RedNIE Red Nacional de Investigadores en Economía

Hate in the Tropics: Political Leaders and the Social Acceptability of Online Hate Speech

Diego Marino-Fages (Durham University)

Agustina Martínez-Pozo (University of Leicester)

DOCUMENTO DE TRABAJO N° 391

Abril de 2026

Citar como:
Marino-Fages, Diego, Agustina Martínez-Pozo (2026). Hate in the Tropics: Political Leaders and the Social Acceptability of Online Hate Speech. Documento de trabajo RedNIE N°391.

Hate in the Tropics: Political Leaders and the Social Acceptability of Online Hate Speech*

Diego Marino-Fages† Agustina Martínez-Pozo‡

March 2026

Abstract

How does the advent of political information influence social norms? This paper examines the impact of Jair Bolsonaro's victory in the 2018 Brazilian presidential election on the prevalence of hate speech. We apply Natural Language Processing techniques to detect hate speech in over 37.6 million tweets, and leverage the electoral surprise of Bolsonaro's victory in a difference-in-differences design. Our findings reveal a substantial increase in online hate speech following the election, particularly in municipalities where Bolsonaro's vote share was lower—where his local and national support diverged most. The increase is primarily driven by the extensive margin of hate speech and is concentrated in homophobic and sexist content—areas in which Bolsonaro's rhetoric was highly controversial. Overall, these patterns suggest that the election outcome reshaped perceptions of the social acceptability of expressing hate.

Keywords: Hate speech; Social Media; Social Norms.
JEL Codes: D72, D83, J15, Z13.

*This paper has benefited from helpful feedback from Eren Arbatli, Adam Brzezinski, Antonio Cabrales, Annalí Casanueva Artís, Agustín Casas, Horacio Larreguy, Warn N. Lekfuangfu, Jaime Marques Pereira, Federico Masera, Juan S. Morales, Miquel Oliver-Vert, Henry Redondo, Margaret Samahita, Johannes Schneider, Carlo Schwarz, and Mateusz Stalinski. We thank the participants of the following events and institutions for their valuable feedback and suggestions: 2nd Workshop on Digital Economics (CCP and University of Cambridge), 6th Monash-Warwick-Zurich Text-As-Data Workshop, Public Governance Working Group (Paris Dauphine University), Durham University, University of Leicester, University of Nottingham, Understanding Offence: (De)limiting the Unsayable Conference (Durham University), 2024 NICEP Conference (University of Nottingham), Text-as-Data workshop (University of Liverpool), CESifo Venice Summer Institute: Economics of Social Media, Development Bank of Latin America and the Caribbean (CAF Buenos Aires), CESifo Conference: Economics of Digitization, European Commission Joint Research Centre (Seville, Spain), 4th Scotland and Northern England Conference in Applied Microeconomics, 2025 EAYE Annual Meeting, 1st Annual Interdisciplinary WZB Conference, and LACEA 2025 Annual Meeting (Recife). This research used the ALICE High Performance Computing facility at the University of Leicester. Ethics approval was obtained from Durham University (Review Reference: DUBS-2026-12416-12209). All errors are our own.

†Durham University, United Kingdom. Email for correspondence: diego.r.marino-fages@durham.ac.uk

‡University of Leicester, United Kingdom. Email for correspondence: a.martinez@leicester.ac.uk

1 Introduction

Social norms are unwritten rules and beliefs governing attitudes and behaviors considered acceptable (or not) in a particular social group or culture. They establish standards on different aspects of life—such as contractual relationships, reciprocity, and fairness—and provide order and predictability in society. Notwithstanding, social norms are not inherently good; examples of harmful social norms are revenge or genital mutilation. An increasingly salient social norm in contemporary societies concerns the acceptability of various forms of public discourse, particularly hate speech. This refers to offensive expressions directed at individuals or groups based on inherent characteristics such as ethnicity, gender, or religion. The prevalence of such discourse undermines social cohesion, fuels polarization, and imposes significant social and psychological costs on affected communities.[1]

While social norms are generally persistent over time (Fernandez, 2007; Giuliano, 2007; Alesina et al., 2013), recent research highlights how specific events can precipitate rapid shifts in their prevalence (Bursztyn et al., 2020a). These findings suggest that social norms, including the acceptability of hate speech, may be more responsive to contextual shifts than traditionally assumed.

In this paper, we study the impact of Jair Bolsonaro's 2018 presidential election on the prevalence of online hate speech in Brazil. Bolsonaro, sometimes called "the Trump of the Tropics," is widely recognized for his contentious viewpoints, encompassing homophobia, racism, and sexism.[2] We argue that Bolsonaro's victory triggered a quick update of the prevailing social norm governing which types of speech are socially acceptable.[3]

Identifying the *causal effect* of election results on hate speech is challenging. A change in hate speech around election time could reflect reverse causality, or the influence of omitted factors that affect both outcomes. We address these concerns by exploiting the unexpected nature of Bolsonaro's performance in the first-round of the elections. He received 46 percent of the vote, far above pre-election polls, which had predicted support

[1] See Madriaza et al. (2025) for a systematic review of the negative effects of hate speech.

[2] To illustrate this point, consider a sample of Bolsonaro's statements: "*I would be incapable of loving a homosexual son*," "*The scum of the earth is showing up in Brazil as if we did not have enough problems of our own to sort out*," and (speaking to a Brazil Congresswoman) "*I would not rape you because you do not deserve it*." Sources: CNBC web portal, `https://www.cnbc.com/2018/10/29/brazil-election-jair-bolsonaros-most-controversial-quotes.html`; Reuters, `https://www.reuters.com/article/us-brazil-politics-bolsonaro-factbox-idUSKCN1II2T3`; AP News, `https://apnews.com/article/1f9b79df9b1d4f14aeb1694f0dc13276`; USA Today, `https://eu.usatoday.com/story/news/world/2018/10/29/jair-bolsonaro-brazils-new-president-has-said-many-offensive-things/1804519002/`. Access date: June 2023.

[3] Figure A9 in the Online Appendix shows how Bolsonaro's Google searches, as well as Twitter mentions, surged around the time of the elections, facilitating the awareness of his earlier controversial statements.

of roughly 35 percent.[4] We therefore interpret the 2018 election as an *electoral surprise*: a common national shock that revealed substantial support for a far-right candidate. Crucially, the local informational content of this electoral surprise varied across municipalities, as the election results revealed substantial spatial variation in support for Bolsonaro. This heterogeneity allows us to examine whether the national electoral outcome had differential effects across local contexts.

To conduct the empirical analysis, we apply two Natural Language Processing (NLP) techniques to detect hate speech in tweets posted from July 2017 to December 2019: a fine-tuned BERT model and a dictionary-based approach using hate targets. We exploit spatial variation in Bolsonaro's support using a difference-in-differences design. Our main specification interacts Bolsonaro's first-round municipal vote share with a post-election indicator. This continuous treatment approach estimates how the post-election change in hate speech varies with local political composition. In a complementary specification, for ease of exposition, we compare municipalities above and below the median of Bolsonaro's vote share. Both specifications identify effects under the assumption that, absent the election, hate speech trends would have evolved in parallel across municipalities with different levels of Bolsonaro's support.

We document a persistent increase in online hate speech following the 2018 presidential election. Strikingly, although the prevalence of hate speech is higher in municipalities where Bolsonaro received greater electoral support, the post-election increase is concentrated in municipalities where Bolsonaro was relatively less popular—precisely where the divergence between local and national outcomes was greatest. This pattern is consistent with a break in the *spiral of silence*, whereby perceived minority status on value-laden issues leads to self-silencing driven by fears of social isolation (Noelle-Neumann, 1974). The election of Bolsonaro constituted a salient national electoral surprise that altered individuals' perceptions of socially acceptable rhetoric. In municipalities where Bolsonaro was less popular, this revelation may have reduced the perceived social cost of expressing hate speech, inducing individuals to voice views that had previously been suppressed. As a result, hate speech increased relatively more in locations where such expressions had been less common prior to the election.

This interpretation is further supported by heterogeneity across hate speech targets and the margins through which the increase in hate speech operates. When disaggregating hate expressions by targets—homophobia, sexism, racism, political hate, political targets, and

[4]Among major polling firms conducting surveys in September-October 2018, predicted first-round vote shares ranged from 32% to 41%, with a mean of 35.2%. Only one firm's final poll placed Bolsonaro above 40%. Source: Wikipedia, `https://en.wikipedia.org/wiki/Opinion_polling_for_the_2018_Brazilian_general_election`, access date: June 2023.

insults—we find a differential post-election impact on homophobia and sexism, consistent with Bolsonaro's controversial views. By contrast, we do not observe comparable patterns for general insults, political hate, or targets. These findings suggest that the results are not driven by growing polarization. Instead, they are driven by harmful speech that targets specific groups, such as the LGBTQ+ community and women, which highly depends on the social acceptability of such behavior.

Regarding the margins through which hate speech increased, we show that both the extensive and intensive margins play important but distinct roles. The post-election increase in the number of users who had not previously posted hateful content and began doing so (i.e., the extensive margin) is larger in municipalities where Bolsonaro was less popular. In contrast, conditional on entry, the post-election increase in the intensity of hateful expression (i.e., the intensive margin) is larger in municipalities with stronger electoral support for Bolsonaro. This pattern is consistent with new entrants converging to different levels of hate speech shaped by their local environments. Together, these findings show how national political shocks interact with local social environments to influence social norms and behavior.

We also document heterogeneity in the extensive margin based on user gender. The post-election increase in hate speech is largest among male users (10.1% rise in the share of male users posting hate speech), followed by female users (6.1%), with the smallest increase among pseudonymous or unidentifiable accounts (3.3%). This differential response supports a norm-based mechanism: male users, and to a lesser extent female users, exhibit the strongest reactions to the perceived reduction in social costs, whereas pseudonymous accounts show minimal change.

Our paper belongs to the literature studying how, for good or bad, social norms drive behavior (e.g., Elster, 2020; Nyborg et al., 2016; Bicchieri, 2016, 2005). Specifically, our paper adds to the literature studying how certain events can trigger rapid updates in social norms and how correcting misperceptions provokes behavioral changes in diverse spheres—for instance, Andre et al. (2024) for climate-friendly behavior, Bursztyn et al. (2020b) for men's support for women's labor, and Morales (2020) for citizens' willingness to express criticism of political leaders.[5]

Large information aggregators, such as election or referendum results, can also shock prevailing social norms. The closest papers to ours are Bursztyn et al. (2020a), which studies the effect of Trump's 2016 election on xenophobic expression in the U.S., and Albornoz et al. (2022), which examines how the Brexit referendum affected hate crime in the United Kingdom. Our paper makes several complementary contributions. First, we provide evi-

[5]See Bursztyn and Yang (2022) for a review of this growing literature.

dence from one of the world's largest developing countries, where Twitter penetration is high and social norms may play a stronger role due to the weak formal institutions (Fergusson Talero et al., 2024), extending Bursztyn et al. (2020a)'s geographic scope from a single U.S. metropolitan area to national-level analysis in a major developing economy. Second, we examine everyday online hate speech—a more common form of hateful expression than the hate crimes studied by Albornoz et al. (2022) and a behavior that may serve as a precursor to more extreme actions. Third, we study a setting in which hateful expression incurs immediate social costs through direct scrutiny from friends and acquaintances, in contrast to the delayed posting on an unfamiliar website in Bursztyn et al. (2020a) or prosecution-dependent costs in Albornoz et al. (2022). This allows us to study norm shifts where reputational consequences are immediate. Fourth, by tracking individual accounts over time, we examine the extensive and intensive margins, showing how they respond differently across local contexts and providing insight into the mechanisms underlying these responses and their policy implications.

Our paper complements recent work by Barros and Silva (2025) on the role of masculinity norms in shaping electoral support. While Barros and Silva (2025) show that men experiencing relative social and economic decline disproportionately supported candidates emphasizing traditional masculine norms prior to the 2018 election, we take these pre-election dynamics as given and focus instead on a distinct research question. Specifically, we show that Bolsonaro's unexpectedly broad electoral support reshaped beliefs about the social acceptability of expressing norm-sensitive views. That is, our contribution is to document how electoral success itself—rather than changes in underlying attitudes—triggered a post-election response in expressive behavior through shifts in perceived social norms.

Furthermore, this paper contributes to the growing literature on the influence of political leaders on social norms (e.g., Ajzenman et al., 2023; Bursztyn et al., 2020a; Farina and Pathania, 2020; Acemoglu and Jackson, 2015, 2017). Our findings suggest that the observed increase in hate speech was not primarily driven by Bolsonaro's campaign rhetoric or new controversial statements during the electoral period—most of his inflammatory remarks preceded his candidacy.[6] Rather, the evidence points to the election outcome itself as a powerful signal: the act of electing a leader with a well-known history of divisive discourse appears to have reshaped individuals' perceptions of what is socially acceptable. This suggests that leaders may not need to actively promote certain behaviors; the mere public validation of their views through electoral success can suffice to shift social norms.

[6]Moreover, his hate speech decreased during the campaign and after his election (see Figure A10 in the Online Appendix).

Our paper adds to the growing literature on social media platforms and their interplay with social norms and behavior (Aridor et al., 2024; Zhuravskaya et al., 2020). A growing body of work links social media to expressions of hate against minority groups. For instance, Trump's tweets about Muslims are related to increases in xenophobic tweets and hate crimes (Müller and Schwarz, 2023), and Trump's "Chinese Virus" tweets contributed to anti-Asian incidents (Cao et al., 2023); social media amplified ethnic hate crimes in Russian cities with stronger pre-existing anti-immigrant sentiments (Bursztyn et al., 2019); social media facilitated the propagation of anti-refugee incidents in Germany (Müller and Schwarz, 2021), and both media and social media contributed to the increase in hate crimes after Brexit (Carr et al., 2020). While this literature focuses primarily on xenophobia and ethnic hate crimes, we examine a broader set of hate expressions across multiple targets. Related work by Beknazar-Yuzbashev et al. (2022) shows that toxicity increases content consumption and is contagious on social media platforms.[7] Taken together, this literature highlights the role of social media in shaping both online and offline expressions of hate.[8] Our setting is particularly relevant because, unlike many countries, Brazil lacked comprehensive content moderation legislation during the period under study.

This paper also contributes to the emerging literature on LGBTQ+ Economics (Badgett et al., 2021). This literature shows that hostile norms can translate into direct psychological harm (Meyer, 2003, 1995), labor-market discrimination (Neumark, 2018), and work-related violence (Tampellini, 2024). By documenting a large, norm-driven shift in anti-LGBT+ expressions, our findings highlight how political events can rapidly alter the lived experiences and economic prospects of sexual minorities. This is especially important in the context of Brazil, a country with one of the highest numbers of homicides against LGBTQ+ individuals (Mendes and Silva, 2020), and where sexual minorities experience violence at higher rates than their heterosexual counterparts (Tampellini, 2024).

2 Data

In this paper, we study how Jair Bolsonaro's 2018 presidential election affected the prevalence of online hate speech in Brazil. Brazil is divided into twenty-six states and one federal district. Each subnational entity is further divided into municipalities; Brazil currently has 5,570 municipalities. Our primary data source is Twitter (now X), which we

[7] In addition, Beknazar-Yuzbashev et al. (2024) propose a theoretical argument under which circumstances social media platforms may find it profitable to display harmful content.

[8] In addition to this literature, other research has linked the internet and various forms of traditional media to violence (Dahl and DellaVigna, 2009; Card and Dahl, 2011; Bhuller et al., 2013; Yanagizawa-Drott, 2014; DellaVigna et al., 2014; Ivandic et al., 2024).

use to measure hate speech at both the municipality and individual levels over the period surrounding the presidential election.[9]

With one of the largest Twitter user bases in the world, Brazil provides a particularly well-suited context for studying online speech.[10] Most of the Brazilians who were online in 2022 used social media for news consumption (64%) and political discussion (78%).[11] Additionally, a survey conducted by the Brazilian Senate news agency found that 45% of respondents reported that information on social media influenced their vote in the 2018 elections.[12] Furthermore, unlike many other countries, Brazil lacked a comprehensive legislative framework governing content moderation on social media platforms during our study period.[13] This institutional context reassures us that observed changes in online hate speech are unlikely to be driven by new statutory moderation regimes.

We combine Twitter data with administrative data. First, we use municipality-level results from the 2018 presidential election published by Brazil's Superior Electoral Court (in Portuguese, Tribunal Superior Eleitoral—TSE), the highest authority in the country's electoral justice system. Specifically, we use Bolsonaro's vote share in the first round of the election at the municipality level. We exploit cross-municipality variation in Bolsonaro's electoral support during the 2018 Brazilian presidential election, treating it as a proxy for local exposure to—and endorsement of—his rhetoric and policy positions. Second, we rely on geospatial data from the Brazilian Institute of Geography and Statistics (in Portuguese, Instituto Brasileiro de Geografia e Estatística—IBGE) to geolocate both tweets and election outcomes. IBGE supplies standardized geospatial data at the national, state, and municipality levels.

[9]Figure A8 in the Online Appendix shows Twitter's penetration, measured by tweets and users, across Brazilian municipalities.

[10]In January 2022, Brazil ranked fourth worldwide in terms of the number of Twitter users, with an estimated 19 million active accounts (after the United States, Japan, and India). Source: Statista web portal, https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/, access date: June 2023.

[11]Sources: Digital News Report, 2022, Reuters Institute & University of Oxford, https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/brazil; Statista web portal, https://www.statista.com/statistics/1326518/brazil-social-media-users-political-discussion/; access date: June 2023.

[12]Source: Agência Senado, https://www12.senado.leg.br/noticias/materias/2019/12/12/redes-sociais-influenciam-voto-de-45-da-populacao-indica-pesquisa-do-datasenado.

[13]The *Marco Civil da Internet* (2014) is the legislation in place, which establishes principles, rights, and obligations for internet users and service providers (but not specifically for social media platforms). Over the last few years, there have been three attempts to pass a new bill on content moderation: Bill 2630/2020, Bill 592/2023, and Bill 4717/2025. However, these bills are still under deliberation. Sources: A Review of Content Moderation Policies in Latin America, https://www.techpolicy.press/a-review-of-content-moderation-policies-in-latin-america/; Digital Policy Alert, https://digitalpolicyalert.org/.

2.1 Twitter data and Text Analysis

In the empirical analysis, our main variable of interest is the proportion of tweets classified as hate speech per municipality (or individual) and date. One advantage of using online hate speech as an outcome variable is that it is directly observable and quantifiable, and not subject to underreporting, unlike hate crimes. Moreover, expressions of hate on social media are instantaneous and carry immediate social consequences. In contrast, hate crimes involve delayed costs that depend on reporting, prosecution, and conviction, thereby obscuring empirical analysis.

We rely on Natural Language Processing techniques to detect hate speech.[14] Precisely, we fine-tune a pre-trained *Bidirectional Encoder Representations from Transformers* (BERT) model (Devlin et al., 2019) to generate a binary classification of hate speech, and build a multi-level classification of hate using a dictionary-based method.[15] By utilizing these two NLP techniques, we produce two independent measures of predicted hate speech, which we use to check the robustness of the main results of this paper. Moreover, the dictionary classification enables us to analyze the differential effects of Bolsonaro's popularity across types of hate speech, providing evidence for the mechanisms underlying the main results. We categorize hate speech into six distinct types: *political hate* and *political target* (categories related to the political process), *homophobia, racism,* and *sexism* (categories associated with social norms), and *insults* (a residual category without a specific target). The next paragraphs describe how we collected and processed Twitter data to construct the corresponding variables.

Data collection. We use the Twitter Application Programming Interface v2 (Twitter API v2) to collect our data. Specifically, we rely on the *v2 full-archive search endpoint*, which gives access to the entire history of publicly available (and yet undeleted) tweets. We retrieve all tweets (excluding retweets) that satisfy the three conditions specified in the Twitter query. First, tweets must be written in Portuguese. Second, tweets must provide geolocation information and be located in Brazil. Lastly, tweets must fall between July 2017 and December 2019 (inclusive). As the daily volume of data retrieved by this query is approximately 300.000 tweets, we further restrict the Twitter query to tweets posted on any Monday within the mentioned period. This query imposes two main assumptions on our sample of tweets. We assume that the tweets posted on any Monday and the geolocated tweets constitute representative samples of the universe of tweets. Online Appendix A.1 provides supportive evidence for these assumptions and complementary information to

[14] See Ayo et al. (2020) for a survey of machine-learning approaches to hate speech detection.

[15] For reviews on text analysis for economists, see Gentzkow et al. (2019) and Ash and Hansen (2023).

this section.

Data processing. We pre-process the text of each tweet to construct the input for the hate speech detection task. Specifically, we anonymize user mentions and URL links while preserving hashtags in their original Twitter format, as they often carry relevant information. We exclude tweets that contain only links and/or user mentions, as they do not convey substantive textual content. In addition, we drop tweets posted by accounts created after 2017 to reduce the influence of accounts potentially created for the electoral campaign, including political bots or other coordinated activity. Prior to the classification with our BERT-based model, we apply additional text pre-processing steps, which are described in detail in the Online Appendix A.2.

BERT model fine-tuning. We fine-tune *BERTimbau*, a BERT model for Brazilian Portuguese developed by Souza et al. (2020), using a labeled dataset of Portuguese-language tweets compiled by Fortuna et al. (2019). The training data consist of 5,668 tweets in Portuguese collected via the Twitter API between January and March 2017 and classified as having or not having hate speech (binary classification).

We split the dataset into training (80%), validation (10%), and test (10%) sets. As Fortuna et al. (2019)'s dataset exhibits class imbalance—with hate speech representing the minority class—we apply random oversampling to the training set to equalize class frequencies (Mohammed et al., 2020).[16] This procedure improves classification performance without altering the validation or test sets. The resulting model achieves state-of-the-art performance for hate speech detection tasks, with traditional evaluation metrics (accuracy, precision, recall, and F1-score) in the 77–79% range on both the validation and test sets.

Dictionary-based classification. In addition to the BERT approach, we implement a dictionary-based method to identify hate speech in tweets, with the specific goal of classifying hate expressions by their intended targets. We define six mutually exclusive categories: *political hate*, *political target*, *homophobia*, *racism*, *sexism*, and *insults*. A tweet is classified as hate speech if it contains at least one term associated with any of these categories.

The first two categories—political hate and political target—are less commonly used in the literature but are central to our setting, as they allow us to assess whether changes in hate speech are primarily driven by political polarization or by broader social dynamics. In contrast, homophobia, racism, and sexism capture forms of hate that are closely tied to

[16]Random oversampling augments the training data by randomly replicating observations from the minority class.

social norms and target specific demographic groups (following Bolsonaro's speech). We treat insults as a residual category, encompassing offensive expressions without a clearly identifiable political or social target.

We construct the dictionary by combining high-frequency hate-related terms identified in prior studies. Specifically, we draw on the hate speech classification developed by Fortuna et al. (2019), the multi-level toxicity taxonomy proposed by Leite et al. (2020), and the homophobia-related lexicon introduced by Pereira (2018). To further enrich coverage, we incorporate terms from the Multilingual Offensive Lexicon (MOL) compiled by Vargas et al. (2025). Online Appendix A.2 provides additional details on the dictionary construction and reports the full list of terms used for each hate category.

Dataset construction. After fine-tuning the BERT model and constructing the dictionary, we apply both methods to classify hate speech in the full corpus of tweets. The BERT model yields a binary indicator, *hate speech*, equal to one if a tweet is classified as containing hate content. The dictionary-based approach produces six binary indicators—*political hate*, *political target*, *homophobia*, *sexism*, *racism*, and *insults*—each capturing the presence of hate speech directed at a specific target. Using tweet-level geolocation information, we assign each tweet to a Brazilian municipality based on its latitude and longitude, relying on IBGE's municipal shapefiles. We then aggregate the classified tweets at the municipality (or individual) level and over time to compute the share of tweets containing each type of hate speech. These shares constitute the main outcome variables in our empirical analysis.

2.2 Descriptive statistics

This paper builds on two key empirical observations. The first one concerns the prevalence of online hate speech, which has evolved markedly over time. Figure 1 plots the share of Brazilian tweets classified as hate speech during the sample period.[17] The solid line reports the daily proportion of tweets containing hate speech, while the dotted line shows a smoothed series obtained using a three-week moving average. The shaded areas indicate key political milestones: the period between the two electoral rounds and the beginning of Bolsonaro's term in office.[18]

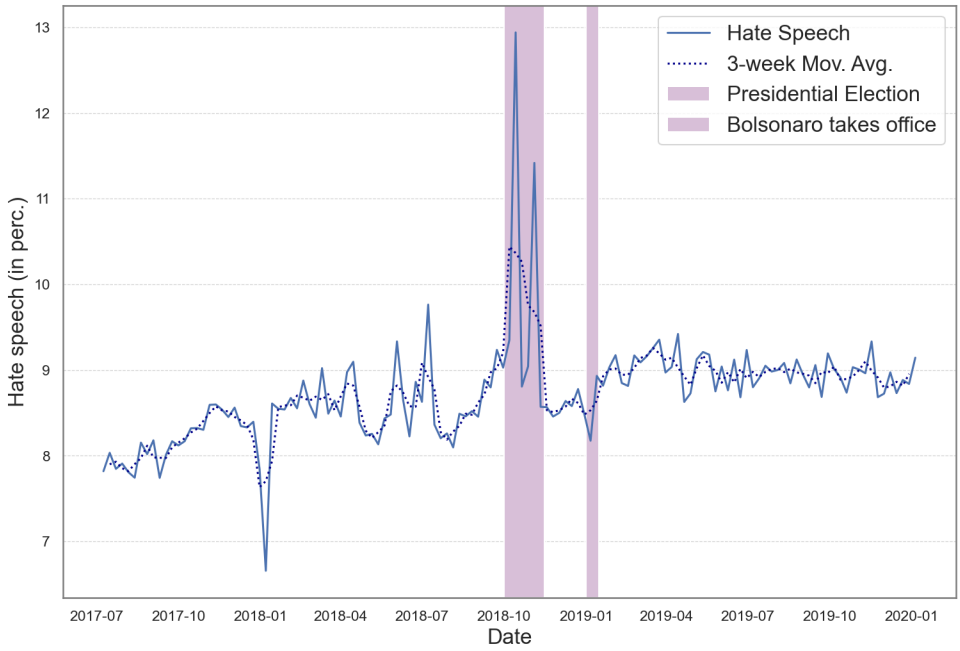
As shown in Figure 1, hate speech exhibits pronounced spikes during the electoral period. The largest peaks occur immediately after the first and second rounds of the presiden-

[17]Figure A11 presents an analogous figure using the dictionary-based classification.

[18]The first and second rounds of the presidential election were held on October 7 and October 28. Bolsonaro took office as Brazil's 38th president on January 1, 2019.

tial election—the closest dates in our sample following each vote.[19] Periods of relatively low hate speech coincide with the 2018 New Year break. Notably, a comparable decline is absent around the 2019 New Year, which coincided with Bolsonaro's inauguration.

Figure 1: Evolution of hate speech in Brazilian tweets, 2017-2019.

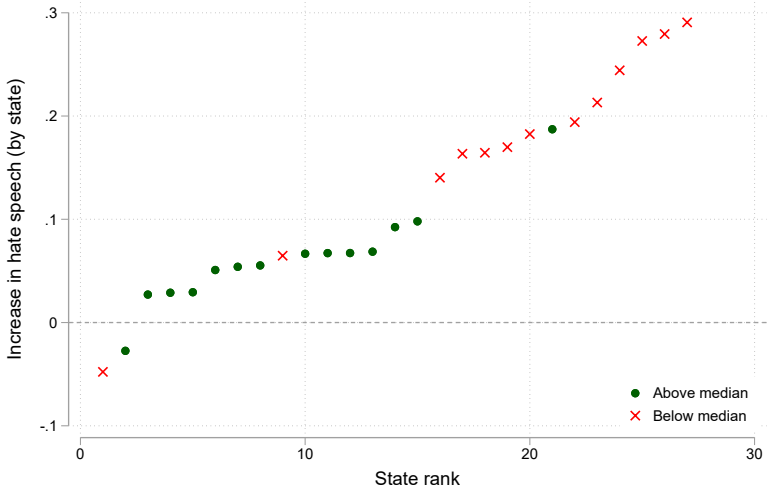


Notes: Hate speech denotes, for each date, the percentage of tweets classified as hate speech by the BERT model. The solid line shows the raw daily series, while the dotted line reports the series smoothed using a three-week moving average.

Beyond these short-run fluctuations, Figure 1 reveals a clear stabilization in hate speech at a higher rate in the post-election period. The average share of tweets classified as hate speech rises from 8% between July 2017 and July 2018 to 9% between January and December 2019—an increase equivalent to roughly 3,000 additional hate-containing tweets on an average day. Figure 2 shows that the increase in hate speech was generalized, with only two out of 27 states not experiencing an increase (i.e., Mato Grosso do Sul and Paraiba). Interestingly, the figure shows that the increase was systematically more pronounced in states where Bolsonaro received lower levels of support.

[19]Two additional, albeit smaller, peaks appear in June and July 2018, coinciding with matches played by Brazil's national football team during the 2018 World Cup. Online Appendix Figure A2 shows that these dates also correspond to sharp increases in overall Twitter activity, with daily tweet volumes approximately 50% higher than average.

Figure 2: Evolution of hate speech in Brazilian tweets, 2017-2019. States by the 2018 election result.



Notes: Increase in hate speech denotes, for each state, the percentage change in hate speech between the pre-period (July 2017 to July 2018) and the post-period (November 2018 to December 2019). Tweets classified as hate speech by the BERT model. States are ranked by the magnitude of this change along the x-axis, and split according to the in-sample median of Bolsonaro's vote share in the first round of the election.

Second, Bolsonaro's vote share—which proxies support for his ideas and rhetoric—varied substantially across Brazilian municipalities. Panel (a) in Figure 3 illustrates the pronounced spatial variation in Bolsonaro's popularity across states and municipalities.[20] In the first round of the 2018 presidential election, his municipal vote share ranged from as low as 3% to as high as 79%. Panel (b) complements this evidence by showing the post-election increase in hate speech (relative to the pre-election period) for the available municipalities, revealing substantial cross-sectional variation that we exploit in our empirical analysis.

3 Empirical strategy

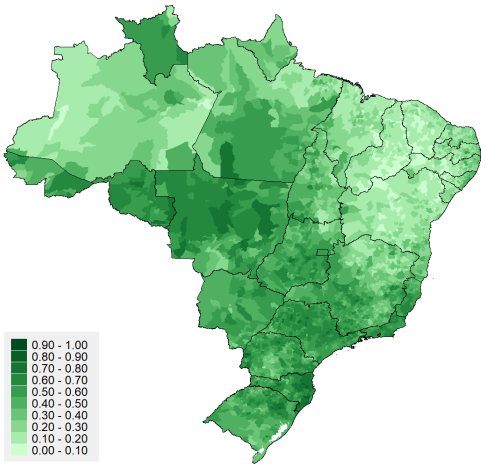
Our objective is to estimate the causal effect of Bolsonaro's election on online hate speech. While Figure 1 documents a marked increase in hate speech at the national level following the 2018 election, aggregate time-series evidence alone cannot establish causality. In

[20]The second round exhibits a similar spatial pattern, see Online Appendix Figure A7. In addition, Online Appendix Figure A6 documents the bimodal distribution of Bolsonaro's municipal vote shares.

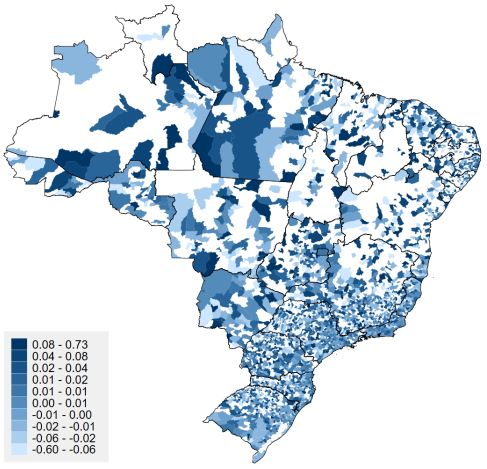
particular, rising hate speech may have contributed to the electoral outcome itself, or both phenomena may reflect broader, contemporaneous social and political trends.

Figure 3: Spatial variation in Bolsonaro's electoral support and post-election increase in hate speech.

(a) Bolsonaro's vote share



(b) Post-election increase in hate speech.



Notes: Panel (a) shows Bolsonaro's vote share (ranging from 0 to 1) at the municipality level in the first round of the presidential election. Darker shades indicate higher vote shares. Panel (b) shows the increase in hate speech, computed as the difference between average hate speech in the pre- and post-election periods. Darker shades indicate a higher increase in hate speech.

Because the presidential election is a nationwide event, it leaves us with no clear control group where Bolsonaro is not elected president. We therefore exploit the substantial heterogeneity in Bolsonaro's electoral support across municipalities.[21] This captures the local magnitude of the national electoral surprise, as well as his popularity and the social norms it may represent. We can then exploit the differential exposure to Bolsonaro's popularity, as proxied by the *unanticipated* election results, to study whether hate speech increased relatively more in municipalities where Bolsonaro was more or less popular as a political leader. We conduct this difference-in-differences analysis by regressing,

$$Hate_{mt} = \beta * Post_t * VoteShare_m + \alpha_t + \pi_m + \epsilon_{mt}, \quad (1)$$

where $Hate_{mt}$ is the share of tweets that contain hate speech in municipality m and month

[21] Specifically, the vote share in municipalities ranges from 3% to 79% in the first round of elections (see Figure 3).

t. $Post_t$ is a dummy variable that takes the value of one after the elections, excluding the election rally (specifically, $Post_t = 0$ from July 2017 to July 2018 and $Post_t = 1$ from November 2018 to December 2019). $VoteShare_m$ is our continuous treatment variable, defined as the share of votes obtained by Bolsonaro in the first round of the election in municipality m. α_t and π_m are month and municipality fixed effects, and ϵ_{mt} is a municipality-month specific error term.

Because our rich dataset enables us to follow Twitter accounts over time, we can further analyze hate speech at the individual level and explore the intensive and extensive margins of hate speech. To study the extensive margin, we replace the outcome variable in Equation (1) with the share of users posting hate speech in municipality m and month t.

To explore the intensive margin of hate speech, we use the share of tweets containing hate speech posted by a specific user as the outcome variable. Formally, we regress,

$$Hate_{imt} = \tilde{\beta} * Post_t * VoteShare_{im} + \tilde{\alpha_t} + \gamma_i + \epsilon_{imt}, \tag{2}$$

where $Hate_{imt}$ is the share of tweets that contain hate speech for the account i in municipality m at month t. $Post_t$ is a dummy variable that takes the value one after the elections, and $VoteShare_{im}$ is the share of votes obtained by Bolsonaro in municipality m (where individual i is located). $\tilde{\alpha_t}$ and γ_i are time and individual fixed effects, and ϵ_{imt} is an individual-municipality-month specific error term.

Our coefficients of interest are $(\beta, \tilde{\beta})$, which capture the *average causal response* (ACR) on the treated to an incremental change in the *treatment dose*, where the dose is the share of votes obtained by Bolsonaro in that municipality. The main identification assumption, in this case, is the strong parallel trends assumption. It requires that, for all doses, the average change in hate speech over time across all municipalities that received a given dose is the same as the average change in hate speech that would have occurred over time for all municipalities that experienced a different dose (Callaway et al., 2024).[22]

To simplify the exposition of the results and check the robustness of our results to a change in the definition of our treatment variable, we also report the regressions using a binary treatment variable. To do so, we assign municipalities to treatment and control groups by considering the in-sample median (at the municipality level) of Bolsonaro's vote share during the first round of the elections. The vote share for Bolsonaro at the median municipality in our sample was 50.7%.[23] Formally,

[22]Formally, let d be the dose and Y_t be the potential outcome in time t. Then, the strong parallel trends assumption implies that for all d in D: $E[Y_t(d) - Y_{t-1}(0)] = E[Y_t(d) - Y_{t-1}(0)|D = d]$.

[23]The results are robust to considering other statistics to assign municipalities into treatment and control groups, for instance, using Bolsonaro's vote share at the national level (46%) or 50%. See Table A4.

$$Hate_{mt} = \delta_1 * Post_t * BelowMedian_m + \theta_t + \pi_m + \epsilon_{mt} \quad (3)$$

and,

$$Hate_{imt} = \tilde{\delta_1} * Post_t * BelowMedian_{im} + \tilde{\theta_t} + \gamma_i + \epsilon_{imt}, \quad (4)$$

where $Hate_{mt}$ ($Hate_{imt}$) is the share of tweets that contain hate speech in municipality m and month t (for user i in month t), $Post_t$ is a dummy variable that takes the value one after the elections, $BelowMedian_m$ ($BelowMedian_{im}$) is a dummy variable that takes the value one for the municipalities where Bolsonaro's vote share was lower than the median municipality, θ_t ($\tilde{\theta_t}$) and π_m are month and municipality fixed effects, and ϵ_{mt} (ϵ_{imt}) is a (individual) municipality-month specific error term. In this case, the coefficients of interest are $(\delta_1, \tilde{\delta_1})$,[24] and the identifying assumption is the traditional parallel trends assumption. That is, in the absence of a differential in Bolsonaro's popularity at the municipality level, the difference in hate speech between municipalities where Bolsonaro received lower or higher support during the elections is constant over time.

4 Results

In this section, we present the main results of the paper. First, we document that hate speech increased after the 2018 presidential election, especially in the municipalities where Bolsonaro was less popular. Second, we show that the extensive margin played a central role in this overall increase in hate speech. Third, we show that the effect is driven by hateful content toward groups to whom Bolsonaro's rhetoric was openly hostile. Finally, we present the results at the individual level, which confirm the previous qualitative results and explore the intensive margin of hate speech and heterogeneity by gender.

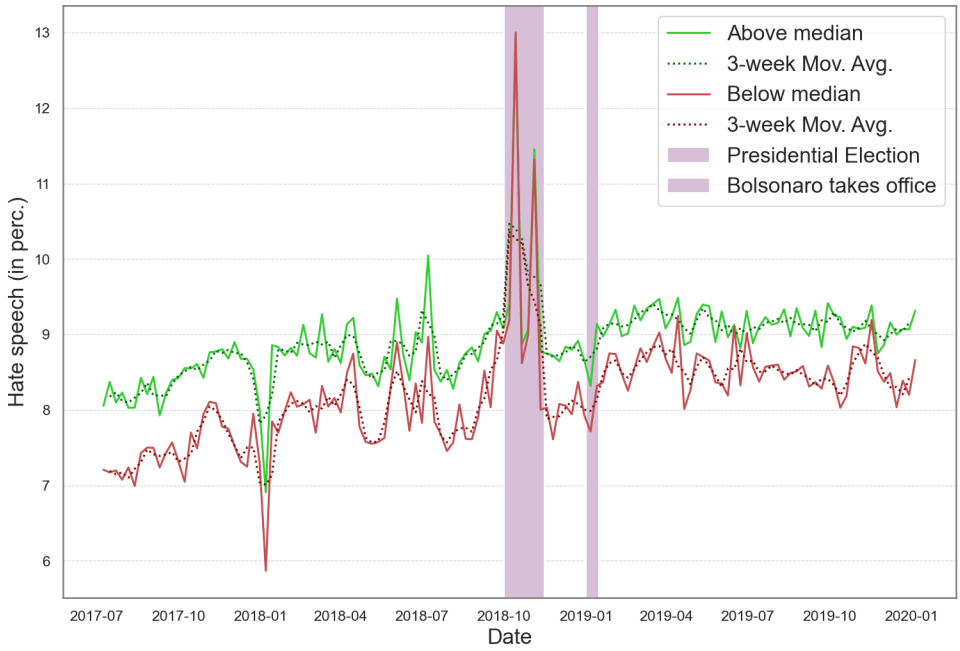
4.1 Municipality level

Before turning to the regression results, we describe the evolution of hate speech in municipalities assigned to the treatment and control groups as defined in equations (3) and (4). Figure 4 mirrors Figure 1, but disaggregates the time series of hate speech, measured using our BERT-based classifier, by treatment status. The figure provides a visual comparison

[24]Notice that, by construction, $(\beta, \tilde{\beta})$ in equations (1) and (2) and $(\delta_1, \tilde{\delta_1})$ in equations (3) and (4) have opposite signs: while the latter capture the effect of $BelowMedian_m = 1$, which depend negatively on Bolsonaro's vote share, the former are proportional to it.

of pre- and post-election trends in hate speech across the two groups.[25] The green lines show the daily proportion and the corresponding three-week moving average of Brazilian tweets classified as hate speech by the BERT model for municipalities in which Bolsonaro received at least 50.7 percent of the vote in the first round of the 2018 presidential election (i.e., $BelowMedian_m = 0$). The red lines display the analogous series for municipalities in which Bolsonaro's vote share was at most 50.7 percent (i.e., $BelowMedian_m = 1$). Shaded areas indicate the periods corresponding to the presidential election and Bolsonaro's inauguration.

Figure 4: Evolution of hate speech in Brazilian tweets, 2017-2019. Municipalities by the 2018 election result.



Notes: The figure plots the percentage of tweets classified as hate speech by the BERT model, separately for municipalities below and above the median of Bolsonaro's vote share in the first round election. The solid lines show the raw daily series, while the dotted lines report the series smoothed using a three-week moving average.

Importantly for our identification strategy, the gap between hate speech pre-trends for treatment and control groups is constant over time, i.e., pre-trends are parallel. Furthermore, the prevalence of hate speech in municipalities with higher or lower support for

[25] Online Appendix A.3 presents the analogous figure using hate speech classified by the dictionary-based method.

Bolsonaro appears to respond similarly to shocks—for example, both decreased around the 2018 New Year's Eve and increased during the 2018 World Cup (in July)—supporting the validity of using the below-median municipalities as the control group. After the elections and the taking up of office by Bolsonaro, the previously constant gap was reduced significantly, with the municipalities where Bolsonaro was less popular increasing the most.

We now turn to the regression results. Table 1 refers to the main question of this paper, *how the 2018 presidential election of Bolsonaro affected online hate speech.* Columns (1) and (2) report estimates based on hate speech identified by the BERT classifier, while columns (3) and (4) use the dictionary-based measure. Columns (1) and (3) present estimates from a difference-in-differences specification with a continuous treatment, as in equation (1). Columns (2) and (4) report standard difference-in-differences estimates, corresponding to equation (3). In all specifications, we include any municipality for which we observe (i) at least 10 tweets daily and (ii) at least 10 times during 2017-2019.[26] Likewise, we exclude the period from August to October 2018 to prevent contamination from hate speech directly associated with the election and the electoral rally.[27]

Columns (1) and (3) present our preferred specifications. These estimates indicate that the post-election increase in hate speech is decreasing in Bolsonaro's local vote share. Under the parallel trends assumption, the coefficients from this difference-in-differences model with continuous treatment can be interpreted as weighted averages of the average causal response $ACR(d)$ across treatment doses. On average across doses, a one–standard–deviation increase in $VoteShare_m$ is associated with a 0.033–standard–deviation smaller increase in hate speech using the BERT classifier and a 0.018-standard-deviation smaller increase using the dictionary-based measure. Columns (2) and (4) report the corresponding binary difference-in-differences estimates, which compare municipalities with above- and below-median Bolsonaro vote shares. Consistent with the continuous treatment results, these specifications show that municipalities below the median experienced significantly larger post-election increases in hate speech—by 0.039 and 0.014 standard deviations—relative to municipalities above the median.[28] Permutation-based p-values (in square brackets), which do not rely on large-sample asymptotic approximations, corroborate these findings. These tests also show no evidence of systematic effects in pre-election periods, supporting the credibility of the parallel trends assumption.

[26]Depending on the specification, this leaves us with approximately 2000-2500 municipalities, which we observe on average for approximately 100 Mondays.

[27]In all specifications, $Post_t$ is an indicator equal to zero from July 2017 through July 2018 and equal to one from November 2018 through December 2019. Table A5 in the Online Appendix shows that our results are robust to alternative definitions of $Post_t$.

[28]Table A4 in the Online Appendix replaces the median cutoff with a 50 percent threshold and the national average (46 percent); the results are qualitatively similar.

Table 1: MUNICIPALITY LEVEL REGRESSIONS.

	Hate Speech_BERT	Hate Speech_BERT	Hate Speech_DICT	Hate Speech_DICT
Post-election × Vote share	-0.033***		-0.018*	
	(0.011)		(0.010)	
Post-election × Below median popularity		0.039**		0.014
		(0.017)		(0.015)
Observations	105,717	105,717	137,271	137,271
R-squared	0.078	0.078	0.098	0.098
Permutation test (p-value)	[0.042]	[0.042]	[0.167]	[0.042]
Municipality FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Municipalities	1,932	1,932	2,490	2,490

Notes: The table reports coefficients of regressing the share of hate speech in the municipality (standardized) on the vote share for Bolsonaro (columns 1 and 3) and the municipalities where Bolsonaro's popularity was below median (columns 2 and 4). Post-election is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. Standard errors clustered at the municipality level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For the permutation test (p-values in squared brackets), we run all possible placebos (23) using either the months pre-election or post-election and assigning random threshold dates.

To illustrate the magnitude of this effect, consider Rio de Janeiro, where Bolsonaro received 58.3 percent of the vote and where approximately 32,000 tweets are posted on a typical weekday, with 9 percent containing hate speech. Our estimates imply that an otherwise similar municipality with a Bolsonaro's vote share 10 percentage points lower (approximately the in-sample mean municipality) would generate 40–50 additional hateful tweets per day. While modest at the individual level, these differences accumulate over time ($\approx$ 1,500 per month or $\approx$ 18,250 per year) and, as we show below, they reflect a broad expansion in participation: more individuals engaging in hate speech, rather than increased activity by a fixed group of users, thereby amplifying the social reach and persistence of hateful content.

The results in columns (1)–(4) provide evidence consistent with a break in the spiral of silence. The divergence between municipal and national election outcomes captures the magnitude of the local electoral surprise. In municipalities with below-median support for Bolsonaro, residents were more likely to underestimate the national prevalence of Bolsonaro supporters and the social norms he represents. The election outcome revealed that such views were more widely held than previously believed, lowering the perceived social cost of expressing them. This mechanism predicts a larger post-election increase in hate speech precisely where Bolsonaro was less popular locally.

We conduct two complementary tests that follow directly from this interpretation. First, if the election lowered the perceived social cost of expressing hate, the post-election increase should operate primarily along the extensive margin—that is, through a larger number of individuals engaging in hate speech, rather than through a higher intensity of posting among pre-existing users. This is a necessary implication of a change in perceived

social acceptability. Hence, instead of using the share of tweets containing hate speech in a given municipality, we use the share of users posting hate speech in that municipality.

Table 2 reports difference-in-differences estimates using the extensive margin of hate speech, measured as the share of users who post at least one hate-containing tweet in a given municipality.[29] Consistent with a break in the spiral of silence mechanism, the results show a strong presence of the extensive-margin channel.

Table 2: MUNICIPALITY LEVEL REGRESSIONS - EXTENSIVE MARGIN.

	Hate Speech_BERT	Hate Speech_BERT	Hate Speech_DICT	Hate Speech_DICT
Post-election × Vote share	-0.016***		-0.012***	
	(0.003)		(0.003)	
Post-election × Below median popularity		0.051***		0.037***
		(0.012)		(0.012)
Observations	47,783	47,783	47,783	47,783
R-squared	0.429	0.429	0.449	0.449
Permutation test (p-value)	[0.042]	[0.042]	[0.083]	[0.083]
Municipality FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Municipalities	2,371	2,371	2,371	2,371

Notes: The table reports coefficients of regressing the share of users posting hate speech on the vote share for Bolsonaro (columns 1 and 3) and the municipalities where Bolsonaro's popularity was below median (columns 2 and 4). Post-election is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. The sample is restricted to municipalities in which we observe at least 10 individuals. Standard errors clustered at the municipality level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For the permutation test (p-values in brackets), we run all possible placebos (23) using either the months pre-election or post-election and assigning random threshold dates.

Following the election, municipalities with lower pre-election support for Bolsonaro experienced a larger increase in the fraction of users engaging in hate speech. In columns (1) and (3), the interaction between the post-election indicator and Bolsonaro's vote share is negative and highly significant for both the BERT- and dictionary-based measures, indicating that municipalities with lower electoral support for Bolsonaro experienced larger post-election increases in the fraction of users engaging in hate speech. Quantitatively, a one–standard–deviation increase in Bolsonaro's municipal vote share is associated with a 1.6 percentage-point smaller post-election increase in the share of users posting hate speech using BERT (and a 1.2 percentage-point smaller increase using the dictionary-based measure). Columns (2) and (4) show complementary results using a binary treatment: municipalities below the median vote share exhibit a 5.1 percentage point larger post-election increase in hate-posting users using BERT (and a 3.7 percentage point larger increase using the dictionary-based measure). This pattern indicates that a substantial

[29]The sample is restricted to municipalities in which we observe at least 10 individuals. As with our individual-level regressions (below), we further restrict attention to users present in no more than three municipalities during the whole period under study. Table A6 shows that the results are robust to keeping only the user accounts observed in a single location.

portion of the aggregate increase in hate speech reflects entry along the extensive margin, a necessary implication of a decline in the perceived social cost of expressing such views.

Second, if this interpretation is correct, the post-election increase in hate speech should be concentrated in forms of expression that are especially norm-sensitive and aligned with Bolsonaro's publicly expressed views. We would expect the increase to target specific social groups—women, the LGBTQ+ community, and different races or cultures—rather than a uniform rise in antagonistic language. To test this implication, we use the dictionary-based classifier to disaggregate hate speech into six categories: homophobia, racism, sexism, political hate, political target, and insults. Figure A12 in the Online Appendix shows the evolution of each category, distinguishing between control and treatment municipalities. It is noteworthy that for homophobia, the gap between the two groups of municipalities entirely vanishes after the election. Although homophobic content is less prevalent than sexist or racist content, its magnitude is comparable to that of political hate, suggesting that norm-sensitive forms of expression account for a meaningful share of the overall increase.

Table 3 reports difference-in-differences estimates by hate-speech target, with treatment intensity measured by Bolsonaro's municipal vote share in Panel A. Consistent with a norm-based interpretation, the post-election response varies substantially across categories. We find a large and precisely estimated effect for homophobic content: a standard deviation increase in Bolsonaro's vote share is associated with a 0.037–standard deviation smaller post-election increase in homophobic hate speech. We also find a negative and statistically significant effect for sexist content, though of smaller magnitude (coeff.=-0.017). By contrast, the estimated effects for racism, insults, and both measures of political hate are small, statistically insignificant, and exhibit no systematic pattern. In particular, the absence of any differential response in political hate suggests that the results are unlikely to be driven by differential changes in political polarization. Panel B presents complementary estimates from a binary specification that compares municipalities below and above the median of Bolsonaro's vote share. These results closely mirror the continuous-treatment estimates: municipalities below the median experienced significantly larger post-election increases in homophobic and, to a lesser extent, sexist content, while no robust differences emerge for political hate or other categories.

To further corroborate the social-norm interpretation of our findings, we examine municipal-level prejudice outcomes from the Americas Barometer in 2017 and 2019. Table A7 reports OLS regressions of standardized measures of homophobia and sexism on Bolsonaro's first-round vote share in the 2018 presidential election (Panel A) and on an indicator for municipalities below the median vote share (Panel B). Consistent with our

findings on Twitter, we see that the homophobia measure in the 2017 wave correlated strongly with Bolsonaro's vote share, but became negative and statistically insignificant in the 2019 wave. The correlation with sexist attitudes is smaller and less systematically related to electoral support.

Table 3: MUNICIPALITY LEVEL REGRESSIONS. RESULTS BY HATE TARGETS.

Panel A: Vote share

	Politics	Pol. Target	Homophobia	Racism	Sexism	Insults
Post-election × Vote share	0.006	0.008	-0.037***	0.003	-0.017**	-0.008
	(0.017)	(0.007)	(0.010)	(0.010)	(0.008)	(0.009)
Observations	137,271	137,271	137,271	137,271	137,271	137,271
R-squared	0.116	0.034	0.045	0.147	0.047	0.076
Permutation test (p-value)	[0.125]	[0.250]	[0.125]	[0.250]	[0.292]	[0.042]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Municipalities	2,490	2,490	2,490	2,490	2,490	2,490

Panel B: Below median popularity

	Politics	Pol. Target	Homophobia	Racism	Sexism	Insults
Post-election × Below median popularity	-0.029	-0.023**	0.050***	-0.017	0.023*	0.011
	(0.021)	(0.011)	(0.014)	(0.015)	(0.013)	(0.014)
Observations	137,271	137,271	137,271	137,271	137,271	137,271
R-squared	0.116	0.034	0.045	0.147	0.047	0.076
Permutation test (p-value)	[0.458]	[0.500]	[0.125]	[0.625]	[0.292]	[0.125]
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Municipalities	2,490	2,490	2,490	2,490	2,490	2,490

Notes: The table reports coefficients of regressing the share of (target-specific) hate speech in the municipality (standardized) on the vote share for Bolsonaro (Panel A) and the municipalities where Bolsonaro's popularity was below median (Panel B). Post-election is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. Standard errors clustered at the municipality level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For the permutation test (p-values in brackets), we run all possible placebos (23) using either the months pre-election or post-election and assigning random threshold dates.

Together, the effects on the extensive margin and the specific targets of hate speech suggest a shift in social norms as the underlying mechanism, as distinct from alternative explanations such as differential increases in polarization or changes in overall online activity.

4.2 Individual level

In the previous section, we document a sharp increase in online hate speech following the 2018 Brazilian presidential election. This increase is highly heterogeneous across space and is disproportionately driven by municipalities in which Bolsonaro received relatively low electoral support.

A key advantage of our data is that it allows us to follow individual Twitter users over time. We therefore extend the aggregate analysis by examining who is driving the post-election increase in hate speech. Conceptually, this increase may arise through distinct channels: (i) entry into hate speech by users who were previously not engaging in hate speech (the extensive margin explored in the previous section), (ii) the strength of hate posted by the new entrants (the intensity of entrants), (iii) an intensification of hateful content among users who were already posting hate speech prior to the election (the intensity of incumbents), (iv) or a combination of the three.

To explore the two intensive margins separately, we partition the sample based on users' pre-election behavior. Users who posted no hate speech prior to the election (N=13,902) identify the intensive margin of the new entrants, while users who had already posted hate speech prior to the election (N=62,370) identify the intensive margin of the incumbents (capturing changes in the share of hateful content). We restrict the analysis to Twitter users whose tweets are geolocated in no more than three municipalities, which limits measurement error in assigning local electoral exposure. In addition, we require users to have posted at least five tweets per month to ensure reliable measurement of individual-level outcomes. For users whose tweets span multiple municipalities, electoral exposure is measured as a weighted average of Bolsonaro's vote share between them, with weights given by the fraction of tweets posted in each municipality.

Table 4 reports individual-level regressions focusing on users who posted hate speech prior to the elections. The effect of Bolsonaro's vote share on the intensive margin for this group is precisely estimated as zero for both the BERT and dictionary-based measures. The corresponding coefficients for *Below median popularity* are likewise small and statistically indistinguishable from zero.[30]

Table 5 reports individual-level regressions for users who posted no hate speech prior to the election and therefore identifies the strength of entry into hate speech. Across both hate speech classifiers, we find large and precisely estimated effects. A one–standard–deviation increase in Bolsonaro's vote share is associated with a significantly larger post-election increase in hate speech, with coefficients of 0.051 using the BERT classifier and 0.017 using the dictionary-based measure. Similarly, users located in municipalities where Bolsonaro's popularity was below the median exhibit substantially smaller post-election increases in hate speech, with effect sizes reaching -0.14 standard deviations for BERT-based hate speech.

[30]Online Appendix Table A8 shows that for the incumbents' intensive margin, there are strong negative and statistically significant effects for sexism and homophobia.

Table 4: INDIVIDUAL LEVEL REGRESSIONS - INCUMBENTS.

	Hate Speech_BERT	Hate Speech_BERT	Hate Speech_DICT	Hate Speech_DICT
Post-election × Vote share	0.002		-0.000	
	(0.004)		(0.004)	
Post-election × Below median popularity		-0.007		-0.004
		(0.012)		(0.011)
Observations	396,321	396,321	508,168	508,168
R-squared	0.256	0.256	0.260	0.260
User FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Users	62,370	62,370	65,386	65,386

Notes: The table reports coefficients of regressing the share of hate tweets by individual (standardized) on the vote share for Bolsonaro (columns 1 and 3) and the municipalities where Bolsonaro's popularity was below median (columns 2 and 4). Post-election is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. The sample is restricted to users with at least 50 tweets over the entire period and with a positive number of hate speech in the pre-election period. Standard errors clustered at the user level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

These estimates imply that, among users who first engage in hate speech after the election, the intensity of hateful expression is higher in municipalities with stronger support for Bolsonaro. This pattern does not contradict our broader interpretation. Rather, it is consistent with the idea that the election surprise relaxed social norms nationwide, while local political support shaped the strength of this response. In municipalities with greater electoral support for Bolsonaro and a higher baseline prevalence of hate speech, users may be willing to express hate speech more intensely once they start engaging in this behavior.[31]

Table 5: INDIVIDUAL LEVEL REGRESSIONS - NEW ENTRANTS.

	Hate Speech_BERT	Hate Speech_BERT	Hate Speech_DICT	Hate Speech_DICT
Post-election × Vote share	0.051***		0.017***	
	(0.007)		(0.005)	
Post-election × Below median popularity		-0.136***		-0.044***
		(0.021)		(0.015)
Observations	65,620	65,620	98,386	98,386
R-squared	0.436	0.436	0.327	0.327
User FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Users	13,902	13,902	15,097	15,097

Notes: The table reports coefficients of regressing the share of hate tweets by individual (standardized) on the vote share for Bolsonaro (columns 1 and 3) and the municipalities where Bolsonaro's popularity was below median (columns 2 and 4). Post-election is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. The sample is restricted to users with at least 50 tweets over the entire period and with no hate speech in the pre-election period. Standard errors clustered at the user level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

For completeness, we also estimate specifications that pool together users who did and

[31] Online Appendix Table A9 shows that the increase seems to be driven by insults, but there is evidence of an opposite effect for politics and racism.

did not post hate speech prior to the election, thereby abstracting from the distinction between the intensive margins for incumbents and entrants. In these pooled regressions, the estimated effects are smaller in magnitude and statistically significant only for the dictionary-based measure, although they consistently point in the same direction as our baseline results. This attenuation is expected, as pooling mechanically averages over two margins that respond differently to the election and masks the substantial heterogeneity documented above. We report the full set of pooled specifications in the Online Appendix Tables A10 and A11, which show that the signs of the coefficients remain stable and similar to those in our municipality regressions across hate speech measures, targets, and specifications.

The availability of individual-level data also allows us to describe which types of users account for the post-election increase in hate speech. In particular, we examine whether the rise in individual hate speech is driven by specific subsets of users—such as those with larger audiences (more followers), greater network engagement (following more accounts), or higher activity levels (tweeting more frequently). Figures A13, A14, and A15 in the Online Appendix show that this is not the case. Each figure plots the average individual-level hate speech (as classified by the BERT model) against the number of followers, the number of accounts followed, and overall tweeting activity. The top panels compare users located in municipalities with above- and below-median support for Bolsonaro before and after the election, while the bottom panels contrast pre- and post-election periods separately within above- and below-median municipalities. Across all dimensions, we observe that post-election differences between users in different types of municipalities narrow substantially: users in below-median municipalities increase their average hate speech by a larger extent than those in above-median municipalities. This convergence indicates that the post-election surge in individual hate speech is broad-based rather than driven by a specific group of users (e.g., political bots).

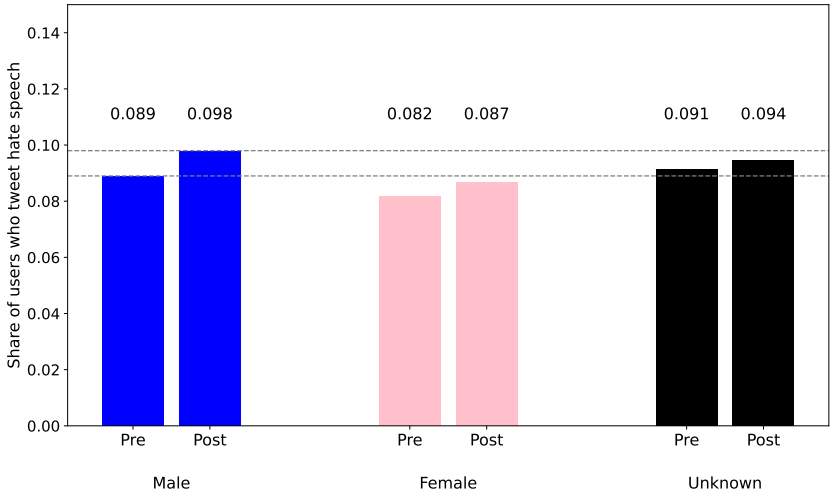
Taken together, our findings suggest that the post-election increase in hate speech operated largely through the extensive margin, with a substantial rise in the number of users engaging in hateful expression. Entry into hate speech was more pronounced in municipalities with lower pre-election support for Bolsonaro. At the same time, conditional on entry, the intensity of hateful expression was higher in municipalities where Bolsonaro enjoyed greater electoral support, indicating convergence toward a higher local norm of acceptability. Overall, this pattern is consistent with our interpretation that the election reshaped prevailing social norms governing the public expression of hateful views.

4.3 Gender heterogeneity

To explore whether the post-election increase in hate speech varies across user types, we classify accounts by likely gender using a comprehensive registry of names from the Brazilian 2010 Census (IBGE) and multilingual naming records. Users are assigned to one of three categories: male, female, or unknown, with the latter including accounts whose names do not match any entries in the reference dataset.

Figure 5 reports the share of users posting hate speech before and after the election by gender category. Before the election, both male and female accounts exhibited lower rates of hate speech than accounts of unknown gender, consistent with the idea that users with non-identifiable or pseudonymous profiles may already face weaker reputational constraints. In both periods, female accounts show lower rates of hate speech, which is consistent with women having a greater concern for social consequences (Bayerl et al., 2025). All three groups exhibit an increase in hate speech in the post-election period, but the rise is largest for accounts classified as male, both in absolute and relative terms. Relative to the pre-election baseline, the share of hate-posting users increases by approximately 10.1% among male accounts, compared with 6.1% among female accounts and 3.3% among accounts of unknown gender. Figure A16 in the Online Appendix shows a similar pattern using the dictionary-based classifier.

Figure 5: Increase in hate speech by gender group.



Notes: The figure shows the share of users posting hate speech in the pre- and post-election periods, separately by gender. Gender is classified as male, female, or unknown, based on the names in the Census. Dashed horizontal lines denote the pre- and post-election averages for males for ease of comparison. Hate speech is measured using the BERT-based classifier.

The comparatively small increase among accounts of unknown gender is also consistent with a norm-based interpretation. To the extent that this category includes a disproportionate share of pseudonymous or anonymous accounts, these users may face lower reputational or social sanctions for expressing hateful views even before the election. If so, a post-election relaxation in the perceived social cost of hate speech should have a weaker effect on their behavior, exactly as the data suggest. Interestingly, the post-election prevalence of hate speech among men surpasses that of unknown-gender accounts, consistent not only with reduced perceived social sanctions but also with higher expressive, identity payoffs as the in-group's status rises (Bénabou and Tirole, 2011; Akerlof and Kranton, 2000).

5 Conclusion

This paper studies how unexpected political events can shape social norms and individual behavior by examining the impact of Jair Bolsonaro's victory in the 2018 Brazilian presidential election on online hate speech. Using Twitter data from 2017–2019 and Natural Language Processing techniques, we document a clear and sustained increase in hate speech nationwide following the election in virtually all states. The fact that hate speech is not uniformly distributed across space is consistent with local differences in prevailing social norms.

Our central finding is that, while the prevalence of hate speech is higher in municipalities where Bolsonaro has stronger support, the post-election increase in hate speech is disproportionately larger in municipalities where Bolsonaro was less popular. In these municipalities, the electoral surprise was stronger, revealing that views associated with Bolsonaro's rhetoric were more widely held nationally than previously believed. This revelation lowered the perceived social cost of expressing hateful content, leading to a sharper post-election rise in hate speech. This pattern is consistent with a break in the spiral of silence.

We further show that this increase operates primarily along the extensive margin. After the election, a large number of individuals who did not post hate content began engaging in hate speech. At the same time, conditional on entering, the intensity of hateful expression is higher in municipalities where Bolsonaro enjoyed greater support. This pattern is consistent with new entrants converging toward stronger local norms in environments, where such views are more socially reinforced. Together, these results highlight how national political shocks interact with local social environments to shape online behavior.

Differentiating hate speech by its targets reveals that homophobia and sexism are key

drivers of the post-election surge. These categories align closely with themes prominent in Bolsonaro's public rhetoric, reinforcing the interpretation that political leadership can influence not only political attitudes but also broader norms governing acceptable public discourse. Using individual-level data, we show that these effects are broad-based: they are not concentrated among users with less followers, who follow less accounts, or with higher tweeting activity. Instead, the post-election increase in hate speech reflects a widespread behavioral response across the user base.

Finally, the rise in hate speech is most pronounced among male users, followed by female users, and smallest among accounts that could not be classified. This unknown-gender group contains a disproportionate share of pseudonymous or anonymous accounts that are less exposed to social scrutiny. Thus, its muted effect supports our interpretation of a shift in the social acceptability of hate speech.

These findings have important implications beyond the online sphere. A growing body of evidence links online hate speech to offline harms, including discrimination, intimidation, and violence. Our results, therefore, highlight that major political events can generate substantial increases in online hate speech through shifts in perceived social norms, a channel that the existing literature links to real-world outcomes. From a policy perspective, they highlight the importance of timely and geographically targeted interventions in the aftermath of elections.

More broadly, by precisely identifying when, where, and through which margins hate speech responds to a major electoral surprise, this paper offers concrete guidance for public policy and platform governance. Our findings suggest that the immediate post-election period represents a critical window during which social norms are revised, and harmful behaviors can diffuse rapidly through new participation. Interventions that are timely and geographically targeted—such as intensified moderation, norm-setting signals, or friction for first-time offenders—may therefore be especially effective in limiting the normalization of hateful discourse. Looking ahead, future research could assess whether similar dynamics emerge in other electoral and institutional contexts and rigorously evaluate which policy and platform interventions are most effective at preventing online hate from translating into offline social harms.

References

Acemoglu, D. and Jackson, M. O. (2015). History, expectations, and leadership in the evolution of social norms. *The Review of Economic Studies*, 82(2):423–456.

Acemoglu, D. and Jackson, M. O. (2017). Social norms and the enforcement of laws. *Journal of the European Economic Association*, 15(2):245–295.

Ajzenman, N., Cavalcanti, T., and Da Mata, D. (2023). More than words: Leaders' speech and risky behavior during a pandemic. *American Economic Journal: Economic Policy*, 15(3):351–371.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753.

Albornoz, F., Bradley, J., and Sonderegger, S. (2022). Updating the social norm: the case of hate crime after the Brexit Referendum. Technical report, Red Nacional de Investigadores en Economía (RedNIE).

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the origins of gender roles: Women and the plough. *The Quarterly Journal of Economics*, 128(2):469–530.

Andre, P., Boneva, T., Chopra, F., and Falk, A. (2024). Misperceived social norms and willingness to act against climate change. *Review of Economics and Statistics*, pages 1–46.

Aridor, G., Jiménez-Durán, R., Levy, R., and Song, L. (2024). The economics of social media. *Journal of Economic Literature*, 62(4):1422–1474.

Ash, E. and Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, 15.

Ayo, F. E., Folorunso, O., Ibharalu, F. T., and Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, 38:100311.

Badgett, M. L., Carpenter, C. S., and Sansone, D. (2021). Lgbtq economics. *Journal of Economic Perspectives*, 35(2):141–170.

Barros, L. and Silva, M. S. (2025). Economic shocks, gender, and populism: Evidence from Brazil. *Journal of Development Economics*, 174:103412.

Bayerl, A., Dover, Y., Riemer, H., and Shapira, D. (2025). Gender rating gap in online reviews. *Nature human behaviour*, 9(3):507–520.

Beknazar-Yuzbashev, G., Jiménez Durán, R., McCrosky, J., and Stalinski, M. (2022). Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN 4307346*.

Beknazar-Yuzbashev, G., Jiménez-Durán, R., and Stalinski, M. (2024). A model of harmful yet engaging content on social media. In *AEA Papers and Proceedings*, volume 114, pages 678–683. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.

Bhuller, M., Havnes, T., Leuven, E., and Mogstad, M. (2013). Broadband internet: An information superhighway to sex crime? *Review of Economic Studies*, 80(4):1237–1266.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bursztyn, L., Egorov, G., Enikolopov, R., and Petrova, M. (2019). Social media and xenophobia: evidence from Russia. Technical report, National Bureau of Economic Research.

Bursztyn, L., Egorov, G., and Fiorin, S. (2020a). From extreme to mainstream: The erosion of social norms. *American Economic Review*, 110(11):3522–3548.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2020b). Misperceived social norms: Women working outside the home in Saudi Arabia. *American Economic Review*, 110(10):2997–3029.

Bursztyn, L. and Yang, D. Y. (2022). Misperceptions about others. *Annual Review of Economics*, 14(1):425–452.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. (2024). Difference-in-differences with a continuous treatment. *NBER Working Paper*, (w32117).

Cao, A., Lindo, J. M., and Zhong, J. (2023). Can social media rhetoric incite hate incidents? evidence from Trump's "Chinese Virus" tweets. *Journal of Urban Economics*, 137:103590.

Card, D. and Dahl, G. B. (2011). Family violence and football: The effect of unexpected emotional cues on violent behavior. *The Quarterly Journal of Economics*, 126(1):103–143.

Carr, J., Clifton-Sprigg, J., James, J., and Vujic, S. (2020). Love thy neighbour? Brexit and hate crime. Technical report, IZA Discussion Papers.

Dahl, G. and DellaVigna, S. (2009). Does movie violence increase violent crime? *The Quarterly Journal of Economics*, 124(2):677–734.

DellaVigna, S., Enikolopov, R., Mironova, V., Petrova, M., and Zhuravskaya, E. (2014). Cross-border media and nationalism: Evidence from Serbian radio in Croatia. *American Economic Journal: Applied Economics*, 6(3):103–132.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186.

Elster, J. (2020). Social norms and economic theory. In *Handbook of Monetary Policy*, pages 117–133. Routledge.

Farina, E. and Pathania, V. (2020). Papal visits and abortions: evidence from Italy. *Journal of Population Economics*, 33(3):795–837.

Fergusson Talero, L., Guerra Forero, J. A., and Robinson, J. A. (2024). Anti-social norms.

Fernandez, R. (2007). Women, work, and culture. *Journal of the European Economic Association*, 5(2-3):305–332.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Giuliano, P. (2007). Living arrangements in Western Europe: Does cultural origin matter? *Journal of the European Economic Association*, 5(5):927–952.

Ivandic, R., Kirchmaier, T., and Machin, S. (2024). International terror attacks and local out-group hate crimes. *The Journal of Law and Economics*, 67(3):589–610.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pages 914–924.

Madriaza, P., Hassan, G., Brouillette-Alarie, S., Mounchingam, A. N., Durocher-Corfa, L., Borokhovski, E., Pickup, D., and Paillé, S. (2025). Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities. *Campbell systematic reviews*, 21(1):cl2–70018.

Mendes, W. G. and Silva, C. M. F. P. d. (2020). Homicide of lesbians, gays, bisexuals, travestis, transexuals, and transgender people (lgbt) in brazil: a spatial analysis. *Ciência & Saúde Coletiva*, 25:1709–1722.

Meyer, I. H. (1995). Minority stress and mental health in gay men. *Journal of Health and Social Behavior*, pages 38–56.

Meyer, I. H. (2003). Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. *Psychological Bulletin*, 129(5):674.

Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248. IEEE.

Morales, J. S. (2020). Perceived popularity and online political dissent: evidence from Twitter in Venezuela. *The International Journal of Press/Politics*, 25(1):5–27.

Müller, K. and Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.

Müller, K. and Schwarz, C. (2023). From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics*, 15(3):270–312.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature*, 56(3):799–866.

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication*, 24(2):43–51.

Nyborg, K., Anderies, J. M., Dannenberg, A., Lindahl, T., Schill, C., Schlüter, M., Adger, W. N., Arrow, K. J., Barrett, S., Carpenter, S., et al. (2016). Social norms as solutions. *Science*, 354(6308):42–43.

Pereira, V. G. (2018). *Using supervised machine learning and sentiment analysis techniques to predict homophobia in portuguese tweets*. PhD thesis.

Souza, F., Nogueira, R., and Lotufo, R. (2020). 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.

Tampellini, J. (2024). Latin american pride: labor market outcomes of sexual minorities in Brazil. *Journal of Development Economics*, 167:103239.

Vargas, F., Carvalho, I., Pardo, T. A., and Benevenuto, F. (2025). Context-aware and expert data resources for brazilian portuguese hate speech detection. *Natural Language Processing*, 31(2):435–456.

Yanagizawa-Drott, D. (2014). Propaganda and conflict: Evidence from the Rwandan genocide. *The Quarterly Journal of Economics*, 129(4):1947–1994.

Zhuravskaya, E., Petrova, M., and Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, 12(1):415–438.

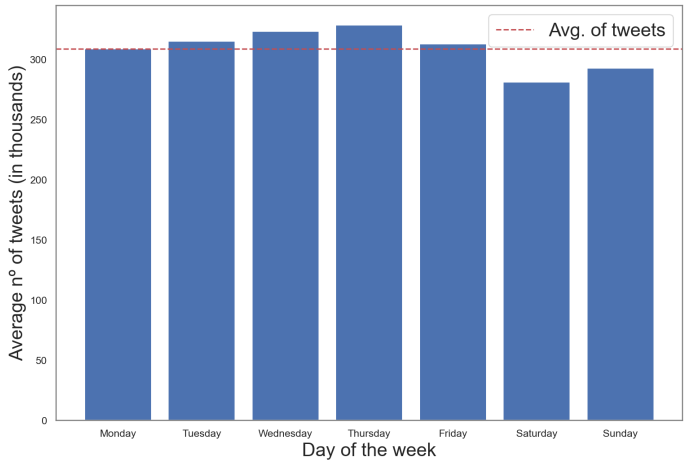
A Online Appendix

A.1 Twitter data

Twitter was an online platform allowing users to publish short messages of a maximum of 140 characters on their profiles (extended to 280 in November 2017). In January 2021, Twitter launched an Academic Research product track, which enabled researchers to access all v2 endpoints. Notably, the *Twitter Search API v2* gave access to the entire history of public conversations and not only recent tweets. To collect the Twitter data used in this paper, we relied on the *v2 full-archive search endpoint*. We collected tweets using the command line tool and Python library, twarc2, `https://twarc-project.readthedocs.io/en/latest/twarc2_en_us/` from June 2022 to May 2023.

Our Twitter query retrieves all publicly available (and non-deleted) tweets written in Portuguese, geolocated in Brazil, excluding retweets, and posted on Mondays between July 2017 and December 2019 (inclusive). We assume the sample of geolocated tweets posted on any Monday are representative of the tweets' universe. The figures below present supportive evidence for this assumption.

Figure A1: Average number of tweets per day in the tweets' sample.

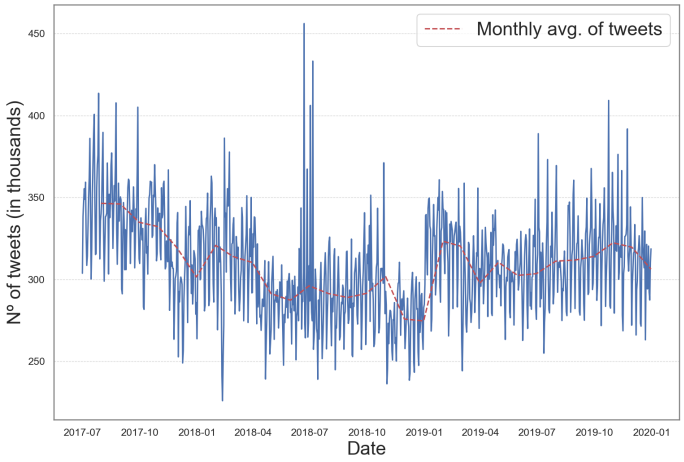


Notes: Average number of tweets per day of the week available through the Twitter API v2 for the period under study (July 2017 to December 2019).

Figure A1 presents the average number of tweets per day of the week for the period under study. The figure shows that the number of tweets is relatively stable on weekdays and slightly decreases on weekends, with Mondays being the closest to the average. The daily average number of tweets is around 309.000. Figure A2 shows the daily amount

of tweets retrieved by the Twitter query used in this paper, but without the restriction of being posted on a Monday. The red dashed line corresponds to the monthly average of tweets.[32] Figure A2 shows that the monthly variation in tweet volume is substantially larger than the variation across weekdays displayed in Figure A1. Taken together, these results suggest that long-run variation in tweet activity dominates short-run fluctuations, and sampling tweets from one day per week over the entire period provides a reliable representation of overall dynamics.

Figure A2: Daily count of tweets retrieved by the Twitter query.



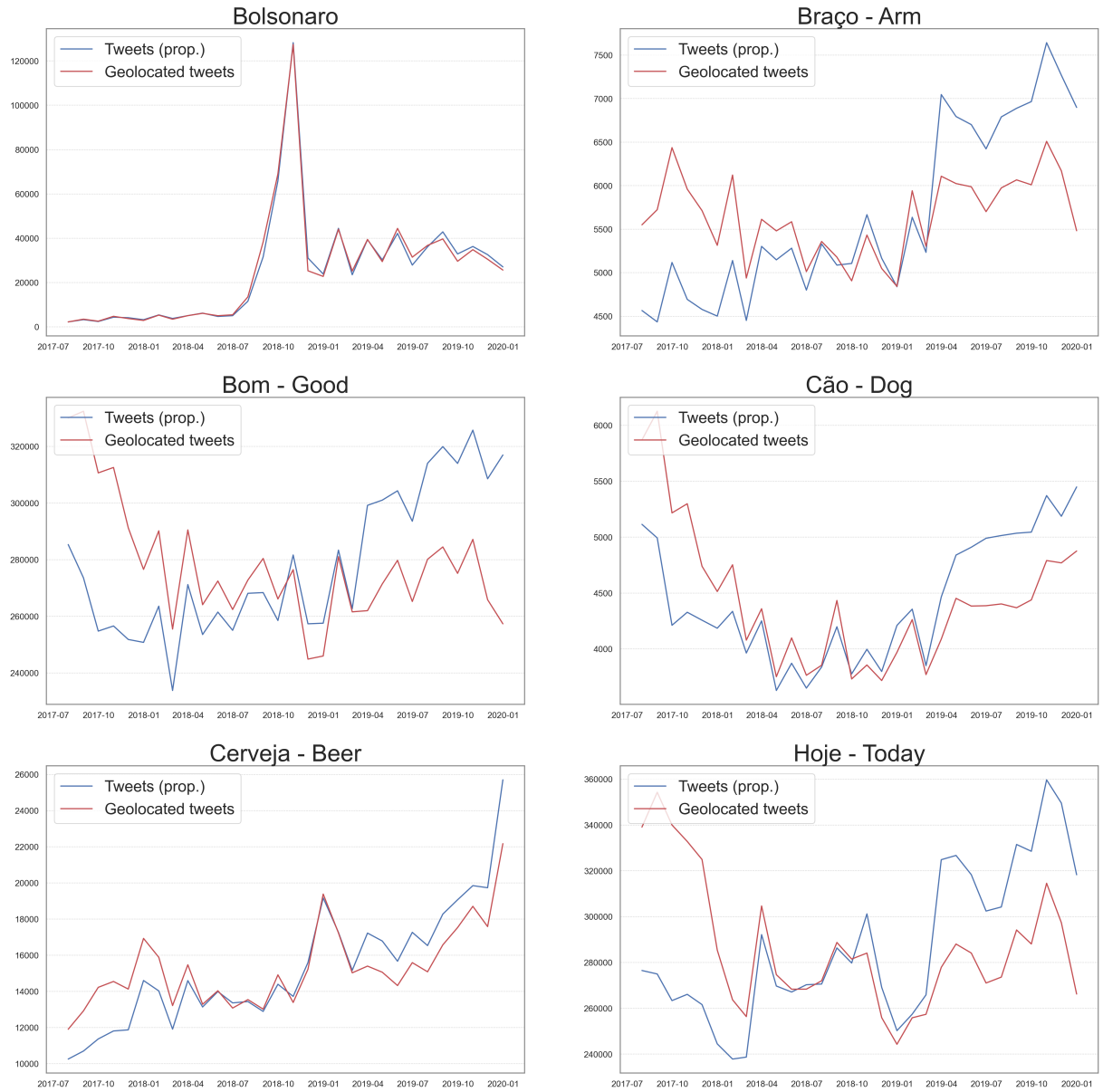
Notes: Evolution of the average number of tweets per day available through the Twitter API v2 for the period under study (July 2017 to December 2019).

Figures A3 and A4 compare trends of tweets containing specific words for two samples: the subset of geolocated tweets and the universe of tweets. In all the sub-graphs of the two figures, the red line corresponds to the number of geolocated tweets, and the blue line to the amount of all tweets multiplied by a *scalability factor*. This factor is the ratio of geolocated tweets to total tweets for each word, ranging from 4% to 8%. Figure A3 considers geolocated tweets containing a set of neutral and frequently used words—*Bolsonaro*, *braço* (arm), *bom* (good), *cão* (dog), *cerveja* (beer), and *hoje* (today). Figure A4 repeats the exercise for sensitive terms classified as hate speech by the dictionary method, including *mariquinha* (derogatory term for a gay man), *sapatão* (derogatory term for a lesbian), *nego* and *preto* (racial terms referring to Black individuals), and *piranha* and *putinha* (derogatory terms for women). Across both sets of words, the patterns observed in geolocated

[32]Peaks during June/July 2018 correspond to dates when Brazil's football team played a match in the 2018 World Cup.

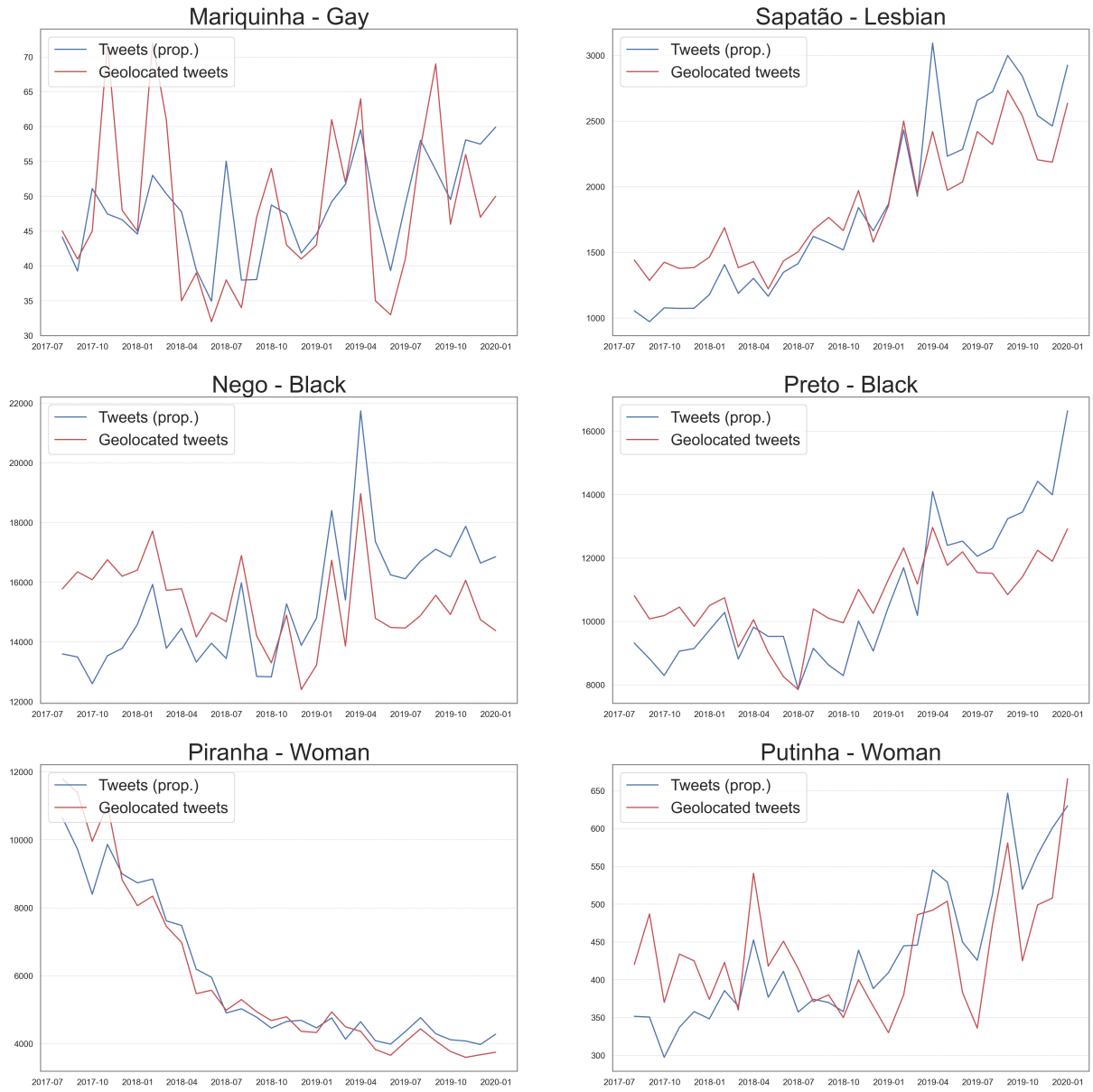
tweets closely track those in the full sample. This similarity suggests that restricting attention to geolocated tweets does not distort underlying usage patterns, particularly for Bolsonaro-related content and dictionary-based hate expressions, which are the core focus of our analysis.

Figure A3: Trend of geolocated and total tweets.



Notes: Each subplot presents the evolution of the number of tweets containing a specific word for two samples: the subset of geolocated tweets and the universe of tweets available through the Twitter API v2.

Figure A4: Trend of geolocated and total tweets.



Notes: Each subplot presents the evolution of the number of tweets containing a specific word for two samples: the subset of geolocated tweets and the universe of tweets available through the Twitter API v2.

A.2 Hate speech detection

In this paper, we implement two NLP techniques for detecting hate speech in tweets: a BERT model and a dictionary-based model. The reason for proposing two NLP techniques is twofold. First, by applying two different methods, we obtain two independent predictions of hate speech, thereby allowing us to assess the robustness of our results. The BERT model achieves state-of-the-art performance on hate speech detection, and the dictionary-based model enables further classification of hate targets. In the next paragraphs, we describe the resources employed and the procedures followed for each classification.

BERT Model. We use *BERTimbau*, a BERT model for Brazilian Portuguese developed by Souza et al. (2020), and fine-tune it for the hate speech detection task. Fine-tuning refers to the technique of training a pre-trained model on a suitable dataset for a new task. Souza et al. (2020) present the model in two sizes, Base and Large, and in this paper, we use *BERTimbau*-Base. The model improves the state of the art on traditional language tasks, outperforming Multilingual BERT models.

Training dataset. To fine-tune the BERT model for hate speech detection, we use the dataset introduced by Fortuna et al. (2019). The dataset consists of 5,668 tweets in Portuguese collected via Twitter's API between January and March 2017. We rely on the binary classification provided in the paper, in which each tweet was independently labeled as either "hate speech" or "not hate speech" by three annotators, with the final label determined by majority vote. Using this procedure, 31.5 percent of tweets are classified as containing hate speech.

Text pre-processing. We largely follow the pre-processing procedure in Fortuna et al. (2019). Specifically, we remove stop words and punctuation marks using the *NLTK* and *re* Python libraries. Unlike the original implementation, however, we retain negative stop words that may alter semantic meaning—namely *mas* (but), *nem* (neither), *não* (no), *sem* (without), and *fora* (out). We anonymize user mentions as *@user* and URLs as *URL*, while preserving *#hashtags* in their native Twitter format, as they may convey relevant information. Finally, to remain consistent with the architecture and tokenization of the Souza et al. (2020) BERT model, we do not convert text to lowercase.

Model fine-tuning. We split the labeled dataset into training (80 percent), validation (10 percent), and testing (10 percent) samples. In NLP applications, model performance

is highly sensitive to the composition of the training data. As is common in hate speech and offensive content datasets, the data exhibit class imbalance: under the binary classification, only 31.5 percent of tweets are labeled as hate speech. To mitigate potential performance issues arising from this imbalance, we apply random oversampling to the training sample to equalize the number of observations in the minority and majority classes. This procedure randomly replicates observations from the minority class with replacement. The validation and test samples remain unchanged.

Training results. Table A1 summarizes the attained evaluation metrics (Precision, Recall, F1-score, and Accuracy), and Figure A5 shows the Receiver Operating Curve (ROC) of our fine-tuned BERT model.

Figure A5: ROC Curve

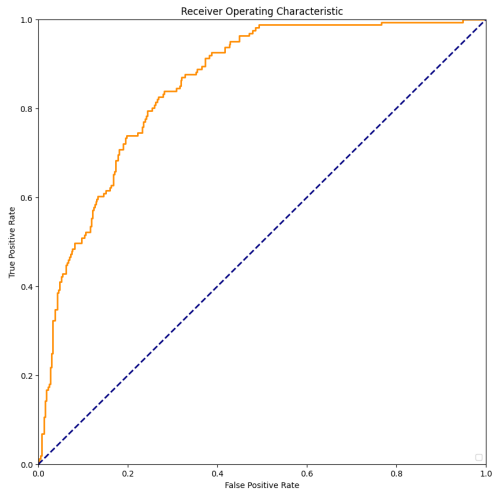


Table A1: Training results

Validation sample (N=567)			
Precision	Recall	F1-score	Accuracy
0.79	0.77	0.77	0.77
Test sample (N=567)			
Precision	Recall	F1-score	Accuracy
0.79	0.77	0.78	0.77

Dictionary method. We complement the BERT-based approach with a dictionary-based method for detecting hate speech, which generates a multi-level classification. The dictionary is designed to distinguish hate speech according to the group it targets, enabling us to analyze the dynamics of different forms of hate speech separately. Specifically, we classify hate speech into six categories: *political hate, political target, homophobia, racism, sexism*, and *insults*.

In addition to these categories, we construct a binary indicator, *hate speech*, which takes the value one if a tweet contains at least one term associated with any of the six hate categories. We use this aggregate measure to benchmark the BERT model and to conduct robustness checks of our results. Table A2 reports the set of words included in the dictionary for each hate category.

Table A2: Words included in the Hate Speech Dictionary

Category	Words			
Political hate	corrupta/o(s) comunista(s) fascista(s)	esquerdista(s) comunistinha(s) feminazi(s)	esquerdopata(s) direitalha militonto(s)	esquerdopatia facha/o(s)
Political target	bolsotario(s) bolsofãs bolsolixo ptfake ptzada	bolsonazi bolsonada bolsomerda ptralhada petezada	bolsofanaticos bolsoburro petesada lulalouco petralha(s)	bolsofurado bolsorato luladrão/ao petzadas
Homophobia	baitola(s) bixa(s) gayzada(s) sapata traveca(s) viadinho(s)	baitolo(s) boiola(s) gayzismo(s) sapatona traveco(s) viado(s)	bicha(s) boiolice(s) marica(s) sapatao viadagem(s)	bichona(s) fufa(s) mariquinha(s) sapatão viadão
Racism	chita nego(s) carioca(s) muçulmano(s)	branquice preto(s) paulista(s) mesquita	macaca/o povo(s) latino(s)	branco(s) sulista(s) islão
Sexism	burra(s) piranha(s) vagabunda(s) ninfeta	gorda(s) puta(s) safadona rapariga	feia(s) putaria safada mulherdeverdade	louca(s) putinha(s) safaneja
Insults	caralho merda(s)	fuder porra(s)	idiota(s)	lixo(s)

For the construction of the dictionary, we draw on high-frequency hate-related terms identified in three prior studies: Fortuna et al. (2019), Leite et al. (2020), and Pereira (2018). First, we rely on the hierarchical hate-speech taxonomy proposed by Fortuna et al. (2019), which provides a structured classification of hate speech in Portuguese. Second, we incorporate evidence from Leite et al. (2020), who introduce a large-scale dataset of Brazilian Portuguese tweets annotated as toxic or non-toxic and, conditional on toxicity, further classified by type. Third, we draw on Pereira (2018) to refine the *homophobia* category. Finally, we further enrich the dictionary using the Multilingual Offensive Lexicon (MOL) introduced by Vargas et al. (2025), which provides additional coverage of offensive terms.

Table A3: Classification of tweets; BERT and dictionary method.

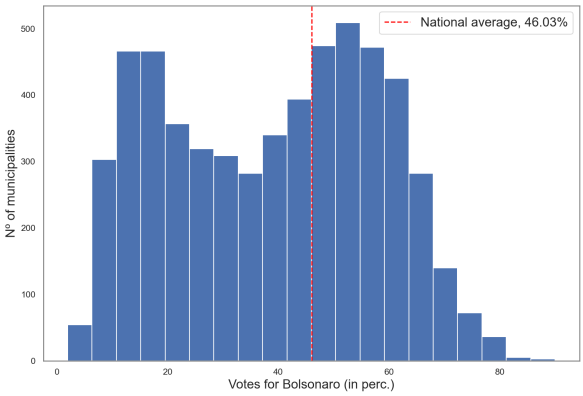
Tweets in Portuguese (in English)	BERT	Dictionary
Direitos humanos? Vai a merda seus porcos URL (Human rights? Fuck you pigs URL)	1	1
Esses caras da @user só falam m..., bando de fdp vermelhos... (These @user guys are full of s**t, a bunch of red S.O.B...)	1	0
Oh seus filhos das putas, a porra da esquerda é livre URL (Oh you sons of bitches, the fucking left is free URL)	0	1

Notes: Example of tweets belonging to our database posted on 12/03/2018 and their classification by the BERT model and the dictionary method. User mentions and web links were anonymized. These examples were manually selected to illustrate cases in which the two classification methods coincide and those in which they do not. Although neither method achieves perfect accuracy, at least one detects instances of hate speech.

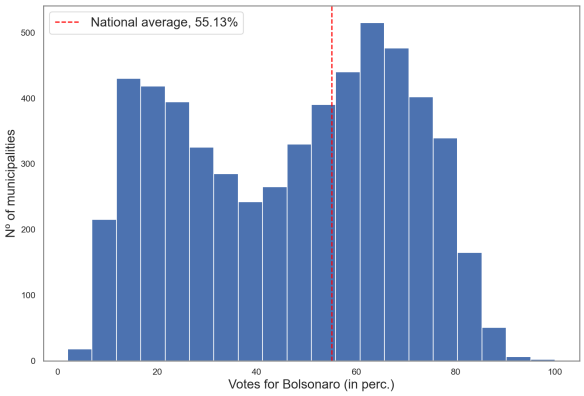
A.3 Tables and Figures

Figure A6: Bolsonaro's vote share at the municipality level. 2018 Presidential Election.

(a) 1º Round, October 7.

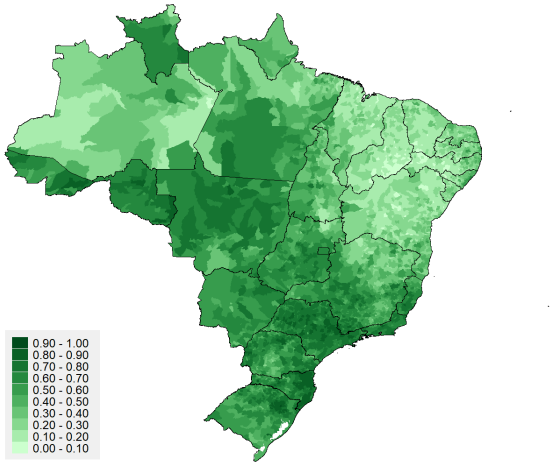


(b) 2º Round, October 28.



Notes: Histograms representing the distribution of Bolsonaro's vote share at the municipality level in the first and second rounds of the 2018 election, respectively.

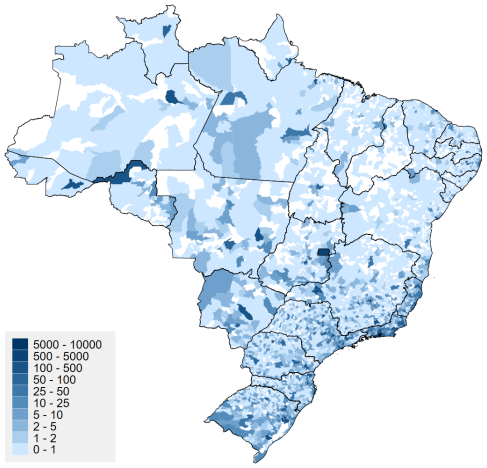
Figure A7: Spatial variation in Bolsonaro's electoral support.



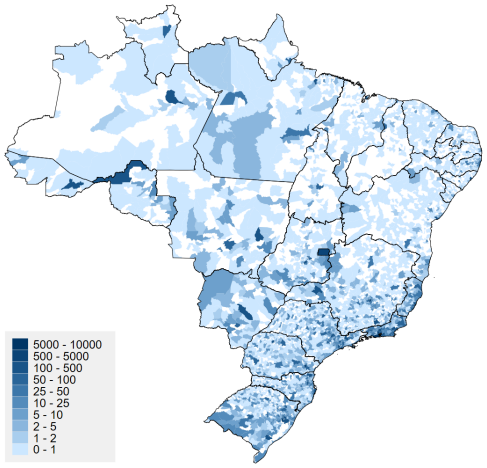
Notes: The map shows Bolsonaro's vote share (ranging from 0 to 1) at the municipality level in the second round of the presidential election.

Figure A8: Twitter penetration at the municipality level.

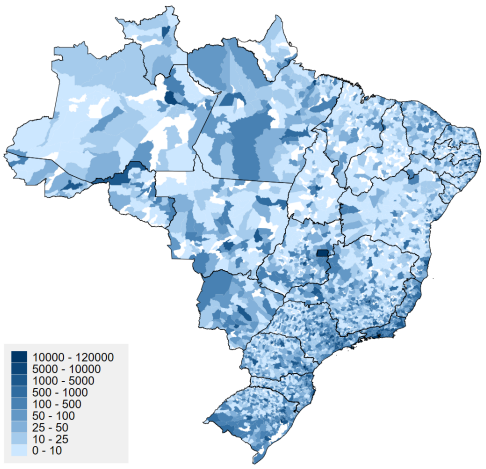
(a) Unrestricted number of tweets.



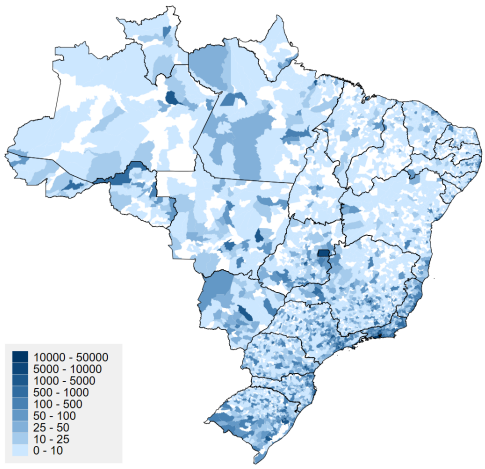
(b) Restricted number of tweets.



(c) Unrestricted number of users.



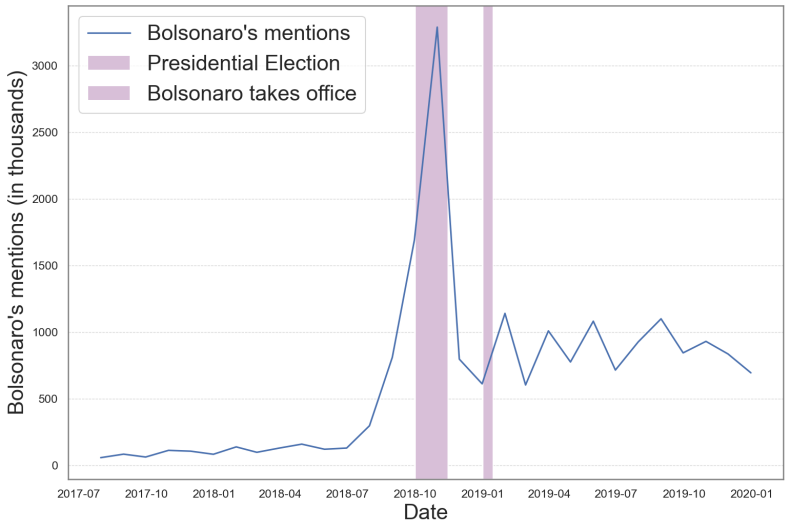
(d) Restricted number of users.



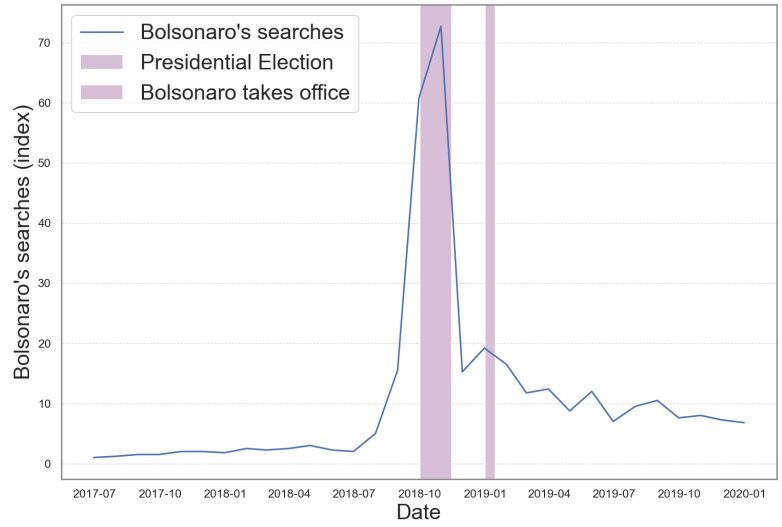
Notes: (a) Total number of tweets (in thousands) per municipality for the period under study; (b) Total number of tweets (in thousands) per municipality for the period under study, restricted to municipalities with ≥ 10 tweets per day (consistent with the restriction imposed for municipality-level regressions); (c) Total number of active users (in the sense of posting tweets) per municipality for the period under study; (d) Total number of active users per municipality for the period under study, restricted to users with ≥ 5 tweets per month (consistent with the restriction imposed for individual-level regressions).

Figure A9: Bolsonaro's mentions and Google trends, 2017-2020.

(a) Tweets including "Bolsonaro."



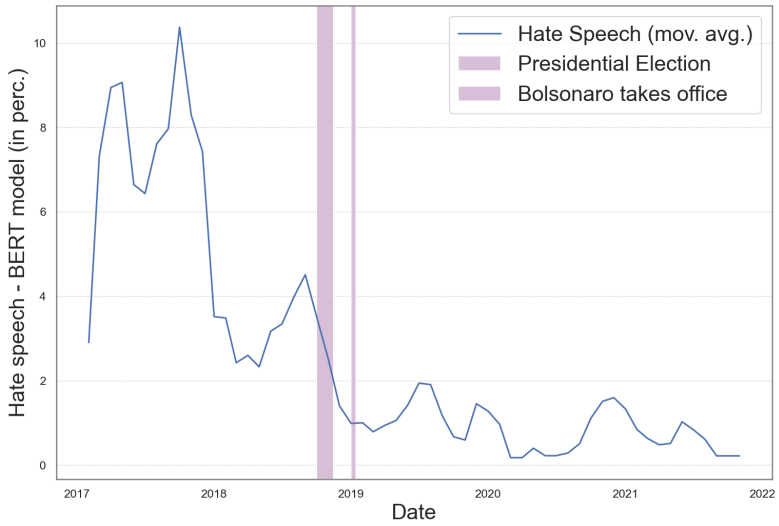
(b) Google trends for "Bolsonaro."



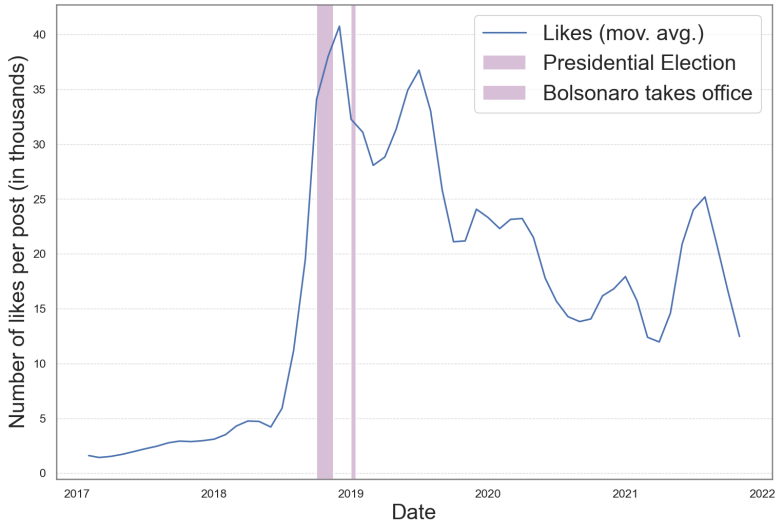
Notes: Panel (a) presents the number of tweets that include the word "Bolsonaro." Panel (b) shows the Google trends for the word "Bolsonaro" (index 0-100). Variables aggregated at the monthly level.

Figure A10: Bolsonaro's tweets, 2017-2021.

(a) Hate Speech.



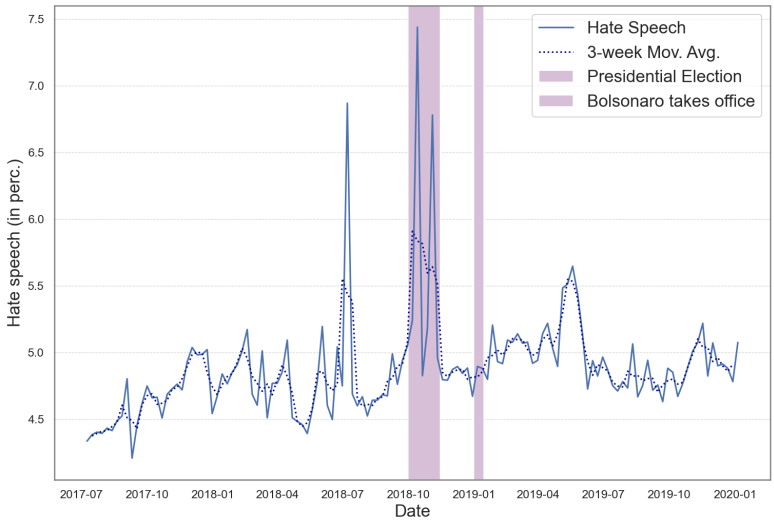
(b) Likes.



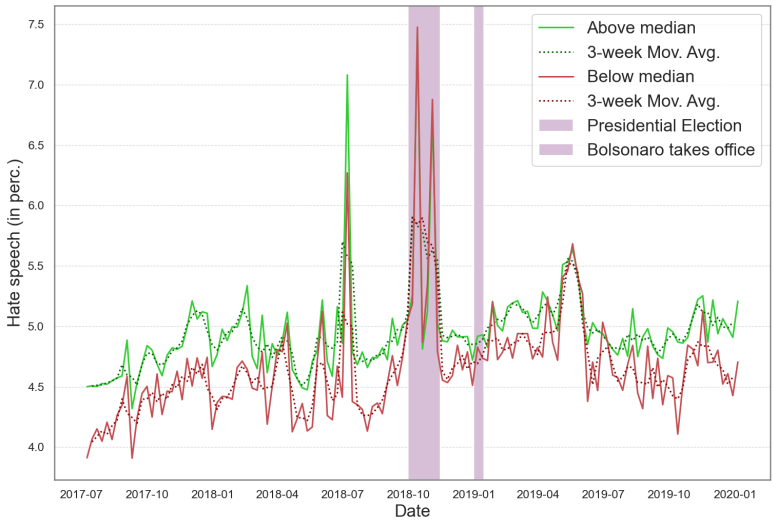
Notes: Panel (a) presents the percentage of Bolsonaro's tweets classified as hate speech by the BERT model. Panel (b) shows the number of likes (in thousands) in Bolsonaro's tweets. Variables aggregated at the monthly level, with a 3-month moving average filter.

Figure A11: Evolution of hate speech in Brazilian tweets, 2017-2019. Dictionary Method.

(a) All municipalities.



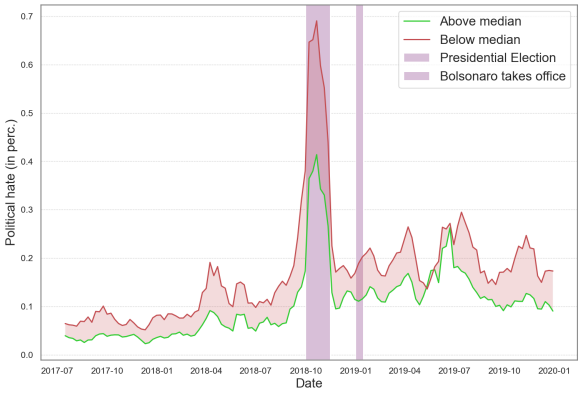
(b) Municipalities, by the 2018 election result.



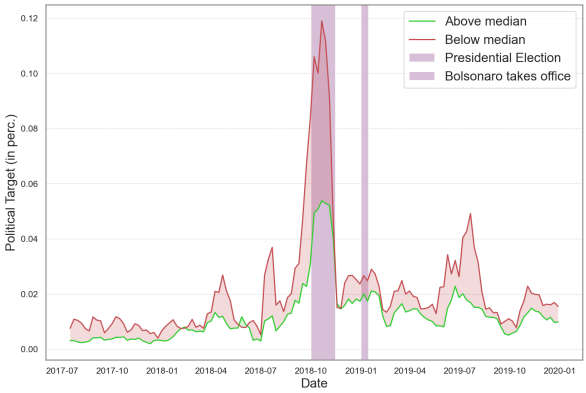
Notes: Percentage of tweets classified as hate speech by the dictionary method. National trends and trends split by Bolsonaro's vote share in the first round of the election.

Figure A12: Evolution of hate speech by targets in Brazilian tweets, 2017-2019. Dictionary method. Municipalities, by the 2018 election result.

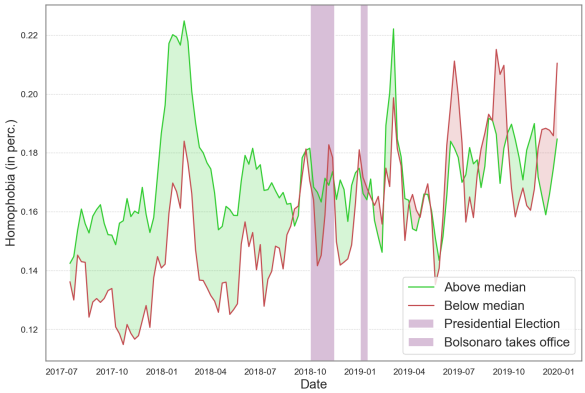
(a) Political hate.



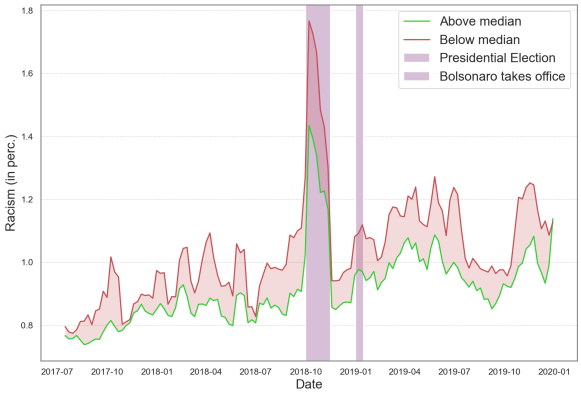
(b) Political Target.



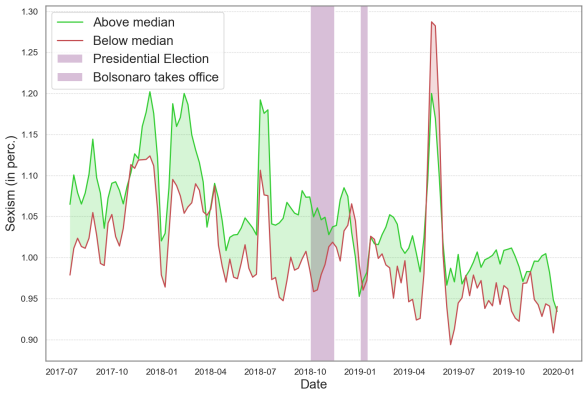
(c) Homophobia.



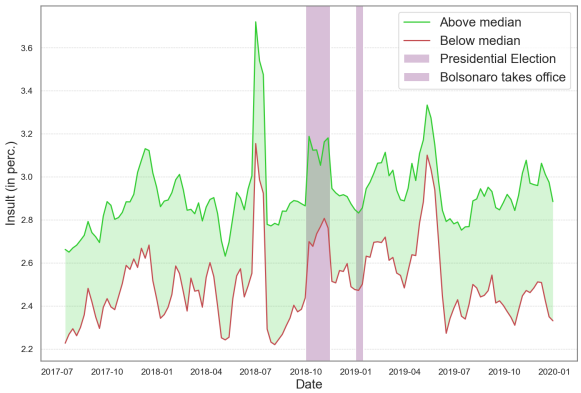
(d) Racism.



(e) Sexism.

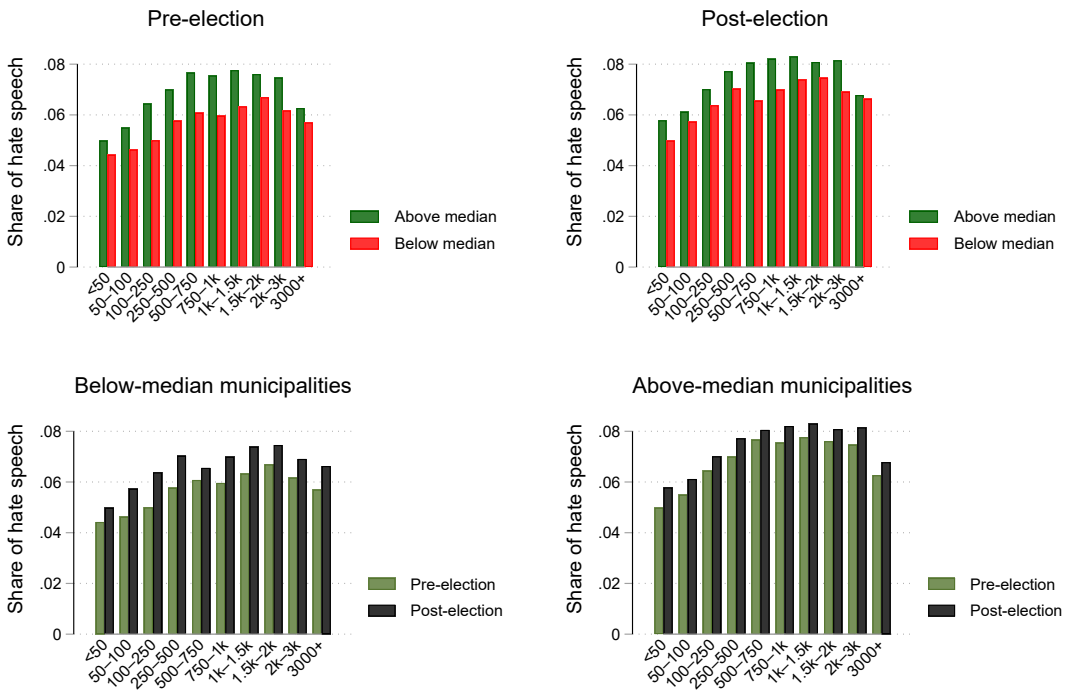


(f) Insults.



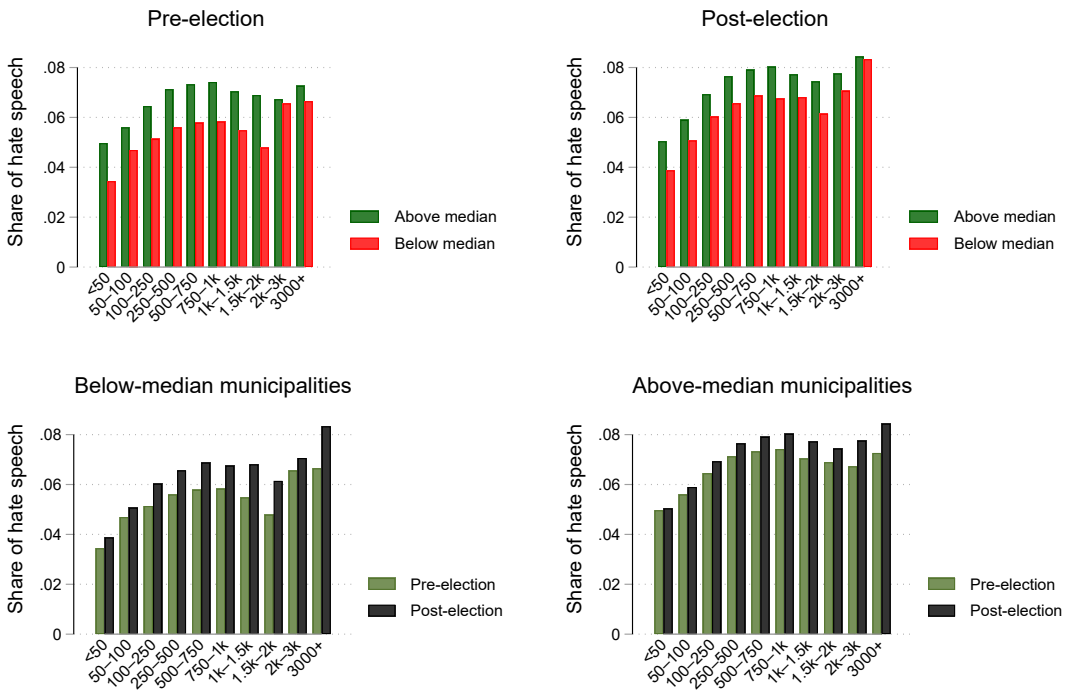
Notes: Percentage of tweets classified as hate speech by the dictionary method and considering hate targets (3-week moving average). Trends split by Bolsonaro's vote share in the first round of the election using the in-sample median vote share.

Figure A13: Share of hate speech by number of followers.



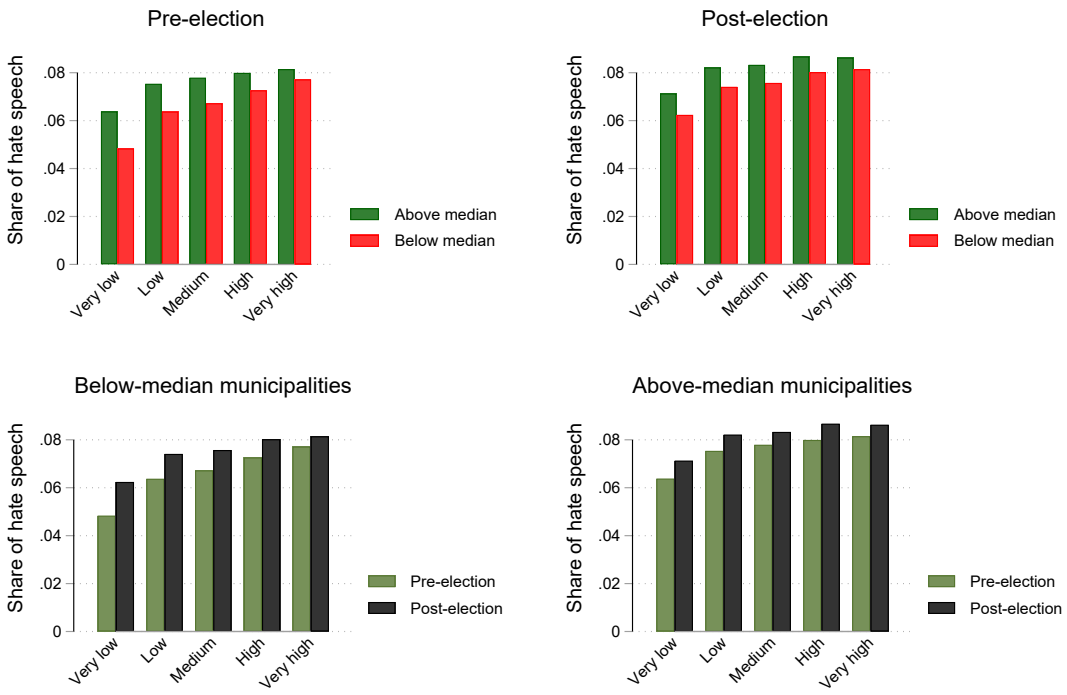
Notes: The figures show average individual-level hate speech (as classified by the BERT model) by the number of followers. The top panels compare users in municipalities with support for Bolsonaro above and below the median, before and after the election. The bottom panels compare the pre- and post-election periods separately for municipalities with above- and below-median support for Bolsonaro.

Figure A14: Share of hate speech by number of following.



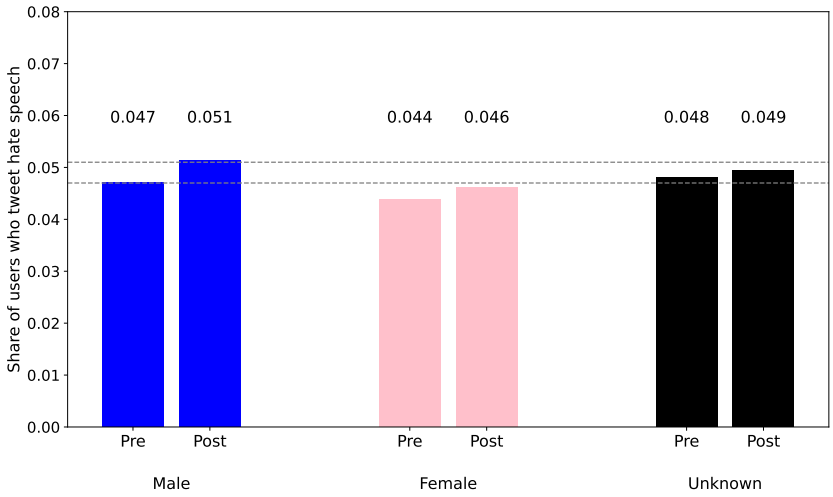
Notes: The figures show average individual-level hate speech (as classified by the BERT model) by the number of users they follow. The top panels compare users in municipalities with support for Bolsonaro above and below the median, before and after the election. The bottom panels compare the pre- and post-election periods separately for municipalities with above- and below-median support for Bolsonaro.

Figure A15: Share of hate speech by tweeting activity.



Notes: The figures show average individual-level hate speech (as classified by the BERT model) by tweeting activity. The top panels compare users in municipalities with support for Bolsonaro above and below the median, before and after the election. The bottom panels compare the pre- and post-election periods separately for municipalities with above- and below-median support for Bolsonaro. Bins are split in quintiles using the number of tweets posted: [0; 9,727], [9,728; 21,770], [21,771; 37,560], [37,561; 63,782], [63,783; 2,923,861].

Figure A16: Increase in hate speech by gender group (Dictionary).



Notes: The figure shows the share of users posting hate speech in the pre- and post-election periods, separately by gender. Gender is classified as male, female, or unknown, based on the names in the Census. Dashed horizontal lines denote the pre- and post-election averages for males for ease of comparison. Hate speech is measured using the dictionary-based classifier.

A.3.1 Regression results at the municipality level

Table A4: MUNICIPALITY LEVEL REGRESSIONS. CUTOFFS FOR DISCRETE TREATMENT.

	Hate Speech_BERT	Hate Speech_BERT	Hate Speech_DICT	Hate Speech_DICT
Post-election × Below national avg.	0.059***		0.035**	
	(0.019)		(0.017)	
Post-election × Below 50%		0.044***		0.016
		(0.017)		(0.015)
Observations	105,717	105,717	137,271	137,271
R-squared	0.078	0.078	0.098	0.098
Municipality FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Municipalities	1,932	1,932	2,490	2,490

Notes: The table reports coefficients of regressing the share of hate speech in the municipality (standardized) on the binary indicator that Bolsonaro's vote share was below the national average (columns 1 and 3) or below 50% (columns 2 and 4). Post_t is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. Standard errors clustered at the municipality level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A5: MUNICIPALITY LEVEL REGRESSIONS. REDEFINING POST.

	Hate Speech_BERT	Hate Speech_BERT	Hate Speech_DICT	Hate Speech_DICT
Post-election × Vote share	-0.031***		-0.018*	
	(0.011)		(0.010)	
Post-election × Below median popularity		0.039**		0.015
		(0.017)		(0.016)
Observations	101,140	101,140	131,328	131,328
R-squared	0.079	0.079	0.098	0.098
Municipality FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Municipalities	1,932	1,932	2,490	2,490

Notes: The table reports coefficients of regressing the share of hate speech in the municipality (standardized) on the vote share for Bolsonaro (columns 1 and 3) and the municipalities where Bolsonaro's popularity was below median (columns 2 and 4). Post_t is a dummy variable that takes a value of 0 from July 2017 to August 2018 and a value of 1 from January to December 2019. Standard errors clustered at the municipality level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A6: MUNICIPALITY LEVEL REGRESSIONS - EXTENSIVE MARGIN. NO MOVERS.

	Hate Speech_BERT	Hate Speech_BERT	Hate Speech_DICT	Hate Speech_DICT
Post-election × Vote share	-0.022***		-0.018***	
	(0.005)		(0.005)	
Post-election × Below median popularity		0.079***		0.062***
		(0.019)		(0.019)
Observations	26,683	26,683	26,683	26,683
R-squared	0.465	0.465	0.482	0.482
Municipality FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Municipalities	1,351	1,351	1,351	1,351

Notes: The table reports coefficients of regressing the share of users posting hate speech on the vote share for Bolsonaro (columns 1 and 3) and the municipalities where Bolsonaro's popularity was below median (columns 2 and 4). Post-election is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. The sample is restricted to users located in a single municipality and to municipalities in which we observe at least 10 individuals. Standard errors clustered at the municipality level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A7: CORRELATIONS BETWEEN PREJUDICE OUTCOMES AND BOLSONARO'S VOTE SHARE.

Panel A

	Homophobia		Sexism	
	2017	2019	2017	2019
Vote share	1.712***	-0.571	-0.441	-0.723
	(0.585)	(0.567)	(0.624)	(0.598)
Observations	107	107	107	107
R-squared	0.124	0.109	0.042	0.224
Controls	Yes	Yes	Yes	Yes

Panel B

	Homophobia		Sexism	
	2017	2019	2017	2019
Below median popularity	-0.536***	-0.170	0.081	0.190
	(0.191)	(0.216)	(0.210)	(0.232)
Observations	107	107	107	107
R-squared	0.105	0.106	0.038	0.218
Controls	Yes	Yes	Yes	Yes

Notes: The table reports coefficients from OLS regressions of standardized prejudice outcomes on municipal-level electoral outcomes. Panel A uses Bolsonaro's first-round vote share in the 2018 presidential election as the main explanatory variable. Panel B uses an indicator equal to one if the vote share was below 50%. All regressions control for age, gender, education and occupation. Robust standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

A.3.2 Regression results at the individual level

Table A8: INDIVIDUAL LEVEL REGRESSIONS - INCUMBENTS - RESULTS BY HATE TARGETS.

Panel A: Vote share

	Politics	Pol. Target	Homophobia	Racism	Sexism	Insults
Post-election × Vote share	0.002	-0.000	-0.011***	0.010**	-0.012***	0.001
	(0.005)	(0.006)	(0.004)	(0.004)	(0.004)	(0.004)
Observations	508,169	508,169	508,169	508,169	508,169	508,169
R-squared	0.390	0.235	0.170	0.284	0.191	0.251
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Users	65,386	65,386	65,386	65,386	65,386	65,386

Panel B: Below median popularity

	Politics	Pol. Target	Homophobia	Racism	Sexism	Insults
Post-election × Below median popularity	0.001	-0.009	0.026**	-0.031**	0.032***	-0.006
	(0.015)	(0.015)	(0.012)	(0.013)	(0.012)	(0.011)
Observations	508,169	508,169	508,169	508,169	508,169	508,169
R-squared	0.390	0.235	0.170	0.284	0.191	0.251
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Users	65,386	65,386	65,386	65,386	65,386	65,386

Notes: The table reports coefficients of regressing the share of (target-specific) hate tweets by individual (standardized) on the vote share for Bolsonaro (Panel A) and the municipalities where Bolsonaro's popularity was below median (Panel B). Post-election is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. The sample is restricted to users with at least 50 tweets in the entire period. Standard errors clustered at the individual level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A9: INDIVIDUAL LEVEL REGRESSIONS - NEW ENTRANTS - RESULTS BY HATE TARGETS.

Panel A: Vote share

	Politics	Pol. Target	Homophobia	Racism	Sexism	Insults
Post-election × Vote share	-0.015***	-0.001	-0.005	-0.010**	0.005	0.032***
	(0.006)	(0.009)	(0.005)	(0.005)	(0.005)	(0.004)
Observations	98,386	98,386	98,386	98,386	98,386	98,386
R-squared	0.286	0.313	0.169	0.232	0.205	0.282
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Users	15,097	15,097	15,097	15,097	15,097	15,097

Panel B: Below median popularity

	Politics	Pol. Target	Homophobia	Racism	Sexism	Insults
Post-election × Below median popularity	0.053***	0.018	0.027*	0.018	-0.021	-0.075***
	(0.021)	(0.021)	(0.015)	(0.016)	(0.013)	(0.012)
Observations	98,386	98,386	98,386	98,386	98,386	98,386
R-squared	0.286	0.313	0.169	0.232	0.205	0.282
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Users	15,097	15,097	15,097	15,097	15,097	15,097

Notes: The table reports coefficients of regressing the share of (target-specific) hate tweets by individual (standardized) on the vote share for Bolsonaro (Panel A) and the municipalities where Bolsonaro's popularity was below median (Panel B). Post-election is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. The sample is restricted to users with at least 50 tweets in the entire period. Standard errors clustered at the individual level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A10: INDIVIDUAL LEVEL REGRESSIONS.

	Hate Speech_BERT	Hate Speech_BERT	Hate Speech_DICT	Hate Speech_DICT
Post-election × Vote share	-0.003		-0.007**	
	(0.004)		(0.003)	
Post-election × Below median popularity		0.010		0.016*
		(0.011)		(0.010)
Observations	570,003	570,003	741,369	741,369
R-squared	0.302	0.302	0.295	0.295
User FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Users	100,161	100,161	105,012	105,012

Notes: The table reports coefficients of regressing the share of hate tweets by individual (standardized) on the vote share for Bolsonaro (columns 1 and 3) and the municipalities where Bolsonaro's popularity was below median (columns 2 and 4). Post-election is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. The sample is restricted to users with at least 50 tweets in the entire period. Standard errors clustered at the user level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A11: INDIVIDUAL LEVEL REGRESSIONS. RESULTS BY HATE TARGETS.

Panel A: Vote share

	Politics	Pol. Target	Homophobia	Racism	Sexism	Insults
Post-election × Vote share	-0.003	-0.001	-0.012***	0.001	-0.014***	0.000
	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(0.003)
Observations	741,369	741,369	741,369	741,369	741,369	741,369
R-squared	0.408	0.281	0.181	0.301	0.202	0.274
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Users	105,012	105,012	105,012	105,012	105,012	105,012

Panel B: Below median popularity

	Politics	Pol. Target	Homophobia	Racism	Sexism	Insults
Post-election × Below median popularity	0.014	-0.003	0.032***	-0.008	0.036***	0.000
	(0.013)	(0.013)	(0.010)	(0.011)	(0.010)	(0.009)
Observations	741,369	741,369	741,369	741,369	741,369	741,369
R-squared	0.408	0.281	0.181	0.301	0.202	0.274
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Users	105,012	105,012	105,012	105,012	105,012	105,012

Notes: The table reports coefficients of regressing the share of (target-specific) hate tweets by individual (standardized) on the vote share for Bolsonaro (Panel A) and the municipalities where Bolsonaro's popularity was below median (Panel B). Post-election is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from November 2018 to December 2019. The sample is restricted to users with at least 50 tweets in the entire period. Standard errors clustered at the individual level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.