Caracterizando Polarização nas Eleições Brasileiras de 2018 e 2022 : Uma Análise das Discussões no Reddit com um Modelo de Regressão para *Stance Detection*

Gustavo F. Cunha[1], **Ana Paula C. da Silva**[1]

[1]Departamento de Ciência da Computação – Instituto de Ciências Exatas – Universidade Federal de Minas Gerais (UFMG)

{gustavocunha,ana.coutosilva}@dcc.ufmg.br

Abstract. *The polarization in political discussions on Reddit during the Brazilian presidential elections of 2018 and 2022 was investigated using a methodology based on a machine learning model for user stance detection in discussion trees. The findings reveal an increase in polarization and a decline in diverse and moderate debates over the analyzed periods. The use of open-source models enabled the analysis of* $27,358$ *discussion trees comprising* $204,296$ *comments. The proposed polarization metrics demonstrated high correlation and consistency, validating the methodology's effectiveness. Qualitative analysis supported the quantitative findings, highlighting the formation of echo chambers and the prevalence of unilateral discussions in 2022 compared to 2018.*

Resumo. *A polarização nas discussões políticas no Reddit durante as eleições presidenciais brasileiras de 2018 e 2022 foi investigada utilizando uma metodologia baseada em um modelo de aprendizado de máquina para posicionamento dos usuários (stance detection) em árvores de discussão. Os resultados indicam o aumento da polarização e uma diminuição dos debates de caráter diverso e ameno ao longo dos períodos analisados. A utilização de modelos com código aberto possibilitou a análise de* 27.358 *árvores de discussão compostas por* 204.296 *comentários. As métricas de polarização propostas demonstraram alta correlação e coerência, validando a eficácia da metodologia. A análise qualitativa corroborou os achados quantitativos, destacando a formação de echo chambers e a predominância de discussões unilaterais em 2022, quando comparado com o ano de 2018.*

1. Introdução

Nos últimos anos, as mídias sociais ganharam importância como canal de influência de comportamento e decisão das pessoas, afetando não apenas o mundo on-line, mas também a vida real, no mundo offline [Kim et al. 2013]. Isto é especialmente evidente no Brasil, com destaque no contexto político, onde as mídias sociais se tornaram um meio efetivo para busca de informações e conhecimento sobre candidatos. Os candidatos, por sua vez, começaram também a fazer uso das plataformas digitais, objetivando disseminar suas ideias e angariar votos, além de atacar seus oponentes [Guimaraes et al. 2022, Caetano et al. 2022].

O debate político se destaca por seu aspecto controverso e cada vez mais polarizado no Brasil e no mundo, como evidenciado por [Ruediger et al. 2014] e

[Heltzel and Laurin 2020], acarretando consequências como perda da diversidade de opiniões e da pluralidade política [Rossini 2020]. Este ciclo vicioso é fortalecido constantemente, dado que conteúdos com alto nível de polarização estão frequentemente relacionados a discurso de ódio e toxicidade [Saveski 2020], além de propagação de desinformação e informações enviesadas [Vicario et al. 2019]. É sabido que a polarização é contributiva para a democracia, quando tida na medida correta, como atestam [Heltzel and Laurin 2020, Mutz 2002, Dahlberg 2007], mas pode trazer malefícios ao coletivo ao atingir níveis perigosos.

Sob esta perspectiva, a caracterização do debate político nas mídias sociais é primordial para a proposta de plataformas mais plurais, com regras eficazes de moderação para casos onde a discussão política ultrapasse os limites democráticos. Vários trabalhos na literatura apresentam este tipo de caracterização em plataformas de interação on-line, em diferentes contextos. No cenário político brasileiro, trabalhos recentes [Recuero et al. 2020, Brum et al. 2022, Franco and Pound 2022] delineiam a polarização em diferentes redes sociais que, por possuírem diferentes características e particularidades, têm metodologias, bem como resultados, mais ou menos acurados.

Neste contexto, como estudo de caso, analisamos as discussões realizadas nas comunidades do Reddit[1], uma rede social organizada em subcomunidades por áreas de interesse (*subreddits*). Nesta plataforma, os usuários discutem diferentes assuntos, através de interações do tipo postagem-comentários, chamadas de *threads*. Nas postagens e comentários também é possível que os internautas deem votos positivos - *upvotes* - e votos negativos, chamados de *downvotes*. Esta rede social foi classificada como o 19° website mais acessado do mundo em 2023 e é composta por milhares de comunidades [Britt et al. 2023]. Além disso, por possuir um limite insignificante de caracteres nas postagens (40.000), a plataforma é rica em discussões políticas, por permitir que os usuários expressem suas opiniões de maneira mais ampla. O conjunto de dados utilizado consiste em atividades de usuários (postagens e comentários) que ocorreram entre os meses de setembro e novembro de 2018 e 2022, período que engloba pré e pós-eleições no Brasil.

Nossos resultados preliminares, apresentados em [Cunha and da Silva 2024], que utilizaram o modelo GPT-4 [Achiam et al. 2023a] para a tarefa de detecção de posição (em inglês, *stance detection*), se mostraram promissores para os comentários coletados do Reddit, independentemente do tópico abordado. Os resultados evidenciaram uma tendência de aumento da polarização e diminuição do caráter de diversidade dos debates realizados por esses usuários entre os períodos analisados, bem como a formação de *echo chambers*, corroborando discussões abordadas em trabalhos de diferentes campos de pesquisa, como [Ortellado et al. 2022, Corrêa 2023, Silva 2023, Lima 2021].

Assim, neste trabalho, aprofundamos as análises apresentadas em [Cunha and da Silva 2024], utilizando a metodologia apresentada para caracterização de discussões polarizadas em mídias sociais, a qual é baseada na combinação e adaptação de técnicas prévias da literatura [Garimella et al. 2018, Alsinet et al. 2021a, de Arruda et al. 2022]. Com o objetivo de ampliar o conteúdo analisado, o GPT-4 foi substituído por um modelo de aprendizado de máquina de código aberto para a execução da tarefa de *stance detection*. A utilização de um modelo aberto possibilitou a análise

[1] http://www.reddit.com/

de um volume de dados maior e mais representativo do conjunto de dados previamente coletado, além da obtenção de maior flexibilidade e independência, uma vez que pode ser refinado para este objetivo específico, e em língua portuguesa.

A estrutura do artigo está organizada como descrito a seguir. Na Seção 2, são apresentados, de maneira não exaustiva, os principais estudos relacionados a esta pesquisa, sendo detalhados aqueles que serviram de base para a metodologia proposta. Na Seção 3, descreve-se o conjunto de dados utilizados, bem como sua modelagem matemática e como é feita a detecção de posição dos comentários e o cálculo das métricas utilizadas. Em seguida, na Seção 4, caracterizam-se os *insights* obtidos com a análise dos resultados. Ainda, na Seção 5, são listadas as principais limitações do estudo. Por fim, a Seção 6 finaliza a monografia resumindo as contribuições do trabalho e tratando de direcionamentos para sua continuidade.

2. Trabalhos Relacionados

Dada a contextualização do problema e sua relevância social, nesta seção serão descritos os trabalhos-base para a metodologia conduzida e, adicionalmente, será explorada uma seleção de trabalhos em maior consonância com o atual desenvolvimento, no que tange à utilização de modelos de aprendizado de máquina para detecção de posição, sua acurácia e aplicabilidade envolvendo linguagem natural e processamento textual e também as diferentes modalidades de entrada utilizadas para modelos com esse propósito, tendo como base referências atuais e inovadoras encontradas na literatura a este respeito.

Em [Garimella et al. 2018], os autores apresentaram uma abordagem completa para quantificar a polarização no X (então Twitter). Foram exploradas três modelagens de grafos distintas, onde as interações entre usuários representavam *retweets*, ou seguidores ou conteúdo compartilhado. Para cada um dos grafos modelados, os autores realizaram o particionamento em dois grupos opostos de discussão, utilizando o algoritmo METIS [Karypis 1997]. Em seguida, introduziram métricas como *Random Walk Controversy* (RWC) e *Betweenness* para quantificação da controvérsia entre os grupos, tendo como base propriedades matemáticas dos grafos. Para isso, foi verificado o quão provável seria a postagem de um usuário em um grupo chegar até o conhecimento de um usuário do grupo oposto. Complementarmente, foi aplicada uma análise de sentimentos ao conteúdo compartilhado e verificou-se que tópicos mais polarizados possuem uma variância maior nos sentimentos dos conteúdos quando comparados àqueles de menor polarização.

Nessa mesma linha, os pesquisadores em [Alsinet et al. 2021a] apresentaram um modelo quantitativo baseado no comportamento do usuário para avaliar o nível de polarização em diferentes tópicos de discussão no Reddit, em língua inglesa. Cada debate foi modelado como uma árvore de discussão de dois lados, na qual a raiz é um comentário inicial de uma discussão e os demais vértices são os comentários da mesma *thread*. As arestas da árvore são modeladas com base na cadeia de respostas do debate e cada comentário é submetido a uma análise de sentimentos que retorna um valor no intervalo $[-4, 4]$. A árvore é então dividida em dois lados de posicionamento, através de uma bipartição. Por fim, é proposta uma medida de polarização que passa por avaliações empíricas, cujos resultados indicam que o modelo é capaz de capturar os diferentes níveis de polarização em assuntos diversos.

Em direção complementar, o trabalho em [de Arruda et al. 2022] utilizou um con-

junto de modelos sintéticos de grafos para representar conexões em mídias sociais. Nos grafos gerados, cada usuário (vértice) possui uma opinião em relação a um tópico em discussão. Esta opinião varia no intervalo $[-1, 1]$, sendo -1 um posicionamento extremamente contrário e 1 um posicionamento a favor do debate estabelecido entre os vizinhos no grafo (que representam os relacionamentos na rede social). Os usuários são divididos em dois grupos: (i) os que possuem posicionamento negativo e (ii) os que possuem posicionamento positivo. A partir da definição da métrica *balance*, que é a razão da cardinalidade destes grupos, os autores caracterizam diferentes cenários onde a discussão entre os usuários é mais aberta a ser realizada por usuários de opiniões diversas e os casos onde os usuários se organizam de forma polarizada.

No cenário brasileiro, [Recuero et al. 2020] examinaram o papel do hiperpartidarismo e da polarização no X (Twitter) durante as eleições presidenciais de 2018, mostrando que há uma forte conexão entre polarização, hiperpartidarismo e desinformação. Em [Brum et al. 2022], os autores quantificaram a polarização política do público no contexto da pandemia de COVID-19. Similarmente, em [Franco and Pound 2022] os autores exploraram, com dados do Instagram, se fatores psicológicos poderiam contribuir para o apelo de uma figura polarizadora como, por exemplo, Jair Bolsonaro.

Já em [Alsinet et al. 2021a], os autores apresentaram um modelo quantitativo baseado no comportamento do usuário para avaliar o nível de polarização em diferentes tópicos de discussão no Reddit, em língua inglesa. Cada debate foi modelado como uma árvore de discussão de dois lados, na qual a raiz é um comentário inicial de uma discussão e os demais vértices são os comentários da mesma *thread*. As arestas da árvore são modeladas com base na cadeia de respostas do debate e cada comentário é submetido a uma análise de sentimentos que retorna um valor no intervalo $[-4, 4]$. A árvore é então dividida em dois lados de posicionamento, através de uma bipartição. Por fim, é proposta uma medida de polarização que passa por avaliações empíricas, cujos resultados indicam que o modelo é capaz de capturar os diferentes níveis de polarização em assuntos diversos.

No que tange ao uso de modelos de aprendizado de máquina para detecção de posição, destaca-se o recente trabalho [Pereira et al. 2023], o qual introduz o UstanceBR, um conjunto de dados multimodal voltado para a análise de *tweets* brasileiros, com foco na predição de posicionamento em relação a alvos específicos. A base inclui $86,8$ mil exemplos anotados, enriquecidos com metadados dos seguidos e seguidores dos autores das postagens. Os autores utilizaram modelos de linguagem pré-treinados, como o BERTimbau, que passou por ajustes finos utilizando os dados do UstanceBR. A metodologia combina texto e *features* de rede social para melhorar a acurácia na tarefa de detecção de posicionamento. Os resultados evidenciam que a integração de dados multimodais potencializa o desempenho do modelo, oferecendo avanços significativos em relação às abordagens convencionais.

Analogamente, em [Santos and Goya 2021], os pesquisadores focam no *stance detection* no X (Twitter) em questões politicamente controversas. O estudo utilizou um conjunto de dados composto por *tweets* em língua inglesa sobre temas polarizados, como mudanças climáticas, controle de armas e políticas de imigração. Para o treinamento dos modelos, características textuais, como n-grams, *embeddings* pré-treinados (como GloVe e BERT), e métricas sintáticas foram combinadas com metadados dos usuários, incluindo informações sobre seguidores, frequência de publicação e histórico de interações. Os mo-

delos testados incluíram classificadores como SVM, Random Forest e BERT ajustado. A metodologia envolveu a segmentação do conjunto de dados em classes de posicionamento favorável, contrário e neutro, utilizando técnicas de validação cruzada para avaliação. Os resultados demonstraram que a abordagem baseada em BERT ajustado obteve o melhor desempenho, com uma F1-score média de 84% nas tarefas de classificação. Além disso, a inclusão de metadados dos usuários aumentou a precisão em debates altamente polarizados, destacando a importância de informações contextuais para uma detecção mais precisa do posicionamento.

Ainda nesta direção, [Liang et al. 2024] exploraram recentemente e de maneira inovadora uma forma de *stance detection* multimodal em *tweets* utilizando um modelo de aprendizado de máquina. A técnica combina texto e imagens, e foram introduzidos cinco novos conjuntos de dados específicos baseados no X (Twitter), cada um contendo pares texto-imagem relacionados a diferentes domínios temáticos. Para abordar o desafio de integrar informações multimodais, os autores propuseram a estrutura de ajuste de *prompt* multimodal direcionada (TMPT), que utiliza *prompts* específicos para alvos, permitindo ao modelo aprender características de posicionamento extraídas tanto do texto quanto das imagens. Essa abordagem inovadora foi avaliada em comparação com métodos unimodais e multimodais tradicionais, demonstrando melhorias significativas em métricas como F1-score, com ganhos de até 15% em cenários onde texto e imagem se complementavam. Além de resultados quantitativos robustos, o TMPT destacou-se na análise qualitativa ao capturar nuances emocionais e reforços argumentativos provenientes das imagens, enquanto o texto fornecia contexto explícito. O estudo não só avançou o estado da arte na detecção de posicionamento multimodal, como também ofereceu uma base sólida para investigações futuras sobre interações entre modalidades em tópicos polarizados nas redes sociais.

Os resultados apresentados neste trabalho diferenciam-se dos demais nos aspectos a seguir. Foram combinados e adaptados os modelos de interação apresentados em [Garimella et al. 2018, Alsinet et al. 2021a, de Arruda et al. 2022]. Para a detecção de posicionamento, substituiu-se a análise de sentimentos que, segundo alguns trabalhos, não é a melhor técnica a ser utilizada [ALDayel and Magdy 2021], pelo uso de um modelo de aprendizado de máquina, a ser detalhado na seção 3.3, que tem se mostrado adequado para a tarefa, como atestado nos trabalhos referenciais relacionados. Também, intencionou-se a proposição de uma metodologia independente de contexto, plataforma e idioma para aferir a polarização ou mesmo verificar evidências de sua presença em discussões nas redes sociais.

3. Metodologia

A seguir descrevemos a coleta do conjunto de dados, a modelagem matemática das discussões, a adaptação e uso do modelo de aprendizado de máquina para *stance detection* e o detalhamento do método utilizado para mensurar o nível de polarização nos debates analisados.

3.1. Conjunto de Dados

Reddit é uma mídia social on-line multilíngue, fundada em 2005. Nesta rede, subdividida em comunidades, debates nos mais diversos contextos e dos mais variados tipos têm lugar.

Selecionamos postagens e comentários entre os meses de setembro e novembro de 2018 e 2022, no contexto das discussões sobre as eleições presidenciais brasileiras.

A seleção dos *subreddits* mais relevantes, em total de publicações para o contexto analisado, foi realizada utilizando a pesquisa por relevância na API Python do Reddit[2]. A partir dos *subreddits* de maior relevância, foram coletadas 95.933 postagens datadas de setembro a novembro de 2018 e 2022 . Um subconjunto destas postagens foi selecionado, com a presença pelo menos de uma das palavras-chave relacionadas ao contexto das eleições, tais como *eleições, política, Lula, Bolsonaro, Haddad, esquerda, direita, PT, PSL, PL*, resultando em 20.515 postagens. Por fim, retiramos também as postagens com conteúdo deletado/removido, escritas em outra língua e/ou constituídas somente por *links*. Finalmente, os comentários realizados nas 10.578 postagens restantes foram coletados. A Tabela 1 apresenta os *subreddits* selecionados, a quantidade de postagens e comentários coletados e o número de usuários envolvidos nos debates analisados.

Subreddit	**#Postagens**	**#Comentários**	**#Usuários**
r/brasil	5.895	139.651	16.273
r/BrasildoB	828	3.357	1.207
r/brasilivre	3.855	61.288	6.346
Total	**10.578**	**204.296**	**23.826**

Tabela 1. Estatísticas dos subreddits selecionados.

3.2. Modelagem Matemática

As discussões realizadas pelos usuários foram modeladas através de árvores, i.e., grafos sem ciclos. Cada postagem inicial, denominada p_j, com $0 \leq j \leq P$ é considerada um contexto de discussão e pode receber R comentários como resposta.

Cada resposta (comentário) r_i, com $0 \leq i \leq R$, associada à uma postagem p_j, gera uma árvore $T = (V, E)$[3], onde V é o conjunto da cadeia de comentários que sucedem r_i (inclusive) e E é o conjunto de arestas direcionadas (v_k, v_l) que conectam um comentário v_k (origem) à sua resposta v_l (destino). A cada uma das arestas direcionadas, associamos um peso $\sigma_{kl} \in [-1, 1]$, calculado a partir da técnica de detecção de posicionamento descrita na Seção 3.3. Resumidamente, se v_l é o comentário resposta à v_k, σ_{kl} assume valores negativos se v_l possui um posicionamento contra ao de v_k e valores positivos se o posicionamento é a favor ao de v_k. Os casos com $\sigma_{kl} = 0$ são os de posicionamento neutro, mas aqui considerados como levemente negativos.

Com os pesos definidos, é realizada a bipartição de cada árvore em dois subgrupos disjuntos de posicionamento A e B, analisando a polarização em torno de dois polos de pensamento. A raiz r_i é atribuída a um dos lados, que convencionamos, sem perda de generalidade, ser o lado A. Para os vértices v_k de nível 1, e para $\sigma_{i,k} > 0$, v_k é adicionado ao lado A e para $\sigma_{i,k} < 0$ ele é adicionado ao lado B. Para o caso de $\sigma_{i,k} = 0$, v_k é interpretado como levemente mais negativo do que positivo, assim, sendo assinalado também ao lado B. Isso porque, segundo a hipótese de [Agrawal et al. 2003], as pessoas tendem a responder mais quando discordam do que quando concordam. Também, em

[2]https://praw.readthedocs.io/

[3]Para facilitar a leitura da notação, iremos desconsiderar o índice que identifica cada árvore associada a cada resposta.

[Alsinet et al. 2021b], é dito que no Reddit, por exemplo, quando o usuário concorda com um comentário, ele pode simplesmente deixar um *upvote* sem responder, mas, por outro lado, quando discorda, é mais provável que exponha suas razões. Este processo se repete a cada novo nível da árvore, considerando pares pai-filho de comentários.

3.3. Detecção de Posição

Um dos pontos mais importantes para a análise da polarização em discussões on-line é definir o posicionamento de um usuário com base na sua resposta a um comentário. Ou seja, o objetivo é medir o quanto uma resposta concorda ou discorda de uma afirmação prévia. A partir dos posicionamentos de todas as respostas, construímos uma árvore bipartida, com dois subgrupos distintos de posicionamento.

O *score* da resposta que representa o seu posicionamento em relação ao comentário inicial varia entre $[-1, 1]$, com -1 para os casos extremamente discordantes e 1 para os casos extremamente concordantes. Este valor é aqui denominado *Stance Score*. A Figura 2 apresenta um exemplo de uma árvore bipartida a partir da definição dos posicionamentos. As arestas são coloridas em três cores: vermelho, amarelo e verde, que representam, respectivamente, peso negativo, peso igual à zero e peso positivo. Já a coloração dos vértices representa os dois lados, A e B, da discussão, de acordo com a bipartição discutida na Seção 3.2.

Experimentalmente, testamos a aplicação de uma análise de sentimentos aos comentários, com a biblioteca VADER [Hutto and Gilbert 2014], cuja adaptação para a língua portuguesa foi realizada por [Almeida 2018], que provê um score no intervalo $[-1, 1]$, onde 1 é extremamente positivo e -1 extremamente negativo. Contudo, este método não se mostrou eficaz ao aplicar a bipartição gráfica, uma vez que, analisando manualmente as árvores, percebemos que havia muitos pares de comentários claramente discordantes que estavam alocados ao mesmo lado do debate e, similarmente, comentários concordantes foram postos em lados distintos na *thread*. Conforme apresentado em [ALDayel and Magdy 2021] e pelos resultados obtidos com os testes realizados no conjunto de dados, o uso de técnicas de análise de sentimentos não se mostrou a melhor técnica para a detecção de posicionamento. Assim, em consonância com os trabalhos em [Zhang et al. 2022, Zhang et al. 2023], havíamos aplicado anteriormente - [Cunha and da Silva 2024] - o modelo GPT-4 [Achiam et al. 2023b] a uma amostra dos dados para realizar esta tarefa, a partir do uso da API da OpenAI.

A proposta do atual trabalho é substituir, no *framework* idealizado, o uso do modelo GPT-4 que, apesar de ter-se mostrado eficaz para *stance detection*, tornou-se um gargalo da metodologia, por ter diversas limitações de requisição [4]. Assim, para a extensão da análise ao universo dos dados coletados, utilizamos o modelo de aprendizado de máquina pré-treinado *Portuguese BERT base cased QA (Question Answering), finetuned on SQUAD v1.1* [Guillou 2021], baseado no BERTimbau [Souza et al. 2020a]. Apesar de ter duas camadas de treinamento em língua portuguesa, realizamos novos ajustes finos ao modelo, antes de submetê-lo à tarefa de detecção de posição, que descrevemos a seguir.

Primeiramente, com um grupo de três pessoas, foram anotados manualmente 2.481 pares de comentários-resposta reais. Cada anotador analisou cerca de 800 co-

[4] https://openai.com/api/pricing/

mentários. Previamente, foi explicado ao grupo que se tratavam de pares de comentário-resposta de *threads* sobre as eleições presidenciais brasileiras de 2018 e 2022 do Reddit. Os anotadores receberam um conjunto com mil pares, contendo postagens de todos os *subreddits* e dos dois períodos analisados. O grupo foi então orientado a atribuir, a cada tupla recebida, um *score* no intervalo $[-1, 1]$, onde o valor mais positivo indica concordância máxima da resposta com relação ao comentário anterior e o valor mais negativo indica que a réplica discorda extremamente do comentário respondido.

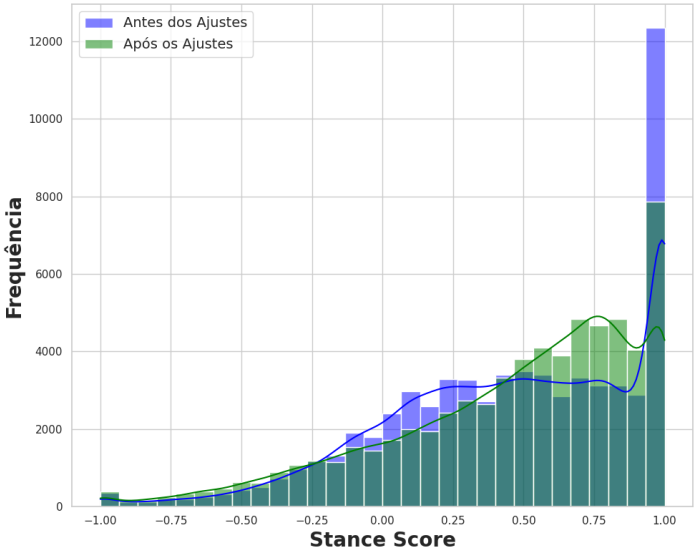
De modo complementar, visando enriquecer o conjunto de treino do modelo, foram geradas, sinteticamente, outras 3.007 tuplas. Para isso, foram redigidas, manualmente, frases iniciais sobre assuntos relacionados ao contexto da análise, incluindo *economia*, *corrupção*, *Lula*, *Bolsonaro* e *fake news*. Cada sentença inicial foi combinada com três conjuntos de respostas, - concordantes, discordantes e neutras - também redigidas manualmente e de maneira coerente ao assunto da frase inicial. Além disso, foram escritas respostas discordantes utilizando termos obscenos, para simular diferentes níveis de concordância e discordância. A atribuição do *score* foi feita aleatoriamente dentro de intervalos pré-definidos: concordantes receberam valores entre 0.3 e 1, discordantes entre -0.7 e -0.3, discordantes com palavrões entre -1 e -0.8, e neutras entre -0.09 e 0.09. Na Tabela 2, estão ilustrados alguns exemplos de tuplas geradas de modo sintético. Ao final, o modelo foi treinado com o total de 5.488 pares de comentários, considerando as tuplas reais e as sintéticas.

Sentença Inicial	**Resposta**	***Stance Score***	**Classificação**
"Ah, claro, a economia tá bombando... só que não."	*"Sim, a economia tá um desastre completo, nem preciso dizer mais nada."*	0.85	Concordante
"O Brasil nunca vai sair dessa crise econômica, não importa o que façam."	*"Vai lá e faz melhor, então"*	-0.55	Discordante
"Lula é o melhor presidente da história, as estatísticas não mentem."	*"P*** que p****, esse papo de mito é uma piada, esse cara é só mais um corrupto, p****!"*	-0.95	Extremamente Discordante
"Bolsonaro? O cara que transformou o Brasil numa maravilha... só que não."	*"Tem quem goste e quem não goste, cada um tem sua opinião."*	0.05	Neutro

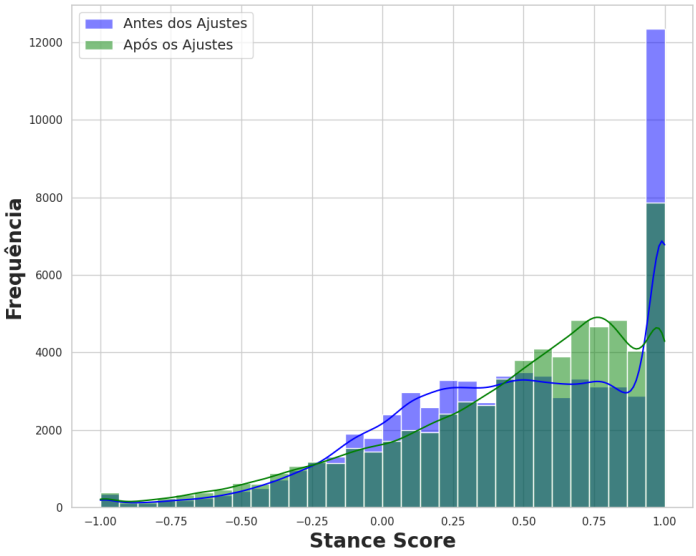
Tabela 2. Exemplos de frases sintéticas geradas com as respectivas classificações atribuídas.

A arquitetura original do modelo, projetada para tarefas de classificação, foi adaptada para regressão. Essa adaptação envolveu a configuração da camada de saída para um único *label* e a aplicação da função *torch.tanh* sobre os *logits* na última camada. Dessa forma, as previsões foram restringidas ao intervalo desejado de $[-1, 1]$. Para assegurar a consistência dos dados, o tokenizador foi ajustado para converter todas as entradas em letras minúsculas. O treinamento foi realizado utilizando um *DataLoader*, responsável

por carregar os dados em *minibatches*. O otimizador escolhido foi o *AdamW*, configurado com uma taxa de aprendizado inicial de 1×10^{-5}. Além disso, utilizou-se um agendador linear para ajustar dinamicamente a taxa de aprendizado ao longo das épocas, considerando o número total de iterações. Para medir a discrepância entre as previsões do modelo e os rótulos reais no conjunto de treino, a função de perda adotada foi o Erro Médio Quadrático (MSE), indicada [Oliveira 2022, Souza et al. 2020b]. O treinamento foi realizado em quatro épocas. Durante cada época, o modelo processou todo o conjunto de dados de treinamento. Para cada *minibatch*, os gradientes foram calculados por meio de *backpropagation* e os pesos atualizados pelo otimizador.



(a) Eleições de 2018.



(b) Eleições de 2022.

Figura 1. Comparação da distribuição do *Stance Score* nos comentários antes e depois dos ajustes do modelo.

Como mostrado na Figura 1, após o treinamento do modelo e a classificação

automática dos 204.296 comentários da base de dados, um volume representativo de instâncias foi classificado como extremamente concordantes pelo modelo, com *Stance Score* no intervalo $[0.9, 1]$. A fim de compreender melhor o que poderia ter influenciado na classificação automática, realizamos uma análise linguística e sintática detalhada de uma amostra de pares de comentário-resposta com *score* neste intervalo.

Nossa análise revelou que a maioria das classificações foi adequada, demonstrando uma alta precisão do modelo em identificar concordância entre as tuplas. Os pares corretamente classificados apresentavam clara concordância semântica, onde as respostas reforçavam ou complementavam os comentários de maneira direta ou implícita. Por exemplo, respostas como "*exatamente*"e "*pois é né*", "*pensei o mesmo*"e "*verdade*"demonstraram uma concordância direta e clara com os comentários originais. Além disso, a análise identificou o uso frequente de expressões coloquiais e informais, que foram corretamente interpretadas pelo modelo, indicando uma boa compreensão do contexto linguístico.

No entanto, uma pequena parcela de comentários (cerca de 20%) apresentou uma concordância moderada ou até questionável. Esses casos foram caracterizados principalmente pela introdução de novos pontos de vista sem abordar diretamente a questão principal do comentário, ou pelo uso de ironia e sarcasmo, que não foram adequadamente capturados pelo modelo. Para as instâncias em que se avaliou incorreta a atribuição do *Stance Score* pelo modelo, fizemos manualmente sua correção. Em seguida, foi realizada mais uma etapa de treinamento do modelo com os casos que passaram por anotação manual corretiva, agora com o valor adequado como parâmetro de entrada para os ajustes. Da Figura 1 pode-se analisar o antes e o depois da distribuição do *Stance Score* atribuído pelo modelo, anterior e posteriormente aos ajustes citados.

3.4. Quantificação da Polarização

Com a definição das árvores bipartidas de discussão, o próximo passo é a verificação da existência da polarização e sua quantificação nas discussões. Para tal, calculamos as seguintes métricas:

- ***Desvio-Padrão*** e ***Média do Stance Score***. Seguindo [Garimella et al. 2018], discursos polarizados possuem maior variância de *Stance Score* das arestas da árvore de discussão. Assim, adaptamos o ponto de corte do desvio-padrão proposto no mesmo trabalho citado, indicando como polarizadas aquelas árvores com desvio-padrão maior que 0.4375. Já a média do *Stance Score*, representa o teor geral da discussão, indicando e quantificando se as *threads* daquele debate têm aspecto mais concordante (média mais positiva) ou discordante (média mais negativa) par a par.
- ***Balance Score*** e ***Balance Side***. Estas métricas avaliam o quão balanceados estão os dois lados (bipartições) da discussão. A primeira, proposta em [de Arruda et al. 2022], calcula a razão entre o mínimo e o máximo do total dos *Stance Scores* positivos e negativos, ou seja, os pesos das arestas positivas e negativas das árvores. Na segunda, proposta neste trabalho, a cardinalidade dos conjuntos gerados pela bipartição é o total de comentários em cada lado da discussão. As medidas têm intervalo de variação $[0, 1]$, sendo 0 indicativo da predominância total de apenas um lado da discussão, indicando pouca ou nenhuma polarização

e 1 evidenciando uma discussão polarizada. A Figura 2, por exemplo, apresenta uma árvore de discussão em que o *Balance Score* é igual a 0.5 (6 arestas vermelhas e 3 verdes) e o *Balance Side* é igual a 0.66 (60% azul claro e 40% azul escuro).

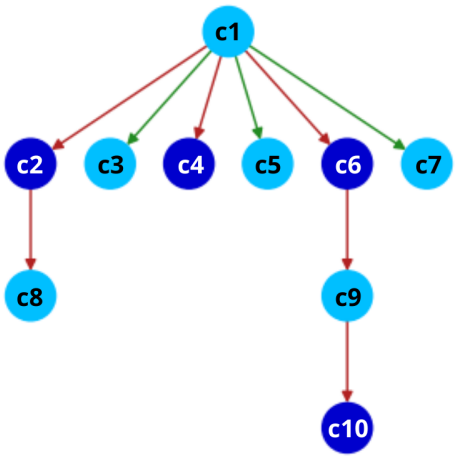


Figura 2. Exemplo de uma árvore bipartida de discussão.

- ***Média do Controversy Level***. Esta métrica, proposta neste trabalho, calcula, primeiramente, o Desvio-Padrão do *Stance Score* das arestas de saída de cada comentário. A média é calculada considerando todos os comentários da árvore em análise. Caso o comentário tenha uma única resposta, o *Controversy Level* é igual ao módulo do *Stance Score* desta resposta. No caso de uma folha, o *Stance Score* é o módulo referente à resposta dada ao comentário de nível superior. Varia no intervalo $[0, 1]$, e valores maiores apontam discussões controversas, com a presença de polarização.

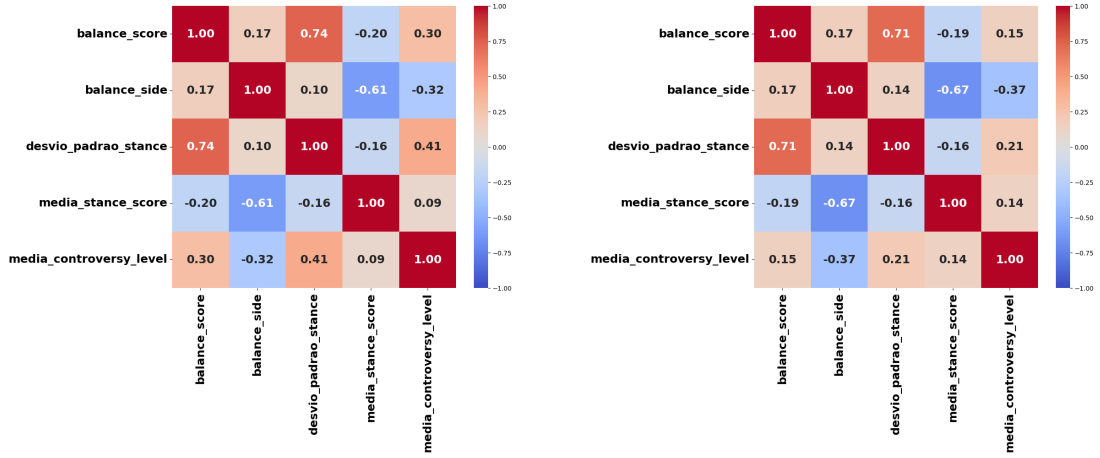
4. Resultados

Nesta seção apresentamos os resultados obtidos com a metodologia proposta aplicada ao contexto das *threads* das eleições brasileiras de 2018 e 2022 no Reddit.

4.1. Correlação e Acurácia das Métricas

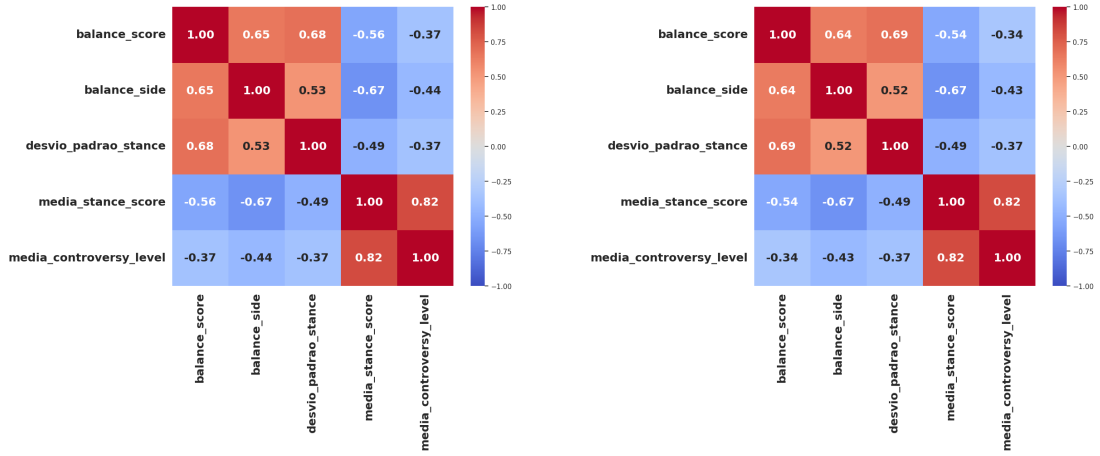
Para avaliar a acurácia e a coerência das métricas de quantificação da polarização nas discussões, verificamos como elas se correlacionam nos dois anos de análise, através dos mapas de calor das Figuras 3 e 4. Através dos mapas é possível comparar como as métricas se correlacionam (i) quando a metodologia utiliza o GPT-4 para detecção de posição e (ii) com o modelo [Guillou 2021] adaptado utilizado para a mesma tarefa.

A partir dos resultados apresentados nas Figuras 3 e 4, nota-se que as métricas apresentaram um aumento expressivo de autocorrelação, corroborando a eficácia e a consistência metodológica na análise da polarização nas *threads*. As correlações nos dois períodos analisados demonstraram resultados robustos, com aumentos e melhorias significativas em relação ao cálculo utilizando o GPT-4. Primeiro, houve aumento na correlação entre o *Balance Side* e *Balance Score* (de 0.17 para 0.65), indicando maior alinhamento entre as medidas, as quais têm propósito similar de medição, diferindo apenas pelo modo



(a) Eleições de 2018. (b) Eleições de 2022.

Figura 3. Matriz de Correlações entre as métricas de polarização - *Stance Score* com GPT-4.



(a) Eleições de 2018. (b) Eleições de 2022.

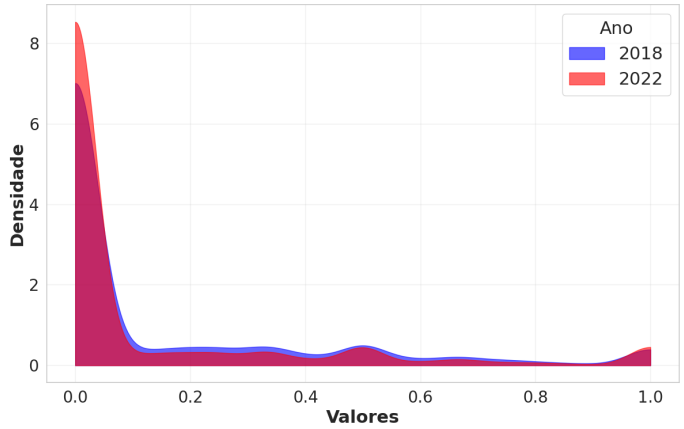
Figura 4. Matriz de Correlações entre as métricas de polarização - *Stance Score* com modelo *Portuguese BERT base cased QA (Question Answering), finetuned on SQUAD v1.1* ajustado para a tarefa.

de cálculo, como detalhado na seção 3.4. Adicionalmente, houve a intensificação da correlação negativa entre *Balance Score* e a média do *Stance Score* (de -0.20 para -0.56), sugerindo que discussões mais divididas estão associadas a desacordos mais intensos. Por fim, houve um aumento substancial na correlação entre a média do *Controversy Level* e a média do *Stance Score* (de 0.10 para 0.82), evidenciando maior variabilidade de opiniões em contextos de onde os usuários têm opiniões moderadas.

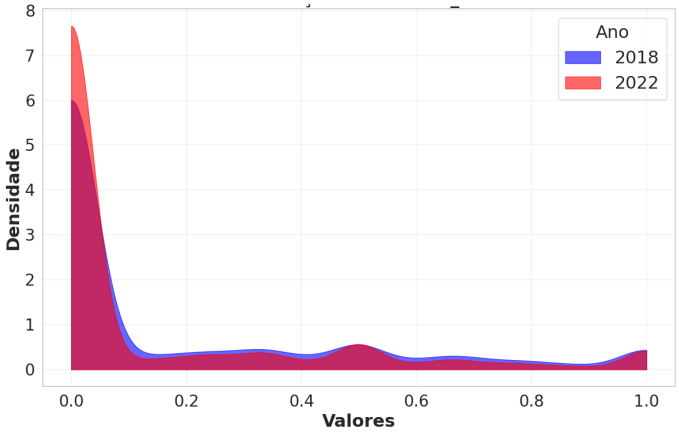
Analisando detalhadamente as correlações obtidas com a adaptação do modelo *Portuguese BERT base cased QA (Question Answering), finetuned on SQUAD v1.1* (Figura 4), notamos que o *Balance Score* e o desvio-padrão do *Stance Score* apresentam uma correlação positiva elevada, de 0.68 em 2018 e 0.69 em 2022, corroborando a conclusão de [Garimella et al. 2018], que afirma que tópicos controversos possuem variância mais elevada em detrimento de tópicos não controversos. O mesmo ocorre com a correlação

entre o *Balance Side* e o desvio-padrão do *Stance Score*, com 0.53 em 2018 e 0.52 em 2022. Também destaca-se a correlação altamente negativa entre o *Balance Score* e *Balance Side* com a média do *Stance Score* e com o *Controversy Level*, evidenciando que, quando a discussão atinge teor de maior discordância (*Stance Score* negativo), os dois lados opostos no debate tendem a estar mais bem formados e balanceados, isto é, a discussão tem participação mais igualitária entre os indivíduos de uma primeira e segunda opiniões.

4.2. Caracterização das Árvores de Discussão em 2018 e 2022



(a) Distribuição do *Balance Score*.

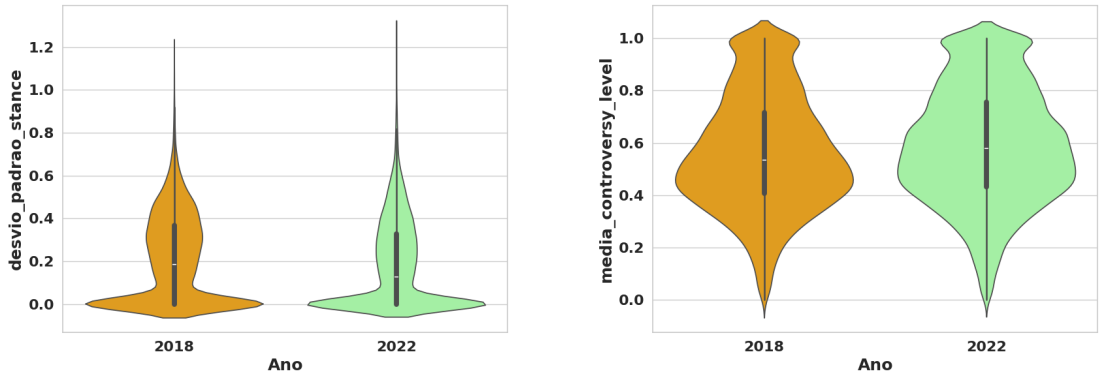


(b) Distribuição do *Balance Side*.

Figura 5. Comparação da distribuição das medidas de balanceamento nas árvores de discussão ao longo dos dois períodos analisados.

Os gráficos da Figura 5 evidenciam que o intervalo entre 0.1 e 0.9 dos valores de *Balance Side* e *Balance Score* contém mais pontos em 2018, o que indica maior frequência de discussões com polarização moderada e caráter mais diverso. Isso é endossado na Figura 6(a), que mostra que, em média, o desvio-padrão do *Stance Score* foi maior em 2018, indicando maior variabilidade nas opiniões expressas.

Já o gráfico da Figura 6(b) aponta o aumento das discussões de caráter altamente discordante (mais polarizadas), uma vez que há, no segundo ano, maior número de pontos com média de *Controversy Level* muito positiva. Também, pelo gráfico da Figura 7, é



(a) Distribuição do desvio-padrão do *Stance Score*.

(b) Distribuição da média do *Controversy Level*.

Figura 6. Distribuição das métricas nas árvores de discussão.

clara a maior ocorrência de *Stance Score* positivamente elevado em 2022, em detrimento de 2018, evidenciando, assim, a predominância, no segundo ano, de discussões com característica fortemente unilateral, e a formação das chamadas *echo chambers* (câmaras de eco), onde os usuários concordam entre si e tendem a reforçar o ponto de vista uns dos outros. É interessante notar que a média do *Stance Score* é complementar às medidas de balanceamento previamente analisadas. Assim, os resultados apresentados na Figura 7 corroboram às análises feitas da Figura 5, de que em 2018 há mais discussões de caráter ameno e de que em 2022 ocorre maior formação de câmaras de eco.

Essas tendências também podem ser inferidas analisando-se o primeiro quadrante dos mapas de calor que ilustram a densidade do *Stance Score* de um comentário pela média do *Stance Score* dos seus filhos (respostas) ao longo dos anos, na Figura 8. Podemos observar que o ano de 2022 possui quantidade de pontos mais elevada comparada a 2018 no primeiro quadrante, principalmente no extremo superior direito, aquelas com forte intra-concordância.

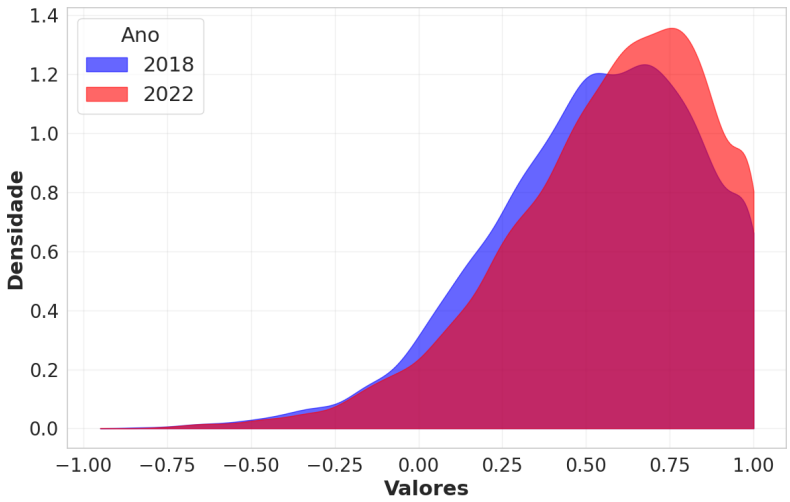
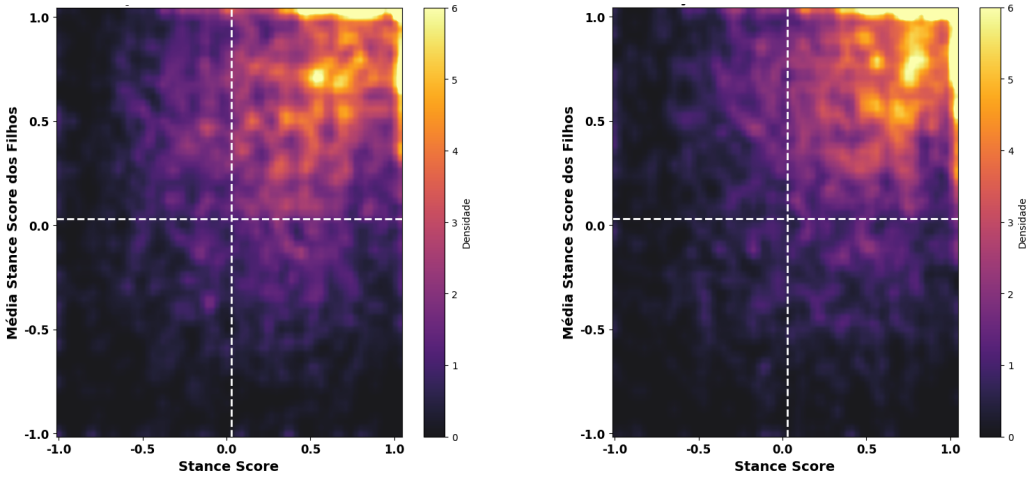


Figura 7. Distribuição da média do *Stance Score* nos dois anos de análise.

Por fim, realizamos algumas análises qualitativas manuais em árvores de dis-

cussão de cada quadrante das Figuras 8(a) e 8(b), verificando o conteúdo discutido e a distribuição dos lados da discussão entre os vértices. Chamaram atenção as árvores cujos vértices estão localizados majoritariamente no primeiro quadrante que, por se tratar de *echo chambers*, são árvores com caráter quase ou totalmente unilateral, com *Balance* muito baixo. Mesmo quando debatem assuntos polêmicos, os comentários dessas *threads* compartilham o mesmo ponto de vista. Também verificamos árvores com vértices no terceiro quadrante que, por outro lado, têm caráter de discussão acalorado, com *Balance* elevado, se tratando normalmente de assuntos polêmicos como, por exemplo, *"Sigilo de 100 anos", "Carona de Lula até a COP27", "Isolamento Político de Bolsonaro às vésperas das Eleições 2022"*. Além disso, nesse grupo, as opiniões são divergentes e as árvores possuem vários ramos do tipo resposta-réplica.



(a) Eleições de 2018.

(b) Eleições de 2022.

Figura 8. Mapa de Calor da Densidade de Distribuição dos Valores de *Stance Score*.

5. Limitações

Restrição de Domínio. Uma vez que nos aprofundamos nas discussões da plataforma Reddit, o escopo fica restrito, dada a infinidade de redes sociais existentes nos dias atuais, cada qual com diferentes aspectos de modelagem e, possivelmente, diferentes teores de discussão.

Confiança na Detecção de Posição. Como a modelagem do grafo depende do processo de *stance detection* para bipartição, a qualidade do método utilizado é fundamental. Assim, a acurácia da aferição da polarização depende diretamente da qualidade do método utilizado para detecção de posição. Para garantir o melhor resultado possível, como explicado na seção 3.3, o modelo utilizado, já pré-treinado em língua portuguesa, passou ainda por diversas etapas de ajuste antes de ser submetido à tarefa de detecção de posição.

Bipartição. É sabido que discussões de assuntos controversos não necessariamente têm apenas dois posicionamentos contrapostos. No contexto deste estudo, ao fazer uma bipartição das árvores de discussão, à luz da contraposição política da esquerda e direita no Brasil, consideramos, assim, apenas dois lados.

6. Conclusão e Trabalhos Futuros

A metodologia proposta neste trabalho para caracterizar discussões polarizadas em mídias sociais, utilizando a modelagem matemática em árvores de discussão e um modelo de aprendizado de máquina para detecção de posicionamento, mostrou-se eficaz e robusta para este e outros contextos. A análise das discussões no Reddit durante as eleições presidenciais brasileiras de 2018 e 2022 revelou uma tendência de aumento da polarização e diminuição da diversidade de opiniões ao longo dos dois períodos analisados. Esses resultados são consistentes com estudos anteriores que destacam os efeitos negativos da polarização extrema na democracia, reforçando a necessidade de plataformas de mídia social que promovam um debate mais plural e equilibrado.

A substituição do modelo GPT-4 pelo modelo de aprendizado de máquina aberto e ajustável permitiu uma análise mais abrangente e representativa do conjunto de dados coletados. As métricas de quantificação da polarização, como o *Balance Score*, *Balance Side* e *Controversy Level*, demonstraram uma alta correlação e coerência, validando a eficácia da metodologia. Além disso, a análise qualitativa das árvores de discussão corroborou os achados quantitativos, destacando a formação de câmaras de eco e a predominância de discussões unilaterais em 2022.

Para futuras pesquisas, sugerimos a expansão do escopo para incluir outras plataformas de mídia social e, talvez, em outros idiomas, a fim de obter uma visão mais completa da polarização nas redes sociais e testar a aplicabilidade e acurácia da metodologia em diferentes contextos. Além disso, a aplicação do *framework* proposto em diferentes cenários culturais e linguísticos pode fornecer *insights* valiosos sobre como a polarização se manifesta em diversas sociedades. Outra direção promissora é a integração de técnicas de processamento de linguagem natural mais avançadas, como modelos multimodais que combinam texto, imagens e metadados de usuários, para melhorar a precisão da detecção de posicionamento. Por fim, o desenvolvimento de ferramentas de moderação de conteúdo baseadas nos achados deste estudo pode contribuir para a criação de ambientes de discussão mais saudáveis e democráticos nas mídias sociais.

Referências

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023a). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023b). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, R., Rajagopalan, S., Srikant, R., and Xu, Y. (2003). Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th international conference on World Wide Web*, pages 529–535.

ALDayel, A. and Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing Management*, 58(4):102597.

Almeida, R. J. A. (2018). Leia - léxico para inferência adaptada. https://github.com/rafjaa/LeIA.

Alsinet, T., Argelich, J., Béjar, R., and Martínez, S. (2021a). Measuring polarization in online debates. *Applied Sciences*, 11(24):11879.

Alsinet, T., Argelich, J., Béjar, R., and Martínez, S. (2021b). Measuring polarization in online debates. *Applied Sciences*, 11(24):11879.

Britt, R. K., Franco, C. L., and Jones, N. (2023). Trends and challenges within reddit and health communication research: A systematic review. *Communication and the Public*, 8(4):402–417.

Brum, P., Cândido Teixeira, M., Vimieiro, R., Araújo, E., Meira Jr, W., and Lobo Pappa, G. (2022). Political polarization on twitter during the covid-19 pandemic: a case study in brazil. *Social Network Analysis and Mining*, 12(1):140.

Caetano, J., Guimarães, S., Araújo, M. M., Silva, M., Reis, J. C., Silva, A. P., Benevenuto, F., and Almeida, J. M. (2022). Characterizing early electoral advertisements on twitter: A brazilian case study. In *Proc. of the SocInfo*.

Corrêa, E. V. B. (2023). Redes sociais, ódio e polarização política: A psicodinâmica da guerra civil digital brasileira. *POLÍTICA EM FOCO: O melhor embate é o debate–Vol. 3*, page 17.

Cunha, G. F. and da Silva, A. P. C. (2024). Caracterizando polarização em redes sociais: Um estudo de caso das discussões no reddit sobre as eleições brasileiras de 2018 e 2022. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web (WebMedia 2024)*. Sociedade Brasileira de Computação-SBC.

Dahlberg, L. (2007). Rethinking the fragmentation of the cyberpublic: from consensus to contestation. *New media & society*, 9(5):827–847.

de Arruda, H. F., Cardoso, F. M., de Arruda, G. F., Hernández, A. R., da Fontoura Costa, L., and Moreno, Y. (2022). Modelling how social network algorithms can influence opinion polarization. *Information Sciences*, 588:265–278.

Franco, A. B. and Pound, N. (2022). The foundations of bolsonaro's support: Exploring the psychological underpinnings of political polarization in brazil. *Journal of Community & Applied Social Psychology*, 32(5):846–859.

Garimella, K., Morales, G. D. F., Gionis, A., and Mathioudakis, M. (2018). Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27.

Guillou, P. (2021). Portuguese bert base cased qa (question answering), finetuned on squad v1.1.

Guimaraes, S., Silva, M., Caetano, J., Araújo, M., Santos, J., Reis, J. C., Silva, A. P., Benevenuto, F., and Almeida, J. M. (2022). Análise de propagandas eleitorais antecipadas no twitter. In *Anais do BrasNAM*.

Heltzel, G. and Laurin, K. (2020). Polarization in america: Two possible futures. *Current opinion in behavioral sciences*, 34:179–184.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Karypis, G. (1997). Metis: Unstructured graph partitioning and sparse matrix ordering system. *Technical report*.

Kim, Y., Hsu, S.-H., and de Zúñiga, H. G. (2013). Influence of social media use on discussion network heterogeneity and civic engagement: The moderating role of personality traits. *Journal of communication*, 63(3):498–516.

Liang, B., Li, A., Zhao, J., Gui, L., Yang, M., Yu, Y., Wong, K.-F., and Xu, R. (2024). Multi-modal stance detection: New datasets and model. *arXiv preprint arXiv:2402.14298*.

Lima, H. F. d. (2021). Polarização política afetiva: São os valores humanos e os traços de personalidade uma explicação?

Mutz, D. C. (2002). The consequences of cross-cutting networks for political participation. *American journal of political science*, pages 838–855.

Oliveira, D. F. N. (2022). *Interpretabilidade de modelos de aprendizado profundo aplicados ao diagnóstico e prognóstico não supervisionado de falhas*. PhD thesis, Universidade de São Paulo.

Ortellado, P., Ribeiro, M. M., and Zeine, L. (2022). Existe polarização política no brasil? análise das evidências em duas séries de pesquisas de opinião. *Opinião Pública*, 28:62–91.

Pereira, C., Pavan, M., Yoon, S., Ramos, R., Costa, P., Cavalheiro, L., and Paraboni, I. (2023). Ustancebr: a multimodal language resource for stance prediction. *arXiv preprint arXiv:2312.06374*.

Recuero, R., Soares, F. B., and Gruzd, A. (2020). Hyperpartisanship, disinformation and political conversations on twitter: The brazilian presidential election of 2018. In *Proceedings of the international AAAI conference on Web and social media*, volume 14, pages 569–578.

Rossini, P. (2020). Beyond toxicity in the online public sphere: understanding incivility in online political talk. *A research agenda for digital politics*, pages 160–170.

Ruediger, M. A. et al. (2014). Redes sociais retratam eleição mais polarizada da história recente do brasil.

Santos, P. D. and Goya, D. H. (2021). Automatic twitter stance detection on politically controversial issues: A study on covid-19's cpi. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 524–535. SBC.

Saveski, M. (2020). Polarization and toxicity in political discourse online.

Silva, A. G. (2023). O insaciável espírito da época: ensaios de psicologia analítica e política, de: Humbertho oliveira, roque tadeu gui e rubens bragarnich. editora vozes, 2021. *Self-Revista do Instituto Junguiano de São Paulo*, 8:e007–e007.

Souza, F., Nogueira, R., and Lotufo, R. (2020a). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Souza, F., Nogueira, R., and Lotufo, R. (2020b). Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Vicario, M. D., Quattrociocchi, W., Scala, A., and Zollo, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):1–22.

Zhang, B., Ding, D., and Jing, L. (2022). How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Zhang, B., Fu, X., Ding, D., Huang, H., Li, Y., and Jing, L. (2023). Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.