

Hate speech and social media: evidence from Bolsonaro's election in Brazil

Diego Marino Fages^{*}

Alejandra Agustina Martínez[†]

July 31, 2023

[Click here for the latest version](#)

Abstract

How does newly available information affect individuals' perception of social norms and, consequently, behavior? We examine the impact of Bolsonaro's victory in the 2018 Brazilian presidential election on the prevalence of online hate speech. This project relies on Twitter data from 2017 to 2019 and employs text analysis techniques to classify tweets with or without hate speech content. To causally identify the impact of Bolsonaro's election on hate speech through Twitter, we follow a difference-in-differences approach, using the election result as an information shock. We estimate two difference-in-differences models, the traditional and another with a continuous treatment variable. In the latter, the election result at each municipality measures the local incidence of this information shock. Our findings reveal that online hate speech experienced a surge following the elections, particularly in municipalities where Bolsonaro's popularity was relatively low. These results are further supported by the individual-level analysis, which suggests that both extensive and intensive margins of individual hate speech contributed to the overall increase. We interpret these findings within the framework of a belief updating mechanism, specifically emphasizing the process of revising social norms that govern what is acceptable to say (or not) in public.

Keywords: Hate speech; Social Media; Social Norms

JEL Codes: D72, D83, J15, Z13

^{*}University of Nottingham, United Kingdom.

[†]Universidad Carlos III de Madrid, Spain.

1 Introduction

Social norms are unwritten rules and beliefs governing attitudes and behaviors considered acceptable (or not) in a particular social group or culture. Social norms give us an expected idea of how to behave and function to provide order and predictability in society. Notwithstanding, social norms are not inherently good - examples of harmful social norms are revenge or genital mutilation. Many scholars have shown that social norms tend to be stable over time - Fernandez (2007), Giuliano (2007), Alesina et al. (2013). However, a growing number of studies show how certain events can trigger quick changes in the prevailing social norms (Bursztyn et al., 2020). These events can be very different in nature, ranging from famines to the arrival of new information, e.g., electoral outcomes.

Social norms establish standards on different aspects of life, ranging from contractual relationships to conceptions of right and wrong, reciprocity, and fairness. A relevant type of social norm regards the acceptability of certain speeches, including hate speech. The latter relates to offensive discourse targeting a group or an individual based on inherent characteristics. Naturally, these speeches destroy social cohesion and generate conflict, with consequent repercussions on citizens' lives and well-being.

In this paper, we study the effect of the Brazilian presidential election of Jair Bolsonaro on the prevalence of online hate speech. Bolsonaro, sometimes called "the Trump of the Tropics," is widely recognized for his contentious viewpoints, encompassing homophobia, racism, and sexism.¹ Therefore, following the Bursztyn et al. (2020)'s argument, we posit that Bolsonaro's victory in the 2018 election may trigger a quick update of the prevailing social norm governing what types of speech are socially acceptable.

Identifying a *causal effect* of the election of Bolsonaro on hate speech is not straightforward. Observing a change in the latter could be a cause of the election results, or other elements might affect both events. Our identification strategy relies on considering the 2018 election outcome as an *information shock* - new and potentially unexpected political information. Following Albornoz et al. (2022), we assume this shock varies by municipality, allowing us to identify its marginal effect. Bolsonaro's victory surprised the Brazilian community, which suggests that considering the election outcome as an information shock is a realistic assumption. Precisely, Bolsonaro got 46% of the votes in the 1^o round of the election and 55% in the 2^o round. The opinion polls conducted by diverse companies in the days before the election estimated that

¹To 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, Reuters, AP News, and USA Today. Access date: June 2023.

Bolsonaro's vote share would be approximately 35% for the 1º round, and only one polling company estimated a vote share above 40%.²

To conduct the empirical analysis, we propose two difference-in-differences design models. First, we split municipalities into control and treatment groups according to the vote share received by Bolsonaro in the 1º round of the election. Specifically, any municipality in which Bolsonaro's vote share is lower (higher) than the national outcome, i.e., 46% of the votes, falls into the treatment (control) group. Second, as in Albornoz et al. (2022), we propose a difference-in-differences design with a continuous treatment variable - see Callaway et al. (2021) for a theoretical reference. In this case, the treatment variable is Bolsonaro's vote share in each Brazilian municipality, which measures the local incidence of the information shock, that is, the 1º round election outcome.

Studying hate speech, as opposed to hate crimes, presents some advantages. Since hate speech is directly observable, it is not subject to underreporting. Furthermore, it differs from hate crimes in cost and timing, as the perpetrator immediately pays the cost of expressing hate. Furthermore, social media content has been shown to affect real-life events, such as the propagation of anti-refugee incidents (Müller and Schwarz, 2021). In practice, we rely on data from the widely used social media platform Twitter in the period spanning between July 2017 and December 2019.³ We construct a measure of hate speech using text analysis techniques - see Gentzkow et al. (2019) and Ash and Hansen (2023). In particular, we fine-tune a pre-trained *Bidirectional Encoder Representations from Transformers* (BERT) model, which allows us to classify tweets as with or without hate content. Our classification model was trained using the Portuguese BERT model introduced by Souza et al. (2020) and the hate speech dataset presented by Fortuna et al. (2019).

We document an increase in online hate speech at the national level following the 2018 presidential election. This increase is mainly driven by municipalities where Bolsonaro *lost*. Furthermore, our findings suggest that the magnitude of the information shock, i.e., the election results, is crucial to explaining the extent of the rise in hate expressions. The largest increase in hate speech is observed in municipalities where Bolsonaro was particularly unpopular. We interpret these findings through the lens of a belief update mechanism. Following the information shock triggered by the 2018 election result, individuals living in a relatively against-Bolsonaro municipality could revise their beliefs regarding socially acceptable speeches. Once the social norm is updated, these individuals may feel justified in expressing controversial and hateful viewpoints through social media platforms, even if they reside in a municipality

²Source: Wikipedia, access date: June 2023.

³This time frame covers approximately one year leading up to the electoral rally and another year following the assumption of office by the 38th Brazilian president.

where the pre-election prevalence of such behavior was relatively low.

Since our rich dataset allows us to follow Twitter accounts over time, we further explore *who* is driving the results. We find that both the intensive and extensive margins of hate speech contributed to explaining this phenomenon. In other words, users posting tweets with hate content before the elections increased the frequency of these tweets after the elections (i.e., intensive margin). Similarly, we also observe some Twitter users who post hate speech tweets only after the elections (i.e., extensive margin), especially in the municipalities where Bolsonaro lost.

Related Literature. We contribute to the economic literature that studies the impact of political information on social norms, particularly the literature documenting that political changes can lead to fast changes in social norms and behavior. Bursztyn et al. (2020) show that Trump’s victory increased individuals’ willingness to express xenophobic opinions. Similarly, Albornoz et al. (2022) argue that the Brexit referendum caused a shift in the social acceptability of xenophobic expressions, such as hate crimes. Compared to the previous studies, our setting presents several differences and advantages. First, our results are representative of one of the largest developing countries, while Bursztyn et al. (2020) focus on one metropolitan area in the U.S. and Albornoz et al. (2022) study a very extreme type of hate expression (i.e., hate crimes). Second, the stakes for expressing hate differ in their and our contexts. Twitter users who post hate speech are immediately available for social scrutiny, especially friends, whereas, in Bursztyn et al. (2020), the information is said to be posted at a later date in a likely unknown website and, in Albornoz et al. (2022), perpetrators only pay a cost if they get caught. Third, since we have an (unbalanced) panel of individuals, we can observe the evolution over time at the individual level and explore intensive vs. extensive margins.

In broader terms, our paper speaks to the economic literature on social norms and conformity. We analyze the effects of a social norm update, departing from the literature that studies its persistence (see Bisin and Verdier (2011) for a survey). In addition, our paper contributes to the literature that examines the interplay between norms and political institutions (Acemoglu and Jackson (2017)) or behaviors (e.g., Gerber et al. (2008), DellaVigna et al. (2016), and Perez-Truglia and Cruces (2017)). Finally, our paper connects with the literature on social norms by studying their geographical variation within a country and analyzing high-frequency individual-level data.

This paper speaks to the literature linking social media and expressions of hate, particularly against minority groups.⁴ Müller and Schwarz (2023) find a positive relationship between

⁴In addition to this literature, other research has linked diverse types of traditional media to violence, for

Twitter usage and ethnic hate crimes since the presidential election of Donald Trump in the US, pointing out that social media may enable people with extreme viewpoints to find a source of legitimacy. Bursztyn et al. (2019) show that social media increased ethnic hate crimes in Russian cities with high pre-existing anti-immigrant sentiments. Müller and Schwarz (2021) find evidence that social media affects the propagation of anti-refugee incidents in Germany. Focusing on sex crime, Bhuller et al. (2013) document an increase in this type of crime associated with the roll-out of broadband internet in Norway. This piece of literature covers social media platforms like Twitter and Facebook but focuses mainly on xenophobia and ethnic hate crimes. This paper considers a wider definition of expressions of hate, encompassing its different targets. In contrast to the existing literature, this paper focuses on hate speech rather than hate crime and online rather offline expressions of hate.

The rest of the paper is organized as follows. Section 2 describes the data. Section 3 presents the identification strategy, and section 4, the results at the municipality and individual levels. Section 5 concludes.

2 Data

In this paper, we aim to understand how the 2018 presidential election of Bolsonaro affected online hate speech in Brazil. Our primary data source is the social media platform Twitter, from which we measure online hate speech at the municipality level and in the period under study. We combine the data we retrieve from Twitter with three types of administrative data. First, we use the 2018 election results at the municipality level, whose data source is the Superior Tribunal Court (in Portuguese, *Tribunal Superior Eleitoral - TSE*), the highest structure within the Brazilian Electoral Justice system. In addition, we rely on geospatial data from the Brazilian Institute of Geography and Statistics (in Portuguese, *Instituto Brasileiro de Geografia e Estatística - IBGE*) to geo-locate tweets and election results. Lastly, we use the 2010 Population Census in Brazil microdata from IBGE to construct demographic variables aggregated at the municipality level.

2.1 Twitter data

Twitter is an online platform that allows users to publish short messages, of a maximum of 140 characters, on their profiles. With one of the largest Twitter user bases in the world, Brazil

example, Dahl and DellaVigna (2009), Card and Dahl (2011), Yanagizawa-Drott (2014), DellaVigna et al. (2014), and Ivandic et al. (2019), among others.

is an appealing case of study for online activity - in this case, related to Twitter users' speech. In January 2022, Brazil ranked fourth worldwide regarding the number of Twitter users, with an estimated 19 million active accounts (after the United States, Japan, and India).⁵ Given our purposes, it is important to note that most of the Brazilians who were online in 2022 used social media for news (64%) and political discussion (78%).⁶

Another advantage of this setting is that online hate speech, as opposed to hate crime, can be directly observed and quantified and, thus, is not subject to changes in reporting. Online hate speech also differs from hate crime regarding its cost and timing. In the former, the perpetrator immediately pays the cost of expressing hateful content. On the other hand, a hate crime must be reported and processed by justice before the perpetrator pays its cost.

In the empirical analysis, our main variable of interest is the proportion of tweets classified as hate speech per municipality (or individual) and date. The next paragraphs describe how we collected and processed Twitter data to construct this variable.

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 the tweets, net of retweets, which satisfy three conditions specified in the Twitter query. First, tweets must be written in Portuguese. Second, tweets must provide geo-location information and be located in Brazil. Lastly, tweets must belong to the period comprised between July 2017 to December 2019, both included. As the daily amount of data retrieved by this query is around 300.000 tweets, we further restrict the Twitter query to retrieve only tweets posted on any Monday belonging to the mentioned period. This query imposes two main assumptions on our tweets' sample. We assume that (i) tweets posted on any Monday and (ii) geo-located tweets are representative samples of the tweets' universe. Appendix A.1 provides supportive evidence for these assumptions and complementary information to this section.

Data processing. We extract relevant content from the tweets' text, which will serve as input for the hate speech detection task. We exclude punctuation marks, stop-words, and multimedia items. We do not remove negative stop-words that may change the statement's meaning: "mas" (but), "nem" (neither), "não" (no), "sem" (without), and "fora" (out). We anonymize user mentions and URL links but keep hashtags in their native Twitter format, as they may contain

⁵Source: Statista web portal, access date: June 2023.

⁶Sources: Digital News Report, 2022, Reuters Institute & University of Oxford, access date: June 2023; Statista web portal, access date: June 2023.

relevant information. We drop all tweets containing only links and/or user mentions and those posted by accounts created after 2018. The reason for the latter is to exclude from the analysis accounts potentially created in the context of the electoral rally.

Hate speech detection. We rely on Natural Language Processing (NLP) techniques to detect hate speech in our tweets' sample. We train a pre-trained *Bidirectional Encoder Representations from Transformers* (BERT) model (Devlin et al., 2018) on a dataset specific to the hate speech detection task. This process is known as *fine-tuning* a pre-trained model. Specifically, we use *BERTimbau*, a BERT model for Brazilian Portuguese by Souza et al. (2020), and train it on a dataset of tweets in Portuguese, by Fortuna et al. (2019).

Souza et al. (2020) present *BERTimbau*, a BERT model for Brazilian Portuguese, in two sizes, Base and Large. In this paper, we fine-tune BERTimbau-Base for the hate speech detection task. Its architecture comprises 12 layers, 768 hidden dimensions, 12 attention heads, and 110M parameters. The authors trained their model on the *brWaC corpus* by Wagner Filho et al. (2018) and two NLP tasks, Masked Language Modeling (MLM) and Sentence Prediction (NSP).

In their paper, Fortuna et al. (2019) collected 5668 tweets in Portuguese through Twitter API from January to March 2017. The authors provide two annotation schemes for the dataset, a binary, and a hierarchical multiple classifications. Each tweet classified as "hate speech" is further split into classes for the hierarchical classification. Its second-level classification relates to the target of hate, and it comprises: "sexism," "body," "origin," "homophobia," "ideology," "religion," "health," and "lifestyle." As a first step, this paper uses the binary classification dataset to fine-tune the mentioned BERT model, in which 31.5% of the tweets were annotated as "hate speech." To construct it, three (Portuguese native) annotators labeled every tweet as "hate speech" or "not hate speech," and the authors applied the majority vote to determine the final annotation of each tweet.

Before fine-tuning, we divide the dataset between 80% for training, 10% for validation, and 10% for testing. In NLP applications, the performance of a model in a given task is directly influenced by the characteristics of the training sample. In the case of Fortuna et al. (2019)'s dataset, as in other datasets on hate speech detection, a class imbalance exists. Tweets annotated as "hate speech" constitutes the minority class. As class imbalance may affect a model's performance in a text classification task, we use a Random Oversampling technique to equalize the number of tweets per class in the training sample (Mohammed et al., 2020).⁷ Our model attains an overall accuracy of 77% in both the validation and test samples. Appendix A.2

⁷Random oversampling involves transforming the existing data to adjust the class distribution. It consists of randomly selecting examples from the minority class and adding them to the original dataset.

provides further details on the hate speech detection task, resources utilized, and the model's training results.

Data classification. After training the BERT model for the hate speech detection task, we use it to detect hate speech in the tweets we collected. We construct a binary variable 0/1, named "predicted hate speech," as a result of this classification. Then, we use the tweet-specific geo-location information to map each tweet to the Brazilian municipalities based on latitude and longitude through IBGE's geospatial shape files. Finally, we compute the proportion of tweets containing hate speech by municipalities (or individuals) over time, which is the main outcome variable of this paper.

2.2 Administrative data

The election result we use as an information shock is the vote share obtained by Bolsonaro in the 1º round of the 2018 Brazilian presidential election. The Superior Tribunal Court (TSE) provides official data at the municipality level on all election results in Brazil since 1994. Given that TSE's records do not contain the geo-coordinates of the electoral districts, we rely on geospatial data from the Brazilian Institute of Geography and Statistics (IBGE) to determine their location. IBGE provides Brazilian geospatial data at country, state, and municipality levels. Lastly, we use microdata from the 2010 Population Census in Brazil, the last available for the pre-Bolsonaro period. Consistently with our analysis unit, we aggregate the census microdata at the municipality level.

2.3 Descriptive statistics

We study how the 2018 Brazilian presidential election influenced online hate speech. To accomplish this, we create two longitudinal datasets of geo-located tweets spanning from July 2017 to December 2019.

In the first dataset, the time unit is a day t - for any Monday included in the tweets' sample - and the cross-sectional unit is a Brazilian municipality m . The main variable is the proportion of tweets classified as hate speech for a given date t and municipality m . Brazil is divided into twenty-six states and one federal district. Each sub-national entity is further divided into municipalities, and Brazil currently has 5570 municipalities. For the empirical analysis, we include any municipality for which we observe (i) at least 10 tweets daily and (ii) at least 10 times during 2017-2019. This leaves us with approximately 1500 municipalities. The

longitudinal dataset at the municipality level is unbalanced, with some municipalities present over the entire period and others for which Twitter data is relatively more scarce. On average, we observe each municipality on approximately 100 Mondays (with a standard deviation of 37 days).

In the second longitudinal dataset, the time unit is a month t , and the cross-sectional unit is a Twitter user i . We include any Twitter user whose tweets are geo-located in no more than two different municipalities. When an individual appears in two locations, we implicitly assume she is engaged in an activity (such as working or studying) in a municipality different from where she lives. The longitudinal dataset at the individual level is also unbalanced, as Twitter activity significantly varies for different individuals. In the regression analysis, we further restrict our attention to the sub-sample of users (i) who posted tweets in the pre and post-election periods and (ii) such that we observe at least 5 tweets per user per month. On average, we observe 190 tweets for each Twitter user distributed over approximately 12 months (6 months before and 6 months after elections).

This paper builds upon two fundamental observations. Firstly, the presidential election, which we consider an information shock, did not uniformly affect all Brazilian citizens. Instead, we observe a geographical variation in Bolsonaro's vote share, which helps us to identify the effect of interest. Secondly, the evolution of online hate speech was not consistently constant throughout the period.

Regarding the first observation, Figure 2 shows that Bolsonaro's popularity varied across states and municipalities. Specifically, Bolsonaro's vote share was between 3% and 79% in the 1º round of the 2018 presidential election, which is the result we use in our empirical strategy to measure the information shock. As can be seen, the corresponding map for the 2º round results shows a similar geographical pattern. In Appendix A.3, we present the (bimodal) distributions of these vote shares at the municipality level.

As for the second observation, Figure 1 shows the proportion of Brazilian tweets classified as hate speech in the period under study. The solid line corresponds to the raw data, consisting of the daily proportion of hate speech tweets, whereas the dotted line corresponds to the monthly average of hate speech. The shadow areas in the graph delimit the periods in which (i) the Presidential Election took place and (ii) Bolsonaro took office.⁸ As can be seen, there was a sharp increase in hate speech during this period. The hate speech peaks on the data correspond to the closest (but later on time) date in our sample to the first and 2º rounds of

⁸Specifically, the first and 2º rounds of the presidential election took place on October 7th and 28th, respectively. Bolsonaro took office as Brazil's 38th president on January 1st, 2019.

the election.⁹

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



Note: The variable Hate speech (in percent) is, for each date, the ratio of tweets classified as hate speech over the total amount of tweets.

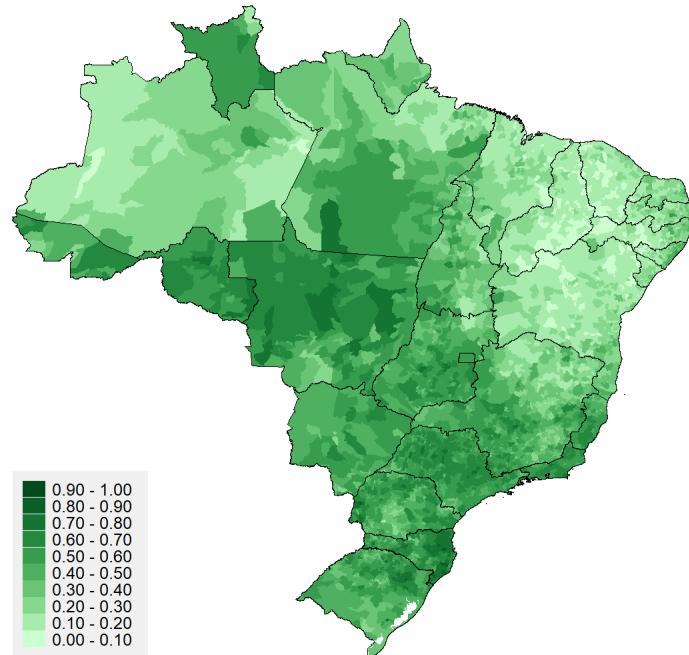
Importantly, Figure 1 reveals that hate speech through Twitter increased post-election. The average proportion of hate speech from July 2017 to July 2018 was 8%, whereas it was 9% from January to December 2019.¹⁰ Note that the above figure is constructed by aggregating hate speech at the national level, so it does not explore the sub-national evolution of hate speech over the period. The rest of this paper aims to answer whether this evolution was uniform (or not) across municipalities and why.

⁹There exist two other (although smaller) peaks in the data, during June and July 2018, corresponding to dates when Brazil's football team played a match in the 2018 World Cup. Figure 7 in Appendix shows that these peaks also correspond to a sharp increase in Twitter activity. Specifically, the daily amount of tweets is around a 50% higher during the period relative to the average. It is also worth noticing that the period with lower levels of hate speech corresponds to dates around the 2018 New Year break. Remarkably, this sharp decrease in hate speech was not observed around the 2019 New Year break, as the date coincides with when Bolsonaro took office.

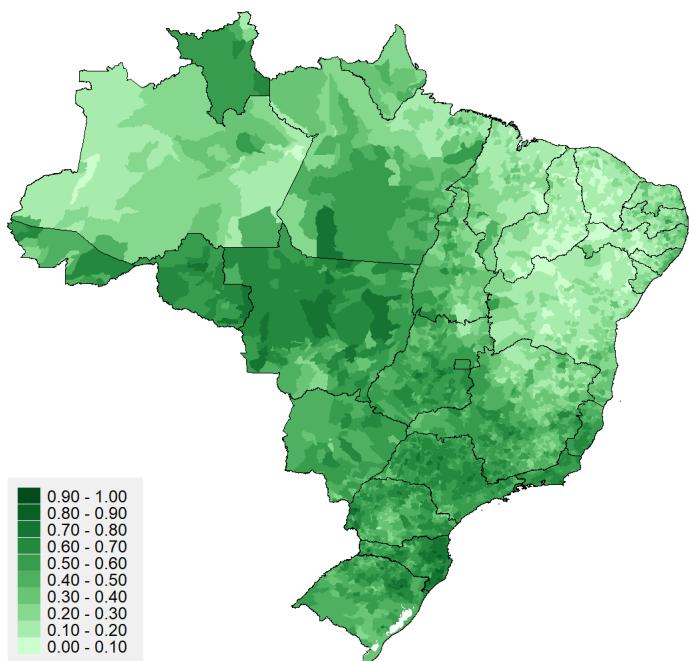
¹⁰Figure 11 in Appendix A.3 supports this observation. The mentioned figure is analogous to the one presented in the main text but with a standardized variable. As can be seen, almost all data points are below zero in the pre-election period and above zero in the post-election period.

Figure 2: Bolsonaro's vote share at the municipality level.

a. 1º Round, October 7th.



b. 2º Round, October 28th.



3 Empirical strategy

We aim to estimate the effect of Bolsonaro's election - and the electoral rally - on hate speech. In the previous section (see Figure 1), we showed that hate speech on Twitter increased after Bolsonaro's election in comparison to the pre-election period at the national level. However, this is not sufficient to conclude that his election is to blame. It is possible that the election result responded to the rise in hate speech or that some other social phenomena are causing both the increase in hate speech and the political movement to the right.

The fact that these are national elections leaves us with no clear control group where Bolsonaro is not elected for president. However, his popularity varies across states and municipalities (see Figure 2). We can then follow Albornoz et al. (2022) and exploit this differential informational shock to study whether hate speech increased relatively more in some places than others. First, we separate the municipalities based on the results of the 1º round of the elections: those where Bolsonaro got *at least* or *at most* the percentage of votes he got at the national level, 46%. For the sake of simplicity, we say that Bolsonaro "lost" the 1º round of elections (or simply, lost) in a municipality if his vote share was lower than 46%. Otherwise, we say that Bolsonaro "won" the election in that municipality. Thus, we perform a difference-in-differences analysis. Formally, we regress,

$$Hate_{mt} = \alpha_0 + \alpha_1 * Post_t * Lost_m + \delta_t + \pi_m + \epsilon_{mt} \quad (1)$$

where $Hate_{mt}$ is the share of tweets that contain hate speech in municipality m and date t , $Post_t$ is a dummy variable that takes the value one after the elections, $Lost_m$ is a dummy variable that takes the value one for the municipalities where Bolsonaro lost the elections (that is, his vote share was lower than 46%), δ_t is a linear time trend, π_m are municipality fixed effects, and ϵ_{mt} is a municipality-time specific error term. In this case, the identifying assumption is the traditional parallel trends assumption. That is, in the absence of the information shock, the difference in hate speech between municipalities where Bolsonaro won and lost the elections is constant over time.

Since our rich dataset allows us to follow Twitter accounts over time, we can further analyze hate speech at the individual level. Indeed, the availability of data at the individual level is an advantage of this paper, compared to Albornoz et al. (2022) and Carr et al. (2020), who studied hate crime at a more aggregate level. The purpose of the individual-level regressions is twofold. Firstly, it allows us to rule out the possibility that the rise in hate speech is driven by a change in the composition of the users before and after the elections. Secondly, individual data

allows us to explore the intensive and extensive margins of hate speech. In other words, is the increase in hate speech driven by people already tweeting hate content before the elections, i.e., intensive margin, or is it caused by people who had not tweeted hate content before, i.e., extensive margin? Formally, we regress,

$$Hate_{imt} = \alpha_0 + \alpha_1 * Post_t * Lost_{im} + \delta_t + \gamma_i + \epsilon_{imt} \quad (2)$$

where $Hate_{imt}$ is the share of tweets that contain hate speech of account i in municipality m at month t , $Post_t$ is a dummy variable that takes the value one after the elections, $Lost_{im}$ is a dummy variable that takes the value one for the accounts located in municipalities where Bolsonaro lost the elections, δ_t is a linear time trend, γ_i are user fixed effects, and ϵ_{imt} is an account-municipality-time specific error term. In both cases, our coefficient of interest is α_1 , which, given parallel trends, captures the *average treatment effect* (ATE).

Finally, we also exploit the continuous variation in Bolsonaro's vote share across municipalities. To do this, we replace $Lost_m$ in equation (1) and $Lost_{im}$ in equation (2) with the actual vote share Bolsonaro received in each municipality, $VoteShare_m$ and $VoteShare_{im}$. Formally,

$$Hate_{mt} = \beta_0 + \beta_1 * Post_t * VoteShare_m + \delta_t + \pi_m + \epsilon_{mt} \quad (3)$$

and,

$$Hate_{imt} = \beta_0 + \beta_1 * Post_t * VoteShare_{im} + \delta_t + \gamma_i + \epsilon_{imt} \quad (4)$$

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

In both cases, our coefficient of interest is β_1 . It captures the *average causal response* (ACR) on the treated to an incremental change in the dose, where the dose is the share of votes obtained by Bolsonaro in the municipality. The main identification assumption, in this case, is the strong parallel trends. 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 - see Callaway et al. (2021).¹¹ Notice that, by definition, α_1 in equations (1) and (2) and β_1 in equations (3) and (4) have opposite signs: while the former captures the effect of $Lost_m = 1$, which depends negatively on Bolsonaro's vote share, the latter is proportional to it.

4 Results

In this section, we present the main results of this paper. First, we document that hate speech increased after the 2018 presidential elections, especially in the municipalities where Bolsonaro lost. Then, we offer the results at the individual level, indicating that both the intensive and extensive margins of hate speech contributed to this explain this phenomenon.

4.1 Municipality level

Before presenting the regression results, let us describe the municipalities that are in the treatment and control groups according to equations (1) and (2). Figure 3 below is an analogous figure to Figure 1, but now splitting the hate speech trends between treatment and control groups.¹²

The green lines correspond to the daily and monthly proportion of Brazilian tweets classified as hate speech for the municipalities in which Bolsonaro got at least 46% of the votes in the 1^o round of the 2018 presidential election, i.e., where $Lost_m = 0$. On the contrary, the red lines correspond to the municipalities where Bolsonaro's vote share was at most 46%, that is, where $Lost_m = 1$. Again, the shadow areas in the graph delimit the periods in which the presidential election took place and Bolsonaro took office.

Importantly for our identification strategy, the gap between hate speech pre-trends for treatment and control groups seems constant over time, i.e., pre-trends are parallel.¹³ Furthermore, ratios of hate speech in municipalities where Bolsonaro won and lost seem to respond similarly to shocks; for example, both decreased around the 2018 New Year's Eve and increased during

¹¹Formally, 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 : $E[Y_t(d) - Y_{t-1}(0)] = E[Y_t(d) - Y_{t-1}(0)|D = d]$.

¹²In Appendix A.3, we present analogous graphs to Figures 1 and 3 but with hate speech aggregated at the monthly level. That is, only plotting the dotted lines in Figures 1 and 3.

¹³Figure 11 in Appendix A.3 supports this observation. The mentioned figure is analogous to the one presented in the main text but with standardized variables. Once differences in levels are canceled, it is easy to see that the two lines move together.

the 2018 World Cup (in July). Nonetheless, the hate speech trends in municipalities where Bolsonaro won and lost changed after the election. As can be seen, the gap between hate speech ratios gets smaller, especially after Bolsonaro took office as the 38th Brazilian president. This evidence suggests a different reaction to Bolsonaro's election in one and another region.

Figure 3: Evolution of hate speech in Brazilian tweets, 2017-2019. Municipalities, by the 2018 election result.

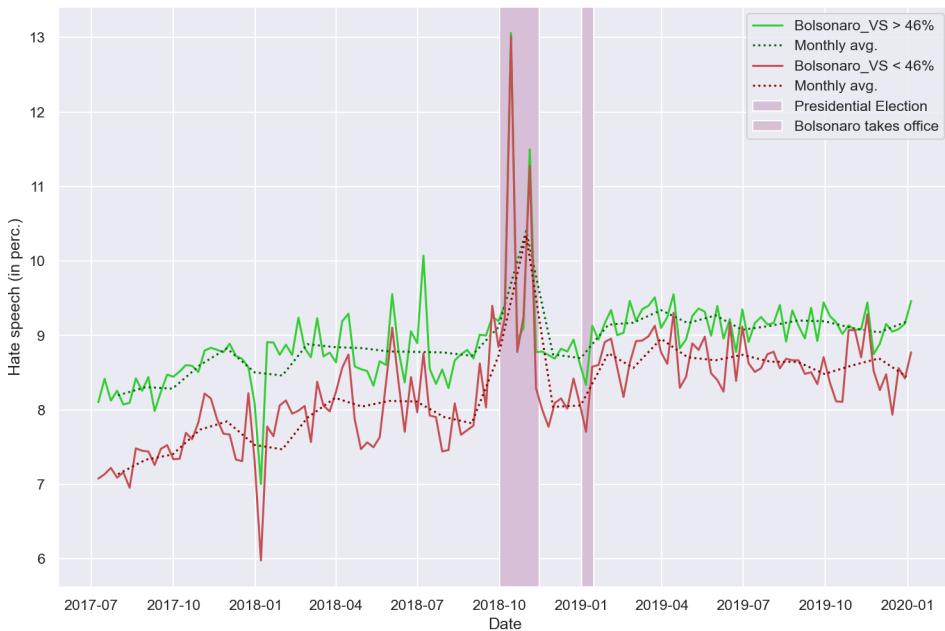


Table 1 presents descriptive statistics for the municipalities that fall into the treatment and control groups.¹⁴ All demographic variables were extracted from the 2010 Brazilian Population Census. The last three variables are the vote share obtained by the Workers' Party (Partido dos Trabalhadores, PT) in the 1º round of the 2006, 2010, and 2014 Presidential Elections, respectively. In 2018, Bolsonaro defeated a candidate affiliated with the Workers' Party, which explains the negative correlation in votes.

As can be seen, demographic characteristics vary for municipalities in the control and treatment groups, but importantly, the variation is relatively small. For example, regarding the availability of the internet at home, closely related to the presence in social media and Twitter, this difference was 0.04p.p. in 2010. The variable that varies the most is the income per capita. However, as we include Municipality FE in our regressions, these differences are not a threat

¹⁴In Appendix A.3, we present an analogous table for all municipalities. The results in one and other tables do not vary substantially.

to identification as long as they are constant over time.

Table 1: Descriptive Statistics, by the 2018 election result.

Variable	Won_m	Lost_m	Difference
urban	0.828	0.738	0.089***
income_pc	755.3	478.3	277.0***
cellphone	0.891	0.810	0.081***
computer	0.427	0.241	0.186***
internet	0.735	0.693	0.042***
primary	0.382	0.368	0.013***
tertiary	0.435	0.474	-0.039***
no_religion	0.058	0.070	-0.012***
catholic	0.678	0.721	-0.043***
pentecostal	0.128	0.117	0.010***
black	0.052	0.082	-0.029***
indigenous	0.003	0.006	-0.003**
brown	0.286	0.489	-0.202***
born_mun	0.567	0.673	-0.106***
born_state	0.673	0.747	-0.075***
vs_pt_2006	0.351	0.528	-0.177***
vs_pt_2010	0.378	0.524	-0.146***
vs_pt_2014	0.310	0.507	-0.197***

Note: N = 1482 (municipalities for which Twitter data is available, after data cleaning). All variables are aggregated at the municipality level. Column "Lost_m" refers to the municipalities where Bolsonaro lost the 2018 election, whereas column "Won_m" refers to where he won. The third column reports the statistical difference between the respective means. Variables "cellphone," "computer," and "internet" are the proportion of households reported to have such goods in the 2010 Population Census. Variables "no_religion," "catholic," "pentecostal," "black," "indigenous," and "brown" are the proportion of individuals registered to have such demographic characteristics in the 2010 Population Census. Variables "primary" and "tertiary" refers to the population with (at most) primary and tertiary education. Variables "bornhere_mun" and "bornhere_state" refer to the proportion of individuals born in the municipality and state where they answered the 2010 Population Census. Variables "vs_pt_2006," "vs_pt_2010," and "vs_pt_2014" are the proportion of votes obtained by the Workers' Party (Partido dos Trabalhadores, PT) in the 1º round of the 2006, 2010, and 2014 Presidential Elections, respectively.

Let us turn to the regression results. Table 2 answers this paper's question, *how the 2018 presidential election of Bolsonaro affected online hate speech*. The first column in the table corresponds to the classic difference-in-differences estimation, presented in equation (1). The second column corresponds to the difference-in-differences model with a continuous treatment variable, i.e., equation (3). In the two models, we define $Post_t$ as a dummy variable taking a

value of one between July 2017 and July 2018 (both included); and a value of zero between January and December 2019 (both included). We drop the period from the election rally to when Bolsonaro took office, as hate speech may behave differently than in regular times. However, in Appendix A.3, we show our results are robust to changes in the definition of $Post_t$.

Column (1) shows the increase in hate speech after the elections that we observe in Figures 1 and 3 was more pronounced in municipalities where Bolsonaro lost (0.4 p.p. higher). Consistent with this evidence, column (2) shows that the proportion of hate speech decreases as the share of votes for Bolsonaro increases. As the estimate in column (2) comes from a difference-in-differences model with a continuous treatment variable, provided the strong parallel trends assumption, the coefficient is a positively weighted average of the average causal response $ACR(d)$ parameters across doses. Thus, on average, across doses, an increase of 1 p.p. in $VoteShare_m$ decreases hate speech in that municipality by 0.01 p.p.

Table 2: Municipality level regressions.

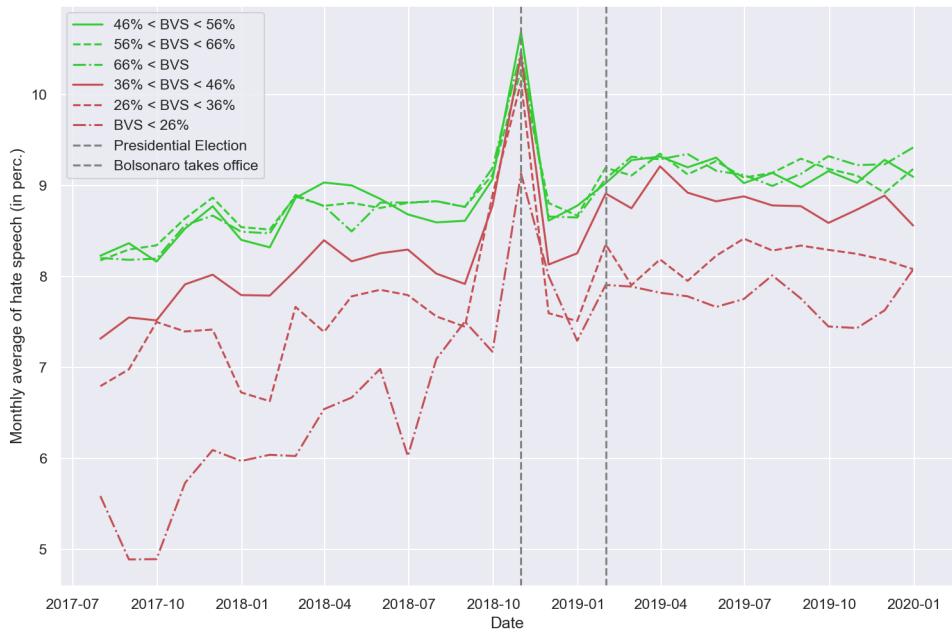
Variables	(1) $Hate_{mt}$	(2) $Hate_{mt}$
$Post_t \times Lost_m$	0.004*** (0.001)	
$Post_t \times VoteShare_m$		-0.010*** (0.003)
Constant	0.084*** (0.000)	0.087*** (0.001)
Municipality FE	Yes	Yes
Date FE	Yes	Yes
Municipalities	1,482	1,482
Observations	89,865	89,865
R-squared	0.074	0.073

Note: Standard errors in parentheses. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from January to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Focusing on the sign of the estimates in columns (1) and (2), we interpret the results in lines of a beliefs' update mechanism. After receiving the information shock, i.e., the 2018 election result, people could update their beliefs about what type of speeches are socially acceptable. The difference in the election result at the municipality and national levels measures the extent

of this update in beliefs. Thus, it is natural to observe that the change in online behavior comes from the individuals who misperceived the social norm before, i.e., those who live in a municipality where Bolsonaro lost. After updating the social norm, they may feel entitled to generate hate speech, even if they live in a municipality with an ex-ante lower level of hate speech.

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



Note: Hate speech trends are constructed separately for each group of municipalities. "BVS" stands for Bolsonaro's vote share in the 1^o round of the 2018 election.

Figure 4 provides further evidence of this mechanism.¹⁵ This figure is analogous to Figure 3, but now splitting hate speech trends between different treatment intensities. For illustrative purposes, hate speech ratios are aggregated at the monthly level. The green lines are the hate speech ratios for the municipalities in which $Lost_m = 0$. The solid, dashed, and dash-dotted lines correspond to municipalities where Bolsonaro got between 46% and 56%, 56% and 66%, and more than 66%, respectively, of the votes in the 1^o round of the 2018 presidential election. On the contrary, the red lines are the monthly proportion of Brazilian tweets classified as hate

¹⁵An important technical comment on this graph is that our data is unbalanced regarding municipalities where Bolsonaro won and lost, so dashed and, especially, dash-dotted red lines are drawn with a relatively low number of data points. This fact leads us to take this figure's interpretation with some caution.

speech for the municipalities in which $Lost_m = 1$. The solid, dashed, and dash-dotted lines correspond to municipalities where Bolsonaro got between 36% and 46%, 26% and 36%, and less than 26% of the votes, respectively.

There are two relevant observations to this figure. On the one hand, hate speech trends in municipalities where Bolsonaro was popular are similar and relatively stable over time. If individuals living in such municipalities perceived the social norm more accurately even before the elections, their resulting behavior change after elections becomes smaller. On the other hand, hate speech trends in municipalities where Bolsonaro was unpopular were very different in the pre-election period. All the hate speech trends have a positive slope, which is negatively correlated with Bolsonaro's popularity. This negative correlation suggests that the size of the information shock is relevant to explain the extent to which people change their behavior.

4.2 Individual level

In the previous section, we have shown that the proportion of online hate speech increased after the 2018 presidential election. At the municipality level, this increase is mainly driven by regions where Bolsonaro *lost* the election. As our Twitter data is at the individual level, we can further extend our main analysis and explore *who* is driving this result. In particular, this increase may be driven by (i) users already posting tweets with hate content before the elections, i.e., intensive margin, (ii) users who start posting hate speech tweets after the elections, i.e., extensive margin, or (iii) both.

Throughout this section, we focus on a sub-sample of Twitter users whose tweets are located in *no more than two* different municipalities. For the regression analysis, we restrict our attention to the sub-sample of Twitter users (i) who appear at least one month before and one month after the election and (ii) such that we observe at least 5 tweets per user per month. When a user's tweets are located in multiple municipalities, we assume the information shock she received is a *weighted average* of Bolsonaro's vote share in the corresponding locations.

Figure 5 shows that the rise in hate speech results from both the intensive and extensive margins. This figure is constructed using all the Twitter accounts in our sample (after restricting it to the upper bound on the number of locations). Panel a shows how the share of Twitter accounts posting zero hate content becomes smaller after the elections. Specifically, 55% of the Twitter users in our sample had never published hate speech content before the 2018 elections, and this number reduced to 53% after Bolsonaro was elected president. This reduction is stronger for the sub-sample of Twitter users who post tweets from a municipality where $Lost_m = 1$ - the corresponding percentages are 51.4% in the pre-election period and 50.1% in

the post-election period.

Panel b focuses on the intensive margin by zooming in on those Twitter accounts that have posted messages with hate speech at least once. We can see that the distribution of individual hate speech has shifted to the right after the elections. On average, the intensity of hate speech increased by 0.9 p.p. in the post-election period (from 15.9% to 16.7% in the sub-sample of users who posted hate speech content).

This observation is further confirmed by the estimates in Table 3, corresponding to the difference-in-differences models at the individual level, equations (3) and (4). We focus on the intensive margin of hate speech, dropping the user accounts that did not publish hateful content during the period under study. Specifically, we delete all Twitter users for whom the proportion of tweets classified as containing hate speech over the period is lower than 5%. Although we lost power in estimating the election's impact on individual hate speech, especially for the continuous treatment differences-in-differences model, the estimates are comparable in sign and magnitude to those previously presented in Table 2. In Appendix A.3, we present supplementary regressions, redefining the intensive margin of hate speech and restricting the sub-sample of users according to their online activity.

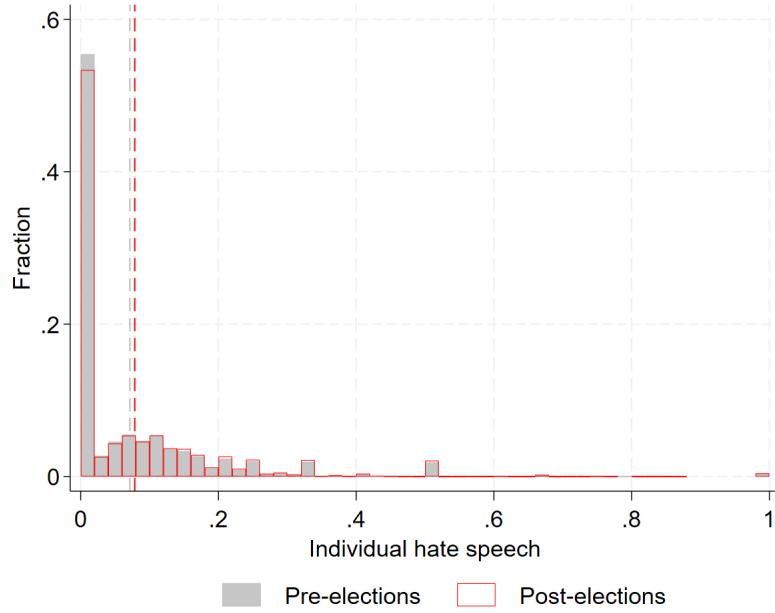
Table 3: Intensive margin of hate speech. Individual regressions.

Variables	(1) <i>Hate_{imt}</i>	(2) <i>Hate_{imt}</i>
<i>Post_t</i> X <i>Lost_{im}</i>	0.003** (0.001)	
<i>Post_t</i> X <i>VoteShare_{im}</i>		-0.009 (0.006)
Constant	0.108*** (0.000)	0.116*** (0.005)
Individual FE	Yes	Yes
Month FE	Yes	Yes
Individuals	85,494	85,494
Observations	418,616	418,616
R-squared	0.257	0.257

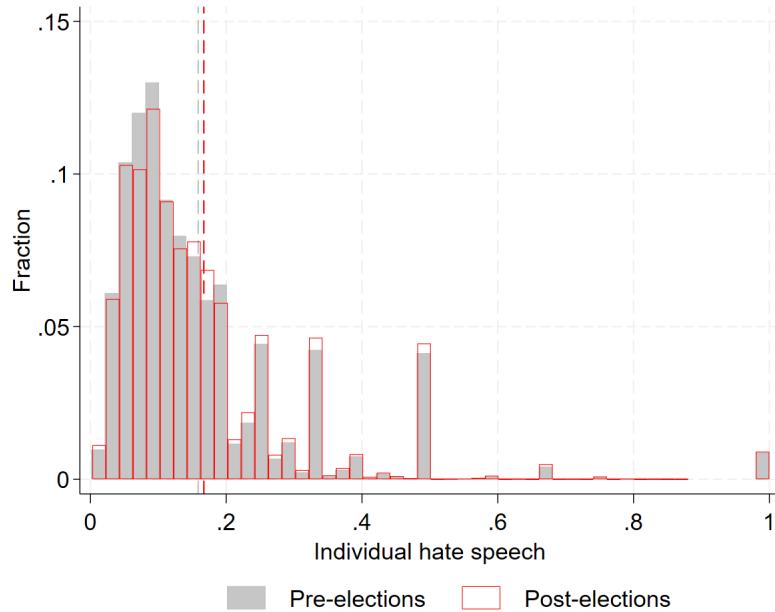
Standard errors in parentheses. Only Twitter users for whom the proportion of tweets classified as containing hate speech over the period is greater than 5%. *Post_t* is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from January to December 2019. *Lost_m* is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Figure 5: Individual hate speech, pre- and post-elections.

a. Extensive margin



b. Intensive margin



Note: Histograms at the individual level. The pre-election period is between July 2017 and July 2018, whereas the post-election period goes from January to December 2019. Panel a: all the Twitter accounts in the sample. Panel b: only Twitter accounts that posted hate speech content at least once. The vertical lines indicate the average hate speech of the accounts before and after the elections.

4.3 Robustness checks

In this section, we check the robustness of the results by relaxing the assumptions we made throughout the paper. For the regressions at the municipality level, we change the variables' definitions and the period under study, among others. For the individual-level regressions, we present results for all Twitter users in the sample, redefine the intensive margin of hate speech, and restrict the sub-sample of users according to their online activity. Appendix A.3 presents the corresponding results, showing that the main results of this paper remain qualitatively unchanged.

5 Conclusion

As social media platforms have proliferated, a new public sphere where individuals share ideas has emerged. Among them are the ones related to hate speech, offensive language, and discrimination. Understanding what factors impact the online spread of these harmful speeches is crucial for modern societies, especially regarding social media content moderation. In this line, we provide novel evidence on how political outcomes impact online expressions of hate.

We document that the 2018 election of Bolsonaro in Brazil, a far-right candidate, increased online hate speech. Interestingly, this impact is more pronounced in regions where Bolsonaro was relatively less popular - according to the regression results at the municipality and individual levels. Then, we propose a beliefs update regarding the social acceptability of hate speech as an underlying mechanism.

There are at least three natural extensions of this project, which are left for future research. Firstly, to go deeply into the underlying mechanism proposed in this paper, we can differentiate types of hate speech and analyze their comparative patterns. For that purpose, we plan to develop a machine-learning model to differentiate hate content by its *target*. In the context of this paper, we are particularly interested in the trajectories of hate speech labeled as "homophobia," "racism," and "sexism."

Secondly, an extension of this paper will look at the persistence of information shocks that potentially trigger both a social norms update and the spread of harmful expressions. In our study case, we did not extend the period under analysis as it would require going over the Covid19 pandemic - a completely different type of shock. Lastly, comparing on and offline expressions of hate, especially analyzing their interdependency, is a policy-relevant question;

closely related to this paper.

References

- Acemoglu, D. and Jackson, M. O. (2017). Social norms and the enforcement of laws. *Journal of the European Economic Association*, 15(2):245–295.
- 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.
- Ash, E. and Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, 15.
- 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.
- Bisin, A. and Verdier, T. (2011). The economics of cultural transmission and socialization. In *Handbook of social economics*, volume 1, pages 339–416. Elsevier.
- 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. (2020). From extreme to mainstream: The erosion of social norms. *American economic review*, 110(11):3522–3548.
- Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. (2021). Difference-in-differences with a continuous treatment. *arXiv preprint arXiv:2107.02637*.
- 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.

- DellaVigna, S., List, J. A., Malmendier, U., and Rao, G. (2016). Voting to tell others. *The Review of Economic Studies*, 84(1):143–181.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- 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.
- Gerber, A. S., Green, D. P., and Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American political Science review*, 102(1):33–48.
- 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. J. (2019). Jihadi attacks, media and local hate crime.
- 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.
- 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.
- Perez-Truglia, R. and Cruces, G. (2017). Partisan interactions: Evidence from a field experiment in the united states. *Journal of Political Economy*, 125(4):1208–1243.
- 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.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: a new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

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

A Appendix

A.1 Twitter data

Twitter is an online platform that allows users to publish short messages, of a maximum of 140 characters, on their profiles. In January 2021, Twitter launched an Academic Research product track, which enables researchers to access all v2 endpoints. Notably, the *Twitter Search API v2* gives 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 from June 2022 to May 2023. For more information about the academic track on Twitter, follow this link.

The Twitter query we create to download tweets restricts our search to all publicly available (yet undeleted) tweets written in Portuguese, geo-located in Brazil, that are not retweets, and belong to any Monday between July 2017 to December 2019, both included. This query imposes two main assumptions on our tweets' sample. We assume the sample of (i) tweets posted on any Monday and (ii) geo-located tweets are representative samples of the tweets' universe. Figures below present supportive evidence for these assumptions.

Figure 6 presents the average number of tweets per day of the week for the period under study. The figure shows that the amount of tweets is quite stable over the weekdays and slightly decreases on weekends. The daily average of tweets is around 305.000. Figure 7 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.¹⁶ It can be seen that the monthly trend in Figure 7 exhibits higher variation than the average number of tweets per weekday in Figure 6. Altogether suggest that (i) long-run variation on tweets is larger than short-run variation and (ii) data for one day per week for the whole period correctly captures how data behaves.

Lastly, Figures 8 and 9 compare the trends of geo-located tweets and the universe of tweets that contain a specific word. In all the sub-graphs of the two figures, the red line corresponds to the amount of geo-located tweets, and the blue line is the amount of all tweets multiplied by a *scalability factor*. This factor is the ratio of geo-located tweets over total tweets in the sample for each word, which is between 4% and 8%.

¹⁶Peaks during June/July 2018 corresponds to dates when Brazil's football team played a match in the 2018 World Cup.

Figure 6: Average number of tweets per day in the tweets' sample.

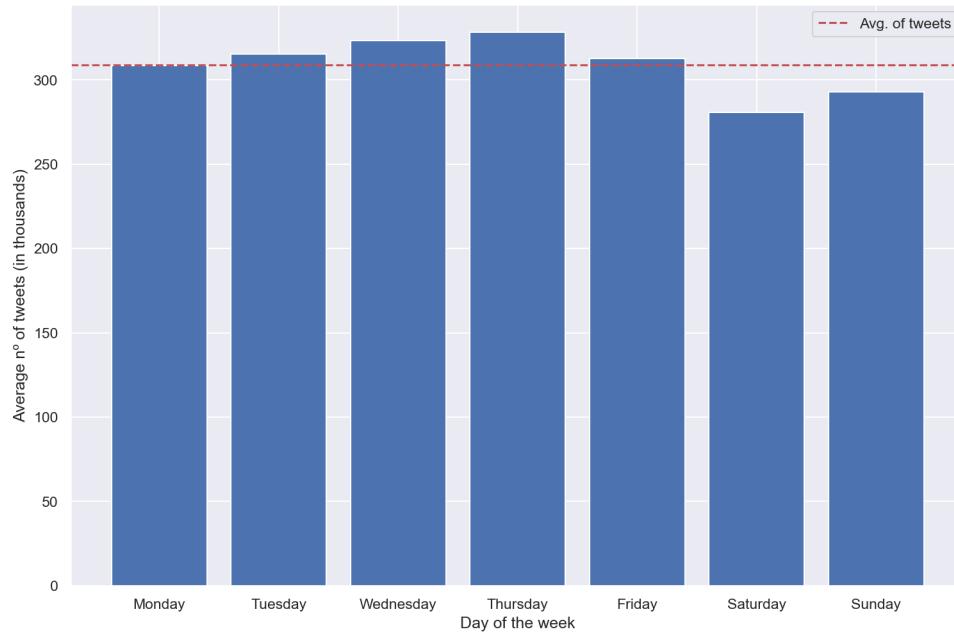
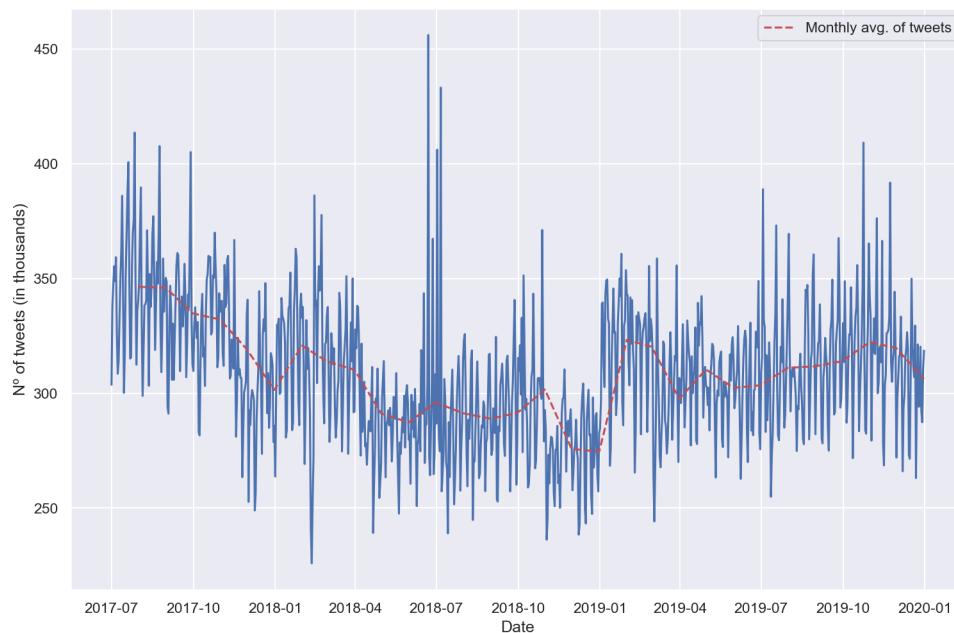


Figure 7: Trend of geo-located and total tweets.



In Figure 8, the words used are: "*Bolsonaro*," "*braço*" (arm), "*bom*" (good), "*cão*" (dog), "*cerveja*" (beer), and "*hoje*" (today). In Figure 9, we use sensitive words - that may reflect hate

speech. Specifically, these words are: "*mariquinha*" (offensive word for a gay man) "*sapatão*" (offensive word for a lesbian), "*nego*" (black), "*preto*" (black), and "*piranha*" and "*putinha*" (offensive words for a woman). As can be seen, both trends behave similarly for each word, suggesting that the sub-sample of geo-located tweets correctly captures how the universe of tweets behaves. This is especially true for the tweets containing "Bolsonaro."

Figure 8: Daily count of tweets retrieved by the Twitter query.



Figure 9: Daily count of tweets retrieved by the Twitter query.



A.2 Hate speech detection

In this paper, we fine-tune a BERT model on a dataset specific to the hate speech detection task. Fine-tuning is the technique of training a pre-trained model on a suitable dataset for a new task. We use a BERT model in Portuguese, by Souza et al. (2020), and a dataset of tweets in Portuguese, by Fortuna et al. (2019). In the next paragraphs, we describe the resources and procedure.

Model. In this paper, we take *BERTimbau*, a BERT model for Brazilian Portuguese by Souza et al. (2020), as a base model and fine-tune it for the hate speech detection task. Souza et al. (2020) present the model in two sizes: Base (12 layers, 768 hidden dimensions, 12 attention heads, and 110M parameters) and Large (24 layers, 1024 hidden dimensions, 16 attention heads, and 330M parameters). The authors train the models in two tasks: Masked Language Modeling (MLM) and Sentence Prediction (NSP). The model training is based on the *brWaC corpus* by Wagner Filho et al. (2018), the largest open Portuguese corpus. After training, they evaluate the model in other traditional NLP tasks, namely, Sentence Textual Similarity (STS), Recognizing Textual Entailment (RTE), and Named Entity Recognition (NER). The model improves the state-of-the-art on these tasks, outperforming Multilingual BERT models. The authors made their models publicly available at these Hugging Face links: Base, and Large.

Dataset. We relied on the dataset presented by Fortuna et al. (2019) to fine-tune the BERT model for the hate speech detection task. It is a dataset of tweets in Portuguese collected through Twitter’s API, and it comprises 5668 tweets in the period from January to March 2017. The authors provide two annotation schemes for the dataset, binary and hierarchical multiple classifications. For the first classification, three annotators classified every tweet. Each of them had to label the tweet as "hate speech" or "not hate speech," and the authors applied the majority vote to determine the final annotation of each tweet. As a result, 31.5% of the tweets were annotated as "hate speech" on the binary classification dataset. For the hierarchical classification, the authors followed a Rooted Directed Acyclic Graph (DAG) in which "hate speech" is the graph’s root. The second level of classes relates to the target of hate, and it comprises: "sexism," "body," "origin," "homophobia," "ideology," "religion," "health," and "lifestyle." As a result, 22% of the tweets were annotated as "hate speech" on the multi-labeled dataset. The authors made their datasets publicly available at this GitHub repository.

Text pre-processing. During text pre-processing, we follow Fortuna et al. (2019) and remove stop-words and punctuation marks using the *NLTK* and *re* Python libraries, respectively. Unlike the authors, we do not remove negative stop-words that may change the statement’s meaning. Explicitly, we keep the words: "mas" (but), "nem" (neither), "não" (no), "sem" (without), and "fora" (out). In addition, we anonymize Twitter mentions as "@user" and links as "URL." We keep "#Hashtags" in the native Twitter format. Finally, we do not transform text to lowercase for consistency with the architecture of Souza et al. (2020) ’s BERT model.

Model fine-tuning. We divide the dataset between 80% for training, 10% for validation, and 10% for testing. In NLP applications, the performance of a model in a given task is directly influenced by the characteristics of the training sample. In the case of Fortuna et al. (2019)'s dataset, as in other datasets on hate speech or offensive comments detection, a class imbalance exists. 31.5% of tweets were annotated as "hate speech" under the binary classification, being this a minority class. As class imbalance may affect the model's performance, we use a Random Oversampling technique to equalize the number of tweets in the minority and majority classes in the training sample (80% of the tweets). The random oversampling approach randomly adds examples from the minority class to the original training dataset, with replacement.

Training results. The hate-speech BERT model we train attains an overall accuracy of 77% in both the validation and test datasets. Table 4 summarizes additional statistics (Precision, Recall, and the F1-score) to characterize the model performance fully.

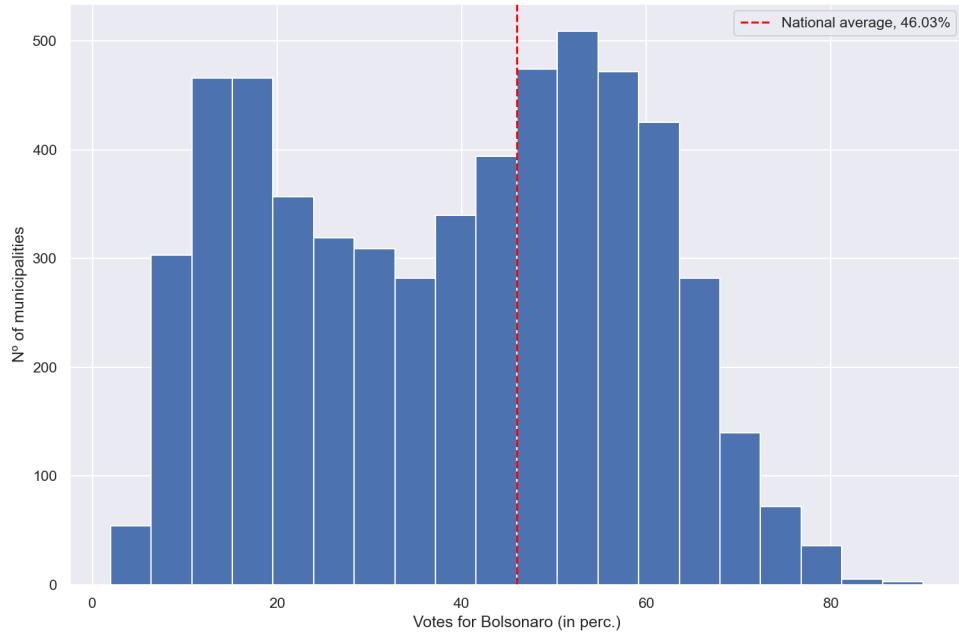
Table 4: Training results

Validation sample				
	Precision	Recall	F1	Support
0	0.86	0.76	0.81	365
1	0.64	0.78	0.71	202
W. Avg.	0.79	0.77	0.77	567
Test sample				
	Precision	Recall	F1	Support
0	0.88	0.79	0.83	406
1	0.58	0.72	0.64	161
W. Avg.	0.79	0.77	0.78	567

A.3 Tables and Figures

Figure 10: Bolsonaro's vote share at the municipality level. 2018 Presidential Election.

a. 1º Round, October 7th.



b. 2º Round, October 28th.

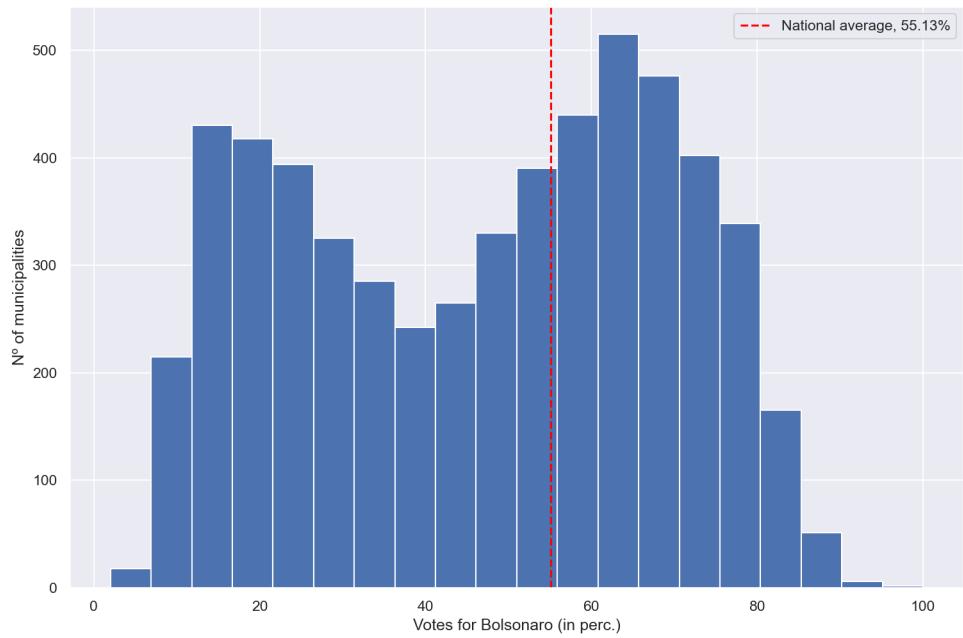
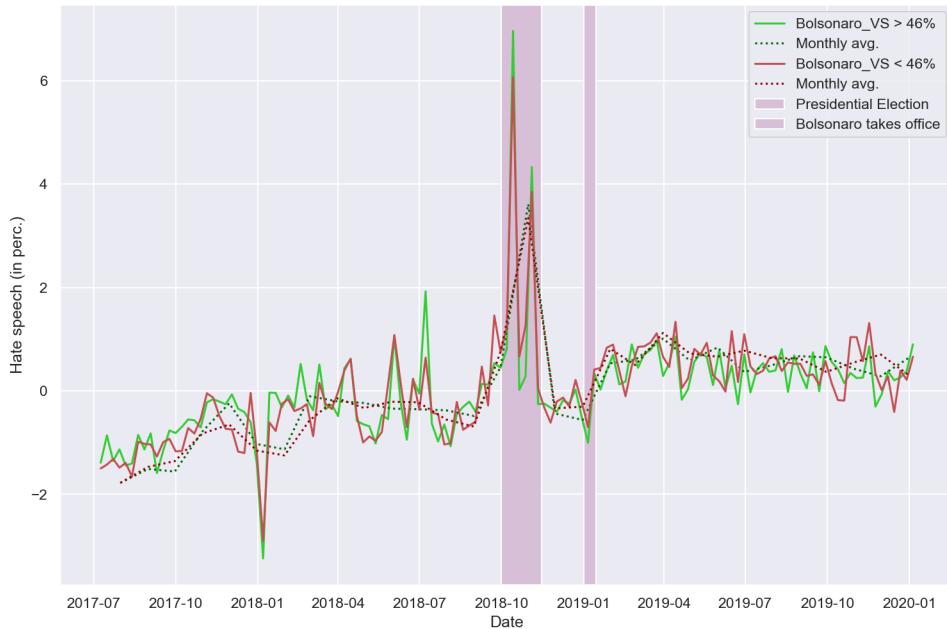


Figure 11: Evolution of hate speech in Brazilian tweets, 2017-2019. Standardized variables.

a. All municipalities.



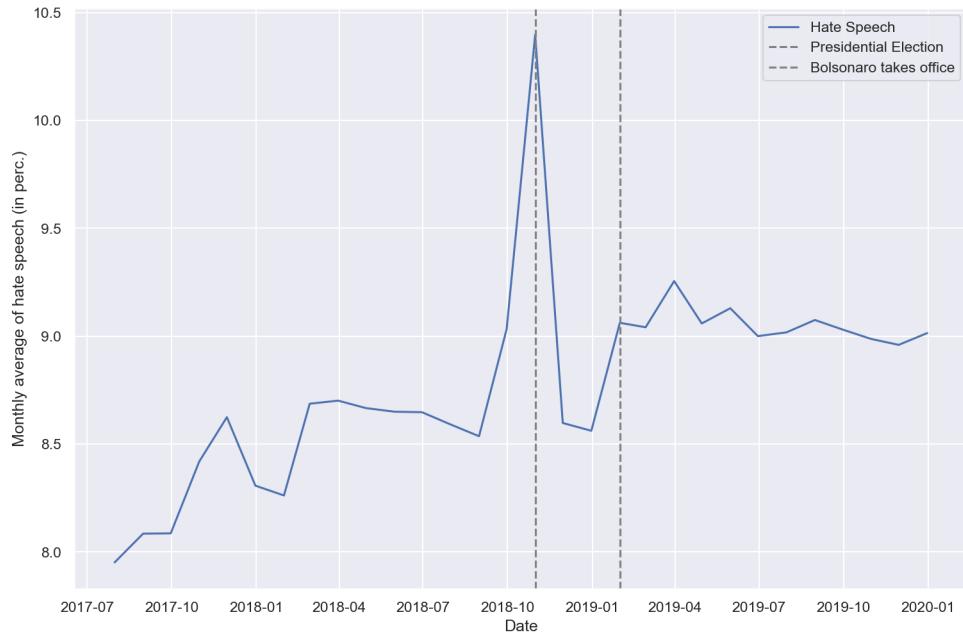
b. Municipalities, by the 2018 election result.



Note: Each hate speech variable was standardized to have zero mean and unit standard deviation.

Figure 12: Evolution of hate speech in Brazilian tweets, 2017-2019. Monthly average.

a. All municipalities.



b. Municipalities, by the 2018 election result.



Note: Hate speech variables were aggregated by month.

Table 5: Descriptive Statistics, by the 2018 election result.

Variable	Lost_m	Won_m	Difference
urban	0.569	0.735	-0.166***
income_pc	346.8	680.4	-333.6***
cellphone	0.697	0.869	-0.173***
computer	0.140	0.363	-0.223***
internet	0.610	0.703	-0.094***
primary	0.383	0.407	-0.023***
tertiary	0.495	0.425	0.070***
no_religion	0.052	0.052	-0.001
catholic	0.793	0.705	0.087***
pentecostal	0.098	0.126	-0.028***
black	0.075	0.048	0.027***
indigenous	0.009	0.005	0.004***
brown	0.558	0.304	0.255***
born_mun	0.710	0.556	0.154***
born_state	0.742	0.675	0.068***
vs_pt_2006	0.553	0.344	0.210***
vs_pt_2010	0.575	0.391	0.184***
vs_pt_2014	0.585	0.331	0.254***

Note: N = 5570 (municipalities). All variables are aggregated at the municipality level. Column "Lost_m" refers to the municipalities where Bolsonaro lost the 2018 election, whereas column "Won_m" refers to where he won. The third column reports the statistical difference between the respective means. Variables "cellphone," "computer," and "internet" are the proportion of households reported to have such goods in the 2010 Population Census. Variables "no_religion," "catholic," "pentecostal," "black," "indigenous," and "brown" are the proportion of individuals registered to have such demographic characteristics in the 2010 Population Census. Variables "primary" and "tertiary" refers to the population with (at most) primary and tertiary education. Variables "bornhere_mun" and "bornhere_state" refer to the proportion of individuals born in the municipality and state where they answered the 2010 Population Census. Variables "vs_pt_2006," "vs_pt_2010," and "vs_pt_2014" are the proportion of votes obtained by the Workers' Party (Partido dos Trabalhadores, PT) in the 1º round of the 2006, 2010, and 2014 Presidential Elections, respectively.

A.3.1 Regression results at the municipality level.

Table 6: Municipality level regressions. Standardized variables.

Variables	(1) $Hate_{mt}$	(2) $Hate_{mt}$
$Post_t \times Lost_m$	0.062*** (0.016)	
$Post_t \times VoteShare_m$		-0.025*** (0.008)
Constant	-0.019*** (0.004)	-0.010*** (0.003)
Municipality FE	Yes	Yes
Date FE	Yes	Yes
Municipalities	1,482	1,482
Observations	89,865	89,865
R-squared	0.074	0.073

Note: Standard errors in parentheses. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from January to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. $Hate_{mt}$ and $VoteShare_m$ are standardized to have zero mean and unit standard deviation. *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Municipality level regressions. Monthly data.

Variables	(1) $Hate_{mt}$	(2) $Hate_{mt}$
$Post_t \times Lost_m$	0.003** (0.001)	
$Post_t \times VoteShare_m$		-0.011** (0.004)
Constant	0.084*** (0.000)	0.087*** (0.001)
Municipality FE	Yes	Yes
Date FE	Yes	Yes
Municipalities	1,482	1,482
Observations	27,324	27,324
R-squared	0.144	0.144

Note: Standard errors in parentheses. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to July 2018 and a value of 1 from January to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Table 8: Municipality level regressions. Redefining $Post_t$.

Variables	(1) $Hate_{mt}$	(2) $Hate_{mt}$
$Post_t \times Lost_m$	0.003*** (0.001)	
$Post_t \times VoteShare_m$		-0.008*** (0.003)
Constant	0.084*** (0.000)	0.086*** (0.001)
Municipality FE	Yes	Yes
Date FE	Yes	Yes
Municipalities	1,482	1,482
Observations	100,375	100,375
R-squared	0.071	0.071

Note: Standard errors in parentheses. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to August 2018 and a value of 1 from November 2018 to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

A.3.2 Regression results at the individual level.

Table 9: Intensive margin of hate speech. Individual level regressions. Sub-sample of Twitter users, restricted by their activity.

Variables	(1) $Hate_{imt}$	(2) $Hate_{imt}$	(3) $Hate_{imt}$	(4) $Hate_{imt}$
$Post_t \times Lost_{im}$	0.002* (0.001)		0.003** (0.001)	
$Post_t \times VoteShare_{im}$		-0.005 (0.006)		-0.005 (0.006)
Constant	0.107*** (0.000)	0.111*** (0.006)	0.111*** (0.000)	0.125*** (0.007)
Individual FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Individuals	52,518	52,518	24,180	24,180
Observations	342,732	342,732	210,565	210,565
R-squared	0.260	0.260	0.226	0.226

Standard errors in parentheses. Columns (1)-(2): all Twitter users who posted at least 25 tweets over the period, from which at least 5 are classified as hate speech. Columns (3)-(4): all Twitter users who posted at least 50 tweets over the period, from which at least 10 are classified as hate speech. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to August 2018 and a value of 1 from November 2018 to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Table 10: Individual level regressions.

Variables	(1) <i>Hate_{imt}</i>	(2) <i>Hate_{imt}</i>
<i>Post_t</i> X <i>Lost_{im}</i>	0.002 (0.001)	
<i>Post_t</i> X <i>VoteShare_{im}</i>		-0.005 (0.005)
Constant	0.090*** (0.000)	0.098** (0.004)
Individual FE	Yes	Yes
Month FE	Yes	Yes
Individuals	113,127	113,127
Observations	523,458	523,458
R-squared	0.342	0.342

Standard errors in parentheses. All Twitter users in the sample. $Post_t$ is a dummy variable that takes a value of 0 from July 2017 to August 2018 and a value of 1 from November 2018 to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.

Table 11: Individual level regressions. Sub-sample of Twitter users, restricted by their activity.

Variables	(1) $Hate_{imt}$	(2) $Hate_{imt}$	(3) $Hate_{imt}$	(4) $Hate_{imt}$
$Post_t \times Lost_{im}$	0.002* (0.001)		0.002** (0.001)	
$Post_t \times VoteShare_{im}$		-0.004 (0.005)		-0.005 (0.005)
Constant	0.091*** (0.000)	0.098*** (0.005)	0.091*** (0.000)	0.097*** (0.005)
Individual FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Individuals	90,355	90,355	50,329	50,329
Observations	475,402	475,402	357,506	357,506
R-squared	0.315	0.315	0.270	0.270

Standard errors in parentheses. All Twitter users in the sample who posted (i) at least 25 tweets over the period in columns (1)-(2) and (ii) at least 50 tweets in columns (3)-(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 November 2018 to December 2019. $Lost_m$ is a dummy variable that takes a value of 1 for the municipalities where Bolsonaro's vote share was lower than 46% and a value of 0 otherwise. *** p<0.01, ** p<0.05, * p<0.1.