Aplicação do Modelo Multilíngue XLM-RoBERTa na Classificação 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 Brazilian Chamber of Deputies members on the same 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 limits the capacity to seek solutions to the problems faced by society. The justification focuses on verifying the capacity of specialized Natural Language Processing models to classify political speeches. Thus, the intention is to create a tool that assists in weighing the debate amidst a polarized environment.*

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.*

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, que serão avaliados por meio da cálculo da distância 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 mundialmente. A desenvoltura de uma ferramenta capaz de analisar a situação do debate proferido em um corpo parlamentar denuncia a qualidade da discussão, como também o potencial da mesma para o alcance de 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

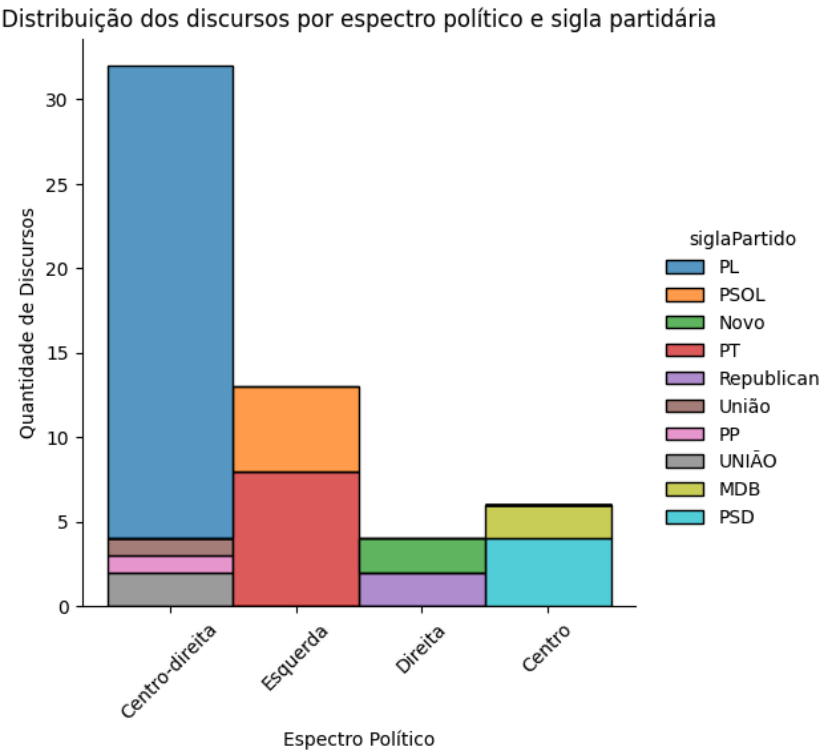
O *dataset* utilizado nesta pesquisa foi coletado utilizando-se da *API* Dados Abertos, da Câmara dos Deputados (2017). Trata-se de uma coleção de *JSONs*, os quais reúnem as seguintes informações:

- ***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 10 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.

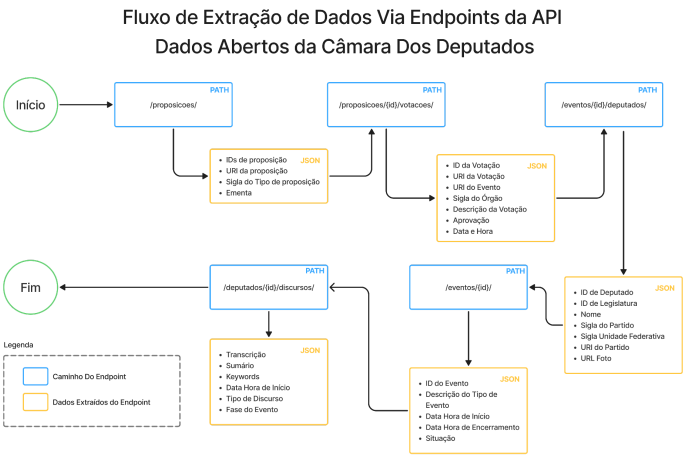
Para compor a base de dados, a fonte de toda informação extraída para esta finalidade provém da *API* Dados Abertos, da Câmara dos Deputados do Brasil. Lançada em 2017, em substituição à primeira versão, de 2011, trata-se de uma *API RESTful* cujo objetivo consiste em disponibilizar publicamente informações 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].

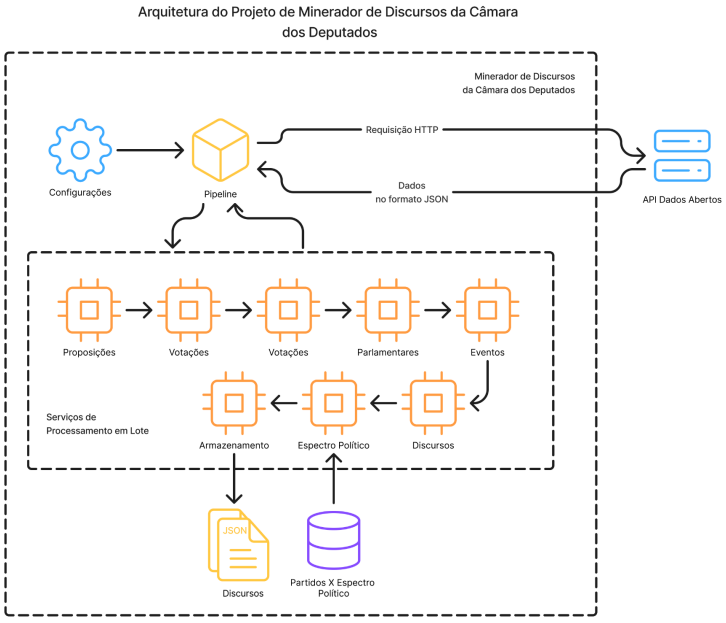
Neste contexto, se fizeram necessários o estudo e o estabelecimento de um fluxo de extração de dados por meio de consultas a diferentes endpoints, de modo a permitir conhecer não somente os projetos de lei (PLs) que possuíam o tema selecionado como pauta central, como também quais parlamentares participaram 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-Distributed Stochastic Neighbor Embedding* (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]. Assim, tornou-se possível, a partir do posicionamento dos discursos no plano visual, observar com clareza a presença de grupos que compartilham 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.

7. Resultados Esperados

Diante da metodologia exposta, espera-se que discursos que compartilhem o mesmo espectro político sejam classificados, a partir da distância de cosseno, como semelhantes (valores, no mapa de calor, que se aproximem de 1,0), enquanto aqueles proferidos por parlamentares de partidos diametralmente opostos em espectro sejam classificados como dessemelhantes (valores que se aproximem de 0).

8. 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.