarity_matrix_oposicao[np.triu_indices
(weighted_similarity_matrix_oposicao.shape[0], k=1)]

mean = np.mean(upper_triangle)
std = np.std(upper_triangle)

plt.figure(figsize=(12, 6))
plt.hist(upper_triangle, bins=50, color='blue', alpha=0.5)
plt.title('Distribuição dos escores de similaridade por
tópico - oposição')
plt.xlabel('Similaridade por cosseno')
plt.ylabel('')
plt.grid(False)

plt.axvline(mean, color='red', linestyle='dashed',
linewidth=1)
plt.axvline(mean - std, color='green', linestyle='dashed',
linewidth=1)
plt.axvline(mean + std, color='green', linestyle='dashed',
```

```
linewidth=1)
plt.axvline(mean - 2*std, color='orange', linestyle='dashed',
linewidth=1)
plt.axvline(mean + 2*std, color='orange', linestyle='dashed',
linewidth=1)

plt.text(mean, plt.ylim()[1]*0.9, f'Mean: {mean:.4f}',
color='red')
plt.text(mean - std, plt.ylim()[1]*0.85, f'-1 SD: {mean-std:.4f}',
color='green')
plt.text(mean + std, plt.ylim()[1]*0.85, f'+1 SD: {mean+std:.4f}',
color='green')
plt.text(mean - 2*std, plt.ylim()[1]*0.8, f'-2 SD: {mean-2*std:.4f}',
color='orange')
plt.text(mean + 2*std, plt.ylim()[1]*0.8, f'+2 SD: {mean+2*std:.4f}',
color='orange')

topic_summaries = []

for year_semester in df_topics_oposicao['Year-Semester'].unique():
    for topic in df_topics_oposicao[df_topics_oposicao
    ['Year-Semester'] == year_semester]['Topic'].unique():
        # Extract top words for this topic
        top_words = df_topics_oposicao[(df_topics_oposicao
        ['Year-Semester'] == year_semester) & (df_topics_oposicao
        ['Topic'] == topic)]['Word'].tolist()

        # Append to the list
        topic_summaries.append({
            'Year-Semester': year_semester,
            'Topic': topic,
            'Top Words': ', '.join(top_words)
        })

df_topic_summaries_oposicao = pd.DataFrame(topic_summaries)
df_topic_summaries_oposicao

def find_similar_topics(target_topic_index, cutoff=0.7923,
top_n=20):
```

```
    similarities = weighted_similarity_matrix_oposicao
    [target_topic_index]

    above_cutoff_indices = [i for i, sim in enumerate(similarities)
    if sim > cutoff]

    most_similar_indices = np.argsort(similarities
    [above_cutoff_indices])[::-1][:top_n]

    similar_topics = df_topic_summaries_oposicao.iloc
    [above_cutoff_indices].iloc[most_similar_indices]
    similar_topics['Cosine Similarity'] = similarities
    [above_cutoff_indices][most_similar_indices]
    return similar_topics

target_topic_index = 41  # Example index
similar_topics_oposicao = find_similar_topics(target_topic_index)
similar_topics_oposicao.to_csv("semelhantes_especificos_opos.csv")
similar_topics_oposicao

from collections import Counter

def count_word_frequency(df):
    word_freq = Counter()
    for words in df['Top Words']:
        word_freq.update(words.split(', '))
    return word_freq

similar_topics_oposicao = find_similar_topics(target_topic_index=41)
word_frequency = count_word_frequency(similar_topics_oposicao)

df_word_freq = pd.DataFrame(word_frequency.items(), columns=
['Word', 'Frequency']).sort_values(by='Frequency', ascending=False)
df_word_freq.to_csv("word_frequency.csv")
df_word_freq

target_topic_index = 41  # Exemplo de índice
similar_topics_oposicao = find_similar_topics(target_topic_index)
```

```
def sum_word_weights_in_topics(topics_list, all_topics_info):
    filtered_topics_info = all_topics_info[all_topics_info.apply
    (lambda x: (x['Year-Semester'], x['Topic']) in topics_list,
    axis=1)]

    word_weight_sum = Counter()
    for _, row in filtered_topics_info.iterrows():
        word_weight_sum[row['Word']] += row['Weight']

    return word_weight_sum

topics_tuples = [(row['Year-Semester'], row['Topic']) for _,
row in similar_topics_oposicao.iterrows()]

word_weights_sum = sum_word_weights_in_topics(topics_tuples,
df_topics_oposicao)

df_word_weights_sum = pd.DataFrame(word_weights_sum.items(),
columns=['Word', 'Total Weight']).sort_values(by='Total Weight',
ascending=False)
df_word_weights_sum

data = []

for index, row in df_topic_summaries_oposicao.iterrows():
    similarities = weighted_similarity_matrix_oposicao[index]

    tot_similar = sum(similarities > 0.7923) - 1  # Subtract 1
    to exclude the topic itself

    semesters_with_similars = set()
    for idx, similarity in enumerate(similarities):
        if similarity > 0.7923:  # Exclude the topic itself
            semester_of_similar_topic = df_topic_summaries_oposicao.
            iloc[idx]['Year-Semester']
            semesters_with_similars.add(semester_of_similar_topic)
    tot_similar_sem = len(semesters_with_similars)

    data.append({
```

```
            'Year-Semester': row['Year-Semester'],
            'Topic': row['Topic'],
            'Top Words': row['Top Words'],
            'tot_similar': tot_similar,
            'tot_similar_sem': tot_similar_sem
        })

df_topic_similarity_count = pd.DataFrame(data)
df_topic_similarity_count_sorted = df_topic_similarity_count.
sort_values(by='tot_similar_sem', ascending=False)
df_topic_similarity_count_sorted.head(30)  # Display the first
few rows for verification

def plot_similarity_across_semesters(target_topic_index,
weighted_similarity_matrix, df_topic_summaries):
    target_topic_words = df_topic_summaries.iloc[target_topic_index]
    ['Top Words']
    print(f"Target Topic {target_topic_index} Top Words:
    {target_topic_words}")

    highest_similarity_scores = {}

    target_semester = df_topic_summaries.iloc[target_topic_index]
    ['Year-Semester']

    for semester in df_topic_summaries['Year-Semester'].unique():
        semester_topic_indices = df_topic_summaries[df_topic_summaries
        ['Year-Semester'] == semester].index

        highest_similarity = max(weighted_similarity_matrix
        [target_topic_index, semester_topic_indices]) if
        len(semester_topic_indices) > 0 else 0

        highest_similarity_scores[semester] =
        highest_similarity

    semesters = sorted(highest_similarity_scores.keys())
    values = [highest_similarity_scores[semester] for semester
    in semesters]
```

```
    plt.figure(figsize=(10, 5))
    plt.plot(semesters, values, marker='o', color='blue')

    if target_semester in highest_similarity_scores:
        plt.scatter([target_semester], [highest_similarity_scores
        [target_semester]], color='red', zorder=5)

    plt.axhline(y=0.8085, color='r', linestyle='--')
    plt.xlabel('')
    plt.ylabel('Maior escore de semelhança com o tópico ' +
    str(target_topic_index))
    plt.title('Maiores escores de semelhança com o tópico ' +
    str(target_topic_index) + ' ao longo do tempo')
    plt.ylim(0, 1.1)  # Set y-axis range from 0 to 1
    plt.xticks(rotation=45)
    plt.tight_layout()
    plt.show()

target_topic_index = 41  # Replace with the desired index
plot_similarity_across_semesters(target_topic_index,
weighted_similarity_matrix_oposicao, df_topic_summaries_oposicao)

def plot_semester_similarity_detailed(weighted_similarity_matrix,
df_topic_summaries, semester):
    semester_topics = df_topic_summaries[df_topic_summaries
    ['Year-Semester'] == semester]
    num_topics = len(semester_topics)

    fig, axs = plt.subplots(5, 2, figsize=(20, 30))
    color_palette = get_cmap('tab10')

    for i, (index, row) in enumerate(semester_topics.iterrows()):
        ax = axs[i // 2, i % 2]
        target_topic_index = index
        target_topic_words = row['Top Words']
        highest_similarity_scores = {}
        for other_semester in df_topic_summaries['Year-Semester'].
        unique():
```

```
                semester_topic_indices = df_topic_summaries
                [df_topic_summaries['Year-Semester'] == other_semester].
                index
                highest_similarity = max(weighted_similarity_matrix
                [target_topic_index, semester_topic_indices]) if
                len(semester_topic_indices) > 0 else 0
                highest_similarity_scores[other_semester] =
                highest_similarity
            semesters = sorted(highest_similarity_scores.keys())
            values = [highest_similarity_scores[s] for s in
            semesters]
            ax.plot(semesters, values, marker='o', color=
            color_palette(i % 10))
            ax.axhline(y=0.7923, color='black', linestyle='--')
            ax.set_title(f'Tópico {index}: {target_topic_words}')
            ax.set_ylim(0.40, 1)
            ax.tick_params(axis='x', rotation=45)

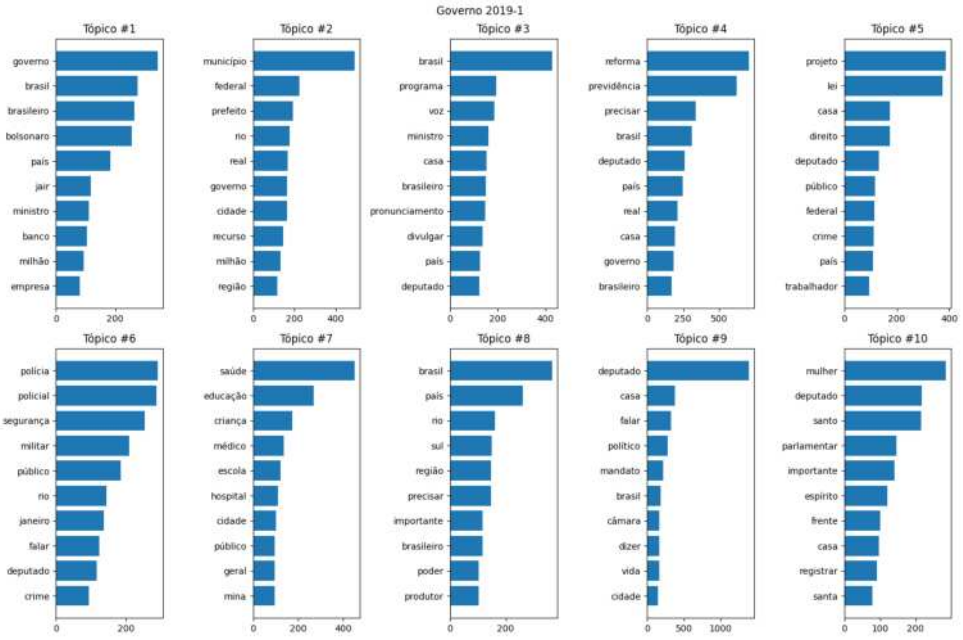
    plt.tight_layout()
    plt.show()

semester = '2022-2'
plot_semester_similarity_detailed(weighted_similarity_matrix_oposicao,
df_topic_summaries_oposicao, semester)
```

ANEXO B – Tópicos detalhados

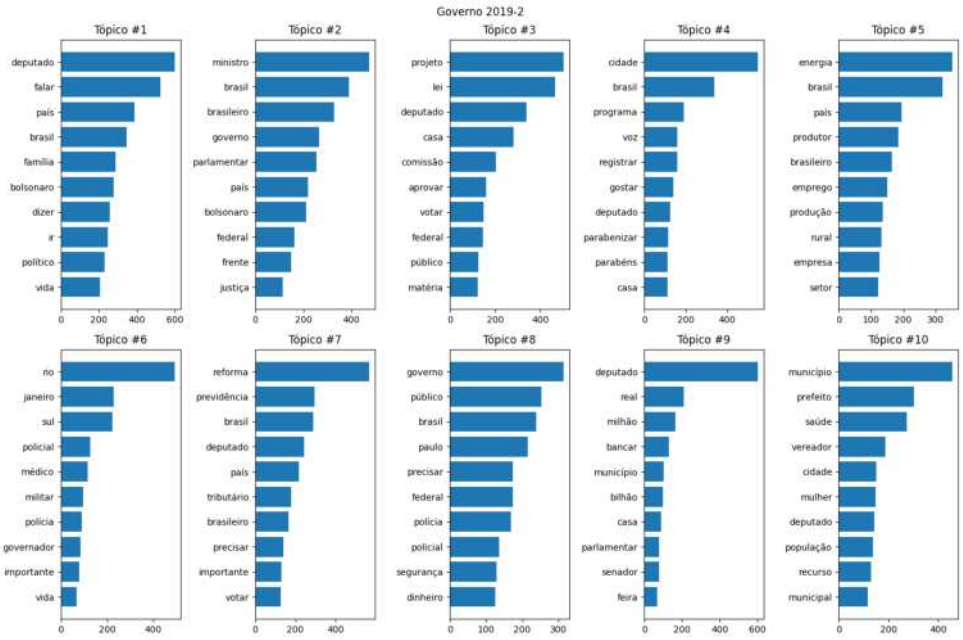
B.1 Governo

Figura 20 – Tópicos extraídos do semestre 2019-1 - governo



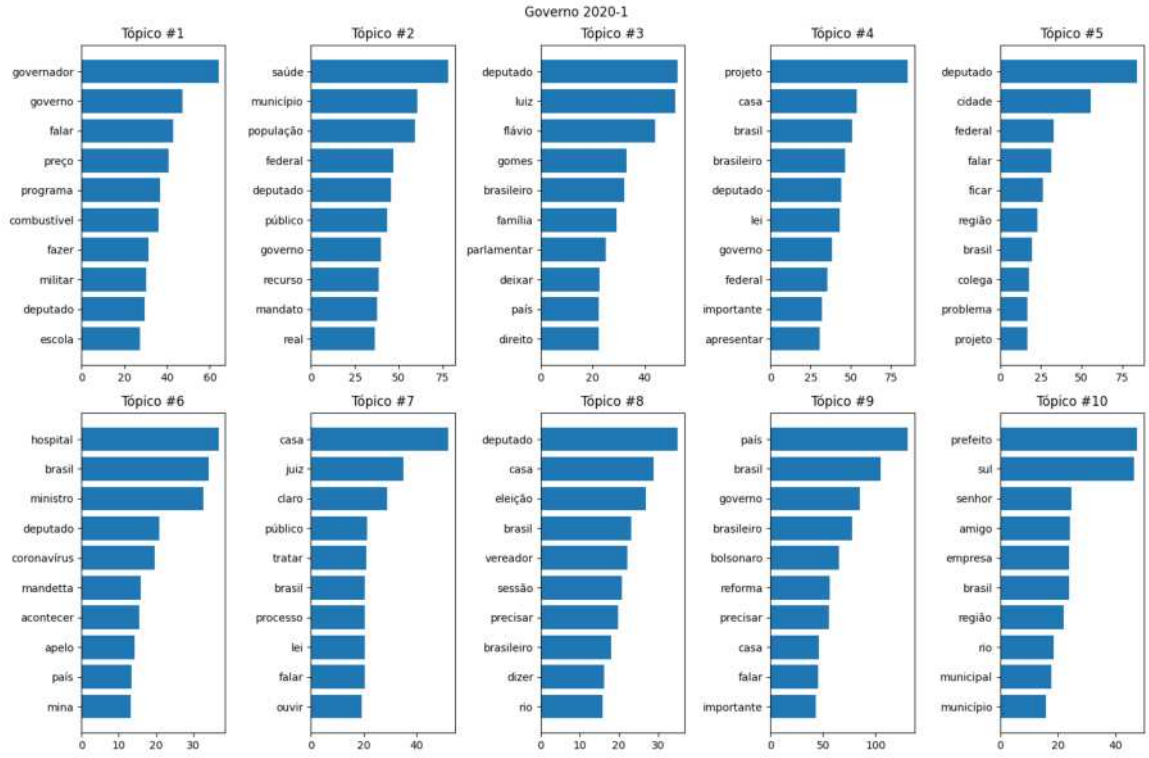
Fonte: Elaboração própria, 2023.

Figura 21 – Tópicos extraídos do semestre 2019-2 - governo



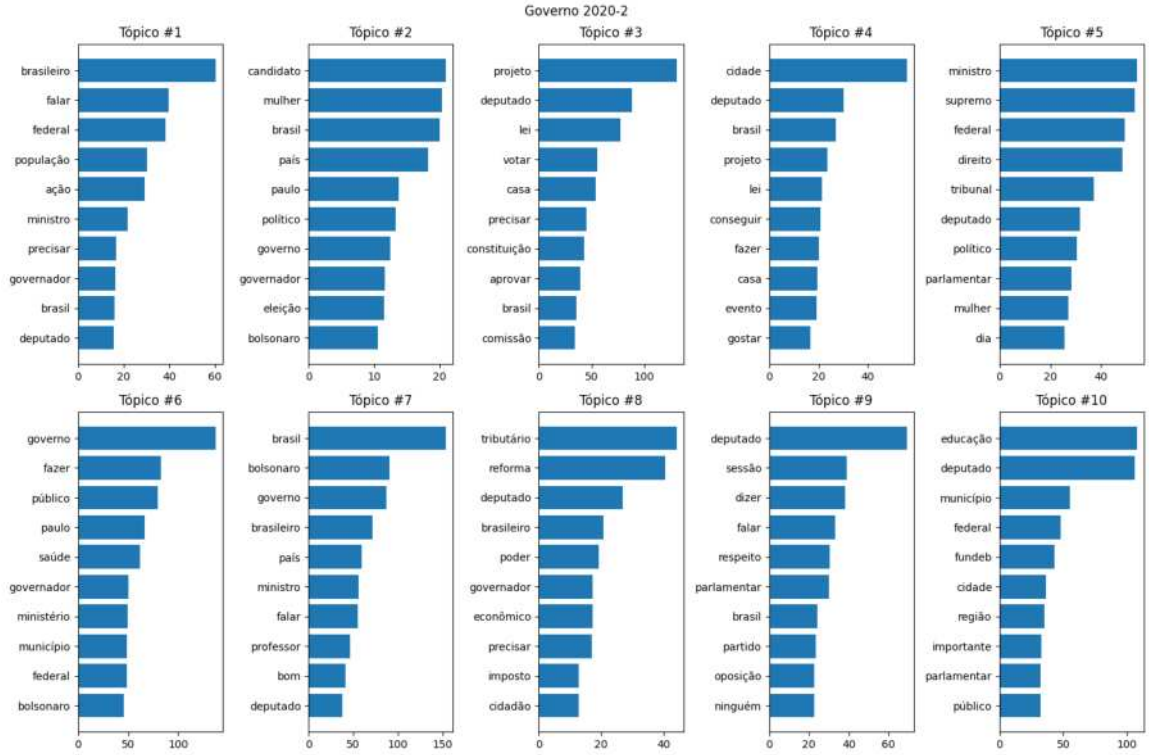
Fonte: Elaboração própria, 2023.

Figura 22 – Tópicos extraídos do semestre 2020-1 - governo



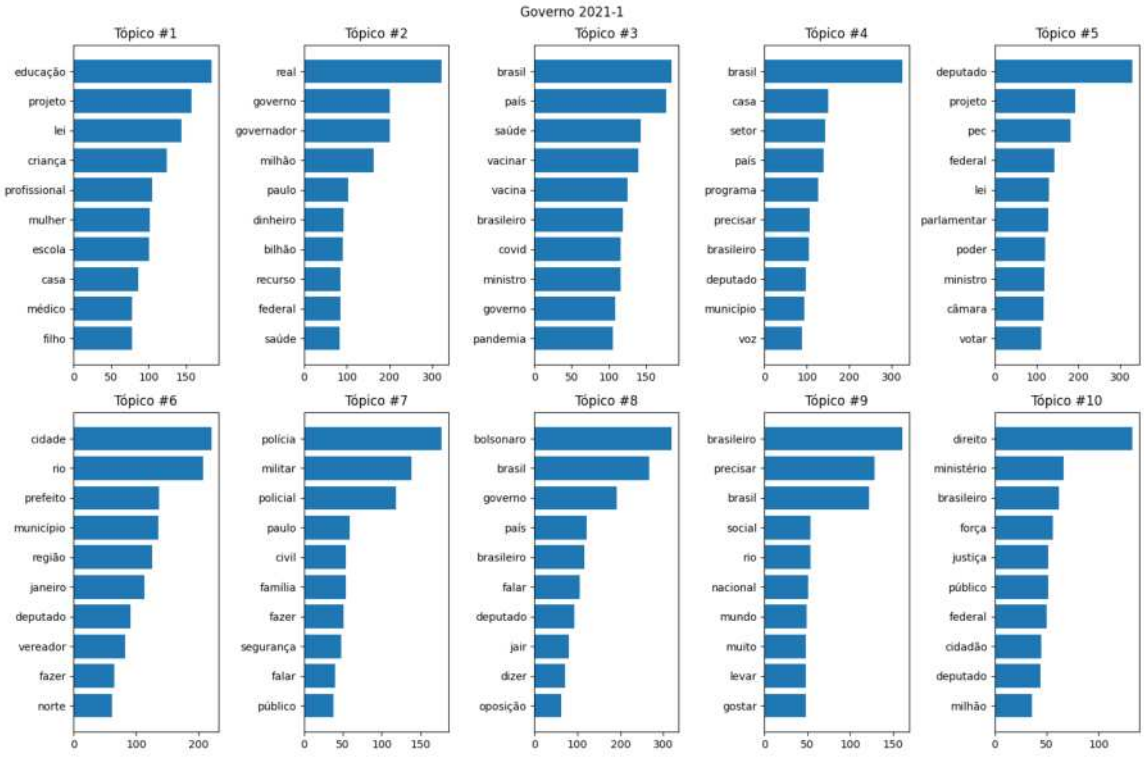
Fonte: Elaboração própria, 2023.

Figura 23 – Tópicos extraídos do semestre 2020-2 - governo



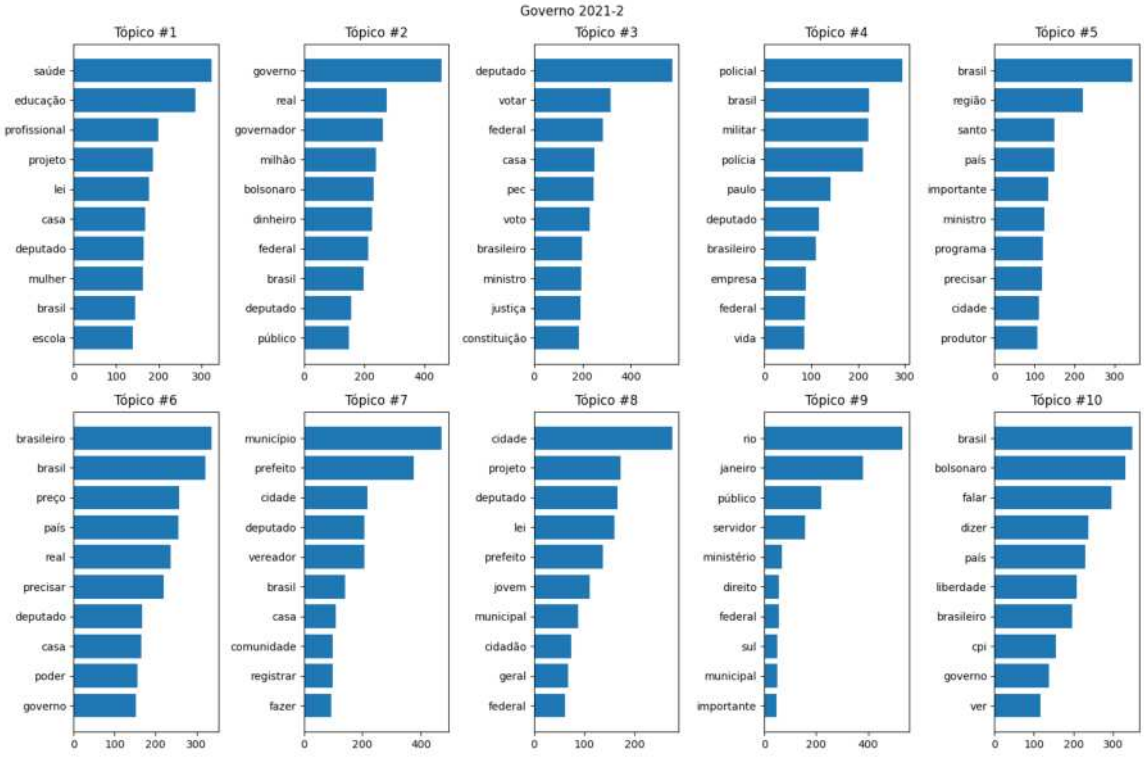
Fonte: Elaboração própria, 2023.

Figura 24 – Tópicos extraídos do semestre 2021-1 - governo



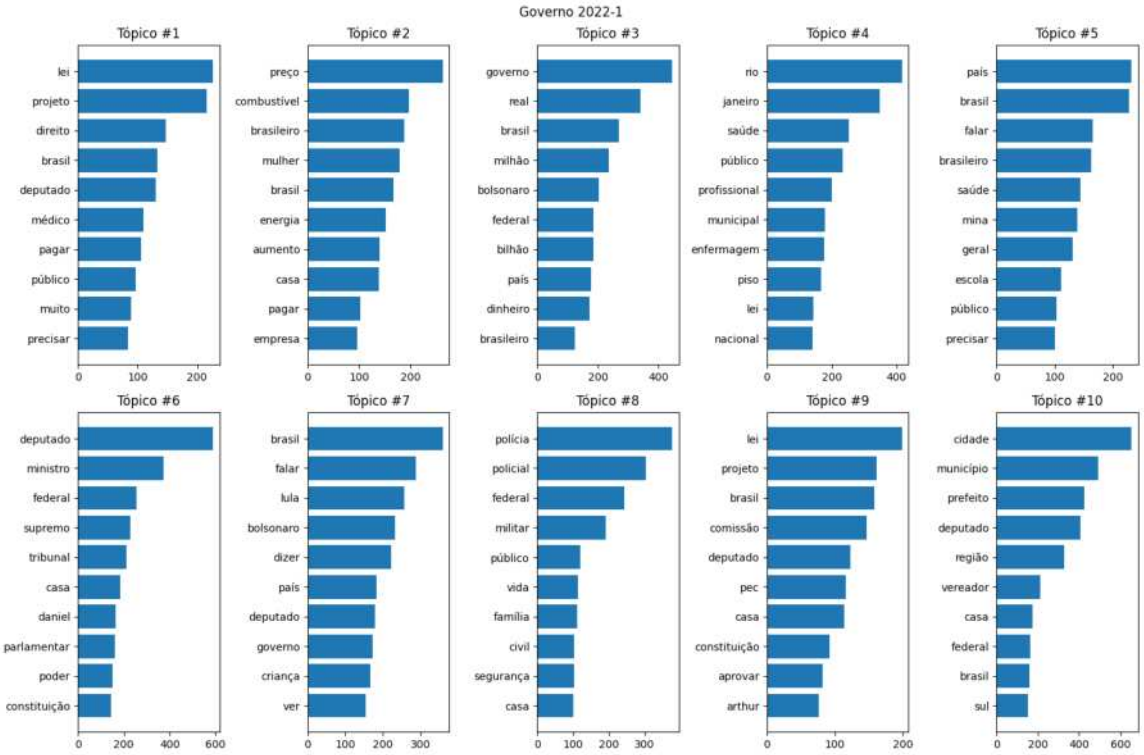
Fonte: Elaboração própria, 2023.

Figura 25 – Tópicos extraídos do semestre 2021-2 - governo



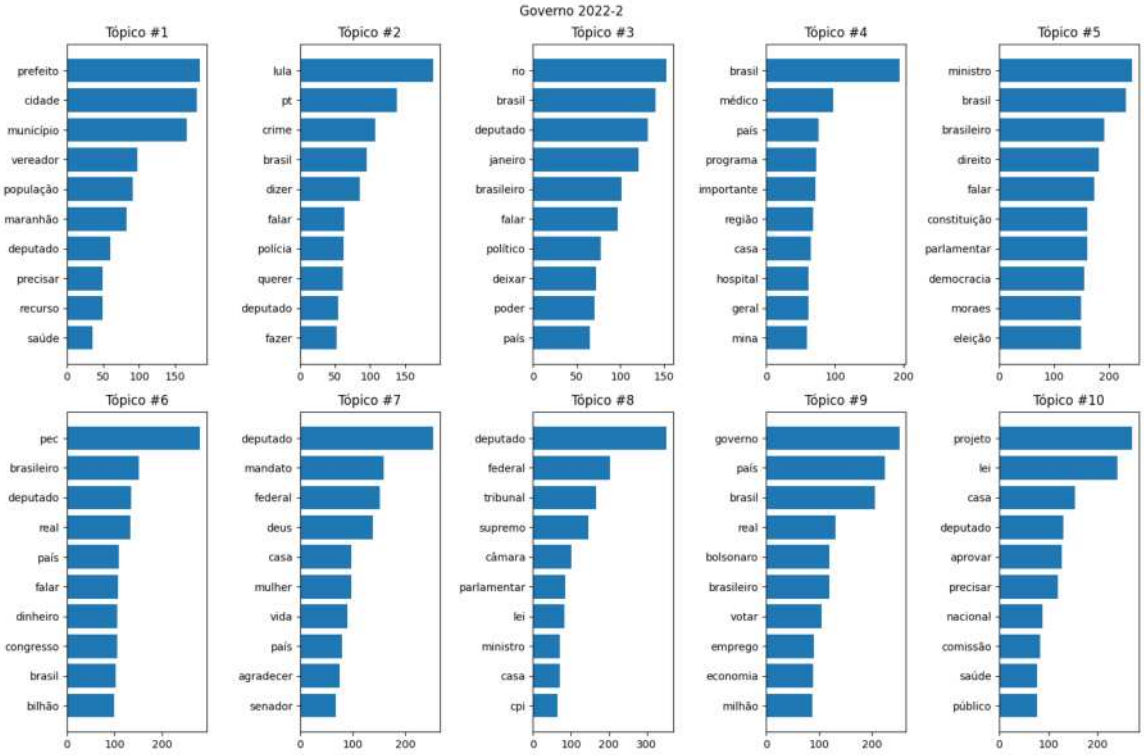
Fonte: Elaboração própria, 2023.

Figura 26 – Tópicos extraídos do semestre 2022-1 - governo



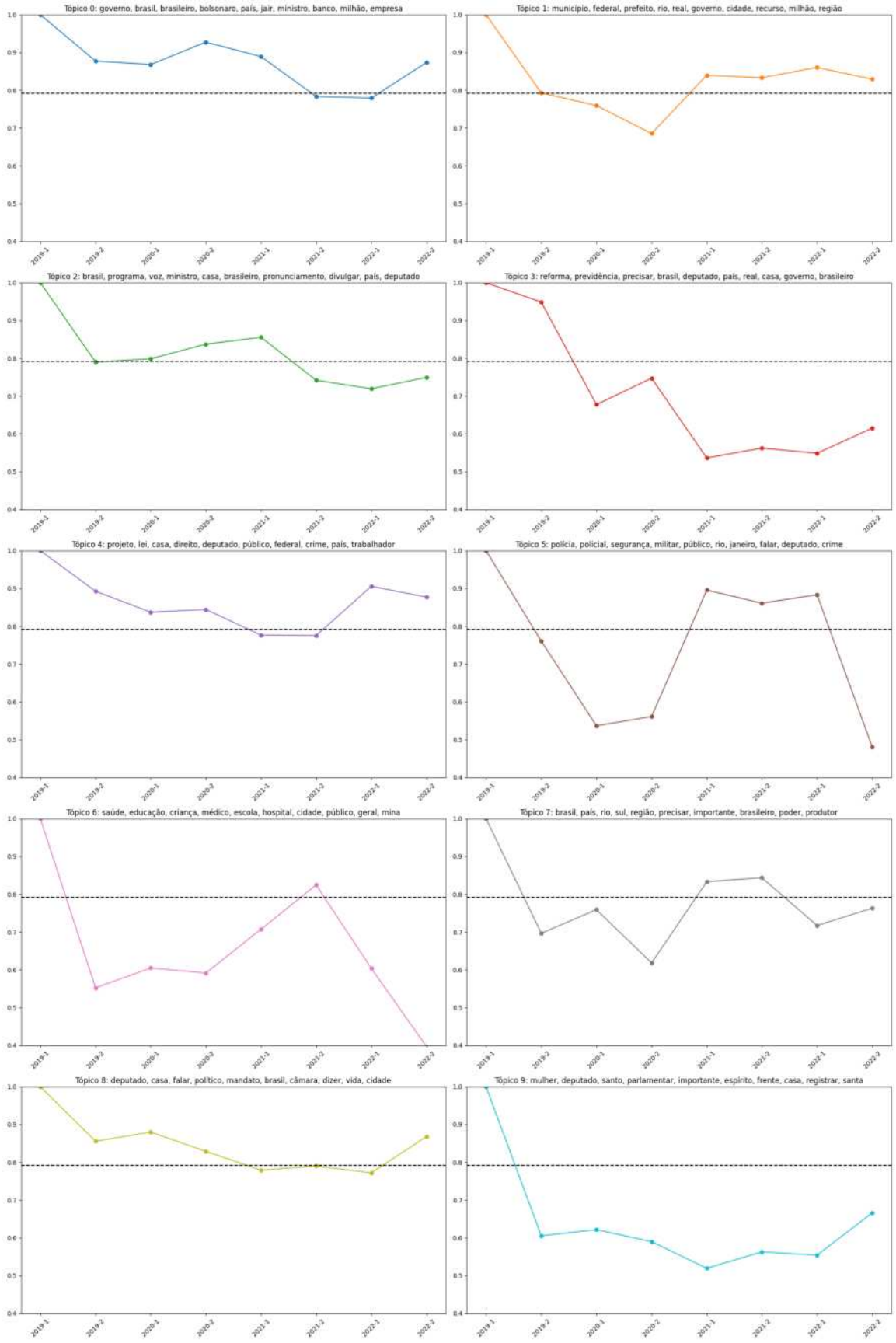
Fonte: Elaboração própria, 2023.

Figura 27 – Tópicos extraídos do semestre 2022-2 - governo



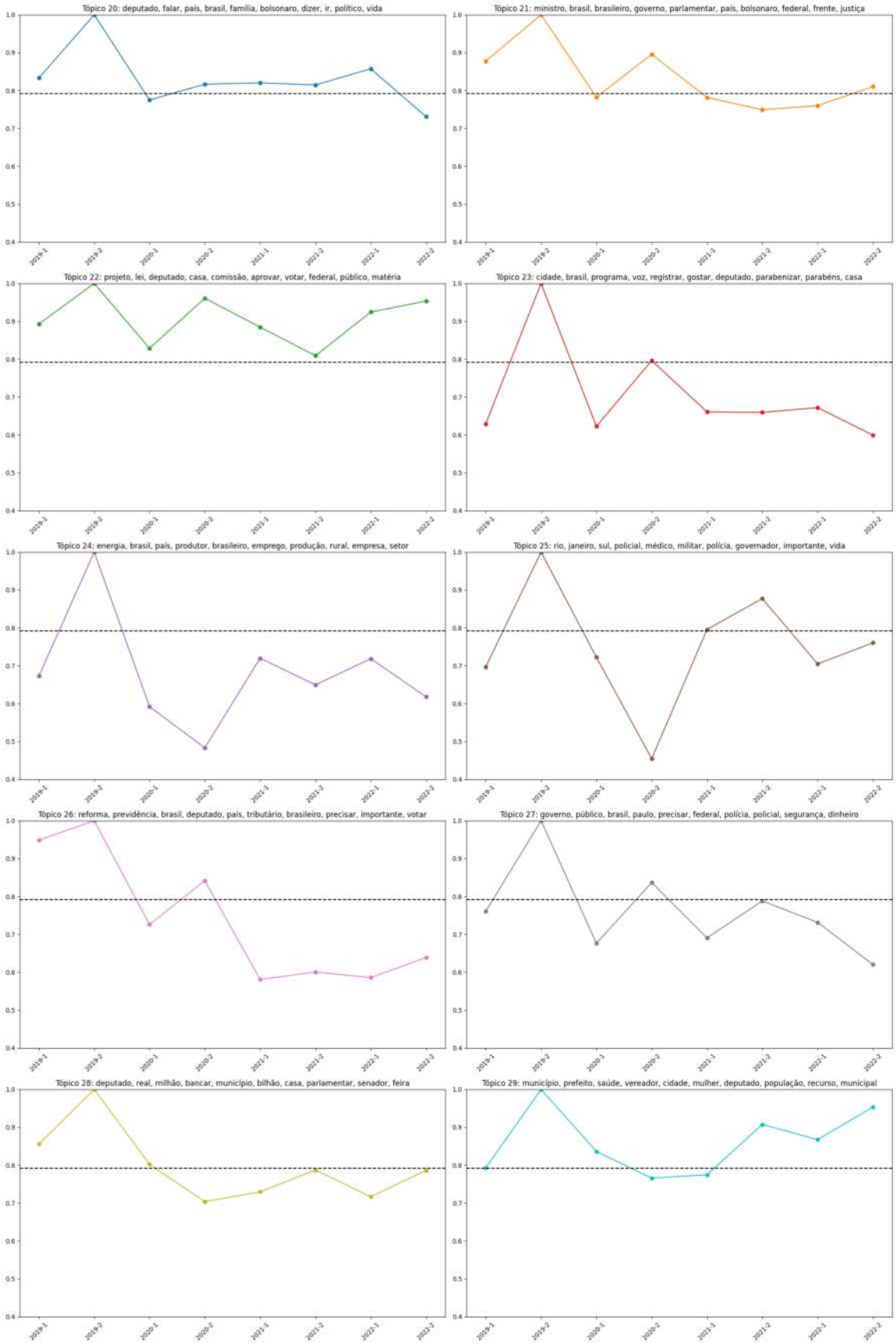
Fonte: Elaboração própria, 2023.

Figura 28 – Semelhança temporal dos tópicos extraídos do semestre 2019-1 - governo



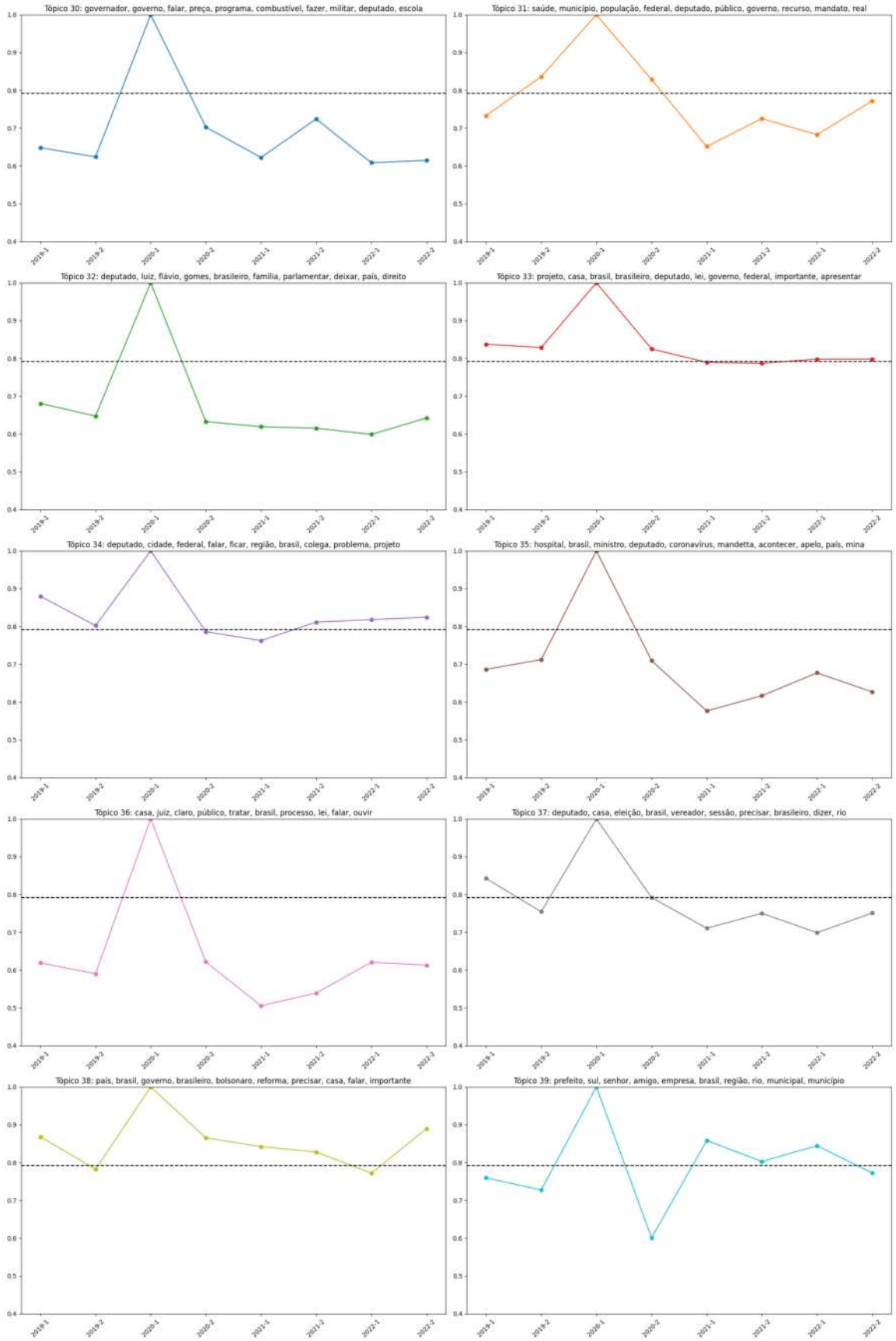
Fonte: Elaboração própria, 2023.

Figura 29 – Semelhança temporal dos tópicos extraídos do semestre 2019-2 - governo



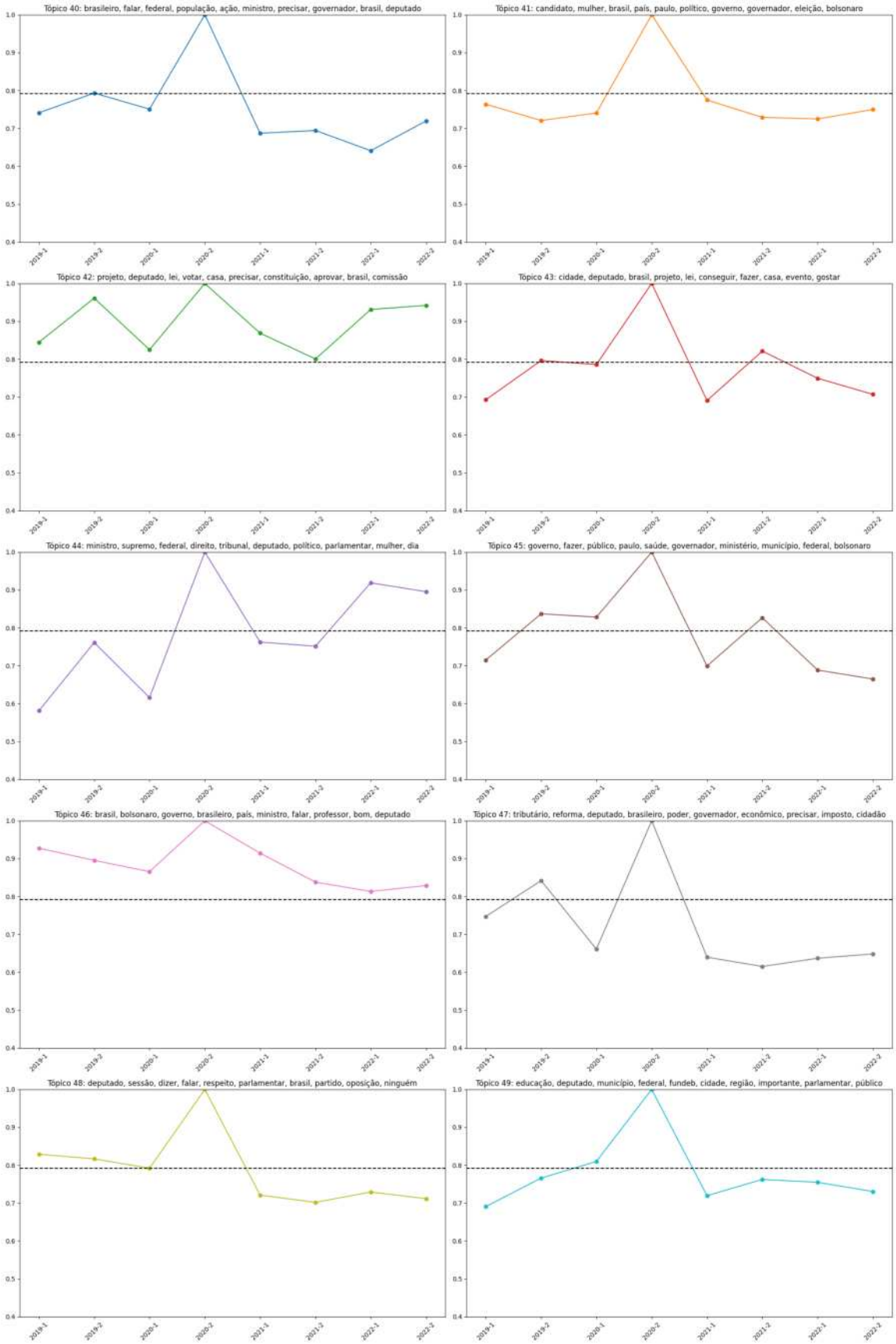
Fonte: Elaboração própria, 2023.

Figura 30 – Semelhança temporal dos tópicos extraídos do semestre 2020-1 - governo



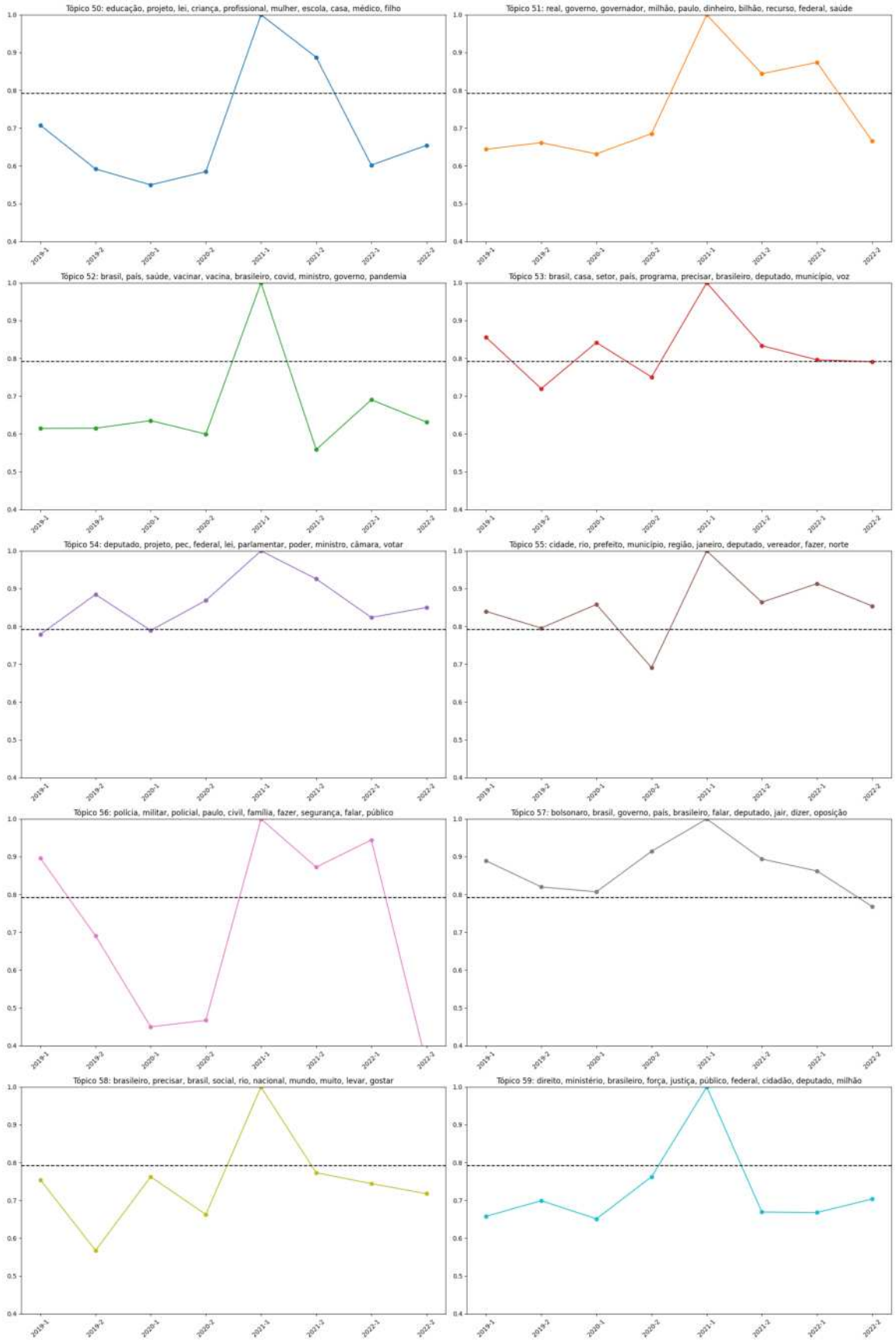
Fonte: Elaboração própria, 2023.

Figura 31 – Semelhança temporal dos tópicos extraídos do semestre 2020-2 - governo



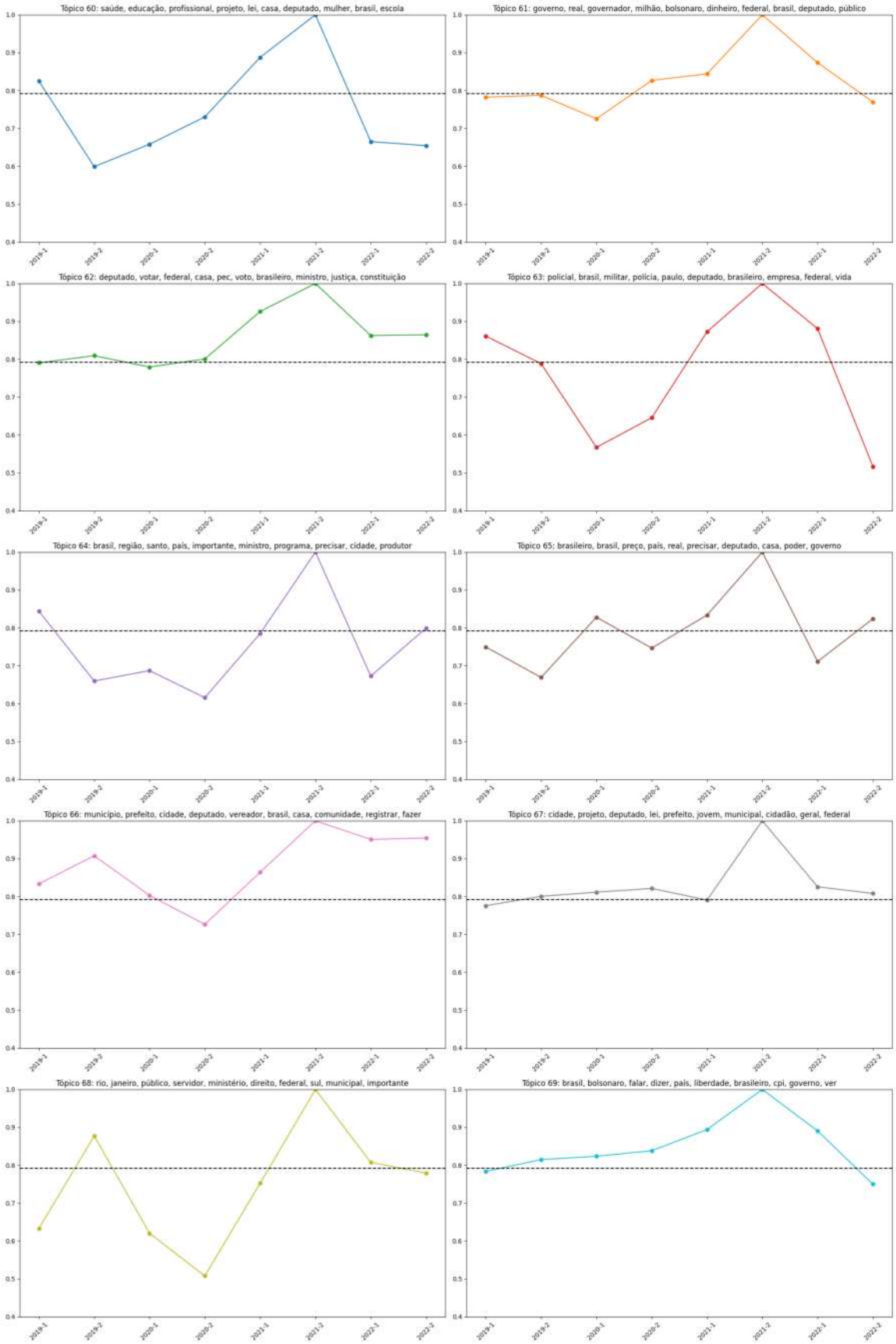
Fonte: Elaboração própria, 2023.

Figura 32 – Semelhança temporal dos tópicos extraídos do semestre 2021-1 - governo



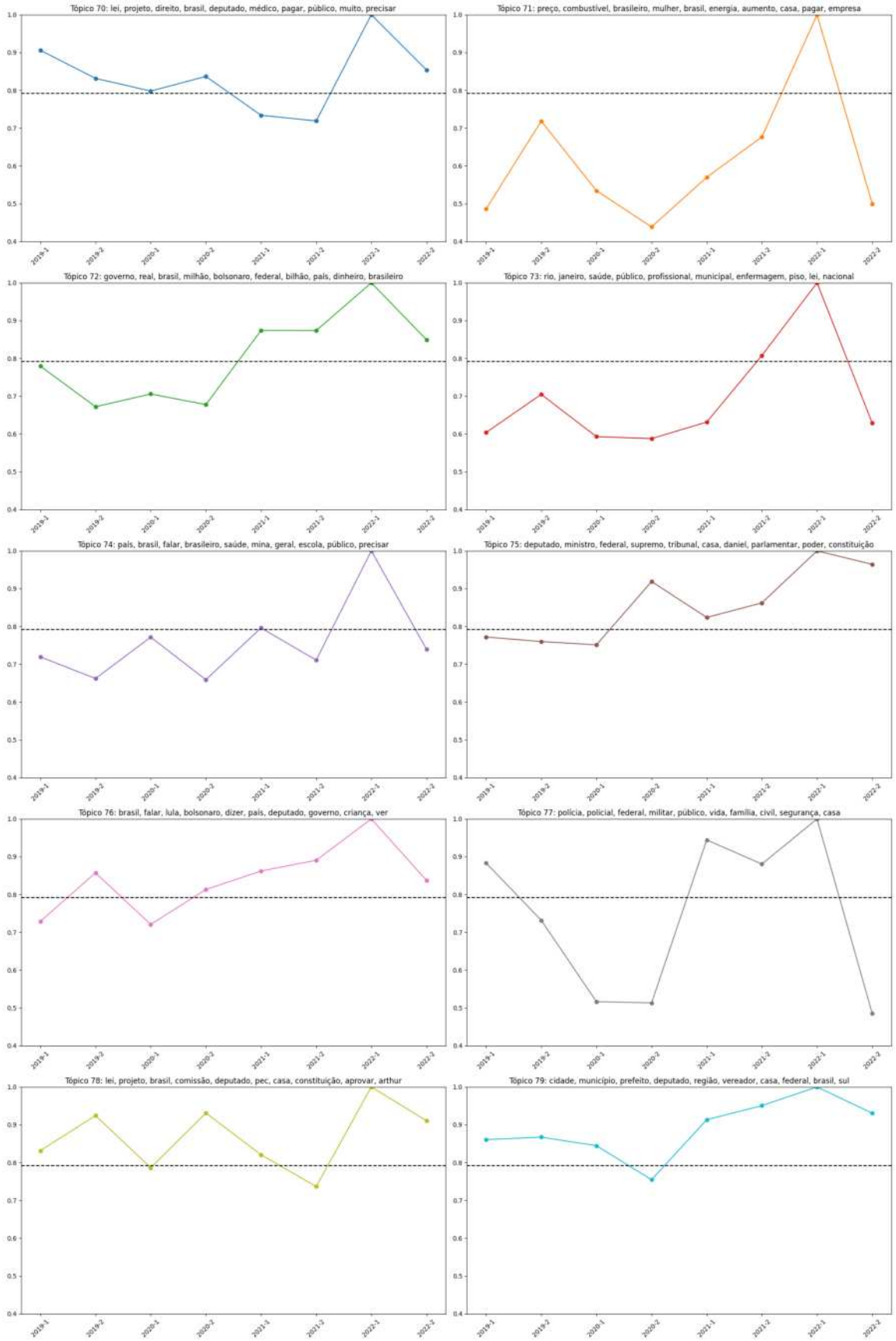
Fonte: Elaboração própria, 2023.

Figura 33 – Semelhança temporal dos tópicos extraídos do semestre 2021-2 - governo



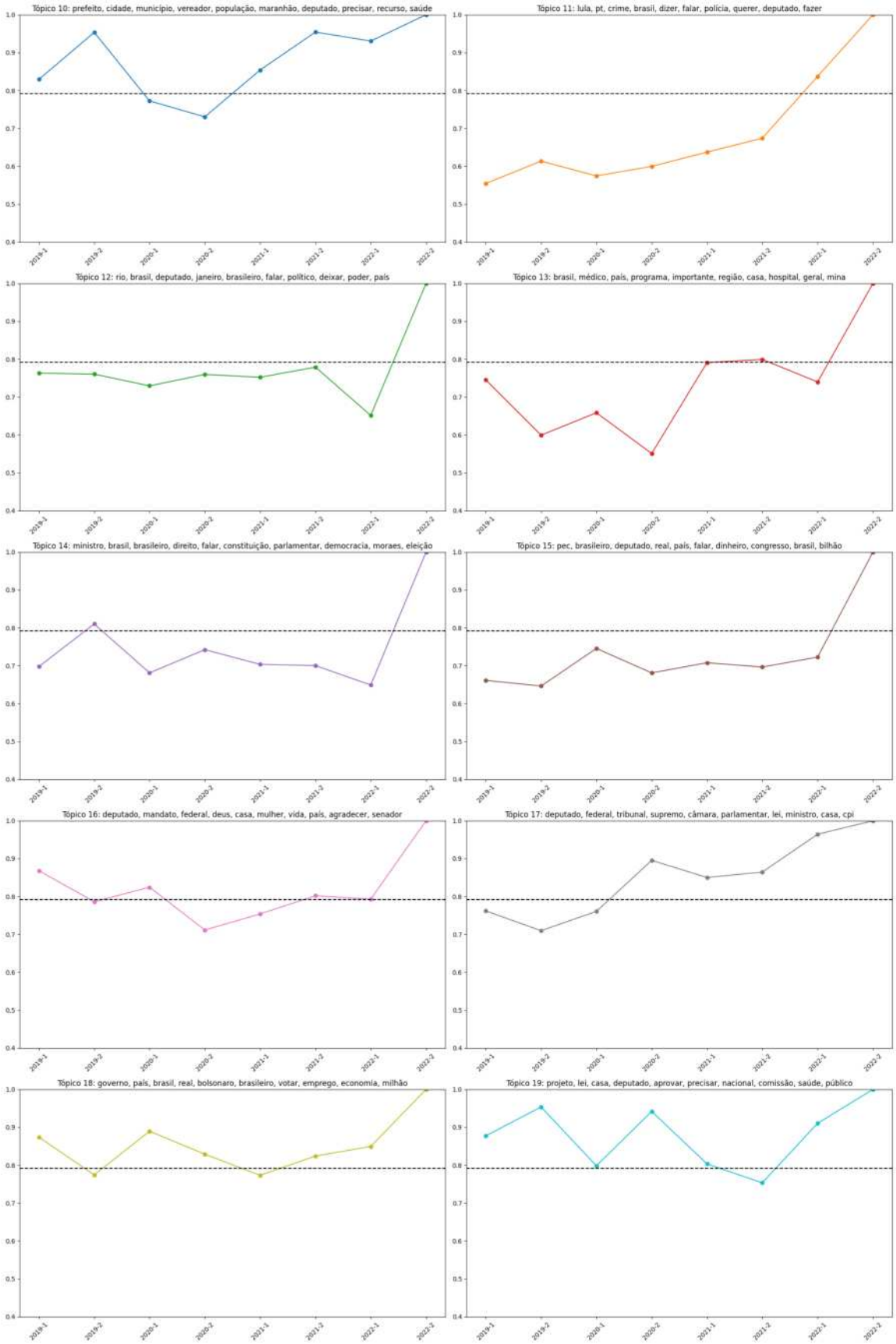
Fonte: Elaboração própria, 2023.

Figura 34 – Semelhança temporal dos tópicos extraídos do semestre 2022-1 - governo



Fonte: Elaboração própria, 2023.

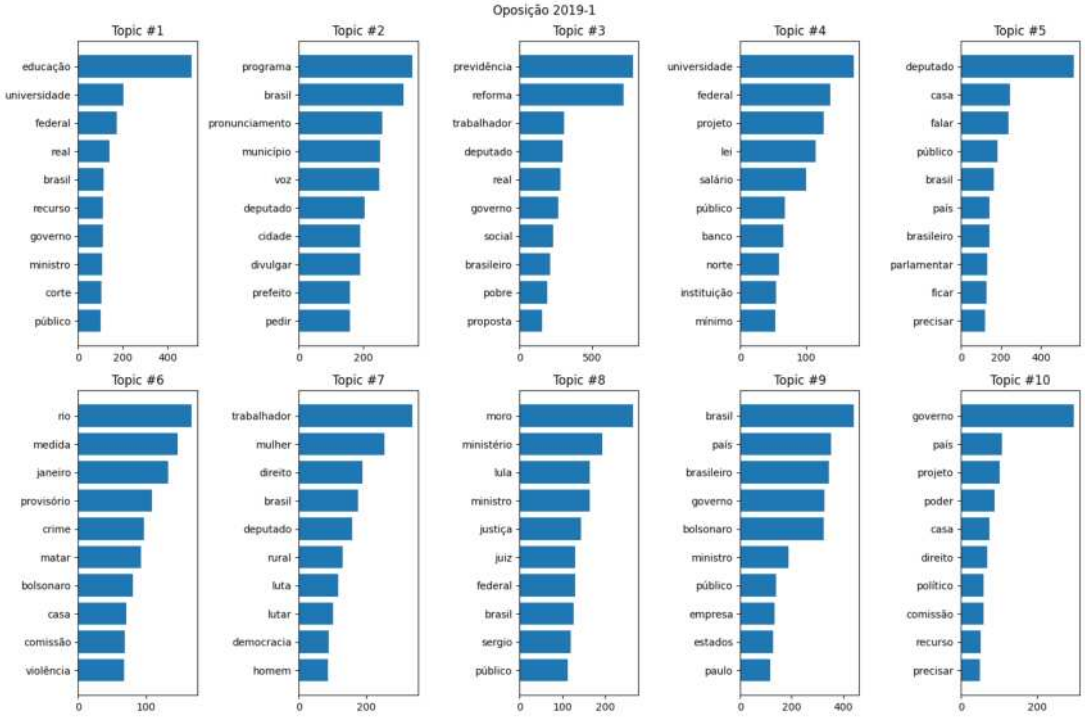
Figura 35 – Semelhança temporal dos tópicos extraídos do semestre 2022-2 - governo



Fonte: Elaboração própria, 2023.

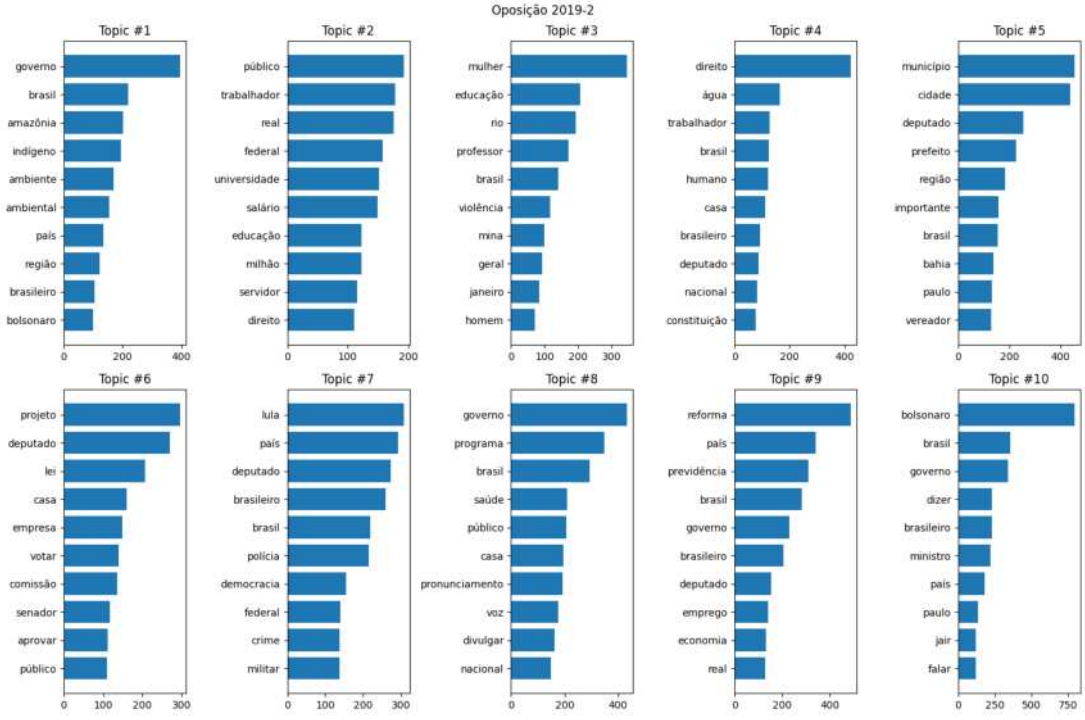
B.2 Oposição

Figura 36 – Tópicos extraídos do semestre 2019-1 - oposição



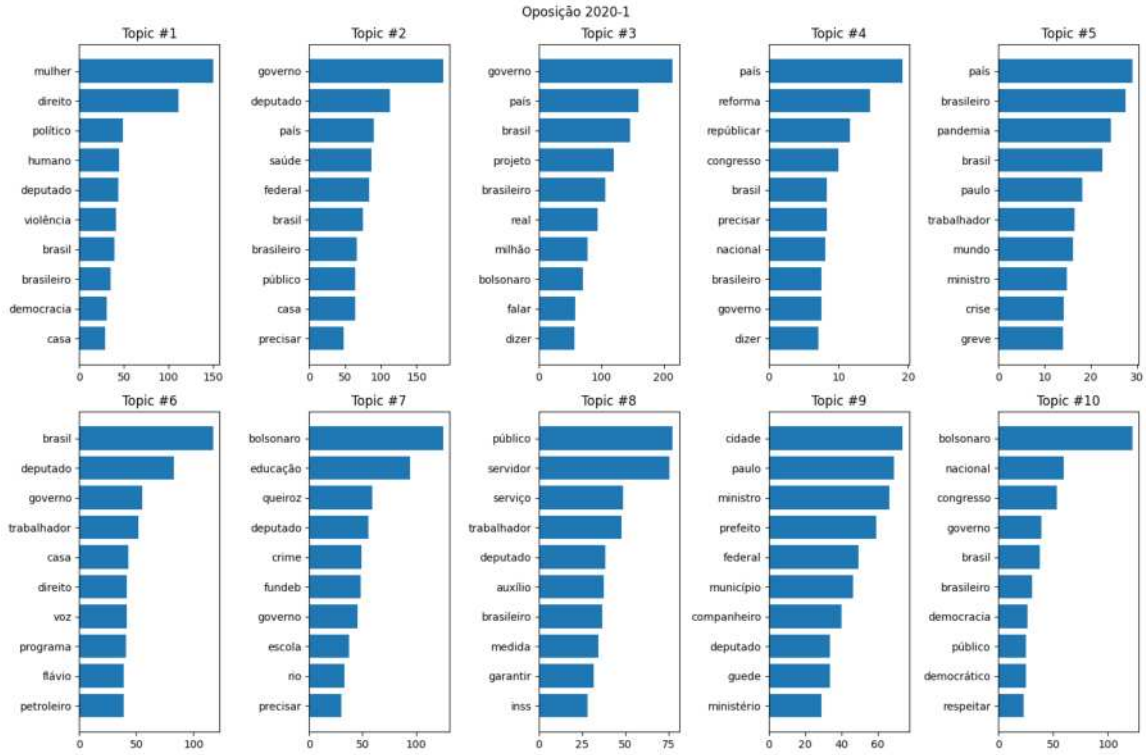
Fonte: Elaboração própria, 2023.

Figura 37 – Tópicos extraídos do semestre 2019-2 - oposição



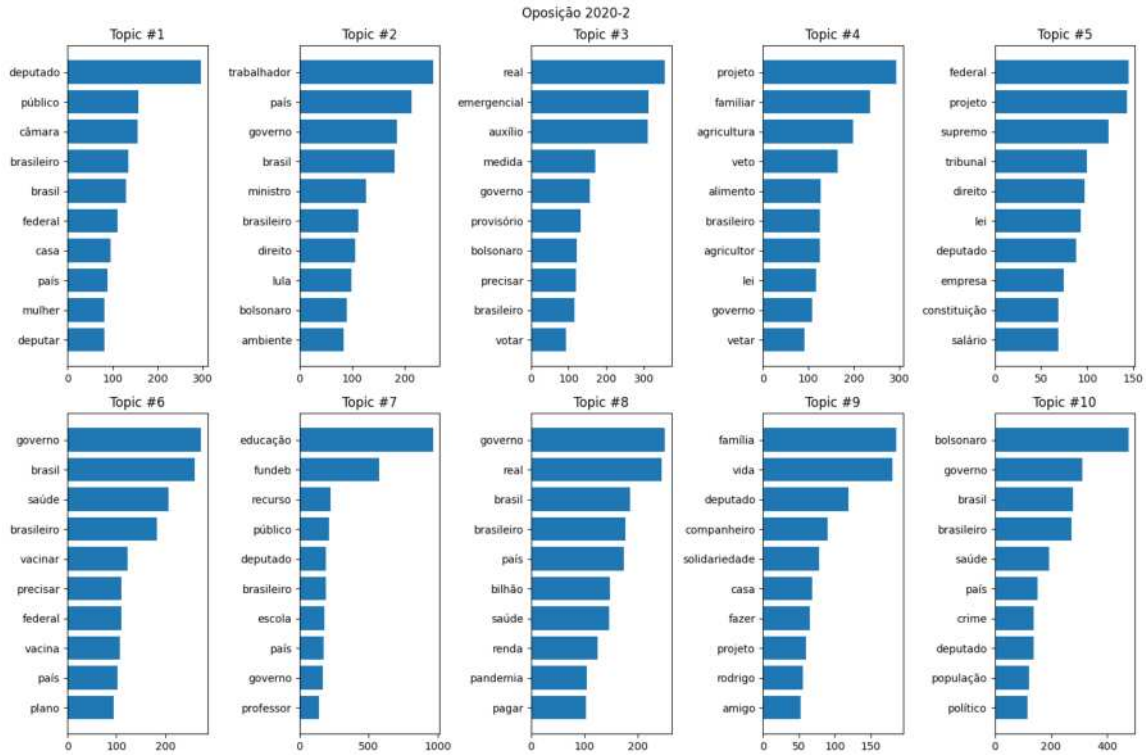
Fonte: Elaboração própria, 2023.

Figura 38 – Tópicos extraídos do semestre 2020-1 - oposição



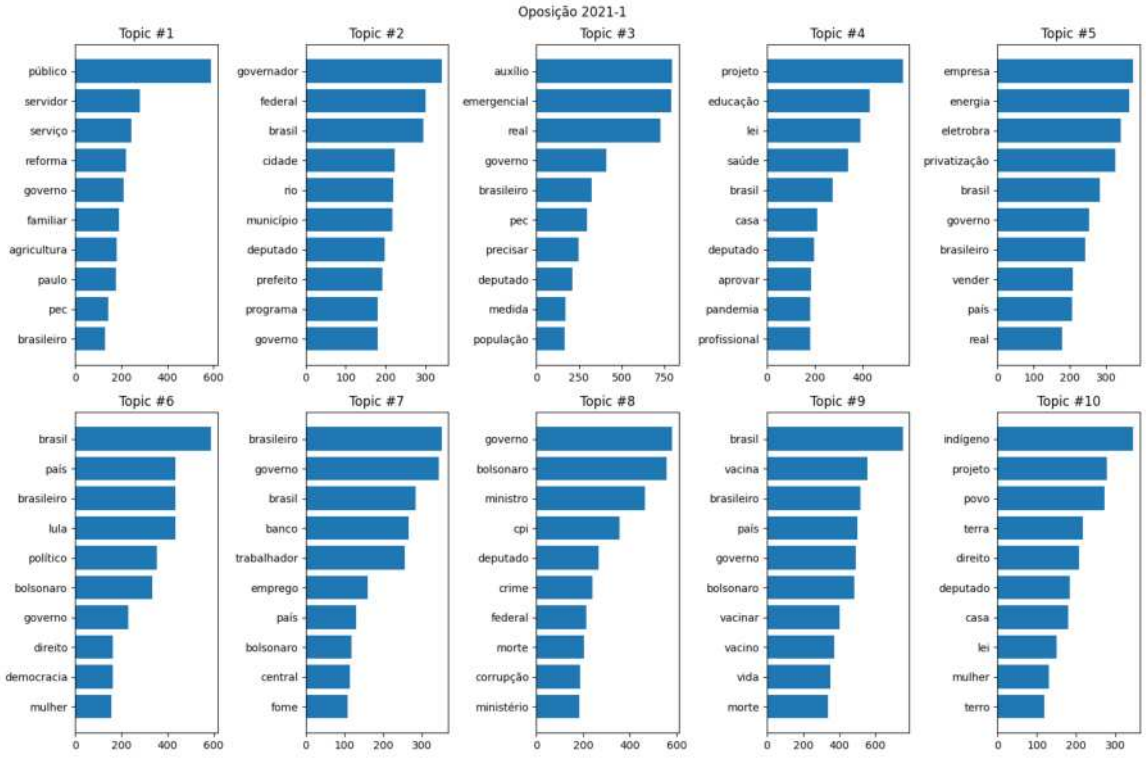
Fonte: Elaboração própria, 2023.

Figura 39 – Tópicos extraídos do semestre 2020-2 - oposição



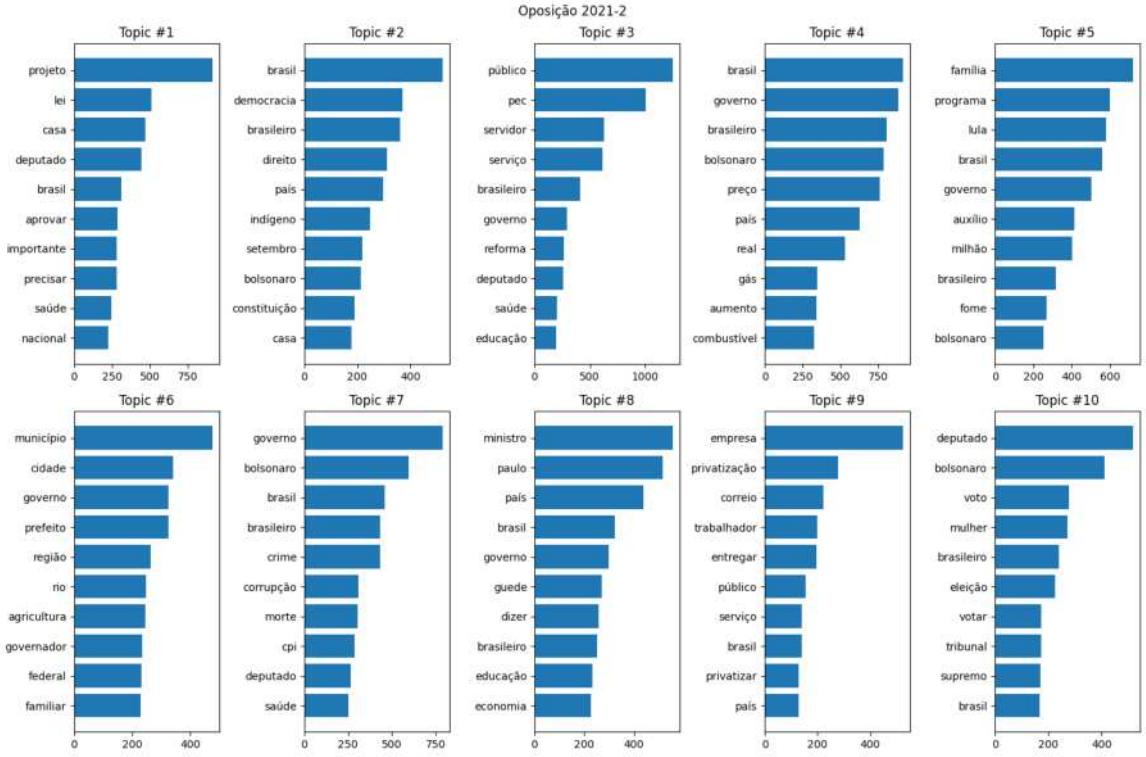
Fonte: Elaboração própria, 2023.

Figura 40 – Tópicos extraídos do semestre 2021-1 - oposição



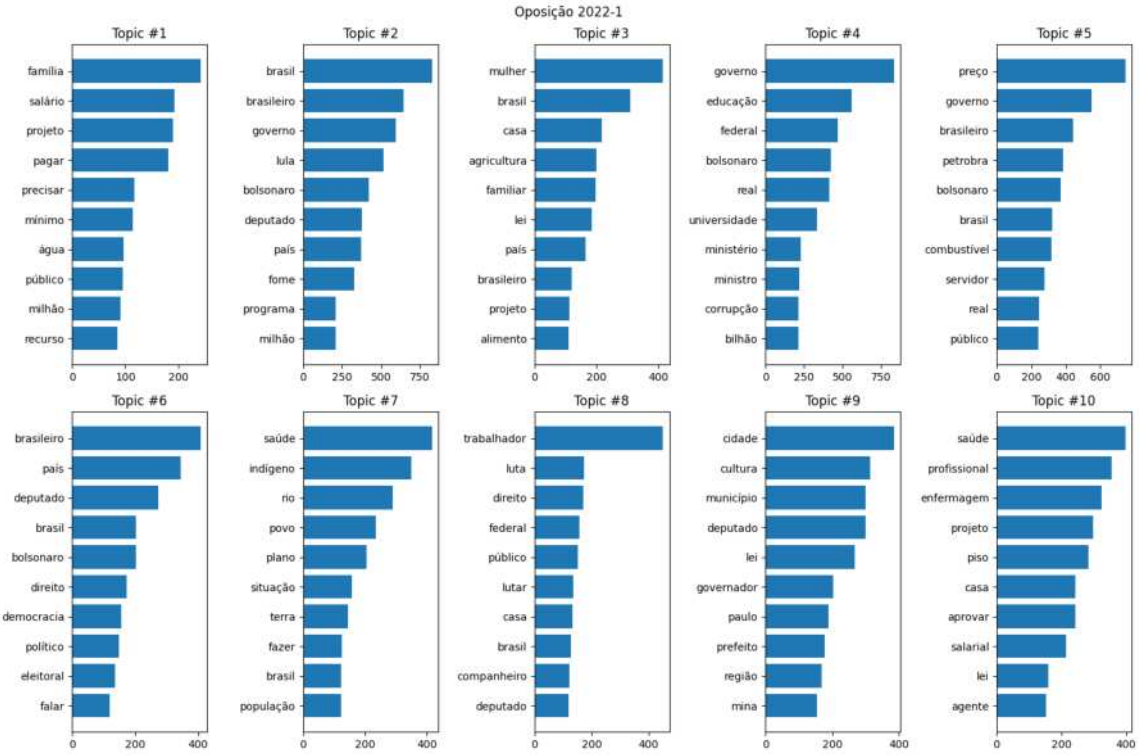
Fonte: Elaboração própria, 2023.

Figura 41 – Tópicos extraídos do semestre 2021-2 - oposição



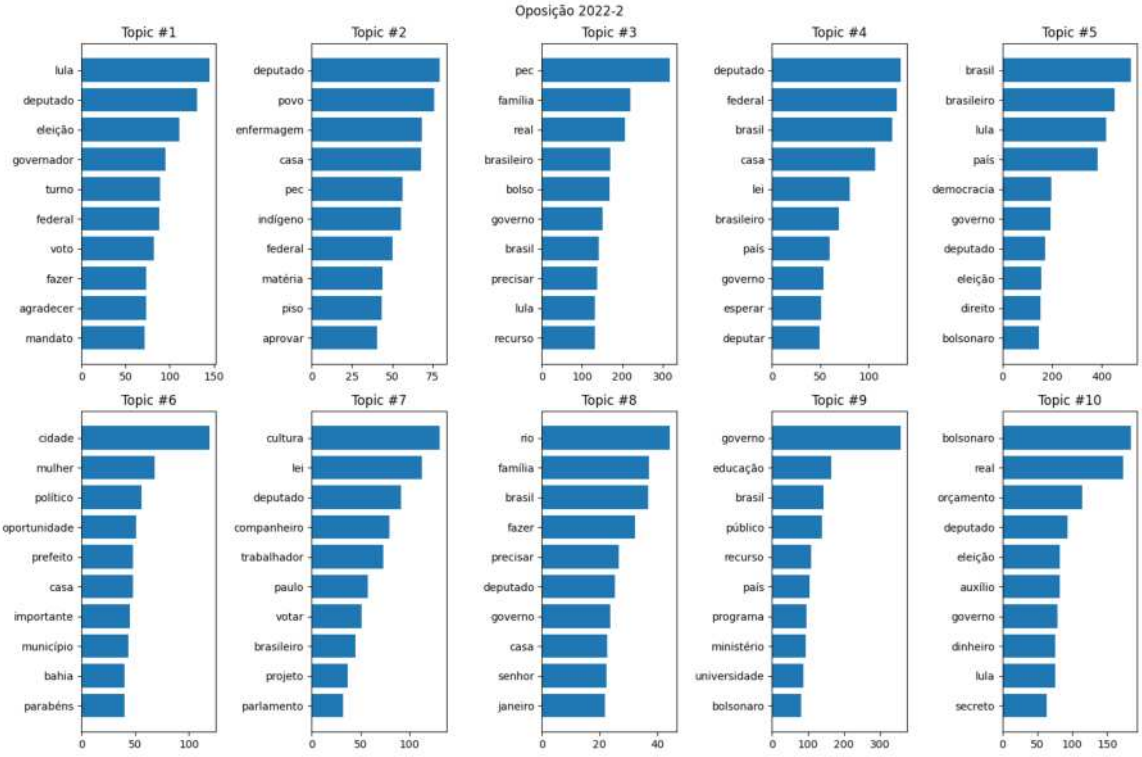
Fonte: Elaboração própria, 2023.

Figura 42 – Tópicos extraídos do semestre 2022-1 - oposição



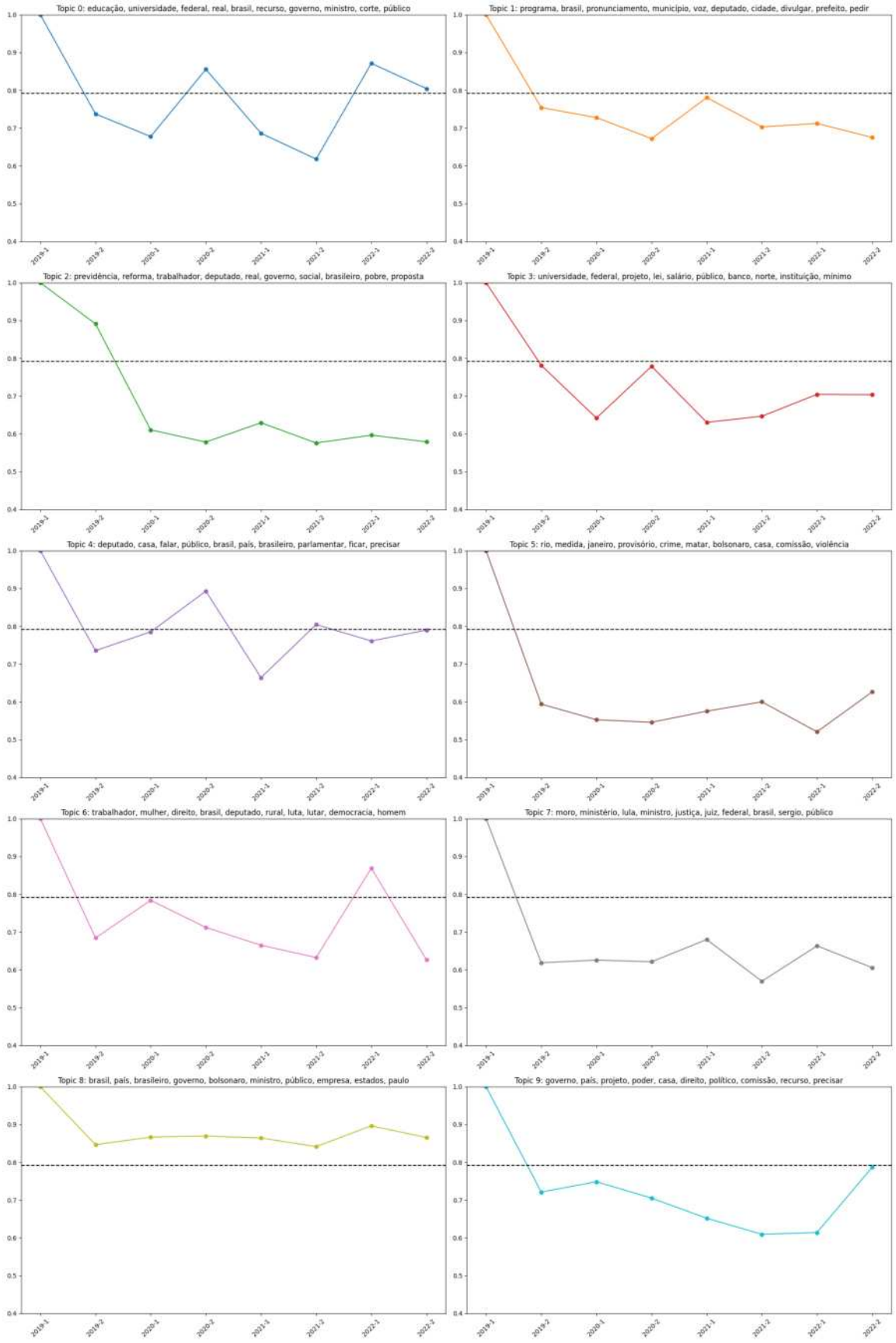
Fonte: Elaboração própria, 2023.

Figura 43 – Tópicos extraídos do semestre 2022-2 - oposição



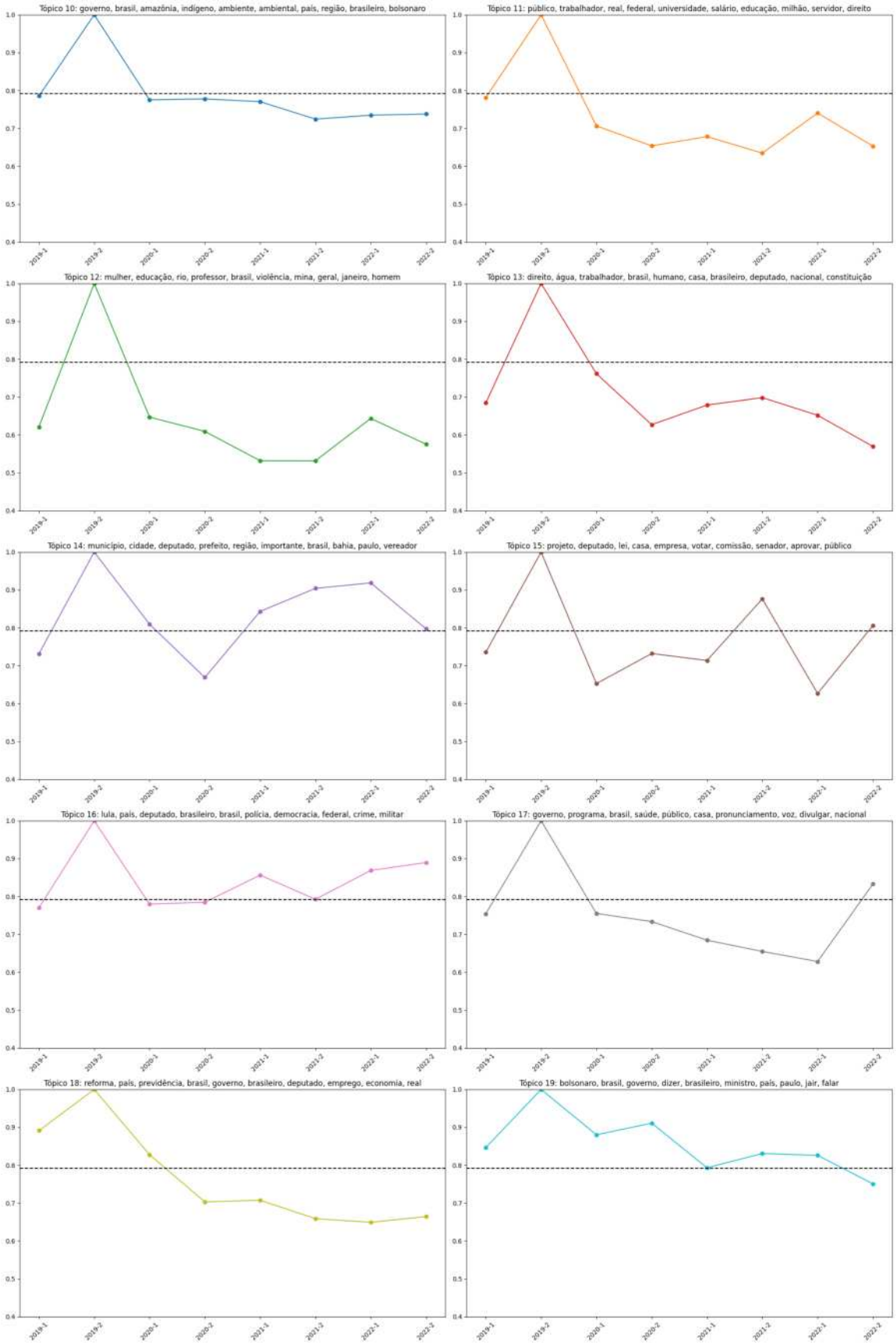
Fonte: Elaboração própria, 2023.

Figura 44 – Semelhança temporal dos tópicos extraídos do semestre 2019-1 - oposição



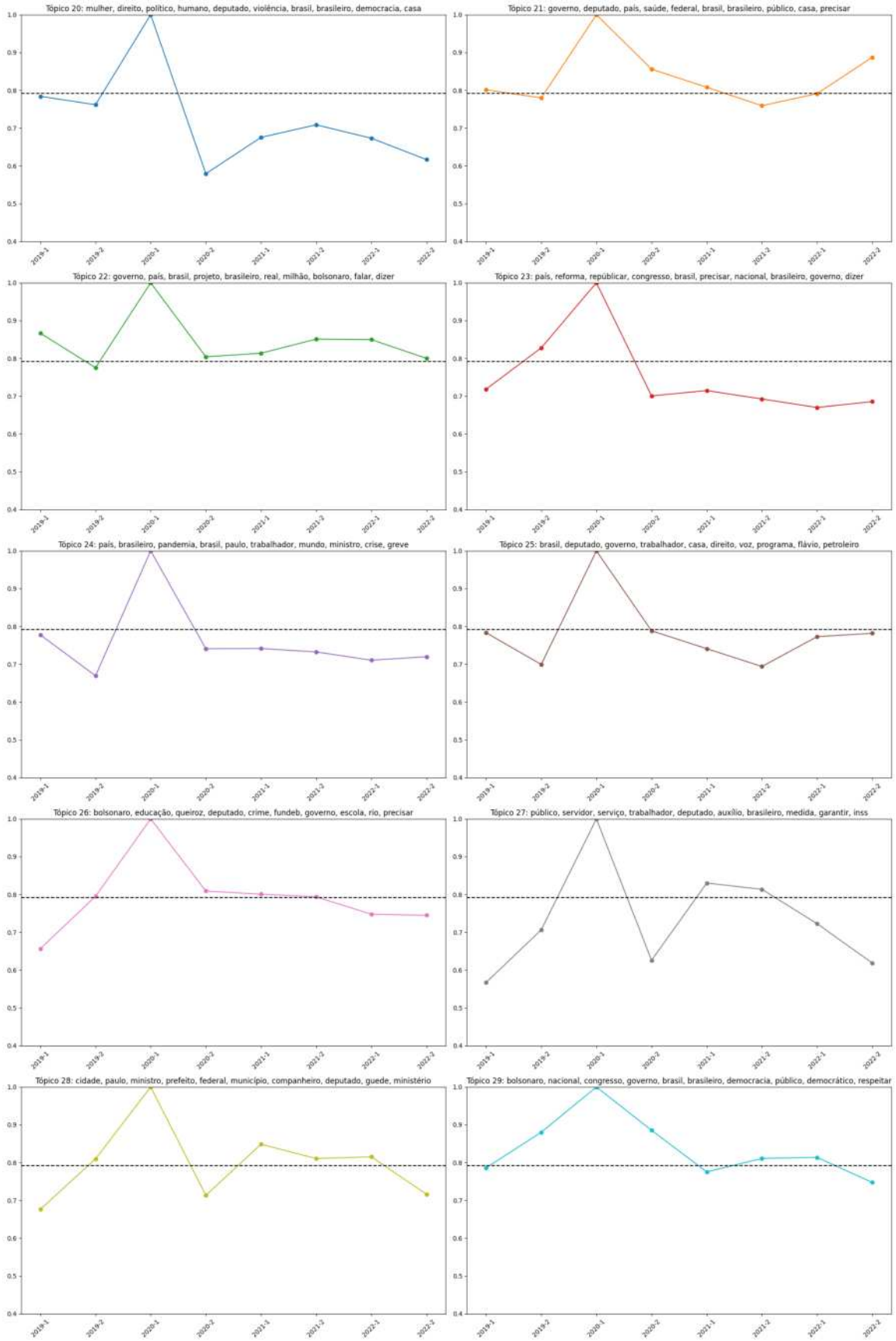
Fonte: Elaboração própria, 2023.

Figura 45 – Semelhança temporal dos tópicos extraídos do semestre 2019-2 - oposição



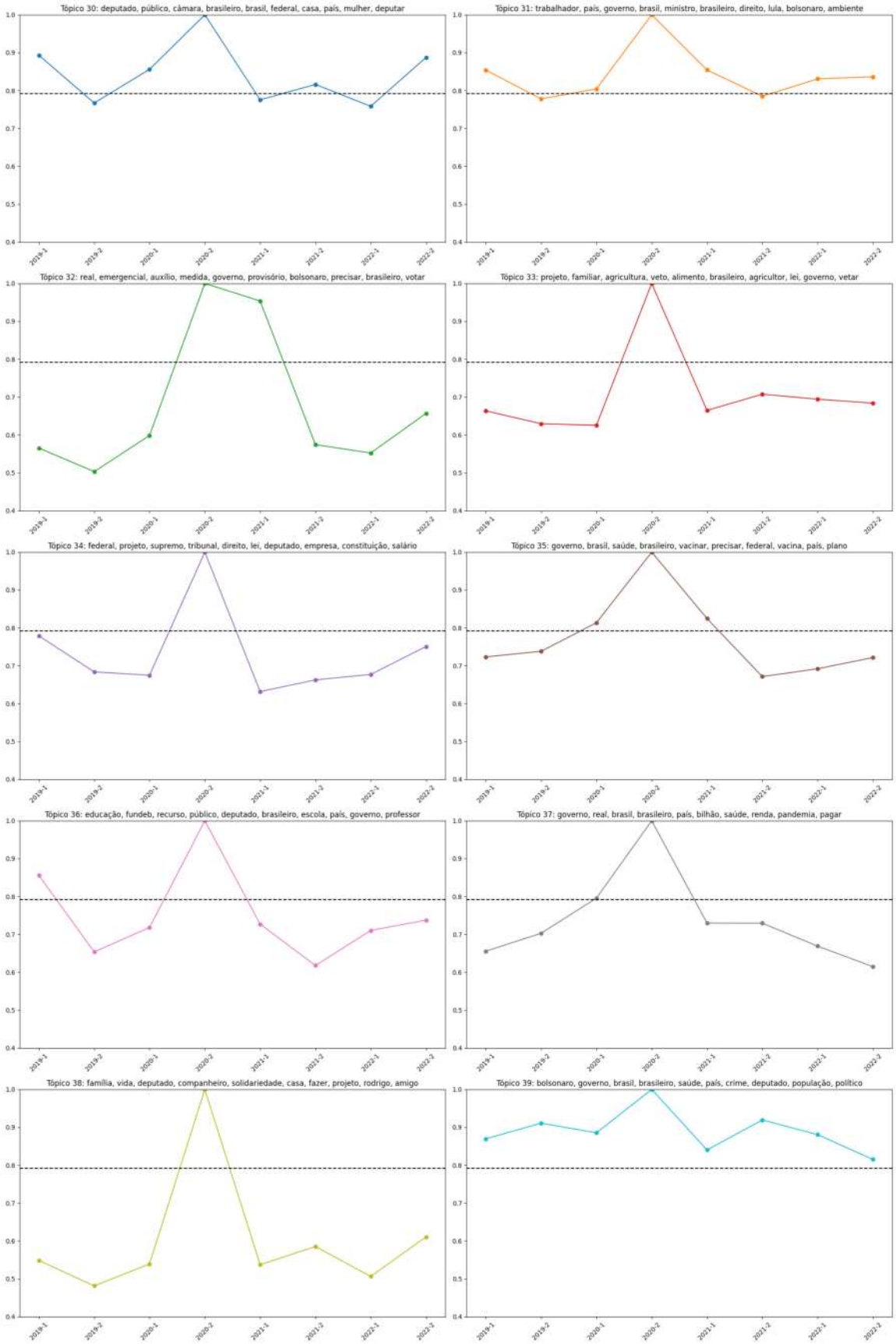
Fonte: Elaboração própria, 2023.

Figura 46 – Semelhança temporal dos tópicos extraídos do semestre 2020-1 - oposição



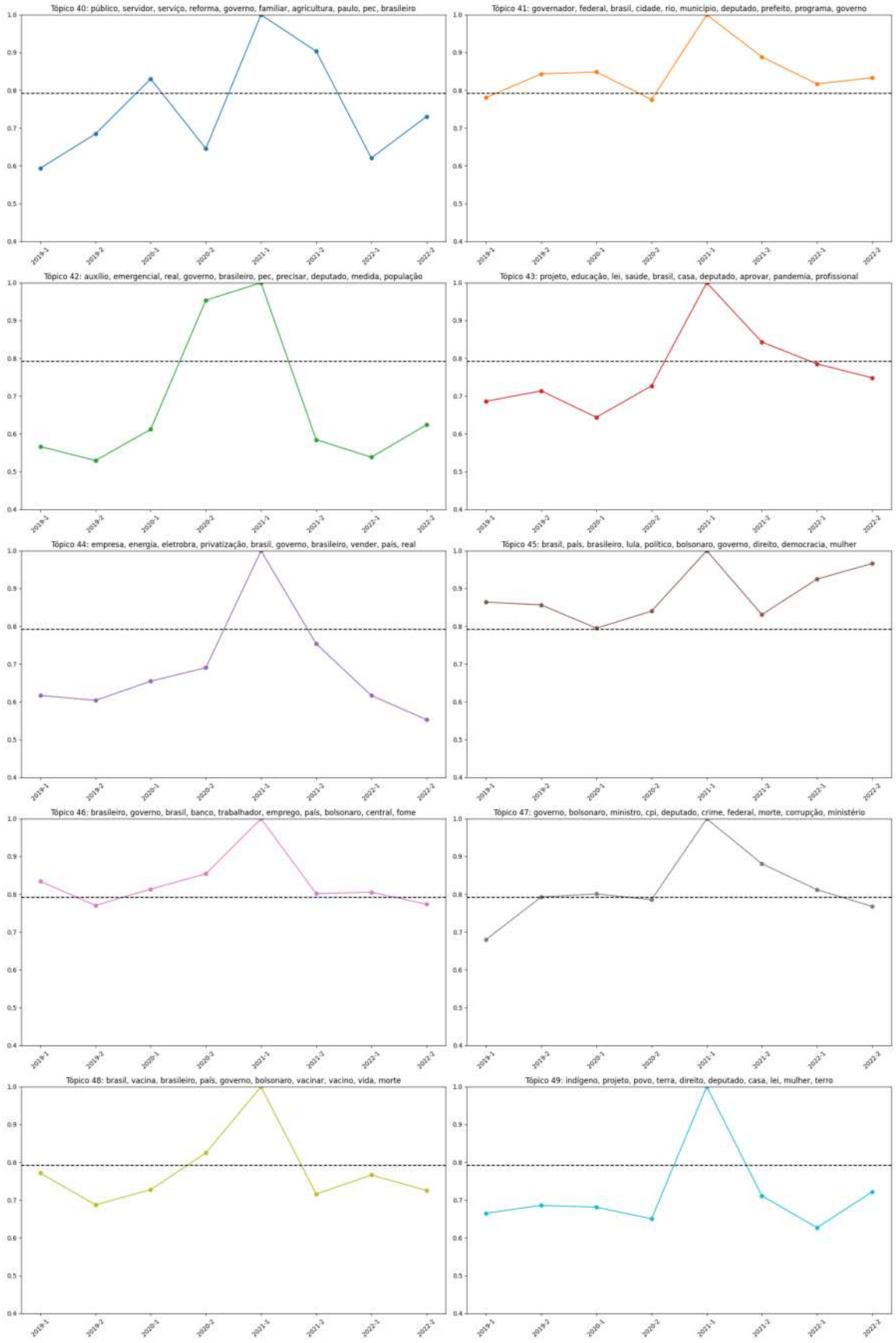
Fonte: Elaboração própria, 2023.

Figura 47 – Semelhança temporal dos tópicos extraídos do semestre 2020-2 - oposição



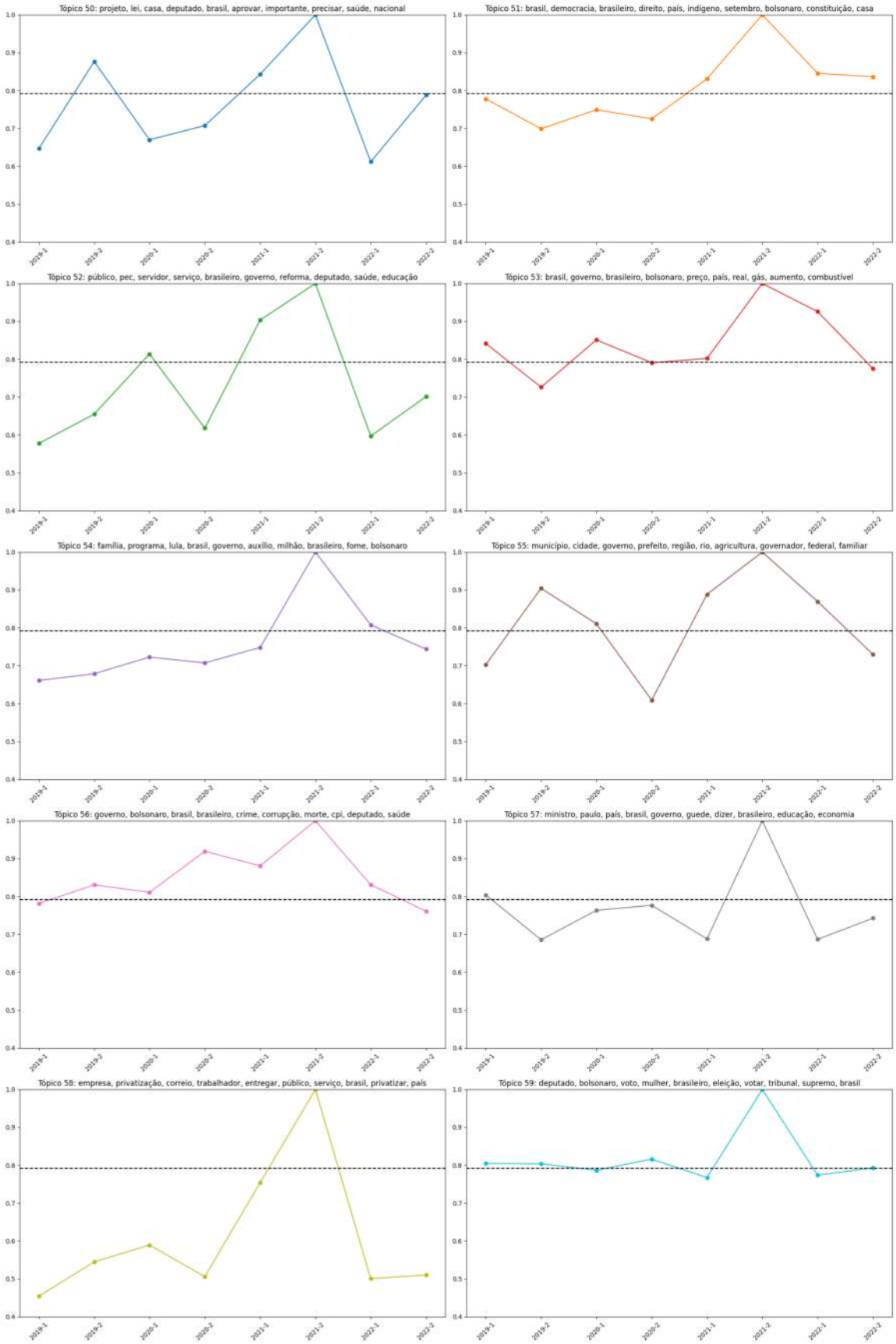
Fonte: Elaboração própria, 2023.

Figura 48 – Semelhança temporal dos tópicos extraídos do semestre 2021-1 - oposição



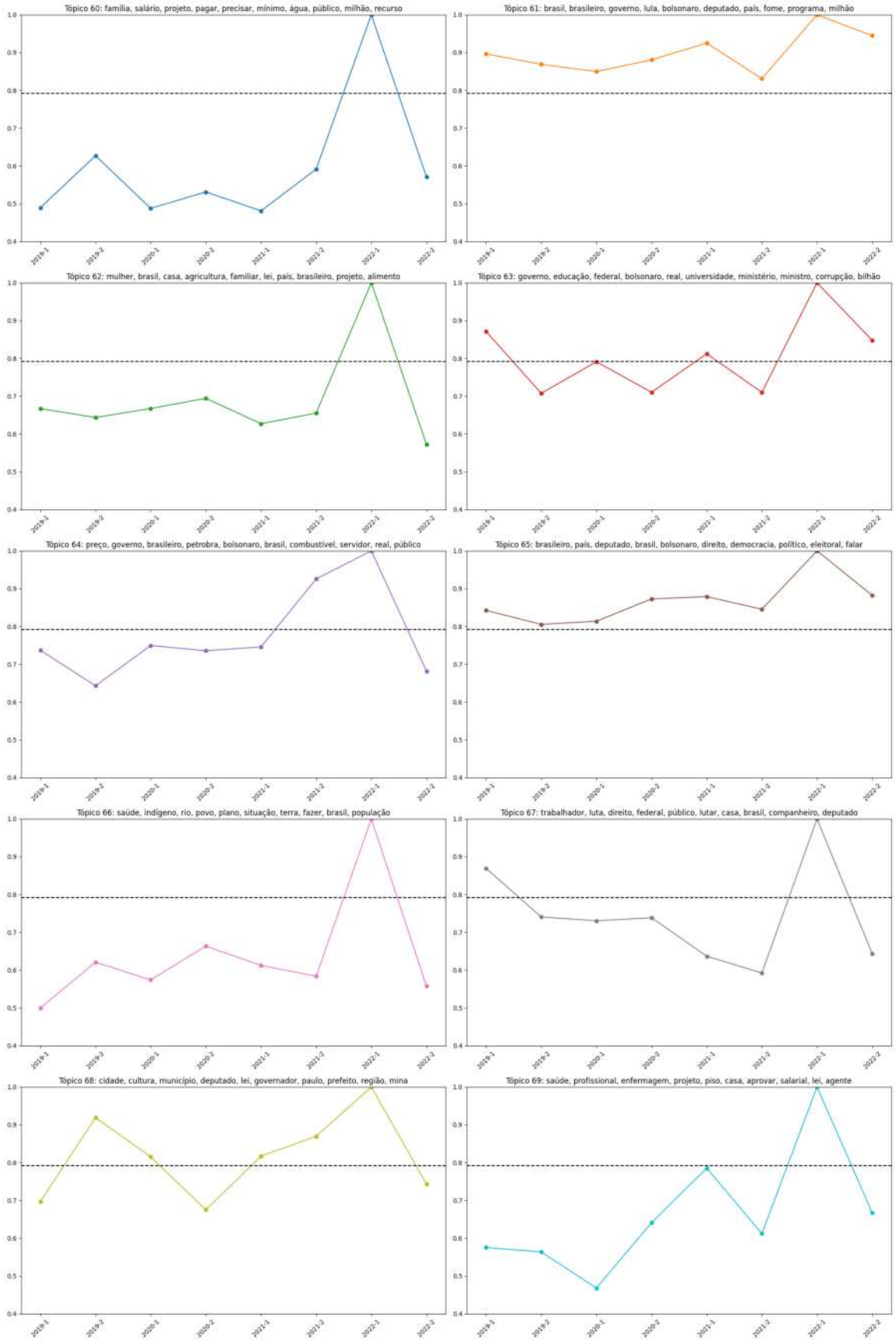
Fonte: Elaboração própria, 2023.

Figura 49 – Semelhança temporal dos tópicos extraídos do semestre 2021-2 - oposição



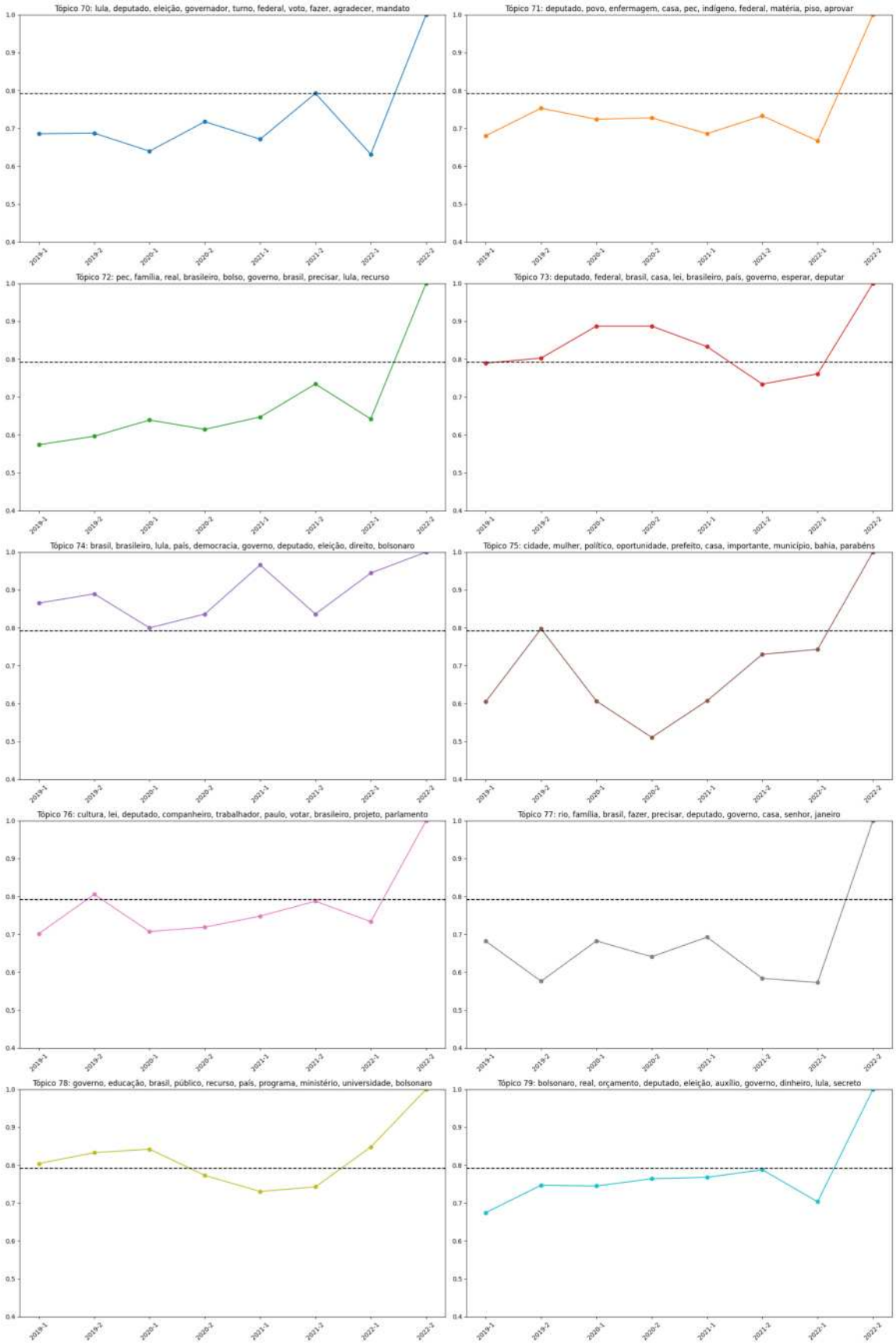
Fonte: Elaboração própria, 2023.

Figura 50 – Semelhança temporal dos tópicos extraídos do semestre 2022-1 - oposição



Fonte: Elaboração própria, 2023.

Figura 51 – Semelhança temporal dos tópicos extraídos do semestre 2022-2 - oposição



Fonte: Elaboração própria, 2023.