Medição e Análise do Discurso Tóxico Contra Deputadas Federais Brasileiras na Plataforma X

Karen Gomes[1], **Jussara M. Almeida**[1], **Marcos André Gonçalves**[1], **Fabrício Benevenuto**[1]

[1] Departamento de Ciência da Computação – Universidade Federal de Minas Gerais

{karengomes, jussara, mgoncalv, fabricio}@dcc.ufmg.br

Abstract. *Toxic discourse in online political environments is often analyzed in an aggregated way, without considering differences across targets. In this work, we investigate toxic discourse directed at Brazilian federal deputies elected in 2022 on the X platform, analyzing how it varies across the ideological spectrum. We combine automatic toxicity detection, topic modeling with BERTopic, language models for labeling, and the analysis of psycholinguistic attributes with LIWC. The results show distinct patterns across political groups. Left-wing deputies receive more attacks associated with emotional and ideological delegitimization, while right-wing deputies are more frequently targeted with attacks related to appearance and sexuality. These findings suggest that online political toxicity varies according to both the target group and the type of attack involved. Warning: This article includes examples of messages containing offensive language for analytical purposes.*

Resumo. *O discurso tóxico em ambientes políticos online costuma ser analisado de forma agregada, sem considerar diferenças entre os alvos. Neste trabalho, investigamos o discurso tóxico direcionado a deputadas federais brasileiras eleitas em 2022 na plataforma X, analisando como ele varia entre grupos ideológicos. Para isso, combinamos detecção automática de toxicidade, modelagem de tópicos com BERTopic, uso de modelos de linguagem para rotulação e análise de atributos psicolinguísticos com LIWC. Os resultados mostram padrões distintos entre os grupos políticos. Deputadas de esquerda recebem mais ataques associados à deslegitimação ideológica e carga emocional, enquanto deputadas de direita são mais frequentemente alvo de ataques ligados à aparência e à sexualidade. Esses resultados sugerem que a toxicidade política online varia conforme o grupo alvo e o tipo de ataque mobilizado. Aviso: Este artigo inclui exemplos de mensagens contendo linguagem ofensiva para fins analíticos.*

1. Introdução

As plataformas digitais tornaram-se arenas centrais para o debate político, nas quais interações em larga escala influenciam a formação da opinião pública e intensificam conflitos ideológicos. Em contextos altamente polarizados, como as eleições brasileiras de 2022, esses ambientes também favorecem a proliferação de discurso tóxico direcionado a figuras públicas. Embora a literatura tenha avançado na quantificação da toxicidade online, grande parte dos estudos ainda trata esse fenômeno de forma agregada, sem investigar como ele varia em função do alvo. Como consequência, a toxicidade frequentemente é tratada de maneira implícita como um fenômeno homogêneo.

Essa questão torna-se ainda mais crítica no caso de mulheres na política, frequentemente alvo de ataques online [Koch et al. 2025]. Os impactos dessas manifestações incluem autocensura, redução da participação pública e, em alguns casos, afastamento da atividade política. Apesar de representarem 53% do eleitorado brasileiro em 2022, mulheres ocuparam apenas 18% dos cargos eletivos [Belisário and dos Reis 2023], evidenciando desigualdades estruturais persistentes. Nesse contexto, parte desses ataques configura *violência política de gênero* [Krook and Restrepo Sanín 2016, Biroli 2018], caracterizada por práticas que buscam deslegitimar a participação política feminina por meio de estereótipos e ataques pessoais [Pinho 2020]. No ambiente digital, essas manifestações frequentemente extrapolam a crítica política e passam a atingir atributos identitários, aparência e vida pessoal.

Diante desse cenário, permanece pouco explorado como conteúdo, atributos psicolinguísticos e posicionamento ideológico interagem na caracterização do discurso tóxico em contextos políticos polarizados. Essa lacuna dificulta compreender como diferentes formas de ataque se distribuem entre grupos e espectros políticos.

Neste trabalho, investigamos o discurso tóxico direcionado a deputadas federais brasileiras eleitas em 2022 na plataforma X. A análise baseia-se em um corpus de 113.511 mensagens públicas coletadas a partir de interações com 17 parlamentares. Para isso, adotamos uma abordagem multidimensional que combina detecção de toxicidade, modelagem de tópicos, rotulagem com modelos de linguagem e análise de atributos psicolinguísticos, permitindo analisar simultaneamente temas e padrões discursivos das mensagens tóxicas.

Os resultados indicam que a toxicidade apresenta padrões distintos entre grupos políticos, variando não apenas em intensidade, mas também na forma como se manifesta. Esses achados contribuem para a compreensão do discurso tóxico em ambientes políticos digitais ao evidenciar diferenças estruturais na forma como ataques são direcionados a diferentes grupos.

Enquanto trabalhos anteriores [Davidson et al. 2017, Tavares and Recuero 2023] frequentemente analisam a toxicidade de forma agregada ou focalizam dimensões específicas do fenômeno, nossa abordagem integra conteúdo temático, atributos psicolinguísticos e posicionamento ideológico em uma análise multidimensional da toxicidade online. As principais contribuições deste estudo são: (i) uma análise multidimensional do fenômeno; (ii) evidências empíricas de padrões diferenciados de ataque entre grupos políticos; e (iii) a caracterização do papel dos atributos psicolinguísticos nesses padrões.

2. Trabalhos Relacionados

A análise de discurso tóxico em redes sociais tem recebido atenção crescente na literatura, com foco na caracterização, detecção e disseminação de conteúdo abusivo em ambientes digitais [Al-Hassan and Al-Dossari 2019, Lima et al. 2020]. Trabalhos recentes também têm investigado especificamente a toxicidade direcionada a mulheres em diferentes plataformas, evidenciando padrões de misoginia e ataques baseados em gênero em contextos online [Martins et al. 2025]. Em particular, a plataforma X consolidou-se como um dos principais objetos de estudo, dada sua centralidade no debate público e seu papel na amplificação de conflitos políticos. No entanto, grande parte desses trabalhos concentra-se na detecção automática ou na quantificação da toxicidade, frequentemente

tratando o fenômeno de forma agregada e independente do alvo [Davidson et al. 2017].

No contexto de gênero e política, estudos mostram que mulheres em posições públicas são frequentemente alvo de ataques online, que apresentam padrões específicos relacionados a gênero [Koch et al. 2025]. No Brasil, pesquisas recentes analisam interações envolvendo deputadas federais no X, identificando padrões de discurso tóxico [Tavares and Recuero 2023], além de evidenciar efeitos de silenciamento durante campanhas eleitorais [Souza et al. 2022]. Embora relevantes, esses trabalhos concentram-se em análises qualitativas ou em dimensões específicas, sem integrar múltiplas perspectivas.

Em âmbito internacional, o abuso online contra mulheres na política também é amplamente documentado, como no relatório *Toxic Twitter* [Amnesty International 2018]. Contudo, esses estudos raramente investigam como diferentes formas de ataque variam sistematicamente conforme características dos alvos, como seu posicionamento ideológico.

Paralelamente, abordagens baseadas em modelagem de tópicos e análise de atributos psicolinguísticos têm sido utilizadas para caracterizar padrões discursivos [Oltmanns et al. 2025, Steinbrenner et al. 2025]. Ainda assim, sua aplicação integrada ao estudo do discurso tóxico em contextos políticos, especialmente em português e sob diferentes espectros ideológicos, permanece limitada.

Neste trabalho, propomos uma análise multidimensional do discurso tóxico que integra conteúdo, forma de expressão e contexto político. Ao incorporar o espectro ideológico como dimensão central, mostramos que a toxicidade não varia apenas em intensidade, mas também na forma como se manifesta, com diferenças estruturais nos tipos de ataque direcionados a cada grupo.

3. Metodologia

Nesta seção, descrevemos a metodologia adotada no estudo, incluindo as etapas de coleta e pré-processamento dos dados, classificação de toxicidade, modelagem de tópicos e análise de atributos psicolinguísticos, organizadas no pipeline ilustrado na Figura 1. Cada uma dessas etapas é detalhada nas subseções a seguir.

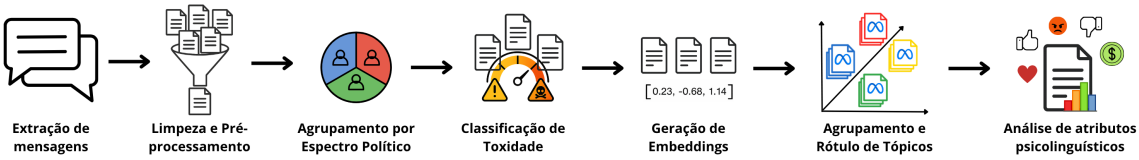

Figura 1. Visão geral da metodologia proposta.

3.1. Coleta de Dados e Pré-processamento

Os dados foram coletados por meio do Nitter, uma ferramenta de código aberto que permite acessar e visualizar conteúdos *públicos* do X sem necessidade de autenticação [Zedeus 2019]. Essa ferramenta foi adotada devido às restrições financeiras da API oficial, especialmente em termos de acesso e custo, e já foi utilizada em estudos anteriores [Alzahrani and AlGhamdi 2025]. A coleta considerou exclusivamente tweets públicos e, como os dados foram analisados de forma agregada, não foi necessário consentimento adicional dos usuários. De acordo com a Resolução CNS nº

510/2016 [Conselho Nacional de Saúde 2016], pesquisas baseadas em informações de acesso público podem ser dispensadas de apreciação por comitê de ética.

Foram selecionados 17 perfis de deputadas federais eleitas entre os 50 candidatos mais votados nas eleições brasileiras de 2022, com base em registros eleitorais e dados públicos do Tribunal Superior Eleitoral[1]. A seleção também incluiu uma etapa de verificação manual, realizada para garantir a correspondência entre as contas analisadas e os perfis oficiais das parlamentares na plataforma X. Essa escolha concentra a análise em parlamentares com maior visibilidade pública e, em conjunto com o recorte temporal, delimita os resultados a esse contexto. Foram analisados tweets publicados por outros usuários que mencionaram ou responderam às deputadas, excluindo retuítes, de modo a considerar apenas interações diretas. As mensagens abrangem o período de 1º de janeiro a 31 de dezembro de 2022, totalizando 113.511 postagens.

Para garantir a reprodutibilidade do estudo, os autores seguiram as diretrizes de compartilhamento de dados da plataforma [X Corp. 2023] e realizaram uma solicitação formal para disponibilização dos identificadores únicos (IDs) dos tweets coletados. No entanto, não houve retorno da plataforma até o momento da submissão deste trabalho.

Para mitigar ruídos típicos de redes sociais, realizou-se um pré-processamento que incluiu a remoção de menções isoladas, URLs e respostas sem conteúdo textual, além da normalização para letras minúsculas. Os emojis foram mantidos e processados implicitamente pelos modelos. Após essas etapas, o corpus final consistiu em 111.473 tweets.

3.2. Agrupamento das Deputadas por Espectro Político

Para classificar cada deputada no espectro político, adotou-se a orientação ideológica do partido pelo qual foi eleita, com categorização em esquerda, centro e direita baseada em [Marques et al. 2014] e [Bolognesi et al. 2022]. Ressalta-se que essa abordagem, por se basear na filiação partidária, pode não refletir integralmente nuances da identidade política individual das deputadas.

A Tabela 1 apresenta as deputadas e seus respectivos espectros políticos. No caso da deputada Carla Zambelli, unificaram-se os perfis devido a alterações de nome de usuário ao longo de 2022. Em agosto, a parlamentar mudou de `@CarlaZambelli38` para `@Zambelli2210` [Poder360 2022], posteriormente desativado por decisão do Tribunal Superior Eleitoral em novembro do mesmo ano [G1 2022]. O perfil atual, `@ZambelliRita_`, herdou parte das interações anteriores. Assim, tweets e respostas das três contas foram agregados em um único perfil analítico.

3.3. Classificação de Toxicidade

As mensagens que mencionavam as deputadas foram classificadas quanto à toxicidade com a biblioteca Detoxify [Hanu and Team 2020], baseada em modelos *Transformer* e amplamente utilizada na detecção de conteúdo tóxico [Melo and Figueredo 2025, Martins et al. 2025]. Como o corpus é composto por textos em português, utilizou-se a variante Multilingual. O modelo produz um escore contínuo entre 0 e 1, representando o nível de toxicidade. Para definir o limiar de decisão, adotou-se um procedimento de calibração baseado em anotações manuais. Inicialmente, todas as mensagens foram processadas, obtendo-se esse escore para cada tweet.

[1]Disponível em: `https://dadosabertos.tse.jus.br/dataset/resultados-2022`.

Tabela 1. Classificação por Partido e Espectro Político

Espectro	Nome	Usuário no X	Partido
Esquerda	Gleisi Hoffmann	@gleisi	PT
	Erika Hilton	@erikakhilton	PSOL
	Sâmia Bomfim	@samiabomfim	PSOL
	Fernanda Melchionna	@fernandapsol	PSOL
	Talíria Petrone	@taliriapetrone	PSOL
	Tabata Amaral	@tabataamaralsp	PSB
	Duda Salabert	@dudasalabert	PDT
Centro	Marina Silva	@marinasilva	REDE
	Alessandra Haber	@draalehaber	MDB
Direita	Carla Zambelli	@carlazambelli38 @zambelli2210 @zambellirita_	PL
	Carol de Toni	@caroldetoni	PL
	Rosana Valle	@deprosanavalle	PL
	Bia Kicis	@biakicis	PL
	Silvye Alves	@silvyealves	UNIÃO
	Rosângela Moro	@rosangelawm	UNIÃO
	Daniela do Waguinho	@danielacarneiro	UNIÃO
	Clarissa Tércio	@clarissatercio	PP

A partir desses escores, foi construída uma amostra estratificada de mensagens, contemplando diferentes faixas de toxicidade, de modo a incluir exemplos representativos tanto de casos não tóxicos quanto potencialmente tóxicos. Essa amostra foi rotulada manualmente por três pesquisadoras, integrantes do grupo de pesquisa, que seguiram uma definição de toxicidade previamente fornecida, baseada em [Jigsaw 2018], segundo a qual uma mensagem é considerada tóxica quando apresenta conteúdo rude, desrespeitoso ou ofensivo, com potencial de prejudicar a interação.

A concordância entre as anotações humanas foi avaliada por meio do coeficiente de Cohen's Kappa, resultando em $\kappa = 0{,}742$, o que indica concordância substancial. Com base nessas anotações, utilizadas como *ground truth*, avaliou-se o desempenho do modelo para diferentes valores de limiar θ, onde uma mensagem é classificada como tóxica se seu escore for superior a θ. Essa avaliação foi realizada por meio de validação cruzada estratificada em cinco *folds*, preservando a proporção entre classes em cada partição. Foram testados valores de θ no intervalo de 0,1 a 1,0, e para cada valor foram calculadas métricas padrão de classificação. O limiar $\theta = 0{,}1$ apresentou o melhor desempenho (Macro-F1=0,809), sendo então adotado para a classificação de todas as mensagens do corpus em tóxicas e não tóxicas.

3.4. Extração de Tópicos com BERTopic

Após a identificação dos tweets tóxicos, a análise temática foi realizada exclusivamente sobre esse subconjunto com o BERTopic [Grootendorst 2022], baseado em modelos Transformer e amplamente utilizado em estudos de discurso político [Carmo et al. 2023]. Em comparação com modelos probabilísticos clássicos, como LDA, trabalhos com textos curtos em redes sociais indicam que o BERTopic tende a produzir tópicos mais coerentes e informativos, especialmente em contextos multilíngues e em línguas morfologicamente ricas, o que reforça sua adequação para o corpus de tweets políticos analisado [de Groot et al. 2022, Medvecki et al. 2024]. Também utilizamos embeddings pré-computados do modelo `text-embedding-3-large`, devido ao seu bom desempenho em cenários multilíngues [OpenAI 2024].

A redução de dimensionalidade foi realizada com UMAP (*n_neighbors*=15, *n_components*=10, métrica do cosseno), seguida do agrupamento com HDBSCAN

(*min_cluster_size*=40, *min_samples*=20). Os parâmetros foram definidos por busca em grade (*grid search*), variando *n_components* em $[5, 10]$, *n_neighbors* em $[5, 12, 15]$, *min_cluster_size* em $[15, 20, 40]$ e *min_samples* em $[30, 40, 50]$, priorizando a obtenção de tópicos semanticamente coesos e evitando fragmentação excessiva ou grande número de outliers.

Foram identificados 88 tópicos iniciais, posteriormente agrupados em 10 macrotemas, definidos manualmente a partir da análise das palavras-chave e dos textos associados. Esses macrotemas cobrem aproximadamente 52% das mensagens analisadas, permitindo uma interpretação mais estruturada do discurso.

3.5. Rotulação Automática de Macrotemas com LLM

Para interpretar os macrotemas, utilizamos o modelo `LLaMA-3.1-8B-Instruct` [Grattafiori et al. 2024], uma LLM pública e de código aberto, responsável por gerar descrições concisas a partir de palavras-chave e exemplos representativos de textos, por meio de um *prompt* estruturado (Figura 2).

Diferentemente da rotulação direta de tópicos, aplicamos a LLM sobre macrotemas, permitindo a geração de descrições mais gerais e representativas dos padrões discursivos. Essa abordagem tem sido adotada na literatura para melhorar a interpretabilidade de modelos baseados em BERTopic [Kozlowski et al. 2024, Khandelwal 2025].

A rotulação possui caráter interpretativo e não altera a estrutura dos agrupamentos. As descrições geradas foram verificadas manualmente, apresentando consistência com os textos analisados.

System:
Você rotula tópicos de discussões políticas em português. O conteúdo pode conter ofensas ou linguagem agressiva, e isso é esperado.
Seu trabalho é APENAS propor um rótulo curto (máx. 30 palavras) que descreva detalhadamente o tema principal, mesmo que os dados sejam ruidosos.
Não diga que não pode responder; sempre proponha um rótulo.

User:
Tópico <topic_id>
Palavras-chave: <keywords_txt>
Exemplos de mensagens: <exemplos_txt>
Responda apenas com um rótulo curto para esse tópico.

Figura 2. Prompt utilizado para rotular tópicos com a LLM.

3.6. Utilização do LIWC

Além da modelagem de tópicos, realizamos uma análise dos atributos psicolinguísticos das mensagens tóxicas com o Linguistic Inquiry and Word Count (LIWC), um sistema baseado em dicionários que associa palavras a categorias linguísticas, emocionais e cognitivas [Pennebaker et al. 2007]. Utilizamos a versão para o português brasileiro (*LIWC2007_PT*) [Balage Filho et al. 2013]. Foram consideradas categorias afetivas, cognitivas e sociais, incluindo emoções positivas e negativas, com foco em raiva, ansiedade e tristeza, além de dimensões como corpo, sexualidade, poder, risco, dinheiro, religião, morte, xingamentos e família.

A análise foi realizada no nível de cada tweet. Para cada mensagem, foram obtidas as contagens das categorias do LIWC, posteriormente agregadas por deputada

e por posicionamento político. Em seguida, calculamos *z-scores* a partir das médias de cada grupo, permitindo comparar a intensidade relativa das categorias entre deputadas e entre espectros ideológicos.

4. Resultados

Esta seção apresenta os principais achados da análise do discurso tóxico direcionado às deputadas federais eleitas em 2022, com foco na distribuição da toxicidade entre grupos ideológicos, nos tópicos recorrentes e nos padrões psicolinguísticos identificados.

4.1. Distribuição geral da toxicidade pelo corpus

Após a aplicação do limiar de toxicidade, foram identificadas 37.115 mensagens tóxicas, correspondendo a cerca de 33% do corpus pré-processado. A Tabela 2 apresenta exemplos do corpus com seus respectivos escores. A maior parte foi direcionada a deputadas de esquerda (49,8%), seguida por deputadas de direita (29,9%) e de centro (20,3%), correspondendo a 18.479 e 11.108 mensagens, respectivamente. No espectro de centro, Marina Silva concentrou 7.528 mensagens, sendo a única representante com volume suficiente para análise, uma vez que Alessandra Haber (@draalehaber) não apresentou mensagens classificadas como tóxicas e foi excluída das etapas posteriores. Esses resultados indicam que a exposição à toxicidade não é homogênea entre os espectros políticos, com maior concentração nas parlamentares de esquerda no conjunto analisado.

Tabela 2. Exemplos de Tweets Tóxicos e Não-Tóxicos

Classificação	Exemplo de Tweet	Score de Toxicidade
Não-Tóxico	“os futuros ministros investigados você se calou. está com medo também?”	0.09974
Tóxico	“essas comunistas do pt devem ser extermida da face da terra.”	0.99199

4.2. Análise de embeddings e proximidade entre deputadas

Para explorar padrões semânticos no discurso tóxico, construímos uma representação vetorial por deputada a partir da média dos embeddings dos tweets a elas direcionados. Essas representações foram projetadas em duas dimensões com UMAP, preservando relações de proximidade do espaço original. Os eixos não possuem interpretação semântica direta.

Na Figura 3, cada ponto representa uma deputada, com cores indicando o espectro ideológico e tamanho proporcional à toxicidade média. Observa-se ausência de separação clara entre os espectros, embora haja proximidades locais, sugerindo que diferentes tipos de ataque atravessam grupos políticos.

De modo geral, deputadas próximas no mapa tendem a receber ataques semanticamente semelhantes, enquanto aquelas mais distantes enfrentam padrões distintos de violência discursiva, sem separação rígida entre espectros ideológicos.

Dentro dos grupos, observam-se variações relevantes: à esquerda, deputadas como Duda Salabert e Erika Hilton se agrupam, enquanto Sâmia Bomfim aparece mais isolada, indicando heterogeneidade nos tipos de ataque; à direita, há comportamento semelhante, com concentração de algumas parlamentares e maior dispersão em casos como Silvye Alves e Carla Zambelli. Em termos de intensidade, deputadas como Sâmia Bomfim, Duda Salabert e Erika Hilton apresentam maiores níveis médios de toxicidade, enquanto, à direita, casos como Silvye Alves também se destacam. No centro, Marina Silva apresenta um padrão mais isolado.

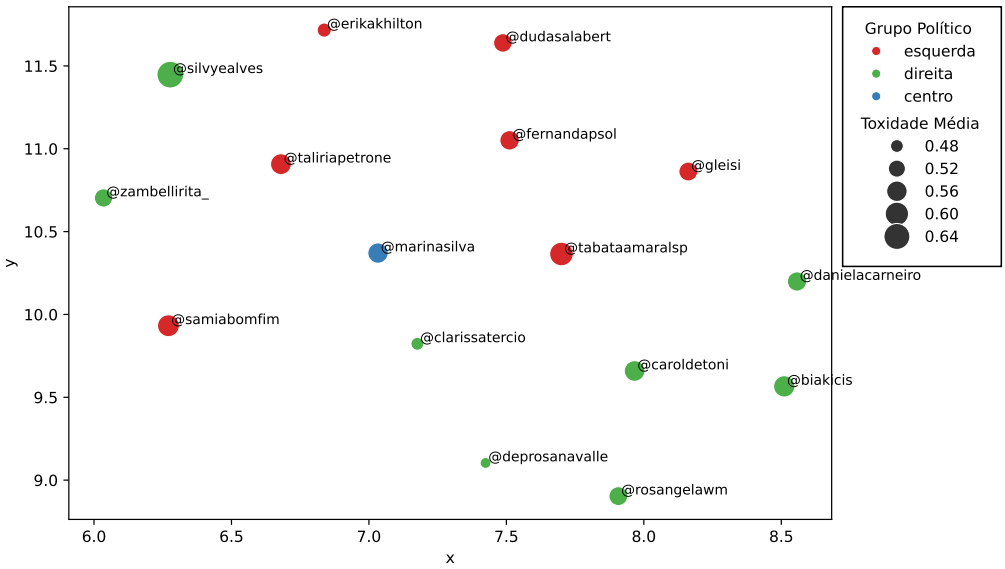

Figura 3. Projeção UMAP dos embeddings médios (tweets tóxicos) por deputada.

Esses resultados sugerem que a toxicidade não se organiza apenas por alinhamento ideológico, mas também por padrões específicos de ataque direcionados a cada parlamentar. Embora exploratória, a projeção oferece evidência qualitativa consistente com os padrões temáticos e quantitativos observados.

4.3. Macrotemas e Tópicos do Discurso Tóxico

Os tópicos identificados e organizados em macrotemas mostram que o discurso tóxico direcionado às deputadas não é uniforme, mas se distribui em diferentes temas, com padrões distintos entre os espectros ideológicos. A Tabela 3 apresenta a descrição dos macrotemas, suas frequências e exemplos representativos, enquanto a Figura 4 mostra sua distribuição ideológica.

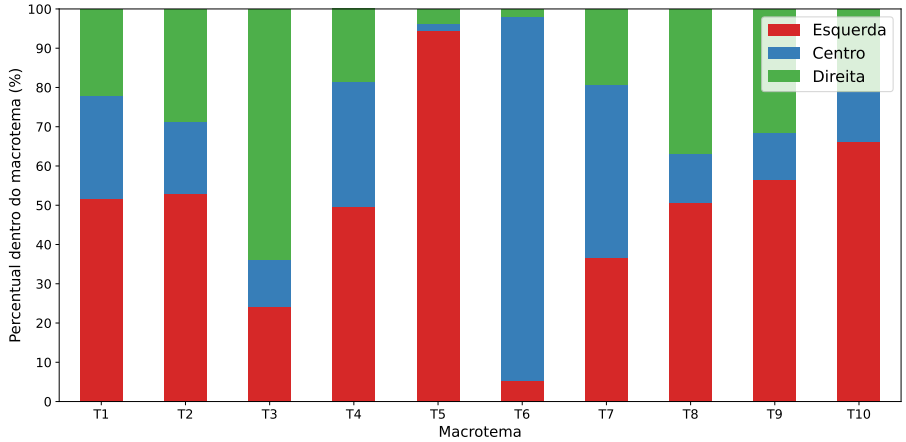

Figura 4. Distribuição ideológica dos macrotemas.

Os macrotemas mostram que a toxicidade varia entre os espectros ideológicos, em vez de se distribuir de forma homogênea. Temas associados à polarização partidária (T1) e à desqualificação pessoal (T2) aparecem de forma transversal, com maior concentração

	Descrição da LLM	Freq.	Exemplos
T1	Polarização política e violência partidária no Brasil.	6207	*Esq.*: "estúpida e amiga de bandido. soltaram o ladrão e vc aplaudiu? amante de vagabundo!" *Centro*: "mais uma coaptação! vai virar corrupta junto com os ladroes" *Dir.*: "deia de ser maluca. a culpa de tudo isso é do bolsonaro"
T2	Insultos e desqualificação pessoal em discussões políticas.	4282	*Esq.*: "mulher insuportável, não sabe falar e quando fala só sai merda!" *Centro*: "é outra incompetente, mas fazer o que quando o que importa é quem ajuda a roubar mais?" *Dir.*: "uma lunática como você, com cargo público, é um perigo pra nação!!!"
T3	Crítica às instituições políticas e corrupção no Congresso Nacional.	1827	*Esq.*: "são teus parceiros! mas querem mamar mais. já está em 250 bilhões" *Centro*: "sua desocupada que só quer mamar na teta da vaca gorda dos brasileiros que trabalham e geram riquezas para o brasil." *Dir.*: "ofala a verdade, sua pilantra. vc e seu presidente genocida destruíram o orçamento da saúde e educação com esse populismo fiscal deveriam estar juntos na cadeia"
T4	Ataques à aparência e desumanização de políticos com linguagem agressiva e ofensiva.	1769	*Esq.*: "cala a boca, g0rd@. vai comer um big mac" *Centro*: "opa o "et" que só aparece de quatro em quatro anos? desculpa et tu até é bonitinho e é amigo de criança e não de ladrão." *Dir.*: "o que tu sabe de ser homem, te recolhe na tua insignificancia velha dementada."
T5	Discussão sobre aborto, estupro e direitos reprodutivos em um contexto político polarizado.	1557	*Esq.*: "se o aborto fosse liberado quando sua mãe engravidou, talvez nós tivéssemos nos livrado de vc, mulher doente." *Centro*: "e vc deve ser uma assassina abortista né? a escória da humanidade!" *Dir.*: "assassina de pobres!!!"
T6	Reações políticas à nomeação de Marina Silva e outros para Ministérios no governo Lula.	884	*Esq.*: "passar por essa vida chegar ao fim tipo como marina simone que vendem até a alma ao diabo só pra não perder a teta dos cofres públicos (...)" *Centro*: "a pior ministra do meio ambiente do planeta. parabéns aos envolvidos" *Dir.*: "uma típica ministra do lulismo..."
T7	Condenação moral e religião: uso de linguagem agressiva e ameaças em debates políticos.	774	*Esq.*: "vc é um ser desprezível ....pagará por isso de deus vc não escapa!!" *Centro*: "maldita! que o inferno pegue você viva. inútil." *Dir.*: 'natal em cristo??? tá de brincadeira, a senhora é do capeta"
T8	Desinformação e mentira política em redes sociais.	691	*Esq.*: "eu tenho medo de gente como você, despreparada e mentirosa. . . " *Centro*: "inaceitável a tua turma fazer isso e vc culpar os patriotas velha mentirosa" *Dir.*: "mentira sua sua ridícula"
T9	Punição e encarceramento: retribuição e controle social através da prisão.	592	*Esq.*: "a cadeia é quem te aguarda, comunista" *Centro*: "a punição tem que começar por você!!!e todos que estão contigo!!! desvairada!!!!!!" *Dir.*: "já deveria está é presa a senhora deputada."
T10	Golpismo e contestação eleitoral: críticas à liberdade de expressão e fraude eleitoral no Brasil.	346	*Esq.*: "golpismo é tirar vagabundo ladrão da cadeia e fraudar uma eleição, esquerdopata." *Centro*: "marina desejo a vc e toda seus amigos esquerdistas e do stf e tse que deus pese a mão em vcs se não pagarem em vida que seja no outro plano" *Dir.*: "ten uma bandida safada travestida de deputada que continua pedindo golpe de estado."

Tabela 3. Macrotemas identificados, com descrição gerada pela LLM, quantidade de mensagens e exemplos representativos por espectro ideológico.

à esquerda, indicando que o confronto político direto e os ataques pessoais são centrais nesse tipo de discurso.

Em contraste, alguns temas apresentam forte especialização por grupo. A crítica às instituições (T3) concentra-se majoritariamente na direita, enquanto discussões sobre aborto e moralidade sexual (T5) e política ambiental (T6) mostram alta concentração em espectros específicos, indicando que certos tópicos funcionam como eixos de disputa ideológica mais delimitados.

Outros macrotemas, como ataques à aparência (T4), religião e condenação moral (T7), desinformação (T8) e punição (T9), apresentam distribuição mais equilibrada, sugerindo que determinadas formas de ataque atravessam diferentes grupos políticos. No caso de T4, observam-se ataques frequentes à aparência e à identidade das parlamentares, muitas vezes ligados à desumanização. Termos como *múmia*, *dinossauro* e *ET*, direcionados a Marina Silva, aparecem associados a comentários depreciativos sobre sua aparência. Já em relação a Sâmia Bomfim, há menções que combinam estereótipos sobre o corpo e o consumo com tentativas de silenciamento, como mensagens que a mandam se calar ou "ir comer hambúrguer", reforçando a desqualificação baseada na aparência.

Em conjunto, os resultados mostram que a toxicidade varia não apenas em intensidade, mas também na forma como se organiza. Diferentes grupos mobilizam padrões distintos de ataque no debate político online.

4.4. Análise dos Atributos Psicoinguísticos do Discurso Tóxico (LIWC)

A análise das categorias psicolinguísticas revela padrões distintos de ataques conforme o espectro político da parlamentar. Como ilustrado no mapa de calor (Figura 5) e nos

Z-scores da Tabela 4, a violência discursiva não é uniforme: deputadas de esquerda concentram maiores níveis de emoções negativas, especialmente raiva e ansiedade, enquanto deputadas de direita apresentam maior incidência de categorias relacionadas ao corpo, à sexualidade e a aspectos concretos do cotidiano. Já a representante do centro apresenta um perfil mais equilibrado, com menor concentração em categorias extremas.

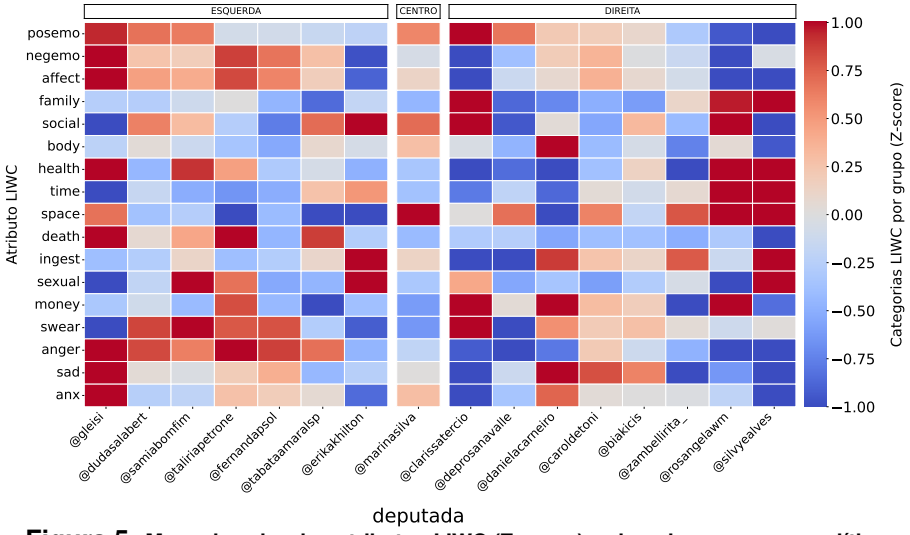

Figura 5. **Mapa de calor dos atributos LIWC (Z-score) ordenados por grupo político.**

Tabela 4. **Deputadas com os maiores e menores valores de Z-score em cada categoria LIWC.**

Categoria LIWC	Maior Z-score	Valor	Menor Z-score	Valor
posemo	@clarissatercio	1.01	@silvyealves	-3.21
negemo	@gleisi	2.40	@rosangelawm	-2.06
affect	@gleisi	2.35	@rosangelawm	-2.14
family	@silvyealves	3.09	@deprosanavalle	-0.87
social	@rosangelawm	1.63	@silvyealves	-2.25
body	@danielacarneiro	3.55	@silvyealves	-0.94
health	@rosangelawm	2.28	@clarissatercio	-1.27
time	@rosangelawm	2.88	@gleisi	-1.16
space	@silvyealves	1.50	@danielacarneiro	-1.47
death	@gleisi	2.62	@silvyealves	-1.75
ingest	@silvyealves	1.90	@clarissatercio	-2.50
sexual	@silvyealves	2.12	@gleisi	-1.63
money	@clarissatercio	2.42	@zambellirita_	-1.80
swear	@clarissatercio	1.47	@deprosanavalle	-2.16
anger	@gleisi	2.16	@silvyealves	-1.39
sad	@danielacarneiro	2.02	@silvyealves	-1.73
anx	@gleisi	3.08	@clarissatercio	-1.39

No espectro da esquerda, observa-se predominância de categorias associadas a emoções negativas, indicando um padrão de ataques marcado por alta carga emocional e agressividade verbal. Em particular, deputadas como `@gleisi` apresentam elevados escores em dimensões como *negemo*, *anger* e *anx*, refletindo mensagens que combinam deslegitimação política com ataques à sanidade e à moralidade. Além disso, em casos como `@dudasalabert`, os ataques frequentemente mobilizam dimensões identitárias, como em mensagens que negam sua identidade de gênero, deslocando o conflito do campo político para o identitário. Esse padrão sugere uma violência discursiva fortemente reativa e baseada em deslegitimação simbólica [Krook and Restrepo Sanín 2016].

Em contraste, no espectro da direita, observa-se maior incidência de categorias relacionadas ao corpo, à sexualidade e a aspectos da vida pessoal, indicando um

deslocamento da violência discursiva para dimensões mais íntimas. Por exemplo, deputadas como `@danielacarneiro` e `@silvyealves` apresentam altos escores em categorias como *body* e *sexual*, refletindo ataques centrados em aparência e sexualização. Ainda que existam variações, como no caso de `@rosangelawm`, associada a categorias mais contextuais, o padrão predominante indica uma forma de ataque orientada à objetificação e exposição pessoal.

No espectro de centro, observa-se um perfil mais equilibrado, sem concentração em categorias extremas. A deputada `@marinasilva` apresenta distribuição mais homogênea entre dimensões emocionais e discursivas, sugerindo ataques menos especializados e mais difusos.

De forma geral, os resultados indicam que a toxicidade incide com maior frequência sobre dimensões pessoais e identitárias do que sobre aspectos de desempenho profissional, favorecendo a objetificação moral e estética das parlamentares [Tavares and Recuero 2023]. Esse padrão sugere que a violência discursiva em ambientes políticos digitais se organiza em diferentes eixos, como os emocionais, identitários e corporais, que variam conforme o grupo alvo.

4.5. Discussão e Implicações

Os resultados deste estudo estão alinhados com trabalhos anteriores sobre discurso tóxico direcionado a mulheres na política em ambientes digitais, como os de [Tavares and Recuero 2023] e [Souza et al. 2022]. Assim como nesses estudos, observa-se que parte das mensagens vai além da crítica política e inclui ataques a atributos pessoais, identidade e aparência. No entanto, diferentemente da literatura existente, os achados deste trabalho indicam que essas manifestações não ocorrem de forma uniforme, mas se organizam em padrões distintos conforme o grupo político alvo. Em particular, observa-se que diferentes grupos concentram tipos específicos de ataque, o que sugere que a toxicidade deve ser compreendida como um fenômeno orientado ao alvo, estruturado em múltiplos eixos de violência discursiva.

De forma mais geral, os resultados revelam três padrões principais: (i) a toxicidade não é homogênea, variando entre grupos políticos; (ii) diferentes grupos concentram diferentes tipos de ataque, indicando uma especialização temática da violência discursiva; e (iii) essas diferenças se manifestam tanto no conteúdo quanto na forma linguística das mensagens. A forte presença de macrotemas como polarização partidária (T1) e insultos e desqualificação pessoal (T2) indica que o discurso tóxico está fortemente associado ao confronto político direto e ao uso de ataques pessoais, em linha com o aumento de agressividade em contextos de alta polarização [Zannettou et al. 2018].

Além disso, observa-se que diferentes temas concentram formas específicas de ataque. Temas como aborto e moralidade sexual (T5) e ataques à aparência e desumanização (T4) evidenciam ofensas voltadas à identidade e ao corpo das deputadas, enquanto temas como crítica às instituições (T3) e contestação eleitoral (T10) refletem ataques à legitimidade política. A presença de temas como desinformação (T8) e punição (T9) indica ainda formas indiretas de violência discursiva, baseadas em acusações, suspeitas e enquadramentos punitivos. Esses resultados mostram que a toxicidade não se limita a insultos explícitos, mas inclui diferentes estratégias discursivas de ataque.

A análise dos atributos psicolinguísticos complementa esses achados ao

evidenciar diferenças nos padrões emocionais e linguísticos das mensagens. Ao integrar modelagem de tópicos com análise lexical baseada no LIWC, torna-se possível caracterizar simultaneamente o conteúdo e a forma de expressão, oferecendo uma visão mais abrangente do fenômeno do que abordagens unidimensionais.

Esses resultados têm implicações diretas para a modelagem e moderação de conteúdo tóxico. Abordagens que tratam a toxicidade como um fenômeno uniforme podem falhar em capturar variações relevantes no tipo de ataque, especialmente em contextos políticos polarizados. Em particular, sistemas de detecção precisam ser sensíveis ao contexto e ao alvo, sendo capazes de diferenciar entre distintas formas de violência discursiva, como deslegitimação ideológica, ataques identitários e objetificação. Em síntese, a combinação entre análise temática e atributos psicolinguísticos permite compreender a toxicidade como um fenômeno estruturalmente diferenciado em ambientes políticos digitais.

5. Conclusão

Este trabalho investigou o discurso tóxico direcionado a deputadas federais brasileiras na plataforma X, com o objetivo de compreender como esse fenômeno se organiza em um contexto de alta polarização política.

Os resultados mostram que a toxicidade não se distribui de maneira uniforme, mas se estrutura em padrões distintos conforme o grupo político alvo. Em particular, observam-se diferenças não apenas na frequência das mensagens, mas também nos tipos de ataque mobilizados, envolvendo dimensões ideológicas, emocionais e identitárias.

Ao integrar detecção automática de toxicidade, modelagem de tópicos e análise psicolinguística, este estudo oferece uma abordagem multidimensional que permite caracterizar simultaneamente o conteúdo e a forma das mensagens. Essa combinação contribui para uma compreensão mais refinada da toxicidade em ambientes políticos digitais, evidenciando a importância de considerar o contexto e o alvo na análise desse fenômeno.

Do ponto de vista prático, os resultados sugerem que abordagens de detecção e moderação que tratam a toxicidade de forma agregada podem ser insuficientes para capturar suas diferentes manifestações. Nesse sentido, modelos mais sensíveis ao contexto podem contribuir para uma identificação mais precisa das dinâmicas de violência discursiva. Por fim, destacamos que os resultados estão associados ao contexto específico das eleições brasileiras de 2022 e ao conjunto de deputadas analisado, não sendo diretamente generalizáveis para outros cenários.

Como trabalhos futuros, propõe-se ampliar a análise para outros períodos e contextos políticos, investigando a evolução desses padrões ao longo do tempo. Também sugerimos explorar a estrutura das interações entre usuários e parlamentares, bem como incorporar outras formas de expressão, como emojis, para aprofundar a compreensão das dimensões emocionais e discursivas do conteúdo tóxico.

Agradecimentos. Agradecemos o apoio institucional do INCT-TILDIAR (processo nº 408490/2024-1), da CAPES, do CNPq, da FAPEMIG e da FAPESP.

Referências

Al-Hassan, A. and Al-Dossari, H. (2019). Detection of hate speech in social networks: a survey on multilingual corpus. In *6th international conference on computer science and information technology*, volume 10, pages 10–5121. ACM.

Alzahrani, M. and AlGhamdi, F. (2025). Social media sentiment analysis for sustainable rural event planning: A case study of agricultural festivals in al-baha, saudi arabia. *Sustainability*, 17(9):3864.

Amnesty International (2018). Women's experiences of violence and abuse on twitter. Acesso em: 23 jan. 2026.

Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *STIL*.

Belisário, K. M. and dos Reis, R. d. C. (2023). Representação feminina na cena política brasileira: Estereótipos e preconceitos. *Teoria & Pesquisa Revista de Ciência Política*, pages e023011–e023011.

Biroli, F. (2018). *Gênero e desigualdades: limites da democracia no Brasil*. Boitempo Editorial.

Bolognesi, B., Ribeiro, E., and Codato, A. (2022). Uma nova classificação ideológica dos partidos políticos brasileiros. *Dados*, 66:e20210164.

Carmo, I., Rêgo, A. L., Barreto, M., Schuler, M., Heine, A., Villas, M. V., and Lifschitz, S. (2023). Gerenciamento de dados de redes sociais com análise de redes e modelagem de tópicos. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 64–70. SBC.

Conselho Nacional de Saúde (2016). Resolução nº 510, de 7 de abril de 2016.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

de Groot, M., Aliannejadi, M., and Haas, M. R. (2022). Experiments on generalizability of bertopic on multi-domain short text. *arXiv preprint arXiv:2212.08459*.

G1 (2022). Carla zambelli tem contas retidas nas redes sociais. Acesso em: 24 jan. 2026.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794.

Hanu, L. and Team, U. (2020). Detoxify. `https://github.com/unitaryai/detoxify`.

Jigsaw (2018). Toxic comment classification challenge. `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`.

Khandelwal, T. (2025). Using llm-based approaches to enhance and automate topic labeling.

Koch, L., Russo Riva, M. P., and Steinert, J. I. (2025). Technology-facilitated gender-based violence against politically active women: A systematic review of psychological and political consequences and women's coping behaviors. *Trauma, Violence, & Abuse*.

Kozlowski, D., Pradier, C., and Benz, P. (2024). Generative ai for automatic topic labelling.

Krook, M. L. and Restrepo Sanín, J. (2016). Género y violencia política en américa latina. conceptos, debates y soluciones. *Política y gobierno*, 23(1):127–162.

Lima, L., Reis, J. C., Melo, P., Murai, F., and Benevenuto, F. (2020). Characterizing (un) moderated textual data in social systems. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 430–434.

Marques, F. P. J. A., Aquino, J. A. d., and Miola, E. (2014). Parlamentares, representação política e redes sociais digitais. *Opinião Pública*, 20:178–203.

Martins, V., Serafim, S. E. V., Pereira, L. C. R., Alves, A. F., Ferreira, C. H., and Almeida, J. M. (2025). A misoginia no youtube brasileiro: Um estudo de caso sobre o conteúdo produzido pela comunidade red pill. In *WebMedia*, pages 28–37. SBC.

Medvecki, D., Bašaragin, B., Ljajić, A., and Milošević, N. (2024). Multilingual transformer and bertopic for short text topic modeling: The case of Serbian. In *Disruptive Information Technologies for a Smart Society. ICIST 2023*, volume 872 of *Lecture Notes in Networks and Systems*, pages 161–173. Springer, Cham.

Melo, G. and Figueredo, F. (2025). O que torna uma frase tóxica? uma análise crítica de modelos especialistas em detecção de toxicidade. In *WebMedia*, pages 376–384. SBC.

Oltmanns, J. R., Khandelwal, R., Ma, J., Brickman, J., Do, T., Hussain, R., and Gupta, M. (2025). Language-based ai modeling of personality traits and pathology from life narrative interviews. *Journal of psychopathology and clinical science*.

OpenAI (2024). New embedding models and API updates. Blog post, January 25, 2024. Disponível em: `https://openai.com/index/new-embedding-models-and-api-updates/`.

Pennebaker, J. W., Booth, R. J., and Francis, M. E. (2007). Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135.

Pinho, T. R. d. (2020). Debaixo do tapete: a violência política de gênero eo silêncio do conselho de ética da câmara dos deputados. *Revista Estudos Feministas*, 28(2):e67271.

Poder360 (2022). Carla zambelli muda usuário no twitter e perde verificação. Acesso em: 10 jan. 2026.

Souza, L., Koch, L., Riva, M. P. R., and Gawi, R. (2022). Mensagens de ódio recebidas por candidatas negras e brancas durante as eleições no brasil de 2022 e suas implicações. *Estudos Eleitorais*, 16(2).

Steinbrenner, T., Lalk, C., Targan, K., Schaffrath, J., Eberhardt, S., Haberkamp, A., Lutz, W., and Rubel, J. (2025). Explaining anxiety prediction in psychotherapy transcripts: The role of patient linguistic features and theoretical constructs. *Behaviour Research and Therapy*, page 104857.

Tavares, C. Q. and Recuero, R. (2023). Toxicidade e violência discursiva contra deputadas federais no twitter. *Galáxia (São Paulo)*, 48:e62122.

X Corp. (2023). Developer platform policy. `https://docs.x.com/developer-terms/policy`. Acesso em: 28 fev. 2026.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G., and Blackburn, J. (2018). What is gab: A bastion of free speech or an alt-right echo chamber. In *The Web Conference*, pages 1007–1014.

Zedeus (2019). Nitter: A free and open source alternative twitter front-end. `https://github.com/zedeus/nitter`. Acesso em: 12 mar. 2026.