Submission date: 01/07/2025
Resubmission date: 16/10/2025
Camera ready submission date: 19/12/2025

1st round notification: 17/09/2025
2nd round notification: 16/12/2025
Available online: 18/03/2026
Publication date: 18/03/2026

Section: Regular article

Development and Application of Lexicons for Identifying Political Biases in Portuguese Texts
Desenvolvimento e Aplicação de Léxicos para Identificação de Vieses Políticos em Textos em Português

Isaac Araújo[1] **Jonas Oliveira** [1], **Samila Rodrigues** [1], **Yhasmim Tigre** [1], **Yuri Malheiros** [1]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa, Paraíba – Brasil

Abstract. *The Internet and social networks play a fundamental role in shaping political opinions, which makes it important to identify and analyze this type of positioning in online content. In this context, this study aims to develop two categories of lexicons to identify political biases in Portuguese texts. The first lexicon deals with economic biases (left-right) and the second with social biases (progressive-conservative). Using a corpus of 64,473 speeches by deputies, we created three versions of each lexicon, one using unigrams, other using bigrams, and the last using trigrams. In the experiments conducted, the lexicons demonstrated their effectiveness in capturing biases, revealing consistent patterns for politicians' positioning on social networks and more neutral biases for news websites. Thus, the created lexicons offer a straightforward and lightweight way to analyze online discourses, potentially contributing to studies in political science and media analysis.*

Keywords. *Natural Language Processing, Lexicons, Political Biases*

Resumo. *A Internet e as redes sociais desempenham um papel fundamental na formação de opiniões políticas, o que torna importante identificar e analisar este tipo de posicionamento em conteúdos digitais. Nesse contexto, este estudo tem como objetivo desenvolver duas categorias de léxicos para identificar vieses políticos nos textos em português brasileiro. O primeiro léxico trata dos vieses econômicos (esquerda-direita) e o segundo dos vieses sociais (progressista-conservador). Usando um corpus de 64.473 discursos de deputados, foram criadas três versões para cada léxico, sendo uma utilizando unigramas, outra bigramas e por fim uma utilizando trigramas. Nos experimentos conduzidos, os léxicos demonstraram sua eficácia na captura de vieses, revelando padrões consistentes para o posicionamento dos políticos nas redes sociais e discursos mais neutros dos portais de notícias. Dessa forma, foi construída uma abordagem de fácil uso e de baixo custo computacional para analisar discursos online, potencialmente contribuindo para estudos em ciência política e análise de mídias.*

Palavras-Chave. *Processamento de Linguagem Natural, Léxicos, Vieses Políticos*

Cite as: Araújo, I., Oliveira, J., Rodrigues, S., Tigre, Y., & Malheiros, Y. (2025). Development and Application of Lexicons for Identifying Political Biases in Portuguese Texts (Desenvolvimento e Aplicação de Léxicos para Identificação Vieses Políticos em Textos em Português). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 18(1), 13:1-13:20. DOI: 10.5753/isys.2025.5698

1. Introdução

A Internet consolidou-se como um dos principais meios de acesso à informação, exercendo um papel fundamental na formação e transformação de opiniões políticas. Essa influência se manifesta principalmente através das redes sociais e sites de notícias que disseminam textos, imagens e vídeos que moldam a percepção pública [Boulianne 2019]. Além disso, as redes sociais facilitam a comunicação direta entre usuários, estimulando discussões e participação ativa. A rápida disseminação de conteúdo nessas plataformas pode amplificar certas narrativas, influenciando significativamente a opinião pública em larga escala. [Campante et al. 2018] [Stieglitz and Dang-Xuan 2013] [Vergeer et al. 2013].

Logo, é importante ressaltar que o conteúdo digital é munido dos posicionamentos dos seus autores. Esses posicionamentos podem ser classificados dentro de espectros políticos, como o biaxial, em que existe um eixo econômico dividido entre esquerda e direita e um eixo social dividido em conservador e progressista. Suas combinações (esquerda-conservadora, esquerda-progressista, direita-conservadora e direita-progressista) tentam caracterizar as diferentes facetas de partidos políticos e movimentos sociais [Eysenck 1954] [Jost et al. 2009]. A Figura 1 traz uma representação visual do Espectro Político Biaxial com a presença de partidos políticos.

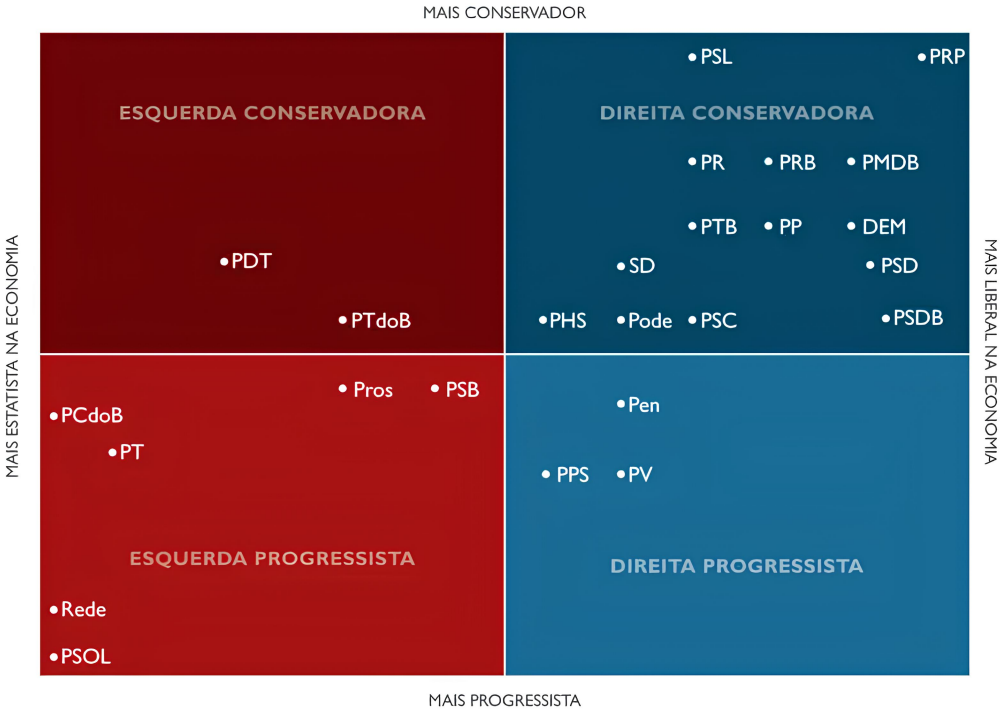

Figura 1. Espectro Político Biaxial com o posicionamento dos partidos brasileiros no período de 2015 a 2017 [Shalders 2017].

Muitas vezes os vieses dos conteúdos das plataformas não são explícitos, fazendo com que os usuários não tenham clareza sobre que tipos de conteúdos estão consumindo, estando à mercê dos sistemas de recomendação [Puglisi and Snyder Jr 2015]. Entretanto,

podemos automatizar a identificação de posicionamentos políticos em textos através de duas abordagens principais: aprendizagem de máquina e uso de léxicos [Taboada 2016] [Liu 2020]. Ambas as técnicas têm sido aplicadas com sucesso em estudos recentes.

A análise do posicionamento ideológico em textos políticos é uma tarefa fundamental para compreender como discursos refletem preferências econômicas e sociais ao longo do espectro político. Nesse contexto, o uso de léxicos específicos, baseados em n-gramas, oferece uma abordagem simples e automática para capturar nuances linguísticas e identificar padrões associados a diferentes orientações ideológicas. Entretanto, o desenvolvimento e validação desses léxicos representam desafios significativos, devido à necessidade de dados previamente classificados dentro do espectro político para a construção do léxico, além do planejamento de experimentos para a validação do léxico. Este estudo tem como objetivos obter e tratar textos políticos previamente classificados dentro do espectro político biaxial, construir léxicos sociais e econômicos a partir dos dados coletados, realizar experimentos que validem os léxicos construído e por fim usá-los para analisar os posicionamentos políticos de usuários em redes sociais e sites de notícias brasileiros.

Este trabalho tem como principal contribuição a construção de léxicos separados em duas categorias, a primeira delas busca analisar vieses econômicos, abrangendo o eixo esquerda-direita e a outra busca modelar visses sociais, cobrindo o eixo conservador-progressista. Para cada uma das categorias, são construídos três tipos de léxicos: o primeiro deles é composto por unigramas, o segundo é composto por bigramas e o terceiro é composto por trigramas. Dessa forma, é possível analisar não só a contribuição individual de cada termo para o vieses dos textos, mas também verificar a variação dos seus comportamentos na presença de outros termos. Todos os léxicos fornecem pontuações para n-gramas que indicam os vieses, assim permitindo uma análise detalhada do conteúdo textual. Para a construção destes recursos, utilizamos um *corpus* composto por discursos de deputados, garantindo assim uma base de dados representativa e relevante para o contexto político. Por fim, os léxicos estão disponíveis em *https://huggingface.co/datasets/yurimalheiros/br-political-biases-lexicons* para futuras pesquisas e aplicações sobre vieses econômicos e sociais.

2. Trabalhos Relacionados

Nesta seção a abordagem proposta será comparada com outras existentes na literatura, explicando a viabilidade da solução e ressaltando sua relevância, originalidade e diferenciais, especialmente no que se refere à simplicidade no manuseio dos léxicos e à interpretação dos resultados gerados.

A relação entre palavras e posicionamento político tem sido amplamente discutida na literatura. [Shiroma et al. 2005] argumenta que a linguagem em documentos oficiais não apenas reflete ideologias, mas também as constitui. Os autores destacam que o discurso pode ser uma prática política capaz de estabelecer, manter e transformar relações de poder. Por exemplo, os textos educacionais frequentemente incorporam significados sociais que se inserem em disputas políticas mais amplas, contribuindo para a construção de hegemonias discursivas. Termos como “flexibilidade”, “gestão” e “inovação” são frequentemente empregados em documentos educacionais, demonstrando a influência da ideologia neoliberal ao alinhar o discurso educacional às demandas do mercado. Dessa

forma, conclui-se que as palavras presentes em discursos políticos carregam os vieses de seus locutores.

No contexto da mídia tradicional, [Mundim 2018] analisa o viés presente nos discursos dos veículos de informação no Brasil. O autor discute a percepção de que os proprietários dos meios de comunicação sacrificam seu lucro por ideologia e pelo desejo de influenciar eleições, uma visão que encontra respaldo em estudos críticos sobre o tema. Para investigar essa questão, foi formulada a hipótese de que o espaço concedido aos candidatos nos jornais poderia estar em desacordo com sua posição na disputa eleitoral. Testes estatísticos, como Granger, ANOVA e HSD de Tukey, foram aplicados para avaliar a existência de padrões sistemáticos de viés na cobertura midiática. Embora tenha sido identificado viés no conteúdo opinativo, não se observou um padrão claro e sistemático de preferência por um determinado partido ou candidato. Assim, análises posteriores podem contribuir para compreender se há, de fato, um posicionamento neutro nessas plataformas.

As redes sociais desempenham um papel cada vez mais relevante na disseminação de discursos políticos. [Resende et al. 2018] destaca que o *WhatsApp* oferece um espaço para discussões públicas de grande alcance, incluindo grupos voltados ao ativismo político e à organização de movimentos sociais. A fim de investigar esses ambientes, foi realizada uma busca por grupos públicos com temáticas políticas, classificando-os em categorias como Notícias, Debates, Direita, Ideologias, Pró-Lula, Pró-Bolsonaro e Partidos, com base no nome do grupo e no conteúdo das mensagens. Essa categorização permite avaliar os vieses presentes nos discursos de grupos e personalidades nas redes sociais, ressaltando a importância do estudo dessas dinâmicas.

Para aprofundar a análise dos vieses em plataformas digitais, técnicas de aprendizagem de máquina têm sido empregadas, desde que existam conjuntos de dados previamente classificados. Um exemplo dessa abordagem é apresentado por [Bestvater and Monroe 2023], que analisa textos em redes sociais utilizando classificadores e análise sentimento e posicionamento político. O estudo revela a complexidade da tarefa, demonstrando que sentimentos positivos nem sempre indicam apoio político. Em muitos casos, a oposição a um protesto pode se manifestar como aprovação ao *status quo* ou como crítica a um movimento específico.

Para língua portuguesa, o trabalho de [Cerqueira et al. 2025] investiga como extrair conhecimento da participação cidadã no portal E-democracia da Câmara dos Deputados brasileira através de análise de comentários sobre projetos de lei. O objetivo é compreender como a opinião pública sobre assuntos legislativos evolui ao longo do tempo, combinando técnicas de detecção de posicionamento (*stance detection*) e modelagem de tópicos. O trabalho empregou o modelos treinados em um *corpus* híbrido que combina dados traduzidos com conteúdo específico de projetos brasileiros. Os resultados demonstraram que o modelo alcançou 73% de F1-*score* e as visualizações temporais revelaram padrões de concentração de comentários em períodos específicos e a predominância de um tópico principal por projeto.

Utilizando análise de tópicos em discursos de senadores brasileiros, o trabalho de [Pinheiro and Faleiros 2022] tem como objetivo identificar padrões temáticos ao longo do tempo e verificar se existe correlação entre a evolução temporal dos tópicos e eventos

históricos, políticos e econômicos. O método empregado utilizou a técnica de *Latent Dirichlet Allocation* sobre uma base de dados de 81.858 discursos de senadores coletados entre 1995 e 2019. Os resultados demonstraram que a abordagem foi capaz de identificar correlações temporais claras em temas populares como corrupção, esportes, eleições e aviação, validando parcialmente a hipótese inicial, embora alguns tópicos não tenham apresentado padrões temporais interpretáveis.

Outra abordagem para detecção de viés envolve a construção e utilização de léxicos especializados. [Hu et al. 2019] exemplifica essa metodologia ao desenvolver um léxico voltado à identificação de vieses em n-gramas relacionados aos partidos Republicanos e Democratas nos Estados Unidos. Esse léxico foi utilizado para analisar o viés presente em *snippets* (pequenos resumos) exibidos abaixo dos links em páginas de busca de notícias políticas. Os resultados apontaram que o *Google Search* tende a amplificar o viés original da notícia em seus *snippets*, influenciando potencialmente a seleção de conteúdo pelos leitores.

A presente pesquisa diferencia-se das anteriores ao modelar os vieses nos eixos social e econômico, proporcionando uma análise abrangente das dimensões ideológicas contidas nos conteúdos políticos. Para isso, utiliza n-gramas do português brasileiro, permitindo uma abordagem linguística adaptada ao contexto local. Além disso, caracteriza-se por ser uma metodologia automática e de baixo custo computacional, viabilizando sua aplicação em diversos cenários sem necessidade de recursos computacionais robustos. Diferentemente das abordagens tradicionais focadas na classificação de textos, esta pesquisa prioriza a quantificação dos vieses, oferecendo uma maneira objetiva de medir e compreender as tendências ideológicas em conteúdos políticos escritos em português brasileiro.

3. Metodologia

Esta seção detalha a metodologia empregada no estudo, que se divide em três etapas principais. Primeiramente, realizamos a obtenção dos dados de discursos dos deputados, coletando um *corpus* representativo para análise. Em seguida, procedemos à construção dos léxicos, desenvolvendo um conjunto de termos relevantes para a identificação de vieses sociais e econômicos nos textos. Por fim, aplicamos os léxicos construídos em três experimentos, analisando seu desempenho em diferentes contextos, como em postagens em redes sociais e em artigos de portais de notícias brasileiros.

3.1. Obtenção dos Dados

O governo brasileiro disponibiliza a transcrição completa de todos os pronunciamentos ocorridos na câmara dos deputados desde 1946 até os dias atuais[1]. No banco de discursos, é possível encontrar informações como data e texto na íntegra, além do nome e partido do locutor do discurso.

Dessa maneira, foi possível fazer um *web scraping* no banco de discursos, com o objetivo de obter dados para construção dos léxicos sociais e econômicos. Nesse estudo,

[1] https://www2.camara.leg.br/atividade-legislativa/discursos-e-notas-taquigraficas

foram coletados textos que datam de janeiro de 2020 até dezembro de 2023, com o objetivo de modelar os vieses atuais dos temas mais discutidos na Câmara dos Deputados. Ao todo foram coletados 64.473 pronunciamentos. Os dados do período analisado estavam no formato *HTML*, por isso as bibliotecas *Selenium* e *Scrapy* da linguagem *Python* foram usadas para a obtenção dos dados.

Nem todas as informações presentes nos discursos eram relevantes para a construção do léxico, portanto foi necessário realizar um pré-processamento nos textos. Entre os principais problemas estava a presença de expressões não associadas aos discursos dos parlamentares, como as interrupções sonoras ocorridas no momento da fala, por exemplo, "(Soa a campainha.)" e "(Palmas.)", ou debates entre ouvintes do pronunciamento.

Assim, a etapa de pré-processamento dos textos teve como principal objetivo a remoção de palavras que não estavam ligadas às falas dos locutores, além disso caracteres numéricos e sinais de pontuação foram retirados dos dados, já que não possuem semântica relevante para os léxicos. Adicionalmente, *stopwords* da língua portuguesa foram desconsideradas apenas para a construção dos léxicos com unigramas, já que se tratam de palavras comuns que não devem possuir vieses relevantes quando analisadas de forma isolada.

3.2. Construção do Léxico

Com os dados dos discursos obtidos, a próxima etapa foi a classificação dos discursos dentro do espectro político biaxial de acordo com o posicionamento do partido que o locutor pertence. Vale ressaltar que o posicionamento político dos partidos foi obtido a partir do compasso político da BBC Brasil[2], que foi construído através da análise de dez votações relevantes na Câmara dos Deputados [Shalders 2017]. O gráfico da Figura 1 representa o posicionamento social e econômico demonstrado por cada um dos partidos políticos entre 2015 e 2017.

O conjunto de dados coletado possui locutores que estão em 29 partidos distintos, dos quais 10 foram classificados no âmbito econômico como partidos de esquerda e 19 como partidos de direita. Além disso, 17 deles foram classificados no âmbito social como conservadores e 12 como progressistas. Assim, para a construção dos léxicos, foram usados 64.473 discursos dos quais 33.505 (51.97%) foram proferidos por partidos de esquerda e 30.968 (48.03%) por partidos de direita. No âmbito social, 32.276 (50.06%) foram de partidos progressistas e 32.197 (49.94%) de partidos conservadores.

Vale lembrar que no período de 2017 até 2023 alguns partidos passaram por mudanças, como alterações de nomenclatura ou fusão com outros partidos, por exemplo, o partido União Brasil (UNIÃO) surgiu da fusão entre o Democratas (DEM) e o Partido Social Liberal (PSL) em 6 de outubro de 2021. Logo, parte dos partidos políticos do conjunto de dados não está listada no compasso da BBC Brasil. Nesses casos, suas páginas oficiais foram utilizadas como referência para classificá-los nos âmbitos social e econômico. Adicionalmente, consultamos as páginas da *Wikipedia* correspondentes a

[2]Subsidiária da *British Broadcasting Corporation* (BBC) no Brasil - https://www.bbc.com/portuguese

cada partido para verificar possíveis mudanças recentes em suas denominações ou eventuais fusões partidárias, garantindo assim a precisão e atualidade das informações utilizadas na classificação. As classificações nos eixos social e econômico dos partidos políticos que possuem discursos no conjunto de dados podem ser vistas na Tabela 1.

Tabela 1. Classificações nos eixos social e econômico de alguns partidos políticos brasileiros

Partido	Eixo Econômico	Eixo Social
PT	Esquerda	Progressista
PSOL	Esquerda	Progressista
PCDOB	Esquerda	Progressista
PSB	Esquerda	Progressista
PDT	Esquerda	Conservador
REDE	Esquerda	Progressista
SOLIDARIEDADE	Esquerda	Progressista
PCdoB	Esquerda	Progressista
PROS	Esquerda	Progressista
AVANTE	Esquerda	Conservador
PL	Direita	Conservador
NOVO	Direita	Conservador
PSD	Direita	Conservador
PP	Direita	Conservador
PSL	Direita	Conservador
REPUBLICANOS	Direita	Conservador
MDB	Direita	Conservador
CIDADANIA	Direita	Progressista
UNIÃO	Direita	Conservador
PSDB	Direita	Conservador
DEM	Direita	Conservador
PV	Direita	Progressista
PSC	Direita	Conservador
PTB	Direita	Conservador
PATRIOTA	Direita	Progressista
PR	Direita	Conservador
PODEMOS	Direita	Conservador
PPS	Direita	Progressista
PMDB	Direita	Conservador

Por conseguinte, as frequências de todos os unigramas, bigramas e trigramas presentes no conjunto de dados foram verificadas, possibilitando o cálculo das probabilidades para cada uma delas de acordo com o eixo analisado. Vale ressaltar que todos os n-gramas que apareceram menos de 20 vezes foram retirados da análise, já que na maior parte dos casos correspondiam a coloquialismos ou erros gramaticais. Outrossim, os n-gramas que possuíam poucas aparições no conjunto de dados tornaram-se termos altamente enviesados para um dos eixos sociais ou econômicos, dessa forma muitos *outliers* se fariam presentes nos experimento posteriores caso os termos pouco frequentes não fossem desconsiderados.

Foram calculadas quatro probabilidades $P_j(p)$ para cada um dos n-gramas, onde j representa os grupos de esquerda, direita, conservador ou progressista. Cada uma das

probabilidades indica a chance do n-grama aparecer no respectivo grupo. Dessa maneira seus cálculos são feitos de forma semelhante, como pode ser vista na Equação 1. Onde $n(p_j)$ indica a frequência de um n-grama nos discursos do grupo j, enquanto $n(S_j)$ indica a quantidade total de n-gramas nos discursos do grupo.

$$P_j(p) = \frac{n(p_j)}{n(S_j)} \tag{1}$$

Assim, é possível finalizar a construção dos léxicos através dos cálculos dos vieses dos n-gramas nos âmbitos econômico (Equação 2) e social (Equação 3).

$$V_{economico}(p) = \frac{P_{direita}(p) - P_{esquerda}(p)}{P_{direita}(p) + P_{esquerda}(p)} \tag{2}$$

$$V_{social}(p) = \frac{P_{conservador}(p) - P_{progressista}(p)}{P_{conservador}(p) + P_{progressista}(p)} \tag{3}$$

É importante ressaltar que os léxicos foram separados de acordo com a quantidade de termos em observação, dessa forma existe para cada eixo um léxico apenas com unigramas, outro apenas com bigramas e outro com apenas com trigramas, totalizando 3 versões para cada eixo do espectro político biaxial. Tendo em vista que os léxicos sociais e econômicos foram construídos a partir do mesmo conjunto de dados, existem um total de 21.872 unigramas, 95.107 bigramas e 79.823 trigramas. Note que os valores numéricos resultantes das Equações 2 e 3 estão contidos no intervalo $[-1, 1]$, logo quanto mais próximo de -1, mais a palavra está enviesada a categoria progressista ou esquerda e quanto mais próximo de 1, a palavra tende para a categoria conservador ou direita.

Usando os valores dos vieses econômicos e sociais dos n-gramas, é possível calcular os vieses de textos inteiros, como visto nas Equações 4 e 5 em que S é representa o conjunto de n-gramas do texto que possuem *score* definido no léxico, $|S|$ é a quantidade de n-gramas em S e s representa um n-grama que pertence ao conjunto S.

$$V_{economico}(S) = \frac{\sum_{s \in S} V_{economico}(s)}{|S|} \tag{4}$$

$$V_{social}(S) = \frac{\sum_{s \in S} V_{social}(s)}{|S|} \tag{5}$$

3.3. Experimentos

Para avaliar os léxicos criados, foram realizados três experimentos. O primeiro deles consistiu em utilizar os léxicos em textos de postagens do *X* (antigo *Twitter*) das 10 personalidades mais seguidas da plataforma conforme listado por [Superlistas 2023]. A partir dos resultados obtidos foi possível classificar os conteúdos dentro do espectro político biaxial e quantificar os seus vieses.

Para esse experimento foram coletadas as 1.000 postagens mais recentes de cada um dos perfis analisados. Depois disso, os textos de cada uma das postagens foram extraídos para que pudessem passar pela análise léxica. Dessa forma, foi possível calcular os vieses sociais e econômicos dos textos através das Equações 4 e 5.

O segundo experimento foi feito com perfis de figuras políticas escolhidas a partir dos dados dos discursos obtidos no site da câmara dos deputados. Foram analisados os perfis do *X* dos 10 oradores com mais falas de cada divisão econômica (esquerda e direita) e social (progressista e conservador). Esse experimento teve como principal objetivo a verificação da coerência do léxico, pois os posicionamentos sociais e econômicos desses perfis são previamente conhecidos, logo basta observar se os vieses fornecidos pelo léxico estão de acordo com o posicionamento do partido que cada perfil pertence. É importante ressaltar que até a data da coleta dos dados nem todas as personalidades políticas possuíam 1.000 publicações, nesses casos foram analisadas todas postagens do perfil.

Por fim, no terceiro experimento, os léxicos foram testados em notícias políticas dos portais de notícias brasileiros G1[3] e UOL[4], com o objetivo de verificar se seus vieses tendem à algum dos lados ou se estão mais próximos da neutralidade. Os textos das notícias foram coletados via *scraping* e em ambos os casos foram obtidas as notícias mais recentes da temática política. Foram obtidos 440 textos do G1 Brasil e do UOL em 10 de novembro de 2024.

4. Resultados

A Tabela 2 exibe a quantidade de termos em cada um dos léxicos e suas distribuições por eixo político, vale lembrar que n-gramas com *Score* Social maior do que 0 foram classificados como Conservadores e caso contrário como Progressistas, além disso n-gramas com *Score* Econômico maior do que 0 foram colocados no grupo da Direita e os demais no grupo da Esquerda. A Tabela 3 traz a média, mediana e desvio padrão dos *scores* dos léxicos.

Tabela 2. Quantidade de Unigramas, Bigramas e Trigramas por Eixo Político

Eixo	Unigramas	Bigramas	Trigramas
Conservador	12.752 (58,30%)	51.720 (54,38%)	42.055 (52,69%)
Progressista	9.120 (41,70%)	43.387 (45,62%)	37.768 (47,31%)
Direita	12.589 (57,56%)	51.525 (54,18%)	41.607 (52,12%)
Esquerda	9.283 (42,44%)	43.582 (45,82%)	38.216 (47,88%)
Qtd. Total	21.872 (100,00%)	95.107 (100,00%)	79.823 (100,00%)

Os vieses sociais e econômicos dos textos das postagens das dez personalidades brasileiras mais seguidas no *X* foram calculados usando os léxicos com unigramas, depois

[3]Portal de notícias do grupo Globo - https://g1.globo.com
[4]Universo Online - https://uol.com.br

Tabela 3. Tabela com a média, mediana e desvio padrão dos *Scores* Econômicos e Sociais

Unigramas		
	Score Econômico	**Score Social**
Média	0.040315	0.048106
Mediana	0.055528	0.054524
σ	0.333741	0.326938
Bigramas		
	Score Econômico	**Score Social**
Média	0.016823	0.022710
Mediana	0.039241	0.033240
σ	0.324634	0.318952
Trigramas		
	Score Econômico	**Score Social**
Média	0.001185	0.007312
Mediana	0.014946	0.009629
σ	0.362681	0.356307

os léxicos de bigramas e por fim os léxicos de trigramas. A Figura 2 mostra o comportamento dos vieses resultantes do uso do léxicos com unigramas em cada uma das personalidades, enquanto a Figura 3 traz o comportamento dos vieses calculados usando os léxicos com bigramas e a Figura 4 retrata os vieses resultantes do léxicos com trigramas.

De forma análoga, foram analisadas os vieses sociais de deputados assumidamente conservadores e progressistas e os vieses econômicos das falas dos deputados pertencentes a partidos de direita e a partidos de esquerda. A Tabela 4 traz os comportamentos dos vieses sociais das postagens do *X* de dez personalidades conservadoras calculados separadamente usando os léxicos de unigramas, bigramas e trigramas, enquanto a Tabela 5 faz o mesmo com personalidades previamente classificadas como progressistas. Outrossim, a Tabela 6 exibe os comportamentos dos vieses econômicos dos conteúdos publicados por personalidades de esquerda e a Tabela 7 retrata os vieses econômicos das postagens de dez deputados de direita.

Por conseguinte, foram verificados os vieses sociais e econômicos das notícias coletadas do G1 Brasil e do UOL para os léxicos com unigramas, bigramas e trigramas. A Figura 5 ilustra os comportamentos dos vieses calculados com os léxicos de unigramas, bigramas e trigramas.

Tabela 4. Viés social de políticos conservadores no X

Usuário	Unigramas		Bigramas		Trigramas	
	Média	σ	**Média**	σ	**Média**	σ
@DraManato	0,0451	0,0922	0,0524	0,1047	0,0642	0,1680
@GenPeternelli	0,0717	0,1128	0,0830	0,1361	0,0940	0,1937
@PompeodeMattos	0,0595	0,0513	0,0313	0,0588	-0,0294	0,1645
@RochaHildorocha	0,0533	0,0728	0,0652	0,0780	0,0756	0,1491
@TiagoMitraud	0,0235	0,0665	0,0238	0,0676	0,0221	0,1449
@adriventurasp	0,0382	0,0773	0,0449	0,0863	0,0667	0,1617
@bibonunes1	0,0589	0,0776	0,0698	0,1180	0,0988	0,2197
@cabogilberto	0,0456	0,0872	0,0763	0,1176	0,0995	0,2268
@marcelvanhattem	0,0707	0,1081	0,0761	0,1131	0,0950	0,1931
@pauloganime	0,0572	0,0783	0,0595	0,0820	0,0817	0,1468

Tabela 5. Viés social de políticos progressistas no X

Usuário	Unigramas		Bigramas		Trigramas	
	Média	σ	**Média**	σ	**Média**	σ
@Alice_Portugal	-0,0256	0,0557	-0,0229	0,0948	-0,0745	0,1274
@BiradoPindare	0,0125	0,1270	0,0089	0,1609	-0,0011	0,2606
@HenriqueFontana	-0,0168	0,0388	-0,0131	0,0799	-0,0068	0,2505
@IvanValente	-0,0375	0,0930	-0,0564	0,1073	-0,0644	0,2097
@JoeniaWapichana	-0,0885	0,1058	-0,0969	0,1190	-0,1301	0,2083
@erikakokay	-0,0358	0,0820	-0,0517	0,0952	-0,0725	0,1831
@fernandapsol	-0,0534	0,1135	-0,0809	0,1280	-0,0913	0,2203
@jandira_feghali	-0,0293	0,0821	-0,0389	0,0975	-0,0491	0,1601
@perpetua_acre	0,0365	0,0951	0,0086	0,0658	0,0333	0,1362
@taliriapetrone	-0,0312	0,0895	-0,0428	0,0963	-0,0692	0,1905

5. Discussão

Inicialmente, é possível constatar uma tendência dos léxicos sociais para o categoria conservador e dos léxicos econômicos para o lado da direita, apesar do número de discursos progressistas e de esquerda ser levemente maior no conjunto de dados usado para construir os léxicos. A Tabela 2 mostra que a maior parte dos n-gramas presentes nos léxicos possuem *scores* sociais e econômicos positivos, ou seja, são conservadores e de direita.

A Tabela 3 retrata as nuances entre os comportamentos dos *scores* dos léxicos de unigramas, bigramas e trigramas. Observando os parâmetros de média e desvio padrão, é possível constatar que a tendência econômica para a direita e social para o lado conservador é mais fraca nos léxicos constituídos por trigramas, já que as médias de ambos os *scores* estão muito próximas de zero. Além disso é possível verificar uma grande quantidade de *outliers*, já que os trigramas normalmente possuem baixas frequências relativas, fazendo com que seus *scores* tendam aos extremos, assim como os vieses dos textos analisados por esse tipo de léxico. Dessa forma, conforme o número de termos na sequência aumenta, os vieses observados nos experimentos tendem a se acentuar em direções mais extremas. Isso ocorre porque sequências mais longas apresentam menor frequência relativa no *corpus*, tornando a distribuição dos dados mais esparsa e suscetível a flutuações estatísticas.

Tabela 6. Viés econômico de políticos de esquerda no X

Usuário	Unigramas		Bigramas		Trigramas	
	Média	σ	**Média**	σ	**Média**	σ
@Alice_Portugal	-0.0318	0.0553	-0.0134	0.1216	-0.0695	0.1215
@BiradoPindare	0.0110	0.1285	0.0092	0.1597	-0.0060	0.2617
@HenriqueFontana	-0.0258	0.0386	-0.0149	0.0688	-0.0121	0.2156
@JoeniaWapichana	-0.0818	0.1094	-0.0924	0.1235	-0.1276	0.2120
@PompeodeMattos	-0.0030	0.0769	-0.0162	0.0977	-0.0853	0.2199
@erikakokay	-0.0364	0.0854	-0.0525	0.0990	-0.0744	0.1879
@fernandapsol	-0.0623	0.1087	-0.0821	0.1352	-0.0781	0.2300
@jandira_feghali	-0.0322	0.0839	-0.0404	0.1045	-0.0523	0.1650
@perpetua_acre	0.0372	0.1006	0.0101	0.0671	0.0299	0.1551
@taliriapetrone	-0.0358	0.0869	-0.0450	0.0952	-0.0717	0.1942

Tabela 7. Viés econômico de políticos de direita no X

Usuário	Unigramas		Bigramas		Trigramas	
	Média	σ	**Média**	σ	**Média**	σ
@AlexManente23	0,0265	0,0682	0,0279	0,0752	0,0231	0,1713
@ArnaldoJardim	0,0667	0,1097	0,0562	0,1034	0,0594	0,1648
@GenPeternelli	0,0695	0,1225	0,0759	0,1371	0,0832	0,2023
@RochaHildorocha	0,0562	0,0726	0,0601	0,0783	0,0589	0,1534
@TiagoMitraud	0,0215	0,0684	0,0232	0,0706	0,0236	0,1497
@adriventurasp	0,0354	0,0778	0,0438	0,0887	0,0650	0,1656
@bibonunes1	0,0529	0,0787	0,0646	0,1193	0,0925	0,2233
@cabogilberto	0,0399	0,0861	0,0686	0,1144	0,0918	0,2291
@marcelvanhattem	0,0694	0,1089	0,0742	0,1147	0,0921	0,1935
@pauloganime	0,0560	0,0780	0,0562	0,0792	0,0766	0,1462

Os resultados obtidos através da aplicação dos léxicos desenvolvidos revelam padrões no posicionamento social e econômico dos dados analisados. Nos 10 perfis brasileiros mais seguidos no *X*, podemos perceber um viés conservador e de direita nas Figuras 2, 3 e 4, já que a maior parte dos *scores* foram positivos em todos as personalidades e em todos os léxicos. Mesmo perfis geralmente associados a posicionamentos mais progressistas ou de esquerda apresentaram *scores* positivos tanto no eixo econômico quanto no social. Por exemplo, @*felipeneto*, conhecido por suas críticas ao governo de direita, apresentou vieses econômicos e sociais médios positivos, entretanto o postagens também possuem vieses de esquerda e progressista. Um dos possíveis motivos para isso é que mesmo com um posicionamento progressista e de esquerda, um perfil pode escrever sobre temas conservadores e de direita com a finalidade de criticá-los. A seguinte fala exemplifica esse fenômeno, pois foi publicada pelo influenciador brasileiro Felipe Neto no *X* e apresentou tanto viés social quanto viés econômico positivo, conforme indicado pelo léxico de unigramas.

> "A revista 'oeste', de extrema-direita, lançou esta capa, implorando por anistia.
> É de uma falta de vergonha na cara impressionante. A turma q diz "bandido bom é bandido morto", agora pede anistia aos seus criminosos, de joelhos e chorando.

Paguem por seus crimes!
SEM ANISTIA”

Assim, uma interpretação para o resultado com os perfis mais seguidos do *X* é que os temas conservadores e direita estão sendo mais abordados que os progressistas e de esquerda. É importante ressaltar que grande parte dos postagens coletadas e analisadas não estavam associadas a temática política, por exemplo, o perfil do *@neymarjr* tem diversos conteúdos voltados à publicidade de produtos.

Ademais, é possível constatar a alta variabilidade dos *scores* com trigramas em comparação com os bigramas pela marcante presença dos *outliers* na Figura 4. Entretanto, além da comportamento mais extremos previsto para o léxico de trigramas, outra explicação para variância dos *scores* das análises dos trigramas é o tamanho dos textos das postagens no *X*, pois normalmente esses conteúdos não possuem muitas palavras, fazendo com que poucos trigramas possam definir o viés de todo o conteúdo.

As Tabelas 4, 5, 6 e 7 demonstram que os léxicos foram capazes de capturar, em grande parte, os posicionamentos econômicos e sociais esperados das postagens do *X* dos políticos analisados. Políticos de esquerda e progressistas tenderam a apresentar *scores* negativos ou próximos de zero, enquanto políticos de direita e conservadores apresentaram *scores* predominantemente positivos. Por exemplo, *@fernandapsol*, conhecida por suas posições de esquerda, apresentou um viés econômico de médio negativo em todos os léxicos, enquanto *@marcelvanhattem*, associado à direita, econômico de médio positivos em todos os léxicos. De modo similar, *@JoeniaWapichana*, reconhecida por suas pautas progressistas, apresentou um viés social médio negativo em todos léxicos, já *@GenPeternelli* identificado com posições mais conservadoras, obteve um *score* social médio positivo em todos os léxicos. Esta coerência sugere que o léxico desenvolvido possui boa capacidade de discriminação entre diferentes posicionamentos políticos.

A análise dos portais de notícias G1 e UOL, apresentadas na Figura 5, revela *scores* com uma inclinação positiva tanto no eixo econômico quanto no social. Esses resultados podem indicar uma tendência conservadora e de direita destes portais em sua cobertura política ou podem ser justificado pela tendencia positiva dos léxicos construídos. No entanto, é importante notar que as médias dos valores são próximas de zero e que os desvios padrões também são próximos de zero. Assim, estes resultados podem sugerir uma preocupação com a neutralidade e um posicionamento mais moderado destes sites. Outrossim, também é importante observar a baixa presença de *outliers* nesses gráficos, isso pode ser justificado pelo fato de que as notícias políticas normalmente são textos com mais n-gramas do que o as postagens no *X*, de forma que muitos termos devem ser levados em consideração para o cálculo dos vieses sociais e econômicos. Além disso, apesar de estar próximo da neutralidade, os léxicos com trigramas geraram dados com variância maior do que os unigramas e bigramas, constatando o comportamento previsto.

6. Conclusão

Este trabalho apresentou a construção e aplicação de léxicos para análise de posicionamento político nos eixos econômico e social. Os léxicos foram desenvolvidos utilizando um *corpus* de 64.473 discursos parlamentares e aplicados em três experimentos: análise de perfis populares no *X*, análise de perfis de políticos, e análise de portais de notícias.

Os resultados demonstraram a eficácia dos léxicos na captura de nuances ideológicas em diferentes contextos. Na análise dos perfis mais seguidos no *X*, observou-se uma tendência para temas conservadores e de direita, mesmo em perfis associados a posições progressistas ou de esquerda. Isso sugere que tópicos conservadores e de direita estão sendo mais discutidos na plataforma. Na análise dos perfis de políticos, os léxicos mostraram-se capazes de discriminar efetivamente entre diferentes posicionamento, alinhando-se às expectativas baseadas nas afiliações partidárias conhecidas. Por fim, a análise dos portais de notícias G1 e UOL revelou uma leve inclinação conservadora e de direita, mas com uma tendência à neutralidade.

Com isso, a principal contribuição deste trabalho é o desenvolvimento de um método para a análise automatizada de posicionamento social e econômico em textos em português. Os léxicos criados oferecem um meio objetivo e escalável de avaliar os vieses em uma variedade de contextos, desde discursos parlamentares até postagens em redes sociais e artigos de notícias. Isso pode ter aplicações significativas em estudos de ciência política, análise de mídia e compreensão da opinião pública.

Para trabalhos futuros, é recomendada a expansão dos léxicos para incluir discursos mais atuais. Além disso, seria valioso aplicar os léxicos em uma análise para rastrear mudanças nos discursos políticos ao longo do tempo e nos vieses das diversas plataformas que interagimos. Seria igualmente relevante incorporar análises de sentimento para identificar possíveis descontentamentos dos locutores em relação às temáticas políticas abordadas em seus discursos. Este trabalho possui algumas limitações que precisam ser destacadas. Os léxicos foram desenvolvidos com base em discursos parlamentares, o que pode não capturar completamente as nuances da linguagem política em outros contextos. Adicionalmente, os léxicos são dependentes de classificações prévias dos partidos que nem sempre são claras, pois os partidos e seus membros podem adotar posicionamentos dinâmicos de acordo com fatores variados.

7. Agradecimentos

Este trabalho foi apoiado pela Universidade Federal da Paraíba (UFPB) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) através do Programa Institucional de Bolsas de Iniciação Científica (PIBIC).

Referências

[Bestvater and Monroe 2023] Bestvater, S. E. and Monroe, B. L. (2023). Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis*, 31(2):235–256.

[Boulianne 2019] Boulianne, S. (2019). Revolution in the making? social media effects across the globe. *Information, communication & society*, 22(1):39–54.

[Campante et al. 2018] Campante, F., Durante, R., and Sobbrio, F. (2018). Politics 2.0: The multifaceted effect of broadband internet on political participation. *Journal of the European Economic Association*, 16(4):1094–1136.

[Cerqueira et al. 2025] Cerqueira, M., da Silva, N. F., Souza, E., Albuquerque, H. O., Dias, M. d. S., and de Carvalho, A. C. (2025). Not only what, but also when: Understanding

brazilian political comments on legislative bills over time through stance detection and topic modeling. In *Conference on Digital Government Research*, volume 1.

[Eysenck 1954] Eysenck, H. J. (1954). *The psychology of politics*, volume 2. Transaction publishers.

[Hu et al. 2019] Hu, D., Jiang, S., E. Robertson, R., and Wilson, C. (2019). Auditing the partisanship of google search snippets. In *The World Wide Web Conference*, pages 693–704.

[Jost et al. 2009] Jost, J. T., Federico, C. M., and Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual review of psychology*, 60(1):307–337.

[Liu 2020] Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

[Mundim 2018] Mundim, P. S. (2018). O viés da cobertura política da imprensa nas eleições presidenciais brasileiras de 2002, 2006 e 2010. *Revista Brasileira de Ciência Política*, (25):7–46.

[Pinheiro and Faleiros 2022] Pinheiro, V. and Faleiros, T. (2022). Aplicação de modelos de tópicos em análises automatizadas de discursos de senadores brasileiros. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 612–623, Porto Alegre, RS, Brasil. SBC.

[Puglisi and Snyder Jr 2015] Puglisi, R. and Snyder Jr, J. M. (2015). Empirical studies of media bias. In *Handbook of media economics*, volume 1, pages 647–667. Elsevier.

[Resende et al. 2018] Resende, G., Messias, J., Silva, M., Almeida, J., Vasconcelos, M., and Benevenuto, F. (2018). A system for monitoring public political groups in whatsapp. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, WebMedia '18, page 387–390, New York, NY, USA. Association for Computing Machinery.

[Shalders 2017] Shalders, A. (2017). Direita ou esquerda? análise de votações indica posição de partidos brasileiros no espectro ideológico. https://www.bbc.com/portuguese/brasil-41058120. Acessado em: 03 de julho de 2024.

[Shiroma et al. 2005] Shiroma, E. O., Campos, R. F., and Garcia, R. M. C. (2005). Decifrar textos para compreender a política: subsídios teórico-metodológicos para análise de documentos. *Perspectiva*, 23(2):427–446.

[Stieglitz and Dang-Xuan 2013] Stieglitz, S. and Dang-Xuan, L. (2013). Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining*, 3:1277–1291.

[Superlistas 2023] Superlistas, A. (2023). Os 100 brasileiros mais seguidos do twitter. https://assuperlistas.com/2024/07/15/os-100-brasileiros-mais-seguidos-do-twitter/. Acessado em: 15 de agosto de 2024.

[Taboada 2016] Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2:325–347.

[Vergeer et al. 2013] Vergeer, M., Hermans, L., and Sams, S. (2013). Online social networks and micro-blogging in political campaigning: The exploration of a new campaign tool and a new campaign style. *Party Politics*, 19(3):477–501.

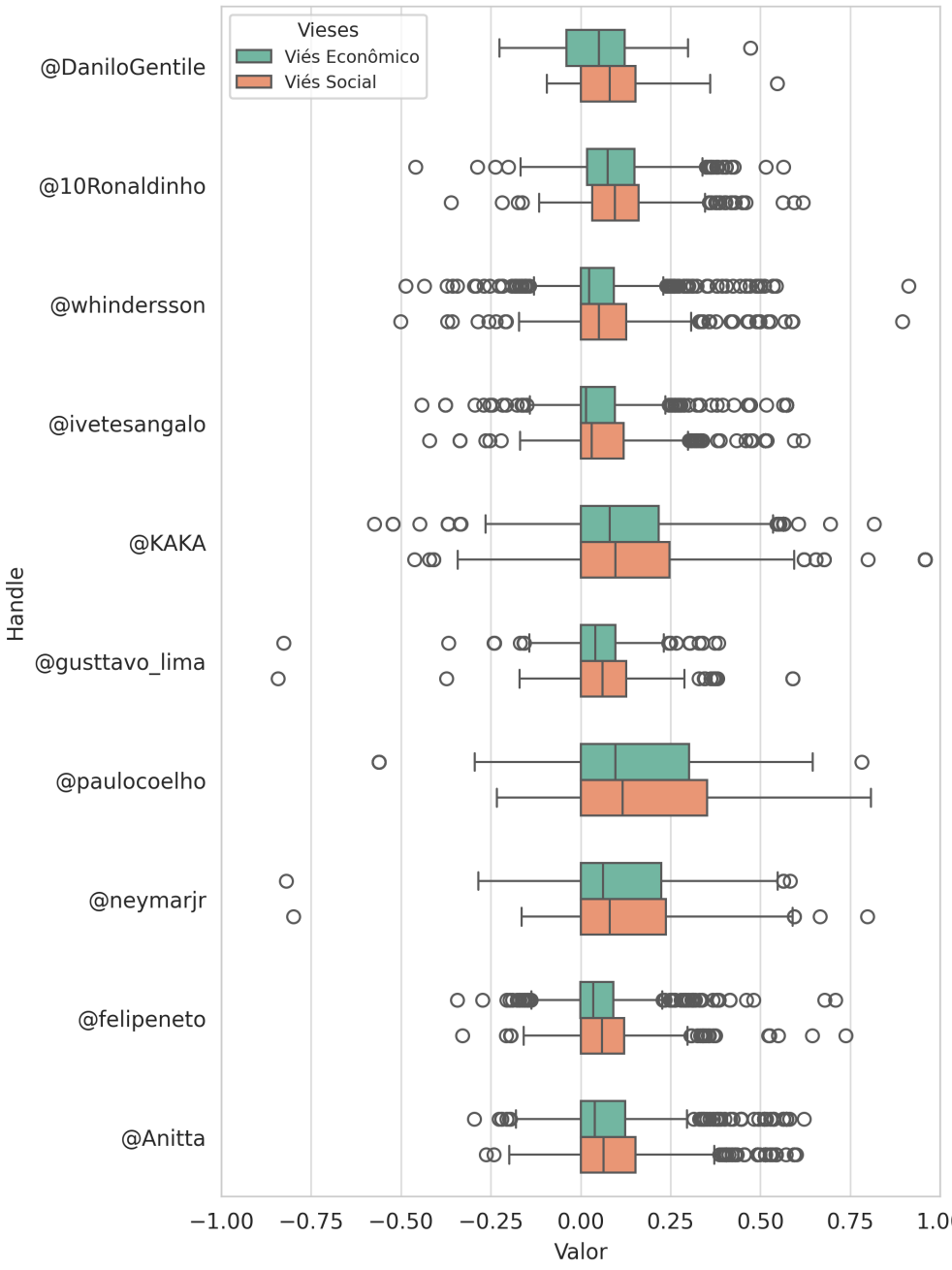

Figura 2. Vieses sociais e econômicos dos textos das postagens das 10 personalidades brasileiras mais seguidas no *X* obtidas pelos léxicos com unigramas.

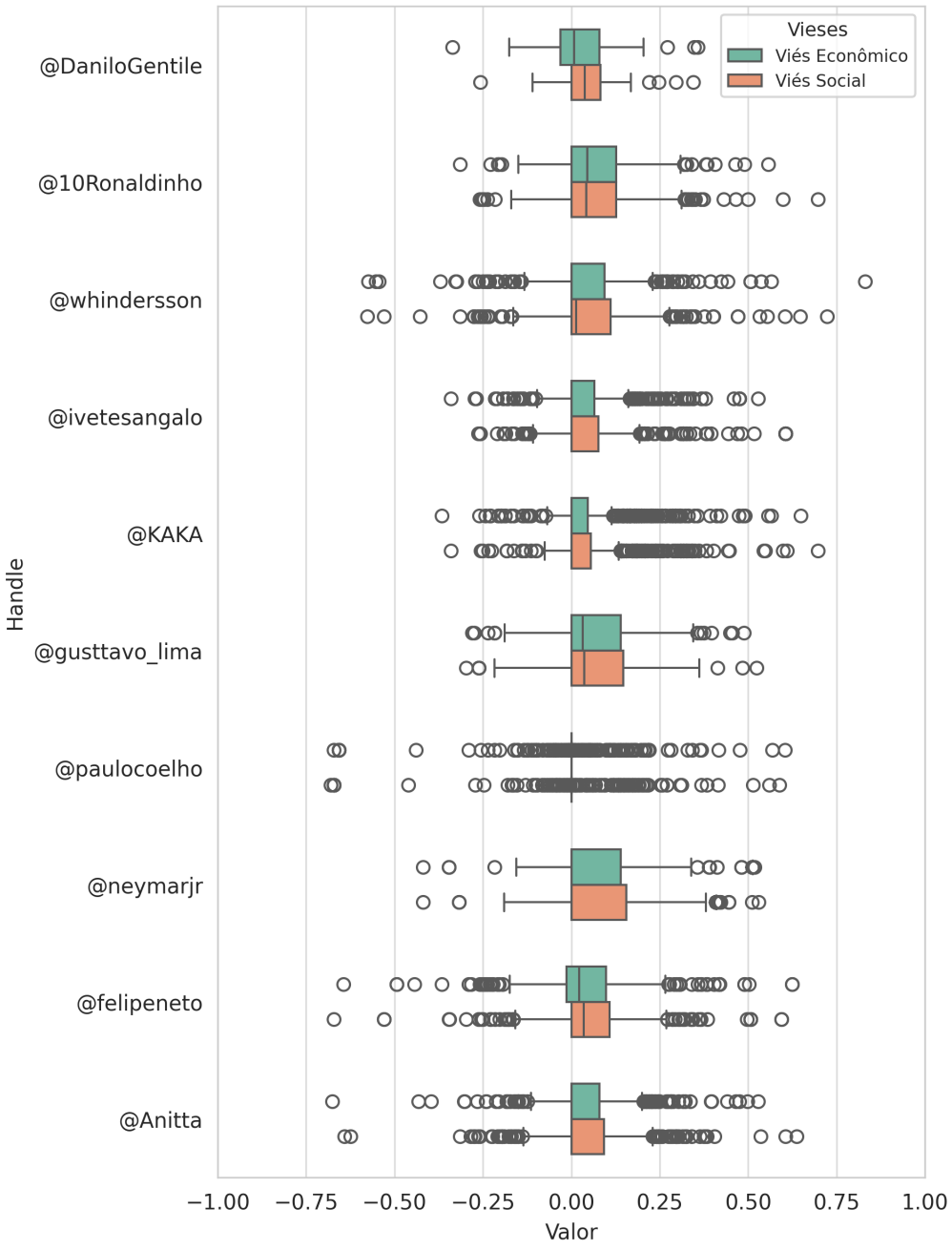

Figura 3. Vieses sociais e econômicos dos textos das postagens das 10 personalidades brasileiras mais seguidas no *X* obtidas pelos léxicos com bigramas.

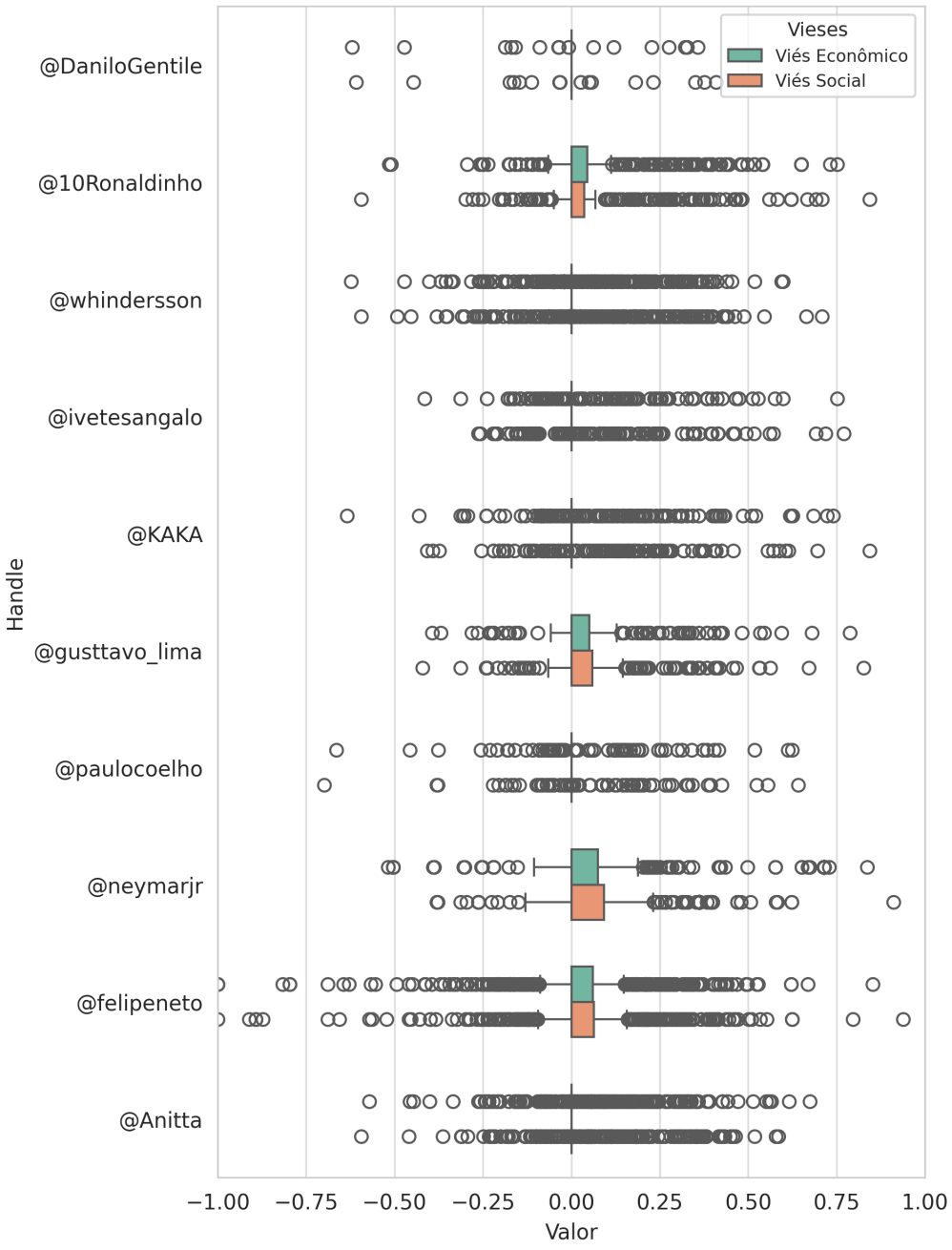

Figura 4. Vieses sociais e econômicos dos textos das postagens das 10 personalidades brasileiras mais seguidas no *X* obtidas pelos léxicos com trigramas.

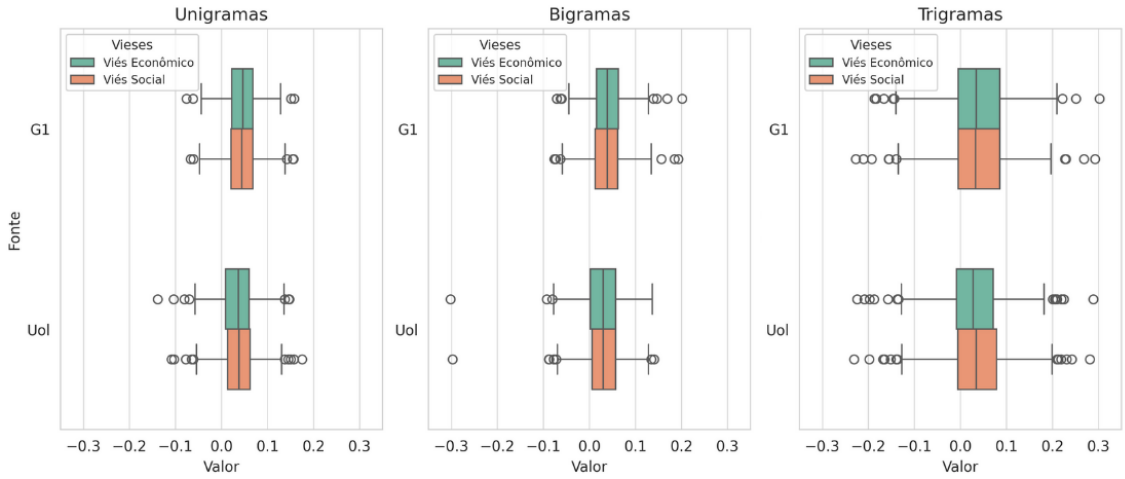

Figura 5. Vieses sociais e econômicos das notícias coletadas dos sites G1 e UOL calculados usando os léxicos de unigramas, bigramas e trigramas.