Aplicação do Modelo Multilíngue XLM-RoBERTa na Análise e Identificação de Grupos de Similaridade de Discursos Políticos

João Pedro Rodrigues Vieira[1]

[1]Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie (UPM)
São Paulo, SP – Brazil

{10403595}@mackenzista.com.br

Abstract. *This project aims to apply the XLM-RoBERTa model to classify political speeches by members of the Brazilian Chamber of Deputies regarding a specific topic, with the objective of visualizing the formation of ideological affinity groups based on the stated discourse. The motivation stems from the growing polarization in political debate, which hinders the capacity to seek solutions to the problems faced by society. The justification focuses on assessing the capability of specialized Natural Language Processing models to classify political speeches. Thus, the intention is to create a tool that assists in analyzing the debate within a polarized environment. Experimental results revealed the model's limitation in distinguishing similarity in texts that share significant lexical overlap or identical vocabulary, even when they convey divergent opinions.*

Resumo. *Este projeto visa aplicar o modelo XLM-RoBERTa na classificação de discursos de parlamentares da Câmara dos Deputados do Brasil a respeito de um mesmo tema, com o objetivo de visualizar a formação de grupos de afinidade ideológica com base no discurso enunciado. A motivação se dá pela crescente polarização em torno do debate político, que limita a capacidade de busca por uma resolução dos problemas enfrentados pela sociedade. A justificativa centra-se na averiguação da capacidade de modelos especializados em Processamento de Linguagem Natural em classificarem discursos de natureza política. Deste modo, pretende-se criar uma ferramenta que auxilie na ponderação do debate em meio a um ambiente polarizado. Os resultados experimentais revelaram a limitação do modelo em identificar semelhança em textos que abordam palavras próximas ou idênticas, mesmo que apresentem opiniões distintas.*

1. Introdução

1.1. Contextualização

O crescimento da polarização no debate político em democracias do mundo todo tem se mostrado um risco a esse regime que, por essência, preconiza a existência da pluralidade de ideias para o alcance de resoluções dos problemas enfrentados por uma sociedade [Levin et al. 2021]. Nesse sentido, como aponta Vasconcelos et al. (2021), espera-se que a heterogeneidade de pensamentos represente um fator natural e benéfico para a cooperação entre partes distintas em prol de um interesse comum. No entanto, como prova o último autor, é a partir do momento em que esta divisão de grupos por pensamento assume uma configuração segregada que a polarização se torna um risco ao modelo democrático.

Diante desta conjuntura, se tornou evidente que, contraditoriamente, no decorrer do tempo, a ampliação da presença política em assuntos de naturezas cada vez mais distintas e complexas, cenário teoricamente favorável à existência da democracia, não enfraqueceu o fenômeno de faccionalização partidária. Sendo assim, é constatado que, embora a pluralidade de cosmovisões e temas em pauta na discussão política estimule a cooperação entre grupos ideologicamente distintos, o grau de partidarismo nestes debates é inversamente proporcional à variabilidade de abordagens aos assuntos tratados e, consequentemente, à capacidade de cooperação entre diferentes atores, o que favorece o fenômeno da polarização política [Kawakatsu 2021].

1.2. Justificativa

Diante do exposto, se mostra benéfico o desenvolvimento de uma ferramenta que auxilie, para o pleno exercício da democracia, na ponderação sobre discursos de cunho político. Deste modo, colaborando para a identificação de um ambiente polarizado, o estudo abriria espaço para a uma leitura sobre o atual estado do debate político e se a conjuntura analisada favorece ou não a cooperação em prol do alcance de um bem comum, razão pela qual os parlamentares foram eleitos.

Além disso, a exploração da desenvoltura de modelos de Processamento de Linguagem Natural (PLN) na análise de discursos políticos é escassa na literatura, sobretudo no que diz respeito a textos em língua portuguesa do Brasil. Neste sentido, o estudo desenvolvido neste trabalho tem potencial de gerar contribuições diretas ao ramo e inaugurar tantas outras possibilidades de estudo.

1.3. Objetivo

Este trabalho, de maneira geral, tem como objetivo identificar a coerência ideológica entre os diferentes espectros políticos aos quais pertencem as siglas partidárias do parlamento brasileiro, ao realizar uma análise discursiva sobre falas de membros da Câmara de Deputados do Brasil. De maneira específica, busca-se entender o desempenho e a consequente aplicabilidade do modelo multilíngue *XLM-RoBERTa* (*XLM-R*) na análise de discursos políticos, avaliando se o mesmo é capaz de gerar *embeddings* que sejam capazes de indicar a similaridade entre os discursos coletados

para estudo. Os *embeddings* serão avaliados por meio da distribuição em plano bidimensional, gerada pela técnica *t-Distributed Stochastic Neighbor Embedding* (*t-SNE*), bem como através do cálculo da similaridade de cossenos, como proposto por Lima (2024).

1.4. Opção do projeto

O presente trabalho é proposto dada a relevância, importância e influência do tema político na atual conjuntura, sobretudo em um contexto em que é possível observar o fenômeno da polarização no debate a nível mundial. A desenvoltura de uma ferramenta capaz de analisar a situação do debate proferido em um corpo parlamentar tem capacidade de denunciar a qualidade da discussão, como também o potencial desta em alcançar soluções que favoreçam o bem social.

2. Fundamentação Teórica

2.1. Análise de discurso

A análise do discurso é um método multidisciplinar que possibilita estudar e analisar um texto, seja ele escrito ou falado. Como o termo "discurso" sugere, o método de análise do discurso se concentra em qualquer texto que possa provocar algum tipo de discurso, uma resposta de qualquer tipo. Dessa forma, amplia o leque de tópicos e assuntos que um analista pode usar, como em revistas médicas, artigos de jornal e até mesmo um discurso do presidente ou uma conversa casual [Johnstone e Andrus 2024; Martínez-Guillem, S., & Toula 2020].

2.2. *Large Language Models*

Large Language Model (*LLM*) é um modelo de *deep learning*, aplicado sobretudo no contexto de PLN. A arquitetura na qual se baseia é a do tipo *Transformer* e o aprendizado do modelo reside no treinamento exposto a um grande conjunto de dados. Deste modo, o modelo pode ser empregado em uma variedade de aplicações: reconhecimento e geração de texto, tradução, predição, dentre outros [Vaswani et al. 2017].

3. Descrição do Problema

O problema da pesquisa em questão consiste em averiguar a aplicabilidade do *LLM XLM-R* na análise de discurso político, pretendendo-se avaliar também elementos textuais que possam conferir viés à classificação. Toma-se, como base, a abordagem adotada por Wynter et al. (2023).

4. Responsabilidade Ética no Contexto de Solução

Por se tratar de uma aplicação de um modelo especializado de Inteligência Artificial (IA) na análise de discursos políticos, é preciso salientar a imparcialidade do modelo para geração das representações vetoriais semânticas, os *embeddings*. O modelo

será responsável, neste contexto, por avaliar semanticamente o conteúdo de cada discurso. A similaridade entre cada texto será calculada em cima das representações semânticas geradas pelo modelo, de modo que esta etapa da pesquisa representa tão somente uma análise sobre o resultado obtido, sem conferir qualidade aos discursos proferidos, mas atestando o quão semelhantes são em termos de conteúdo. Este tipo de investigação, sob o ponto de vista ético, é favorável ao exercício da democracia, ao fornecer um recorte do estado de suas discussões.

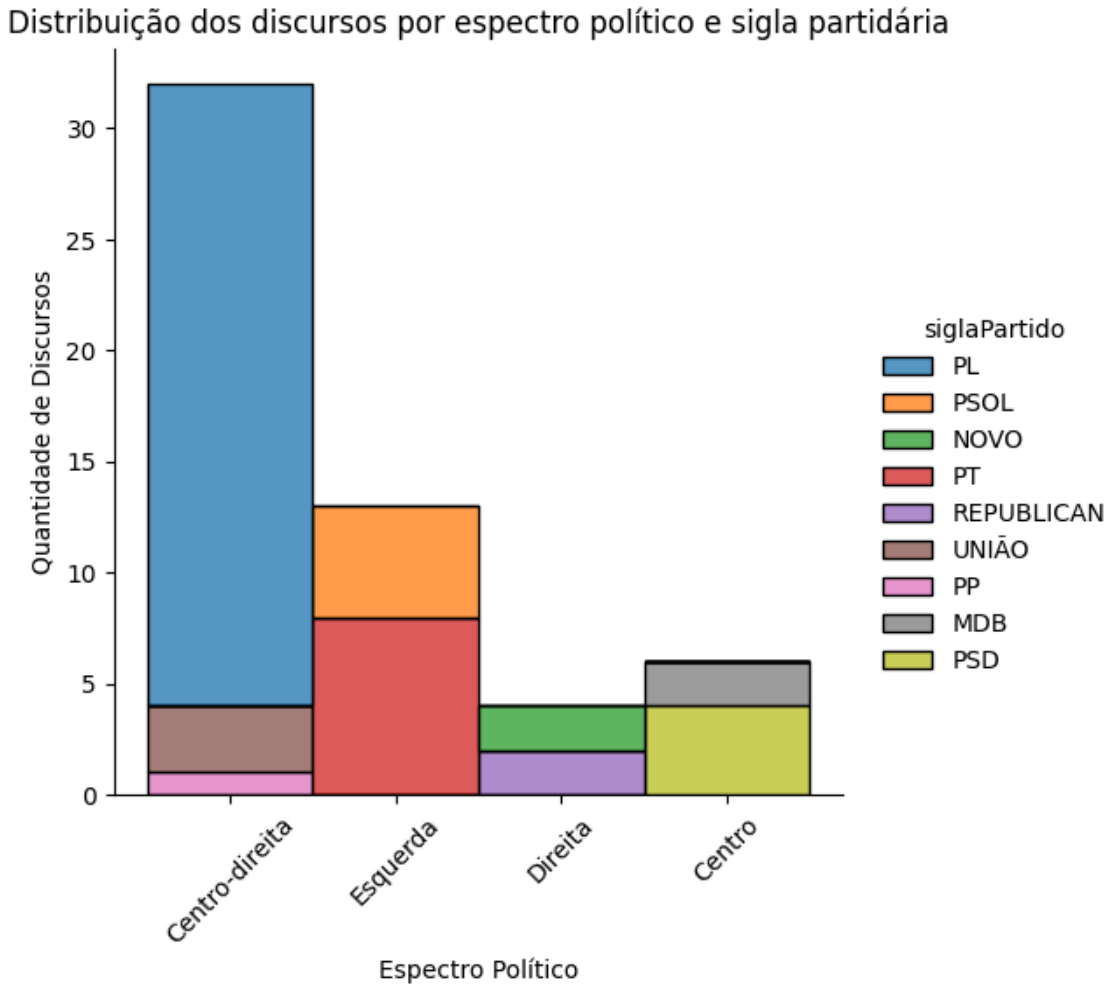
5. Dataset

O *dataset* utilizado nesta pesquisa foi coletado utilizando-se da *API* Dados Abertos, da Câmara dos Deputados (2017). Trata-se de uma coleção de *JSONs*, os quais reúnem as seguintes informações:

- ***ID* de discurso:** *GUID* que identifica cada discurso no *dataset*;
- ***ID* do evento:** um identificador numérico para um evento ocorrido na Câmara;
- **Descrição do tipo de evento:** uma breve descrição sobre o evento;
- ***ID* do deputado:** identificador numérico para um parlamentar na base de dados da Câmara;
- **Nome do Deputado**: nome completo do parlamentar;
- ***URI* Deputado:** uma *URL* que redireciona a uma página que sumariza informações sobre o deputado que discursou;
- **Sigla UF:** Unidade Federativa a qual o deputado pertence;
- ***ID* Legislatura:** identificador numérico da legislatura vigente à época na base de dados da Câmara;
- **Sigla Partido:** sigla partidária a qual o deputado está filiado;
- **Tipo Discurso:** uma categoria dada aos discursos na Câmara, indicando o momento em que tal discurso foi proferido, como também o objetivo;
- **Fase do Evento:** em que momento da sessão tal discurso foi proferido;
- **Data e Hora de Início:** em que momento o discurso se iniciou;
- ***Keywords*:** palavras-chave sobre o que foi tratado no discurso;
- **Sumário:** resumo oferecido pela Câmara sobre o discurso proferido;
- **Transcrição:** discurso transcrito, com ou sem revisão do orador (deputado);
- **Espectro Político:** espectro político no qual o partido do parlamentar se enquadra;
- **Posicionamento Ideológico:** posicionamento ideológico no qual o partido do deputado se enquadra.

Ao todo, foram 55 discursos coletados, que compreendem o período de 2023 a 2025, cujo tema centra-se na posse e porte de armas de fogo. Foi possível observar a presença de 9 partidos diferentes, situados em 4 espectros políticos e 3 posicionamentos ideológicos distintos. Destes discursos, nesta faixa temporal, mais de 30 representam discursos de parlamentares filiados a partidos de direita (4) e centro-direita (32), enquanto aqueles filiados a siglas partidárias de esquerda e centro representaram, respectivamente, 13 e 6 dos discursos reunidos no *dataset*.

Figura 1. Distribuição dos discursos por espectro político e sigla partidária



Fonte: autoria própria

6. Metodologia

A presente pesquisa busca investigar em que medida os discursos refletem proximidade ou divergência entre deputados de espectros políticos semelhantes, considerando a influência de fatores contextuais e ideológicos. Para tanto, serão empregadas estratégias de visualização e análise vetorial, utilizando as representações discursivas metrificadas pelo *LLM* (*embeddings*), a fim de conferir dimensionalidade à análise interpretativa. Assim, será possível não apenas validar a qualidade dos *embeddings* gerados pelo *XLM-R*, mas observar o grau de coerência ideológica entre os discursos parlamentares.

Dada a necessidade de estudar a viabilidade de analisar discursos em língua portuguesa por meio da utilização de modelos de IA, neste trabalho, optou-se pela utilização do modelo *XLM-R*. A motivação da escolha por tal *LLM* reside não apenas em

seu foco voltado à análise textual, mas também em sua característica multilíngue e desempenho notável em tarefas PLN, nas quais estabeleceu novos patamares de estado da arte [Conneau et al. 2020].

Para o fornecimento de dados ao *XLM-R*, foi desenvolvido um projeto de minerador de texto, com o qual, a partir da determinação de um processo de coleta e cruzamento de informações, se tornou possível construir uma base de dados para a pesquisa. Trata-se de uma coleção de discursos proferidos e outras informações pertinentes, que serão apresentadas a seguir, por deputados a respeito de um tema discutido no parlamento, que entrou em pauta como projeto de lei (PL), dentro de uma faixa temporal, dada em anos. Por estar atrelado a uma Iniciação Científica ainda não publicada, o código-fonte do minerador de texto se encontra temporariamente restrito. No entanto, até que seja disponibilizado publicamente, a sua arquitetura é mostrada no esquema da Figura 3.

Para compor a base de dados, a fonte de toda informação extraída para esta finalidade provém da *API* Dados Abertos, da Câmara dos Deputados do Brasil. Lançada em 2017, em substituição à primeira versão, de 2011, trata-se de uma *API RESTful* cujo objetivo consiste em disponibilizar publicamente informações pertinentes à Câmara [Câmara Dos Deputados 2017]. Nesta pesquisa, os dados foram coletados, por meio dos endpoints da *API*, no formato *JSON*.

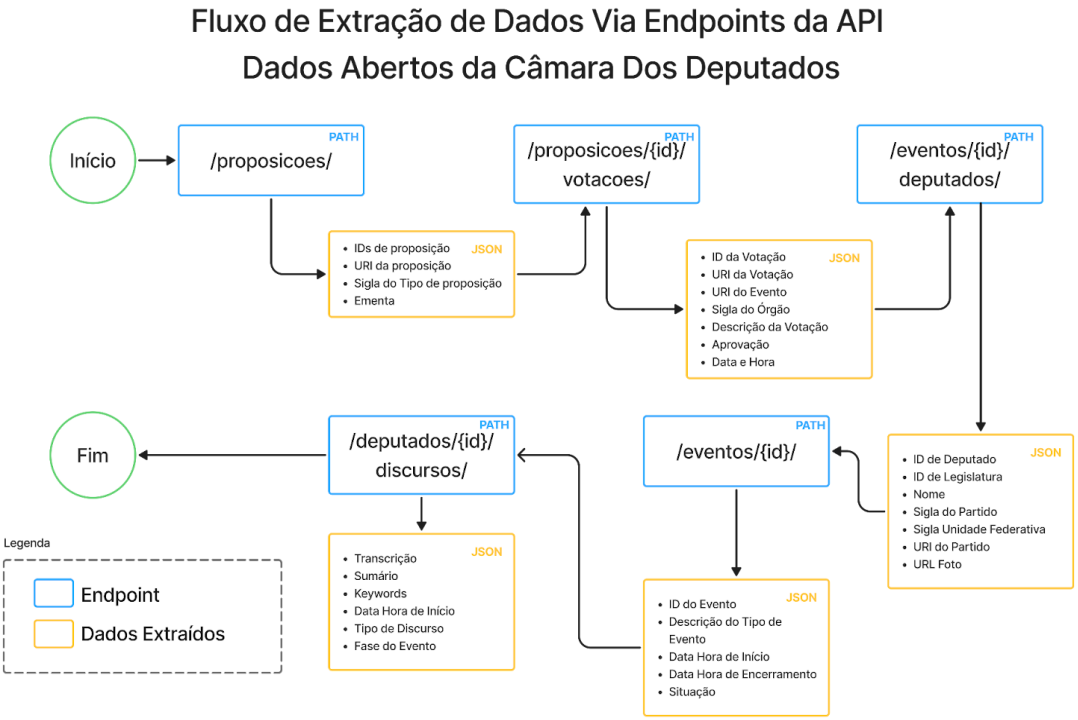
Para a construção do minerador de texto, foi empregada a biblioteca *Apache Beam*, que viabiliza a criação de pipelines de processamento paralelo de dados, permitindo a definição de fluxos bem estruturados para o processamento eficiente de um grande número de informações [Apache 2025]. Com isso, tornou-se viável coletar discursos parlamentares relacionados a um mesmo tema, situados em um período específico. Assim, foi definido, em código *Python* e utilizando-se da biblioteca, um fluxo de extração de dados via endpoints da *API* Dados Abertos, possibilitando a coleta de discursos tanto com base no tema do PL discutido no Plenário quanto em uma faixa temporal definida em anos.

No entanto, a extração destes discursos não constitui uma tarefa trivial: na *API*, cada coleção de discursos só pode ser obtida via endpoint - a saber, o caminho `/deputados/{id}/discursos` -, sendo necessário conhecer o identificador numérico (id) atribuído a cada parlamentar dentro da base de dados da Câmara. Adicionalmente, é possível filtrar o retorno dos dados por meio da definição, nos headers da requisição, de valores no padrão ISO 8601 (AAAA-MM-DD) - correspondente a ano, mês e dia, respectivamente [International Organization For Standardization s.d.] - para os parâmetros `dataInicio` e `dataFim`, que delimitam o recorte temporal em que os discursos se situam [Câmara dos Deputados 2017].

Neste contexto, se fizeram necessários o estudo e o estabelecimento de um fluxo de extração de dados por meio de consultas a diferentes endpoints, de modo a permitir conhecer não somente os projetos de lei (PLs) que possuíam o tema selecionado como pauta central, como também quais parlamentares participaram de suas respectivas

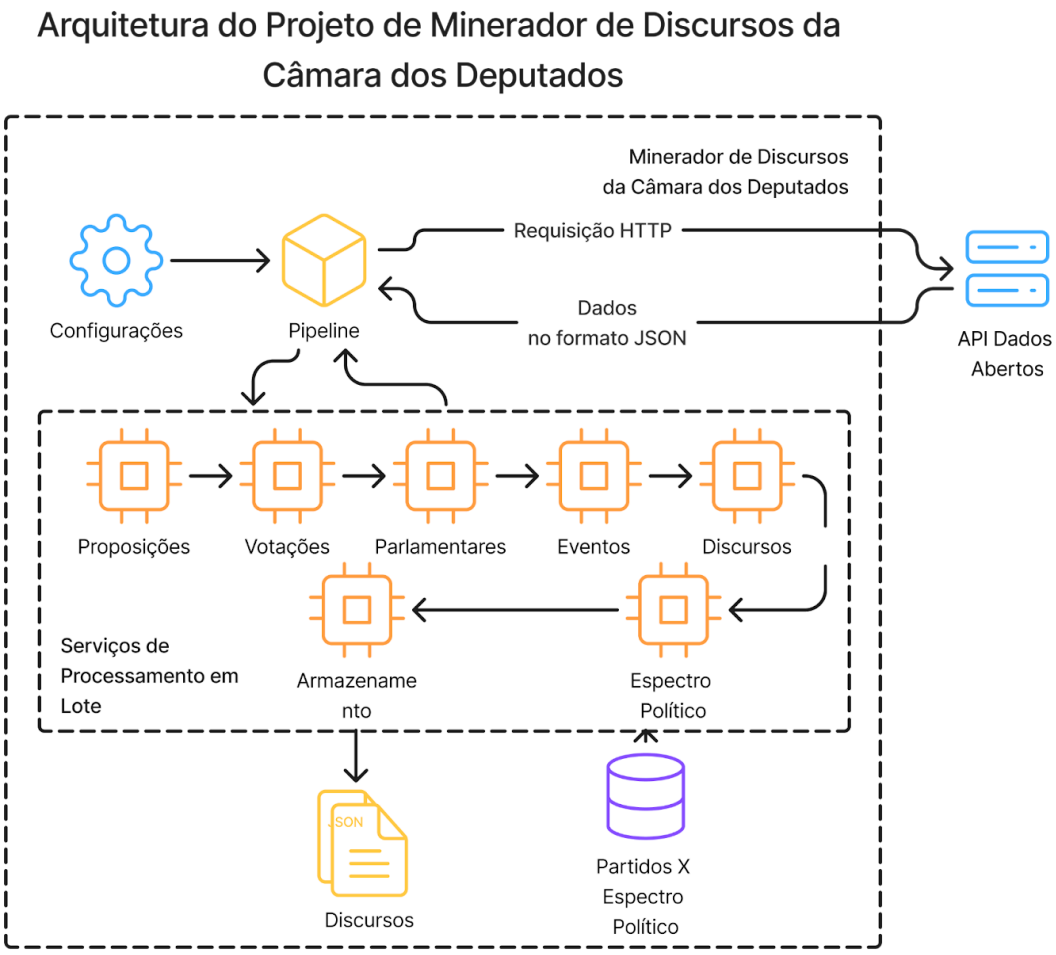
votações e discussões em Plenário. Diante desta necessidade, foi estabelecido um plano, como evidenciado na Figura 2, que foi posto em ação por meio da codificação de minerador de texto em *Python*, que se utiliza da biblioteca *Apache Beam* para auxiliar na construção do fluxo de extração de discursos parlamentares. A arquitetura do projeto do minerador de discursos é apresentada na Figura 3. Nele, para relacionar cada sigla partidária com uma classificação de espectro político, que indica afinidade ideológica, foi utilizado um mapeamento fornecido pela *Wikipedia* (2025).

Figura 2. Fluxo de extração de dados da API Dados Abertos



Fonte: autoria própria

Figura 3. Arquitetura Do Minerador De Discursos da Câmara dos Deputados



Fonte: autoria própria

Neste trabalho, a fim de favorecer a observância de grupos discursivos distintos com base no espectro político, foi escolhido o tema posse e porte de armas de fogo, assunto cujo a adoção de um posicionamento determina também a afinidade ideológica do sujeito que opina sobre, como apontado pela pesquisa IPSOS-IPEC (2025). Para a coleta dos dados, conforme o plano supracitado, definiu-se o período correspondente ao governo vigente do então presidente da república Luiz Inácio Lula da Silva, em um período que compreende o ano de 2023 a março de 2025.

Para possibilitar a análise visual, metrificada e interpretativa da distribuição semântica dos discursos, foi aplicada a técnica de redução de dimensionalidade *t-SNE*, que permite representar os vetores de *embeddings* em duas dimensões, permitindo a sua visualização em um espaço bidimensional. Para tal, foi utilizada a biblioteca *Seaborn* [Waskom, 2021]. A distância utilizada para o cálculo do *t-SNE* foi a similaridade por cosseno. Assim, tornou-se possível, a partir do posicionamento dos discursos no plano

visual, observar a proximidade do posicionamento entre discursos similares, sendo possível diferenciá-los pelo partido ao qual o parlamentar faz parte, ou mesmo a partir do espectro político ao qual a sigla partidária pertence.

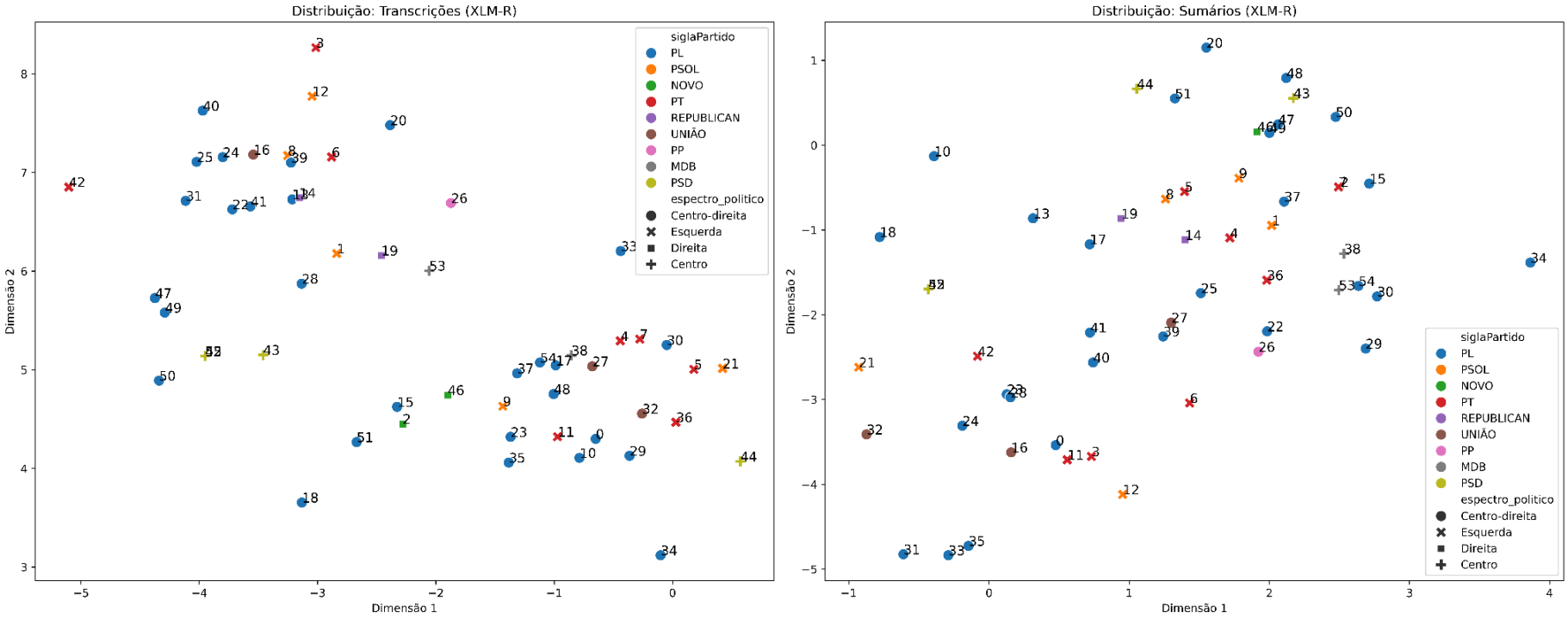
Além disso, com o objetivo de validar os resultados obtidos das classificações realizadas pelo modelo, foi utilizada, como proposto por Lima (2024), a distância de cosseno para o cálculo da similaridade entre discursos. Deste modo, torna-se possível medir a proximidade entre os vetores de *embeddings*, evidenciando a proximidade de discursos ideologicamente semelhantes com base nos dados coletados para o estudo. Para visualizar este cálculo de proximidade, foram utilizados mapas de calor, nos quais as cores mais próximas ao valor 1,0 indicam proximidade semântica, enquanto as mais próximas de 0,0 apontam para o distanciamento entre os conteúdos dos textos comparados.

Foram analisados dois tipos de texto: transcrições, que descrevem a fala integral do deputado, e sumários, resumos ofertados pela Câmara dos Deputados a respeito do conteúdo versado pelos deputados. Sobre o primeiro, foi realizada, como pré-processamento, a remoção do identificador do deputado, com o objetivo de reduzir o viés na geração de *embeddings* via *XLM-R*. O identificador que antecede a fala do deputado, a exemplo, segue o padrão: “O(A) SR(A). NOME DEPUTADO (Bloco/SIGLA PARTIDO - UF. Sem/Com revisão do(a) orador(a).) -” [Câmara Dos Deputados 2017].

7. Resultados

Diante da metodologia exposta, foram obtidas distribuições a partir da aplicação do *t-SNE* sobre os *embeddings* (Figura 4) relativos às transcrições dos discursos e seus respectivos sumários, calculados pelo modelo *XLM-R*.

Figura 4. Distribuição dos textos via *t-SNE*



Fonte: autoria própria

Mediante aplicação da técnica *t-SNE*, foram obtidas duas visualizações, que contrastam na determinação das proximidades entre os diferentes tipos de texto no plano bidimensional. Para entender a natureza da determinação, serão apresentados os discursos de índice 5 e 8, cujas transcrições são dessemelhantes, ao passo que seus respectivos sumários se assemelham. O primeiro, proferido pela deputada Erika Kokay, do PT, diz:

> A SRA. ERIKA KOKAY (Bloco/PT - DF. Sem revisão da oradora.) - A Polícia Federal, em uma operação conjunta entre Brasil, Paraguai e Estados Unidos, apreendeu uma quadrilha que, em 3 anos, movimentou por volta de 1,2 bilhão de reais e disponibilizou 43 mil fuzis e pistolas que chegaram às mãos de criminosos. **Liberar as armas significa colocar mais armas nas mãos dos criminosos**. Aliás, quem defende as forças de segurança do nosso País deveria defender que elas detivessem o monopólio das armas. Dizer que as mulheres têm que defender a si próprias e tirar a responsabilidade do Estado é coisa de quem é vassalo ou sabujo de um Presidente que, por diversas vezes, expressou, sem nenhum pudor, o seu sexismo e a sua misoginia. Nós estamos aqui, nesta noite de hoje, para aprovar uma pauta que nos permita fazer entregas à sociedade sobre mecanismos e instrumentos para se avançar no enfrentamento da violência contra as mulheres. Os 21 dias de ativismo se encerram no próximo dia 10, que é o Dia Internacional dos Direitos Humanos. Esse projeto vem nessa construção, com o intuito de aplicar um formulário ou colher das vítimas de violência em todo o País dados sobre o histórico da violência. É fundamental que nós tenhamos o histórico da violência, que nós saibamos como a violência se dá na sua crueldade, deixando as suas marcas na pele e também na alma de tantas mulheres e do conjunto da sociedade. Uma sociedade que sofre tantas violências contra as mulheres é uma sociedade que deixa marcas; marcas no seu imaginário, marcas na sua cultura, marcas na sua forma de existência. Aliás, neste País, mulheres já foram marcadas, como se gado fossem, pelos homens que se sentiam seus proprietários. Nós já tivemos o Estatuto da Mulher Casada, que a considerava incapaz de ter determinadas deliberações sobre a sua própria vida. É disso que nós estamos falando. Essa proposição, com esse histórico, permite, a partir dos dados, dados estes que foram pisoteados e negados no negacionismo estrutural que foi a tônica do Governo do hoje inelegível Jair Bolsonaro, um efetivo combate à violência. E assim é o projeto [Câmara Dos Deputados 2017].

O segundo, proferido pela deputada Fernanda Melchiona, do PSOL, diz:

> A SRA. FERNANDA MELCHIONNA (Bloco/PSOL - RS. Pela ordem. Sem revisão da oradora.) - Presidente, a Federação PSOL REDE, evidentemente, orienta "sim". O projeto inclui o TCU nesse organismo internacional. Nós achamos que é um projeto tranquilo. Quero também registrar, Presidente, que, para o próximo requerimento de urgência, a nossa orientação é "não". Não sei quanto tempo a votação durará, então já deixo consignado que, em relação ao Reporto, nós não somos a favor da renovação de uma isenção fiscal sem que haja garantia de contrapartidas e empregos. E eu queria dizer aos Deputados da extrema Direita que quem insistiu em votar o requerimento de urgência desse PDL que flexibiliza o porte de armas, que tem gerado tanta violência, como aumento de feminicídio, acidentes com crianças, infelizmente... Hoje, qualquer briga de trânsito pode eventualmente terminar em tiroteio. **Enfim, liberar armas significa mais violência**. Precisamos é de

> políticas públicas efetivas, que enfrentem o problema gravíssimo da segurança pública no Brasil. Eles pediram o regime de urgência para a votação e perderam. O choro é livre [Câmara Dos Deputados 2017].

Relativo aos sumários, já qualificados enquanto próximos pelo *t-SNE*, os mesmos, respectivamente, dizem:

> A Deputada discutiu o Projeto de Lei nº 1.213, de 2022, que altera a Lei nº 14.149, de 5 de maio de 2021, para dispor sobre a aplicação obrigatória do Formulário Nacional de Avaliação de Risco no âmbito das Polícias Civis dos Estados e do Distrito Federal. **Além disso, destacou a importância de focar em medidas efetivas para enfrentar a violência contra as mulheres, criticando a ideia de armar as mulheres como solução para a segurança**. Mencionou a operação conjunta entre Brasil, Paraguai e Estados Unidos, que apreendeu uma quadrilha envolvida na movimentação de bilhões de reais e disponibilização de armas para criminosos [Câmara Dos Deputados 2017].

> A Deputada orientou a bancada na votação do Requerimento de Urgência para apreciação do Projeto de Lei nº 5.711, de 2023, que dispõe sobre a atuação do Tribunal de Contas da União como membro do Conselho de Auditores da Organização das Nações Unidas. Além disso, antecipou a orientação para o suposto próximo requerimento de urgência para apreciação do projeto de lei sobre o Reporto, argumentando contra isenções fiscais sem contrapartidas. **Criticou a extrema Direita por promover a flexibilização do porte de armas, apontando consequências como aumento da violência, feminicídios e acidentes** [Câmara Dos Deputados 2017].

Embora concordem em opinião, evidenciando a coerência ideológica a qual pretende-se identificar, as falas e os resumos foram posicionados de maneiras distintas, quanto à similaridade de seu conteúdo semântico, analisado pelo *XLM-R*. A partir da análise textual, torna-se evidente que, ao resumir cada opinião, evidenciando a concordância entre as duas deputadas quanto a possibilidade de aumento do número de feminicídios - e da violência, de modo geral - ao flexibilizar o porte e porte de armas, os sumários foram posicionados de maneira mais precisa, por extraírem o conteúdo central das falas. As transcrições, como observado, por se valerem de construções textuais complexas, confundiram o modelo quanto ao tema central da fala.

Entretanto, a proximidade ao extrair o conteúdo central da fala não se apresenta apenas entre deputados que ocupam o mesmo espectro político, o discurso proferido pela deputada Mariana Carvalho, do REPUBLICANOS, partido pertencente à direita (diametralmente oposto ao espectro abordado anteriormente), é também posicionado próximo aos dois supracitados. A deputada, que também versa sobre o tema posse e porte de armas de fogo relacionado à violência contra a mulher, diz:

> A SRA. MARIANA CARVALHO (Bloco/REPUBLICANOS - MA. Pela ordem. Sem revisão da oradora.) - Sra. Presidente, parabéns pela condução dos trabalhos. É uma honra estar participando hoje da pauta feminina, com tantos projetos de proteção às mulheres sendo votados nesta Casa. Eu quero dar uma ênfase ao meu Estado do Maranhão, que infelizmente é o segundo Estado do Nordeste com mais agressões e tentativas de feminicídio. Na minha cidade, Imperatriz, isso tem aumentado a cada dia. Faço referência

> aqui à Patrícia Medrado, de João Lisboa, uma cidade ao lado de Imperatriz, que foi assassinada de forma trágica, brutal. E, na minha cidade, uma mulher também chamada Patrícia foi assassinada de forma cruel. Com apenas 34 anos, estava desaparecida por 9 dias e foi encontrada com um tiro na cabeça. Isso é trágico, é brutal. De fato, precisamos nos posicionar nesta Casa para que as mulheres tenham a proteção necessária. E nós, como Parlamentares, faremos a nossa parte aqui. Faço referência ao requerimento de urgência, que está previsto para ser votado hoje, para o Projeto de Decreto Legislativo nº 3, de 2023, que susta o Decreto do Desarmamento, porque eu acredito nessa pauta. **Que as mulheres, se assim quiserem e passarem por todos os processos necessários, tenham direito, sim, a possuir uma arma e a se defender, e tenham o direito à legítima defesa.** Então, eu faço aqui essa referência de que nós lutaremos pela proteção das mulheres [Câmara Dos Deputados 2017].

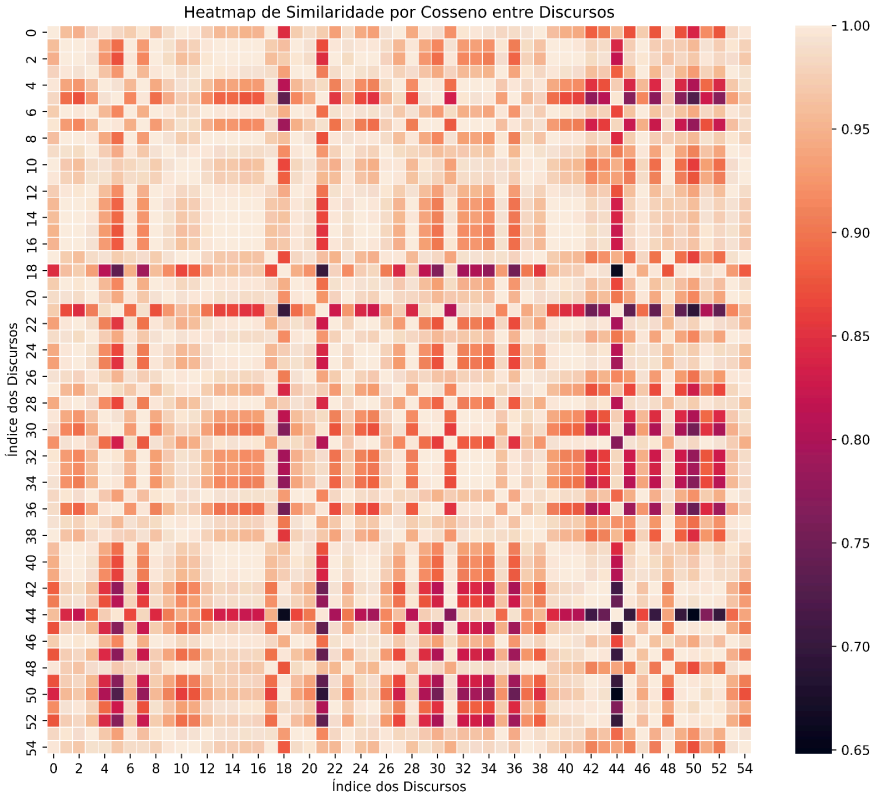
O sumário referente ao discurso o resume da seguinte maneira:

> A Deputada destacou a importância das proposições de proteção às mulheres em votação na Casa. **Ressaltou a trágica situação de violência contra mulheres no Estado do Maranhão, mencionando casos específicos de feminicídio.** Por fim, apoiou o requerimento de urgência para a votação do Projeto de Decreto Legislativo nº 3, de 2023, que susta o Decreto nº 11.366, de 1º de janeiro 2023, que limita o registro e a aquisição de armas de fogos, defendendo o direito das mulheres à legítima defesa [Câmara Dos Deputados 2017].

Assim, torna-se evidente o motivo pelo qual sumário referente à fala da deputada do partido REPUBLICANOS foi posicionado próximo às deputadas do PT e PSOL, pois se utiliza de termos iguais ou semelhantes aos destacados nos resumos apresentados anteriormente.

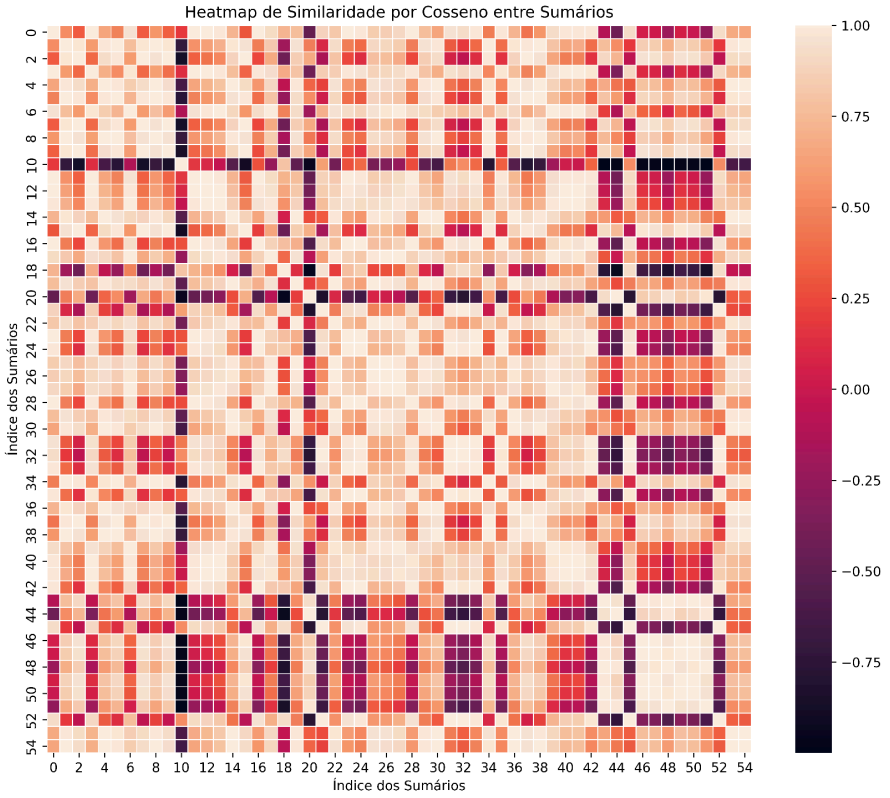
Ao submeter os *embeddings* ao cálculo da similaridade de cosseno, foram obtidos os mapeamentos que elucidam a proximidade entre os discursos integrais (Figura 5) e os resumos relacionados a estes discursos (Figura 6).

Figura 5. *Heatmap* de similaridade de cosseno: transcrições



Fonte: autoria própria

Figura 6. *Heatmap* de similaridade de cosseno: sumários



Fonte: autoria própria

Contrastando com o mapa de calor referente aos sumários, percebe-se ampla similaridade no que aborda as transcrições integrais. Isto, de maneira semelhante ao analisado, pode ser explicado pela utilização de termos comuns, pertinentes ao tema posse e porte de armas de fogo, como “violência” e “armas”, além do emprego de recursos de fala próprios do contexto político. Assim, mesmo que grupos sejam identificados a partir através do cálculo das proximidades pelo *t-SNE*, as transcrições integrais, de maneira geral, não apresentam diferenças semânticas significativas.

Por outro lado, percebe-se a capacidade do modelo em diferenciar com mais clareza o conteúdo presente nos sumários. De fato, analisando o conteúdo do discurso de índice 21, do deputado Henrique Vieira, do PSOL, percebe-se que seu discurso, qualificado como similar ao de índice 5, da deputada Erika Kokay, do PT, possui seu sumário entendido como dessemelhante ao da deputada. Em seu discurso, o deputado diz:

> O SR. PASTOR HENRIQUE VIEIRA (Bloco/PSOL - RJ. Sem revisão do orador.) - Boa tarde a todos e a todas! **Nós temos uma tarefa histórica, que**

> **é defender a democracia e enfrentar a política do ódio, a política da violência, a política que é a expressão da extrema direita no Brasil.** Eu estou convicto de que essa é uma tarefa fundamental do Governo Lula, companheiro Chico Alencar, e uma tarefa fundamental do nosso tempo histórico, porque não estamos falando de um bom debate de ideias e divergências que qualificam e amadurecem a democracia, estamos falando de uma política que é baseada na violência, no ódio, na intimidação, no negacionismo e no fundamentalismo religioso. Por que eu estou trazendo esse tema? **Enfrentar a extrema direita tem a ver, por exemplo, com o que o Governo está fazendo com uma política de controle e de restrição sobre vendas e circulação de armas no Brasil, porque a lógica da extrema direita é basicamente milicializar a sociedade brasileira.** Estamos avançando nesse ponto. Mas há outro ponto fundamental para a democracia no Brasil, que é o projeto de lei cujo regime de urgência vai ser votado hoje. Nós precisamos enfrentar as fake news como política. A extrema direita inclusive chegou à Presidência por meio de crimes na Internet, com disseminação de mentira e incitação à violência. Regular não é censurar. Aquilo que é crime no cotidiano, nas relações, também tem que ser crime na Internet. Não pode ser uma terra sem critério, onde a **violência** é estimulada. Nós, por exemplo, estamos debatendo com muita seriedade, com muita gravidade, a questão dos ataques nas escolas, e tudo indica que muitos ataques começam a ser organizados, orquestrados e estimulados justamente nas redes sociais. É preciso haver critério, com transparência, com cuidado, com a responsabilidade das plataformas, com uma entidade de regulação autônoma, que não vai ser aparelhada por nenhum Governo. Mas debater a regulação da Internet não tem a ver com censura, tem a ver inclusive com o combate à política do ódio, às fake news, ao extremismo e a grupos violentos que se organizam através da Internet. São muitas as frentes para defendermos a democracia, enfrentarmos a política do ódio, reduzirmos a desigualdade e construirmos um País de justiça e de solidariedade. Não é PL da censura. Tem medo quem se utiliza da Internet para praticar crime e incitar a violência [Câmara Dos Deputados 2017].

O sumário da fala do deputado diz:

> Desafio do governo petista de defesa da democracia e combate à política de ódio e violência. **Importância das restrições impostas à comercialização de armas de fogo e munições. Defesa de aprovação do Projeto de Lei nº 2.630, de 2020**, sobre a instituição da Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet (**projeto de lei das fake news**) [Câmara Dos Deputados 2017].

Ao analisar os excertos (transcrições e sumários de índices 5 e 21, dos deputados Erika Kokay e Pastor Henrique Vieira, respectivamente), percebe-se que suas transcrições empregam em abundância o termo “violência”, o que justifica o entendimento do modelo enquanto textos semelhantes. Os sumários, por outro lado, apesar de versarem sobre o tema de posse e porte de armas de fogo, destacam pontos distintos. O primeiro, da deputada do PT, centra-se na questão do feminicídio e da proteção da mulher, enquanto o do pastor menciona diretamente o controle da comercialização de armas, como também destaca a defesa no combate à desinformação na *internet*, tema, inclusive, central em seu discurso.

Ao observar a similaridade entre os discursos previamente analisados, percebe-se que o *heatmap* corrobora com a identificação das similaridades apresentada pela distribuição fornecida via *t-SNE*.

8. Conclusão

O fenômeno da polarização política destacou a importância do debate no âmbito democrático. Em um cenário cuja pluralidade de pensamentos é crítica para a busca da satisfação dos anseios de uma sociedade, a configuração dicotômica crescente nas ideias discutidas se mostra como um risco à manutenção deste regime e, sobretudo, no alcance de resoluções aos problemas enfrentados pela sociedade.

Este trabalho, se propõe a empregar o modelo *XLM-R* na análise discursiva de membros da Câmara dos Deputados do Brasil, de modo a auxiliar na identificação da coerência ideológica entre os diferentes espectros políticos aos quais pertencem as siglas partidárias do parlamento nacional.

Mediante aplicação da metodologia proposta, se torna possível concluir que o modelo *XLM-R* é capaz de identificar a semelhança semântica entre textos. No entanto, trata-se de uma identificação limitada, que não se adequa totalmente ao esperado, uma vez que o mesmo agrupa, a depender da estrutura retórica e do emprego de termos próximos ou idênticos no texto, discursos de opiniões distintas, enquanto separa textos que, embora concordem sobre o mesmo tópico, apresentam construções textuais discrepantes.

Neste sentido, cria-se espaço para a exploração de métodos de mineração de argumentos em estudos posteriores. Assim, com o objetivo de padronizar a representação das falas, espera-se facilitar a identificação da opinião central de cada parlamentar, de modo a favorecer a identificação de grupos de coerência ideológica.

9. *Links* e ambientes

- *Github*: https://github.com/JPedroRodrigues/political-speech-xlm-r
- Vídeo no *Youtube*: https://youtu.be/Xpk5WMcZGe0

Referências

Apache Software Foundation. (s.d.) “Apache Beam: unified programming model”, https://beam.apache.org/, Maio.

Câmara Dos Deputados. (2017) “Dados Abertos – API da Câmara dos Deputados”, https://dadosabertos.camara.leg.br/swagger/api.html, Maio.

Conneau, A. et al. (2020) “Unsupervised Cross-lingual Representation Learning at Scale”, arXiv preprint arXiv:1911.02116.

International Organization For Standardization. (s.d.) “ISO 8601 – Date and time format”, https://www.iso.org/iso-8601-date-and-time-format.html, Agosto.

IPSOS-IPEC (2025) “ÍNDICE DE CONSERVADORISMO BRASILEIRO”, https://www.ipsos.com/sites/default/files/ct/publication/documents/2025-08/Ipsos-indice_de_conservadorismo_1.pdf, Julho.

Johnstone, B. and Andrus, J. (2024) Discourse analysis, 3rd edition, John Wiley & Sons, Hoboken.

Kawakatsu, M. et al. (2021) “Interindividual cooperation mediated by partisanship complicates Madison’s cure for ‘mischiefs of faction’”, Proceedings of the National Academy of Sciences, v. 118, n. 50.

Levin, S. A., Milner, H. V. and Perrings, C. (2021) “The dynamics of political polarization”, Proceedings of the National Academy of Sciences, v. 118, n. 50, p. e2116950118.

Lima, W. P. C. (2024) “Uma análise de partidos políticos baseada em discursos no Congresso Nacional Brasileiro”, Tese de Mestrado, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Rio de Janeiro, https://eic.cefet-rj.br/ppcic/wp-content/uploads/2023/11/62-Willian-Pitter-Cardoso-Lima.pdf, Agosto.

Martínez-Guillem, S. and Toula, C. (Eds.) (2020) Critical Discourse Studies and/in Communication: Theories, Methodologies, and Pedagogies at the Intersections, 1st edition, Routledge, New York. DOI: https://doi.org/10.4324/9781003050353.

Vasconcelos, V. V., Constantino, S. M., Dannenberg, A., Lumkowsky, M., Weber, E. and Levin, S. (2021) “Segregation and clustering of preferences erode socially beneficial coordination”, Proceedings of the National Academy of Sciences of the United States of America, v. 118, n. 50, e2102153118.

Vaswani, A. et al. (2017) “Attention is all you need”, In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17), Curran Associates, Red Hook, NY, p. 6000-6010.

Waskom, M. L. (2021) “seaborn: statistical data visualization”, Journal of Open Source Software, v. 6, n. 60, 3021. DOI: https://doi.org/10.21105/joss.03021.

Wikipedia (2025) “Posicionamentos dos partidos brasileiros”, https://pt.wikipedia.org/wiki/Posicionamentos_dos_partidos_brasileiros, Agosto.

Wynter, A. et al. (2023) “An evaluation on large language model outputs: Discourse and memorization”, Natural Language Processing Journal, v. 4, p. 100024.