Hate in the Tropics: Bolsonaro's Triumph and the Surge of Online Hate Speech in Brazil*

Diego Marino-Fages† Alejandra Agustina Martínez‡

February 2025

Abstract

How does the advent of political information influence social norms and individual behavior? This paper examines the impact of Bolsonaro's victory in the 2018 Brazilian presidential election on the prevalence of hate speech. We leverage Twitter data from 2017 to 2019 and employ Natural Language Processing (NLP) techniques to detect hate speech in tweets. Relying on the election result as an information shock, we propose a difference-in-differences approach to identify the effect of Bolsonaro's triumph in hate speech. Our findings reveal a significant increase in online hate speech after the election, especially in municipalities where Bolsonaro had lower support. Next, we classify tweets based on the targets of hate speech into five categories and find that the surge in hate speech is mainly driven by homophobia, sexism, and racism – areas in which Bolsonaro's rhetoric was highly controversial. Overall, we interpret these results through a belief-updating mechanism, emphasizing the process of revising social norms that determine (un)acceptable public discourse.

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, 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 (University Paris-Dauphine), 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).

†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, e.g., contractual relationships, conceptions of right and wrong, 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.

A relevant social norm concerns 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.

Although social norms tend to be stable over time (Fernandez, 2007; Giuliano, 2007; Alesina et al., 2013), a growing number of studies show how certain events can trigger quick changes in their prevalence (Bursztyn et al., 2020a). These events can be very different in nature, ranging from famines to the arrival of new information, such as electoral outcomes.

In this paper, we study the impact of Jair Bolsonaro's 2018 presidential election on the prevalence of online hate speech in Brazil. Our findings indicate that after the election, there was an overall increase in online hate speech across the country, particularly pronounced in areas where Bolsonaro had relatively little support. Bolsonaro, sometimes called "the Trump of the Tropics," is widely recognized for his contentious viewpoints, encompassing homophobia, racism, and sexism.[1] Therefore, following previous literature, we argue that Bolsonaro's victory prompted a quick update of the prevailing social norm governing what types of speech are socially acceptable.[2]

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. To address this challenge, we rely on the fact that

[1] 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.

[2] Figure A9 in the Appendix shows how Bolsonaro's Google searches, as well as Twitter mentions, surged around the time of the elections, so people may have become aware of his earlier controversial statements.

Bolsonaro's victory surprised the Brazilian community. He got 46% of the votes in the 1º round of the election and 55% in the 2º round. The opinion polls conducted by diverse companies in the days before the election estimated that Bolsonaro's vote share would be approximately 35% for the 1º round, and only one polling company estimated a vote share above 40%.[3] Therefore, our identification strategy relies on considering the 2018 election outcome as an *information shock* – new and unexpected information regarding the support for a far-right candidate at the national level. Following Ajzenman et al. (2023) and Albornoz et al. (2022), we exploit the size of this shock to identify its marginal effect.

To conduct the empirical analysis, we propose two difference-in-differences designs. 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, we propose a difference-in-differences design with a continuous treatment variable.[4] 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.

To measure hate speech, we apply two text analysis techniques[5] to a corpus of tweets[6] posted on the social media platform Twitter (now re-branded as X) during the period spanning between July 2017 and December 2019.[7] First, 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). Second, to better confirm the previous results with a more transparent method, we use a dictionary-based method. This method also allows us to introduce a multi-level classification of hate speech. Specifically, we classify tweets containing hate speech into five subcategories, i.e., political hate, homophobia, racism, sexism, and insult.

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* (i.e., his vote share was smaller than 46%). Furthermore, our findings suggest that the

[3]Source: Wikipedia, `https://en.wikipedia.org/wiki/Opinion_polling_for_the_2018_Brazilian_general_election`, access date: June 2023.

[4]See Callaway et al. (2024) for a theoretical reference.

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

[6]In order to avoid the bots created for the political campaign, we restrict the sample to tweets posted by accounts that were created before 2018.

[7]This 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.

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. The information shock, induced by the election outcome, allowed individuals living in a relatively anti-Bolsonaro municipality to reassess their beliefs regarding socially acceptable speeches. Once the social norm was updated, these individuals may have felt justified in expressing hateful viewpoints through social media platforms, even if they resided in a municipality where the prevalence of such behavior was relatively low before the elections. In other words, this can be understood as a break in the *spiral of silence*.[8]

This interpretation of the results is reinforced when analyzing the differential impact of Bolsonaro's victory on each type of hate speech – namely, homophobia, racism, political hate, insult, and sexism. We do not find such a differential impact for political hate and insults, but we find it for homophobia, sexism, and racism. While finding an effect on political hate might be interpreted as a sign of growing polarization, the effects on homophobia, sexism, and racism underline a social norms channel. That is, having a harmful speech that targets specific groups, such as the LGBT community, women, and different races, highly depends on the social acceptability of such behavior.

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, although with different magnitudes. In other words, we observe some Twitter users who post hate speech tweets only after the elections (i.e., extensive margin), especially in the municipalities where Bolsonaro lost. Similarly, we document that users posting hate content before the elections increased the frequency of these tweets after the elections (i.e., intensive margin).

Our paper belongs to the literature that studies how, for good or bad, social norms drive behavior (e.g., Elster, 2020; Nyborg et al., 2016; Bicchieri, 2016, 2005). More specifically, our paper adds to the literature studying how certain events can trigger rapid changes in social norms. For instance, Andre et al. (2024) finds that correcting misperceptions about the prevalence of climate-friendly behavior drives people to behave more pro-environmentally; Bursztyn et al. (2020b) finds that correcting men's misperceptions about other men's support for women working outside the home increases their willingness to help their wives search for jobs; Morales (2020) finds that a change in the perceived popularity of Maduro in Venezuela affects the willingness to express criticism of the pres-

[8]The spiral of silence theory argues that people often remain silent when they perceive their views on a value-laden issue are in the minority, driven by the fear of social isolation (Noelle-Neumann, 1974).

ident and support for the opposition (see Bursztyn and Yang, 2022, for a review of this growing literature). Unlike these previous examples, we find that an information shock produces an undesirable outcome, i.e., an increase in hate speech.

Large information aggregators, such as the results of elections or referendums, can also shock the prevailing social norms. Examples of these are Albornoz et al. (2022), which studies the effect of the outcome of the Brexit referendum on hate crime, and Bursztyn et al. (2020a), which studies the effect of the Trump 2016 election on the willingness to express xenophobic opinions. While these are the closest papers to ours, we present several differences and advantages. First, our results are representative of one of the largest developing countries, while Bursztyn et al. (2020a) focus on one metropolitan area in the United States, 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 the context of each paper. Twitter users who post hate speech are immediately available for social scrutiny (Metzger, 2009), especially from friends, whereas, in Bursztyn et al. (2020a), the information is said to be posted at a later date on a likely unknown website and in Albornoz et al. (2022), perpetrators only pay a cost if they get caught. Third, since our data allows us to follow a large number of Twitter accounts over time, we can explore the roles of the intensive and extensive margins of hate expressions – essential knowledge for designing effective policies to minimize these effects.

Furthermore, the paper is related to the growing literature on the effect of leaders on social norms (e.g., Ajzenman et al., 2023; Bursztyn et al., 2020a; Farina and Pathania, 2020; Acemoglu and Jackson, 2015, 2017). Our results suggest that the increase in hate speech was not driven by the electoral rally or Bolsonaro's speeches during that period – note that his controversial quotes date before his presidential candidacy, and see the evolution of hate speech in Bolsonaro's tweets in Figure A8 in the Appendix. Furthermore, Barros and Santos (2021) provides a more detailed description of the background and a potential reason for Bolsonaro's success. Instead, we argue that the rise in hate speech has been triggered by the information shock produced by the election results, confirming that the majority of citizens supported Bolsonaro.

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). More specifically, to the literature linking social media and expressions of hate, particularly against minority groups.[9] Müller and Schwarz (2023) find a positive relationship between

[9]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., 2019).

Twitter usage and ethnic hate crimes since the presidential election of Donald Trump in the United States, 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. Carr et al. (2020) demonstrates that the Brexit referendum in the United Kingdom resulted in a rise in hate crimes, providing evidence that both media and social media contributed to this increase. Cao et al. (2023) shows that Donald Trump's "Chinese Virus" tweets contributed to the rise of anti-Asian incidents in the United States. This literature covers a wide range of social media platforms, like Twitter and Facebook, but focuses mainly on xenophobia and ethnic hate crimes. This paper, in contrast, considers a wider definition of expressions of hate, zooming into its different targets.

In addition, we focus on hate speech rather than hate crime and online rather than offline expressions of hate. Beknazar-Yuzbashev et al. (2022) present empirical evidence that toxicity increases content consumption and is contagious on social media platforms, and Beknazar-Yuzbashev et al. (2024) propose a theoretical argument on under which circumstances social media platforms may find profitable to display harmful content. Altogether, this literature suggests that social media platforms play a significant role in understanding online and offline hate expressions.

Analyzing a developing country such as Brazil presents additional advantages. First, it is one of the countries with the highest Twitter penetration. Second, as a developing country with weak institutions, social norms may arguably play a stronger role in driving behavior than in more developed countries (Fergusson Talero et al., 2024). Third, it allows us to document and exploit the geographical variation in social norms within one of the largest developing countries.

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 formerly known as 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 published by 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.

An 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. The perpetrator immediately pays the cost of expressing hateful content in the former. On the other hand, hate crimes must be reported and processed by justice before the perpetrator pays the costs, which may obscure the analyses.

2.1 Twitter data

Twitter (currently re-branded as X) was an online platform allowing users to publish short messages of a maximum of 140 characters on their profiles (extended to 280 characters after November 2017). With one of the largest Twitter user bases in the world, Brazil is an appealing case of study for online activity – in this case, related to Twitter users' speech. 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).[10] 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%).[11]

In the empirical analysis, our main variable of interest is the proportion of tweets classified as hate speech per municipality (or individual) and date. We rely on Natural Language Processing (NLP) techniques to construct this variable. Precisely, we train a pre-trained *Bidirectional Encoder Representations from Transformers* (BERT) model and build a multi-level classification of hate based on a dictionary method. 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 effect of Bolsonaro's victory on each

[10]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.

type of hate speech, providing evidence of the mechanisms behind the main results. 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 the tweets (net of retweets) that 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 between July 2017 and 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 the tweets posted on any Monday and the geo-located tweets constitute 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 anonymize user mentions and URL links but keep hashtags in their native Twitter format, as they may contain 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 the user accounts potentially created in the context of the electoral rally (i.e., political bots). Before classifying tweets with our BERT model, we perform some data pre-processing tasks described in Appendix A.2.

Hate speech detection. We rely on NLP techniques to detect hate speech in our tweets' database.[12] We implement two classification techniques. Firstly, we train a pre-trained 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. Once fine-tuned, our BERT model is able to classify tweets as having or not hate speech, i.e., it leads to a binary classification of tweets. Secondly, we construct a dictionary that enables us to obtain a multi-level classification of hate speech. Precisely, we classify tweets into five categories associated with specific hate targets. In the next paragraphs, we discuss each method.

[12] See Ayo et al. (2020) for a review on hate speech detection via machine learning techniques.

Appendix A.2 provides further details on the hate speech detection task and the resources utilized.

BERT model fine-tuning. 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.

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 classification. 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."

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 Fortuna et al. (2019)'s dataset, a class imbalance exists, with tweets annotated as "hate speech" constituting the minority class. As this 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).[13] Our model attains an overall accuracy of 77% in both the validation and test samples.

Dictionary method. We introduce a dictionary-based method for the detection of hate speech in tweets. We create this dictionary with the specific aim of classifying hate speech according to its targets. Precisely, we consider five distinct categories of hate speech: political hate, homophobia, racism, sexism, and insult. The class "political hate" is less common in the existing literature but highly relevant to the context of our study.

For the dictionary construction, we draw upon top-frequency words associated with each type of hate speech from three different papers. First, we use the hierarchical classification of hate speech presented by Fortuna et al. (2019). Second, we employ the multilevel classification of toxicity in tweets provided by Leite et al. (2020). Lastly, we enrich the class homophobia by using the information in Pereira (2018). In addition, we consider a tweet to contain "hate speech" if it includes at least one of the hate categories listed in the dictionary.

[13] Random 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.

Data classification. After training the BERT model and constructing the dictionary, we use them to detect hate speech in the tweets we collected. As a result of the BERT classification, we construct a binary variable 0/1, named "predicted hate speech." As a result of the dictionary classification, we construct six binary variables 0/1, named "predicted hate speech," "predicted political hate," "predicted homophobia," "predicted sexism," "predicted racism," and "predicted insult." 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 different types of hate speech by municipalities (or individuals) over time, which are the main outcome variables of this paper.

2.2 Administrative data

The election result used as an information shock in this paper is the vote share at the municipality level obtained by Bolsonaro in the 1º round of the 2018 Brazilian presidential election. The Superior Tribunal Court (TSE) has provided 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 Datasets

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. Depending on the specification, this leaves us with approximately 2000-2500 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). Figure A7 in the Appendix shows Twitter's penetration in terms of tweets and users in Brazilian municipalities.

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 three different municipalities. When an individual appears in more than one location, we implicitly assume she is engaged in an activity (such as working or studying) in municipalities different from where she lives. The longitudinal dataset at the individual level is also unbalanced, as Twitter activity significantly varies across 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).

2.4 Descriptive statistics

This paper builds upon two fundamental observations. Firstly, the presidential election, which we consider an information shock, did not uniformly affect all Brazilian citizens. Secondly, the evolution of online hate speech was not consistently constant throughout the period. Regarding the first one, we observe a significant geographical variation in Bolsonaro's vote share, which helps us to identify the effect of interest. Figure 1 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. Figure A6 in the Appendix presents the (bimodal) distributions of these vote shares at the municipality level.

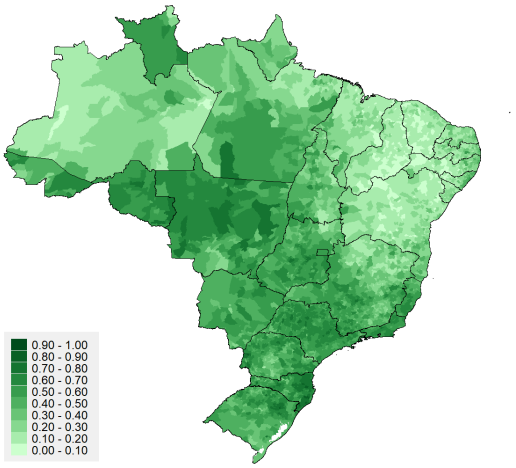
As for the second observation, Figure 2 shows the proportion of Brazilian tweets classified as hate speech in the period under study.[14] The solid line corresponds to the raw data, consisting of the daily proportion of hate speech tweets, whereas the dotted line corresponds to the same data after applying a 5-week moving average filter. The shadow areas in the graph delimit the periods in which (i) the Presidential Election took place and

[14] In Appendix A.3, we present an analogous graph but with hate speech classified by the dictionary method.

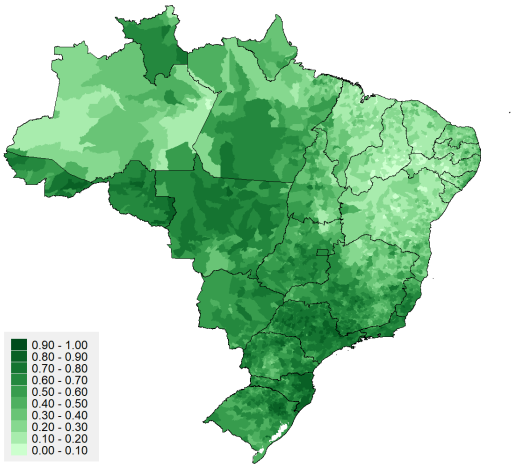
(ii) Bolsonaro took office.[15]

Figure 1: Bolsonaro's vote share at the municipality level.

(a) 1º Round, October 7.



(b) 2º Round, October 28.



Note: Bolsonaro's vote share (0-1) at the municipality level in the 1º and 2º presidential election rounds. The darker the color, the higher the vote share.

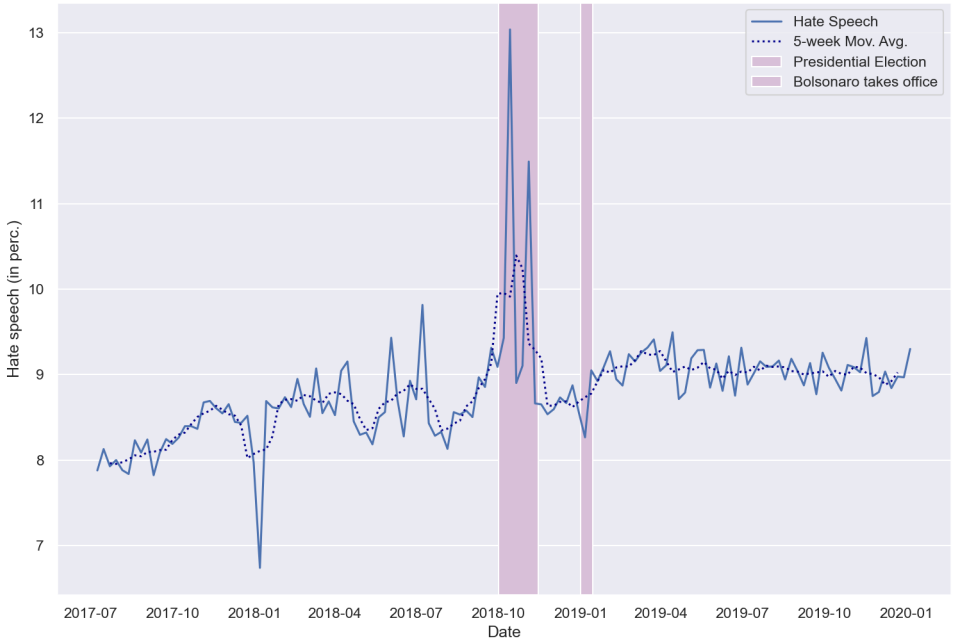
As can be seen, there was a sharp increase in hate speech during this period. The hate speech peaks on the data correspond to the closest (but later on time) date in our sample to the 1º and 2º rounds of the election.[16] It is also worth noticing that the period with lower levels of hate speech corresponds 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.

Importantly, Figure 2 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. This is approximately equivalent to an increase of 3000 tweets containing hate speech on an average day. 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.

[15]Specifically, the 1º 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 1, 2019.

[16]There 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 A2 in the Appendix shows that these peaks also correspond to a sharp increase in Twitter activity. Specifically, the daily amount of tweets is around 50% higher during that period (relative to the average).

Figure 2: Evolution of hate speech in Brazilian tweets, 2017-2019.



Note: The variable Hate speech is, for each date, the percentage of tweets classified as hate speech by the BERT model. Solid line: raw data. Dotted line: processed data, after applying a 5-week moving average filter.

3 Empirical strategy

We aim to estimate the effect of Bolsonaro's election on hate speech. In the previous section, we showed that hate speech increased at the national level after Bolsonaro's election (see Figure 2). 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 1). We can then exploit the differential informational shock, as proxied by the election results, 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.[17] Thus, we perform a difference-in-differences analysis. Formally, we regress,

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

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,[18] $Lost_m$ 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 and π_m are time and 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.[19] 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, we ask the following question: Is the increase in hate speech driven by people already tweeting hate content before the elections (i.e., intensive margin) or caused by people who had not tweeted hate content before (i.e., extensive margin)? Formally, we regress,

$$Hate_{imt} = \tilde{\alpha_0} + \tilde{\alpha_1} * Post_t * Lost_{im} + \delta_t + \gamma_i + \epsilon_{imt} \tag{2}$$

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 and γ_i are time and user fixed effects, and ϵ_{imt} is an account-municipality-time specific error term. Our coefficients of interest are $(\alpha_1, \tilde{\alpha_1})$,

[17] Our municipality-level dataset contains a daily average of 266 (183) tweets in municipalities where Bolsonaro won (lost).

[18] Precisely, $Post_t$ takes a value of 0 from July 2017 to July 2018 and a value of 1 from January to December 2019.

[19] Our individual-level dataset includes 358,029 (115,002) accounts in municipalities where Bolsonaro won (lost).

which, given parallel trends, capture 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,

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

and,

$$Hate_{imt} = \tilde{\beta}_0 + \tilde{\beta}_1 * Post_t * VoteShare_{im} + \delta_t + \gamma_i + \epsilon_{imt} \tag{4}$$

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, π_m and γ_i are time, 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 coefficients of interest are $(\beta_1, \tilde{\beta}_1)$, which capture 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 (Callaway et al., 2024).[20] Notice that, by definition, $(\alpha_1, \tilde{\alpha_1})$ in equations (1) and (2) and $(\beta_1, \tilde{\beta}_1)$ in equations (3) and (4) have opposite signs: while the former capture the effect of $Lost_m = 1$, which depend negatively on Bolsonaro's vote share, the latter are proportional to it.

4 Results

In this section, we present the main results of the paper. First, we document that hate speech increased after the 2018 presidential elections, especially in the municipalities where Bolsonaro lost. Second, we show that the effect is driven by hate towards groups to whom Bolsonaro was openly against. Finally, we present the results at the individual

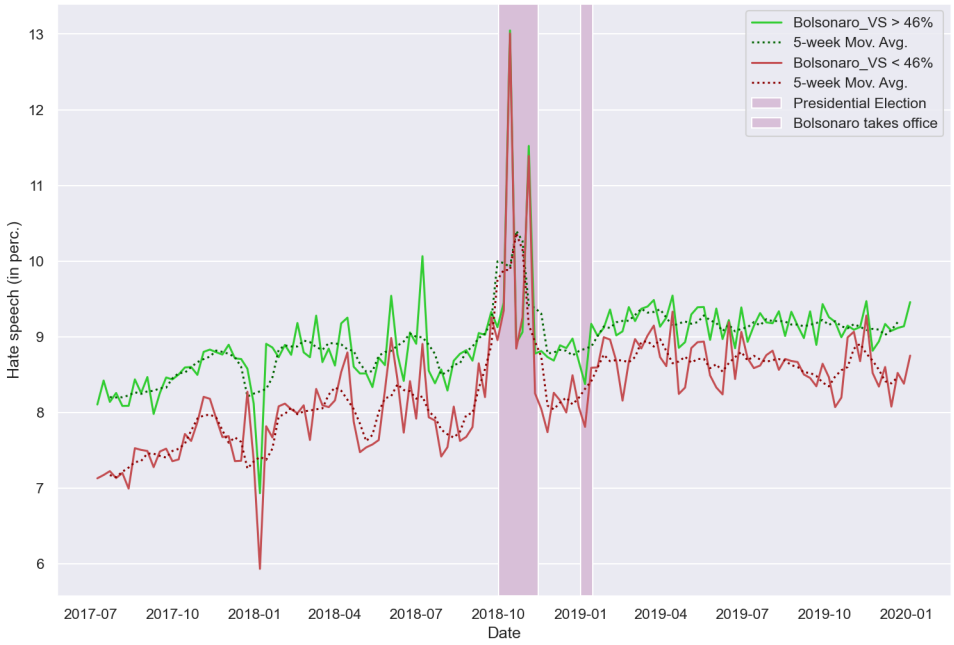
[20]Formally, let d be the dose and Y_t be the potential outcome in time t. Then, the strong parallel trends assumption implies that for all d in D: $E[Y_t(d) - Y_{t-1}(0)] = E[Y_t(d) - Y_{t-1}(0)|D = d]$.

level, indicating that both the intensive and extensive margins of hate speech contributed to 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 2, but now splitting the hate speech trends between treatment and control groups.[21]

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



Note: Percentage of tweets classified as hate speech by the BERT model split by Bolsonaro's vote share in the 1º round of the election. Solid line: raw data. Dotted line: processed data, after applying a 5-week moving average filter.

The green lines correspond to the daily and 5-week moving average filtered proportion of Brazilian tweets classified as hate speech by the BERT model for the municipalities in which Bolsonaro got at least 46% of the votes in the 1º round of the 2018 presidential

[21] In Appendix A.3, we present an analogous graph but with hate speech classified by the dictionary method.

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 is constant over time, i.e., pre-trends are parallel. Furthermore, the prevalence of hate speech in municipalities where Bolsonaro won and lost seems 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) – reassuring the validity of the municipalities acting 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 least popular increasing the most.

Let us turn to the regression results. Table 1 answers the main question of this paper, *how the 2018 presidential election of Bolsonaro affected online hate speech*. Columns (1) and (2) in the table present the results when the BERT model detects hate speech, whereas in columns (3) and (4), hate speech is classified using the dictionary method. The first and third columns in the table correspond to the classic difference-in-differences estimation, presented in equation (1). The second and fourth columns correspond 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 zero between July 2017 and July 2018 (both included) and one between January and December 2019 (both included). In the main specification, we eliminate the period from August to December 2018 to prevent contamination from hate speech that can be directly linked to the election and the electoral rally. Appendix A.3 confirms that our results are robust to changes in the definition of $Post_t$.

Columns (1) and (3) show the increase in hate speech after the elections that we observe in Figures 2 and 3 was more pronounced in municipalities where Bolsonaro lost (0.059 or 0.043 standard deviations higher than the municipalities where Bolsonaro won). Consistent with this evidence, columns (2) and (4) show that the proportion of hate speech decreases as the share of votes for Bolsonaro increases. As the estimates in columns (2) and (4) come from a difference-in-differences model with a continuous treatment variable, provided the strong parallel trends assumption, each 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 standard deviation in $VoteShare_m$ decreases hate speech in that municipality by 0.033 or 0.024 standard deviations.

Table 1: Municipality level regressions.

Variables	(1) $Hate_{mt,BERT}$	(2) $Hate_{mt,BERT}$	(3) $Hate_{mt,dict}$	(4) $Hate_{mt,dict}$
$Post_t$ X $Lost_m$	0.059***		0.043**	
	(0.020)		(0.017)	
$Post_t$ X $VoteShare_m$		-0.033***		-0.024**
		(0.011)		(0.011)
Constant	-0.019***	-0.006***	-0.025***	-0.016***
	(0.002)	(0.002)	(0.003)	(0.001)
Municipality FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Municipalities	1,930	1,930	2,487	2,487
Observations	97,581	97,581	126,766	126,766
R-squared	0.084	0.084	0.105	0.105

Note: Standardized variables. Standard errors clustered at the municipality level are 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.

We interpret our results from columns (1) to (4) as evidence of an update in the beliefs held by the citizens regarding the acceptability of hate speech. The difference in the election results at the municipality and national levels may serve as a proxy for the extent of the information shock, which triggered an update in beliefs about the prevailing social norms. In other words, individuals residing in municipalities where Bolsonaro lost (thus, predominantly surrounded by individuals who do not support Bolsonaro) are more likely to have the highest level of misperception regarding the actual number of people who support Bolsonaro and, potentially, share his viewpoints. Hence, these are the municipalities where hate speech increases the most.

If our interpretation of the results is accurate, we should expect that the types of hate speech experiencing a surge are related to topics influenced by social norms. That is, the expressions of hate should be targeted to specific groups, such as women, the LGBT community, and different races. To study this, we rely on our dictionary-based method to classify tweets into five different categories: homophobia, racism, sexism, political hate, and insults. Figure A13 in the Appendix shows the evolution of each category and hate speech overall during the period. Although homophobia is less prevalent than sexism and racism, it is still similar to the amount of hate present in the political arena.

Table 2 presents the regressions using each variable as outcomes. Confirming our hy-

potheses, we find significant effects for homophobia. The regression results for racism and sexism are consistent in terms of sign and magnitude with this interpretation, although they are not significant at the standard levels. Importantly, we find no differential effects between municipalities of Bolsonaro's election on political hate. The corresponding coefficients are close to zero or fluctuate in signs and are not statistically significant, reassuring that a differential increase in polarization does not drive our results. Figure A14 in the Appendix shows the evolution of each of the categories, differentiating control and treatment groups. It is noteworthy that for homophobia, the two groups of municipalities switch entirely at the beginning of the electoral campaign. The results in the pre-election period support the findings in Barros and Santos (2021), which argues that men gravitate towards a politician who exacerbates masculine stereotypes to compensate for losses in social and economic status in previous years. In this paper, we argue that these trends are reverted due to the information shock induced by Bolsonaro's popularity.

Table 2: Municipality level regressions. Results by hate targets.

	(1)	(2)	(3)	(4)	(5)
Variables	$Political_{mt}$	$Homophobia_{mt}$	$Racism_{mt}$	$Sexism_{mt}$	$Insult_{mt}$
Panel A: Binary Treatment					
$Post_t$ X $Lost_m$	-0.011	0.044**	0.024	0.020	0.022
	(0.022)	(0.018)	(0.021)	(0.015)	(0.017)
Constant	-0.023***	-0.006**	-0.024***	-0.009***	-0.008***
	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)
Observations	126,766	126,766	126,766	126,766	126,766
R-squared	0.093	0.048	0.145	0.046	0.092
Panel B: Continuous Treatment					
$Post_t$ X $VoteShare_m$	-0.001	-0.034***	-0.016	-0.010	-0.009
	(0.017)	(0.011)	(0.013)	(0.010)	(0.010)
Constant	-0.024***	0.004***	-0.019***	-0.005***	-0.004***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	126,766	126,766	126,766	126,766	126,766
R-squared	0.093	0.049	0.145	0.046	0.092

Note: Standardized variables. Standard errors clustered at the municipality level are 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.

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.

Our data allows us to follow users over time; hence, we can further extend our main analysis and explore *who* is driving the results. 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.

In this section, we focus on a sub-sample of Twitter users whose tweets are located in *no more than three* different municipalities. For the regression analysis, we restrict our attention to the sub-sample of Twitter users 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 4 shows that the rise in hate speech results from both the intensive and extensive margins, albeit different orders of magnitude. To construct the figure, we consider that Twitter accounts have posted hate content if at least one of their tweets were classified as hate speech.

Panel (a) shows how the share of Twitter accounts posting zero hate content (as detected by the BERT model) became smaller after the elections. Specifically, 61.5% of the Twitter users in our sample had never published hate speech content before the 2018 elections, and this number reduced to 59.3% 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 69.7% in the pre-election period and 65.9% in the post-election period.[22] The figure also shows the share of Twitter accounts posting zero hate content when we restrict the sample to accounts that appear at least once before and after the elections. In this case, the levels are smaller, but the same pattern emerges, i.e., the number of accounts posting zero hate content is reduced from 56.2% to 53.3%, and from 64.3% to 59.5% in the municipalities where $Lost_m = 1$.[23]

Panel (b) focuses on the intensive margin by zooming in on Twitter accounts that have

[22]The corresponding numbers for hate speech classified by the dictionary method are as follows: 64.8% and 62.8% of users have never published hate content before and after the elections, respectively. In municipalities where $Lost_m = 1$, the percentages are 71.1% and 67.5%, respectively.

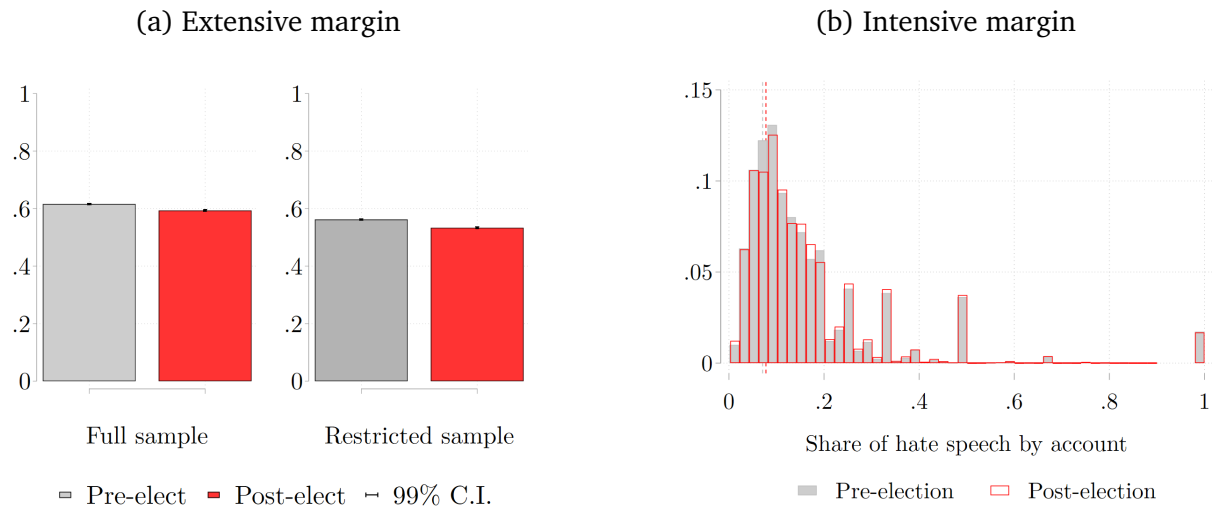
[23]The corresponding numbers for hate speech classified by the dictionary method are as follows: 56.1% and 54.0% of users have never published hate content before and after the elections, respectively. In municipalities where $Lost_m = 1$, the percentages are 62.2% and 58.0%, respectively.

posted messages with hate speech at least once. As can be seen, the distribution of individual hate speech (at the national level) has shifted to the right after the elections. The average (by individual) percentage of tweets containing hate speech has risen from 16.2% to 16.7% in the sub-sample of users who posted hate speech content at least once.[24] In the municipalities where $Lost_m = 1$, the corresponding averages went from 14.5% to 18.0%.[25] It is also noticeable that many accounts only posted hate content. However, these accounts are from users who have posted 1.2 tweets on average and are balanced between municipalities where Bolsonaro was more and less popular. Overall, both the intensive and extensive margins of hate speech played a role in this phenomenon, with the impact at the extensive margin being more relevant in terms of magnitude. A similar pattern emerges when we use the dictionary method (see Figure A15). Figures A16 and A17 show the entire CDFs for the combinations of pre and post elections, and pro and anti Bolsonaro municipalities, and Figures A18, A19 and A20 present the differences by the number of followers, number of accounts following and the tweeting activity.

[24]The corresponding numbers for hate speech classified by the dictionary method are as follows: 10.8% and 11.6% before and after the elections, respectively.

[25]Using the dictionary method, the percentages in the municipalities where $Lost_m = 1$ increased from 12.1% to 13.2%.

Figure 4: Hate speech at the individual level (Twitter users)



Note: Hate speech classified by the BERT model. The pre-election period is between July 2017 and July 2018, and the post-election period is between January and December 2019. Panel (a) presents the share of Twitter accounts posting zero hate speech, using the full sample and the sample restricting to users who posted tweets before and after the elections. Panel (b) shows the distribution of hate speech shared by users in the restricted sample, excluding those users posting zero hate speech in the whole period.

Next, we present the regression results of our difference-in-differences models at the individual level, equations (3) and (4), in Table 3. In this exercise, we further restrict the sub-sample of users according to their online activity in two ways. First, we restrict our attention to those users who posted at least 50 tweets during the entire period. Additionally, we focus on the intensive margin of hate speech by considering users who posted at least 10 tweets classified as hate speech during the pre-election period. As can be seen, the estimates are comparable in sign, magnitude, and statistical significance to those previously presented in Table 1. In Appendix A.3, we present supplementary regressions, relaxing these restrictions and redefining the intensive margin of hate speech.

Table 3: Individual level regressions. The intensive margin of hate speech.

	(1)	(2)	(3)	(4)
Variables	$Hate_{imt,BERT}$	$Hate_{imt,BERT}$	$Hate_{imt,dict}$	$Hate_{imt,dict}$
$Post_t$ X $Lost_{im}$	0.032**		0.019	
	(0.015)		(0.015)	
$Post_t$ X $VoteShare_{im}$		-0.014**		-0.014**
		(0.006)		(0.006)
Constant	0.190***	0.192***	0.242***	0.243***
	(0.001)	(0.000)	(0.001)	(0.000)
Individual FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
Individuals	25,306	25,306	27,461	27,461
Observations	214,990	214,990	262,247	262,247
R-squared	0.236	0.236	0.279	0.279

Note: Standardized variables. Standard errors clustered at the individual level are in parentheses. Sample: all Twitter users who posted at least 50 tweets over the full period and who posted at least 10 hate speech tweets in the pre-period. $Post_t$ 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.

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 variable 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 those 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 mod-

eration. Along these lines, 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 both the municipality and individual levels). Furthermore, we find evidence of a differential impact of the election based on the targets of hate speech. In particular, we observe a differential effect on homophobia and, although not significant, on sexism and racism, but no differential impact on political hate. The evidence is consistent with the proposed mechanism, emphasizing the update on the beliefs held regarding the social acceptability of online hate speech.

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.

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):1422–74.

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.

Barros, L. and Santos, M. (2021). Right-wing populism in the tropics: Economic crisis, the political gender gap, and the election of bolsonaro. *Discussion Papers*, (242).

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.

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. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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. J. (2019). Jihadi attacks, media and local hate crime.

Leite, J. A., Silva, D. F., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. *arXiv preprint arXiv:2010.04543*.

Metzger, M. J. (2009). *The study of media effects in the era of Internet communication*. na.

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.

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.

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.

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