FGV DAPP
DIRETORIA
DE ANÁLISE
DE POLÍTICAS
PÚBLICAS

HATE SPEECH IN DIGITAL ENVIRONMENTS

Definitions, specificities and context of online discrimination in Brazil on Twitter and Facebook

Rio de Janeiro, March 2021

Support

Embaixada
da República Federal da Alemanha
Brasilia

Department of Public Policy Analysis of Fundação Getulio Vargas

HATE SPEECH IN DIGITAL ENVIRONMENTS:

Definitions, specificities and context of online discrimination in Brazil on Twitter and Facebook

Rio de Janeiro
FGV DAPP
2021

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pela Biblioteca Mario Henrique Simonsen/FGV

Hate speech in digital environments : definitions, specificities and context of online discrimination in Brazil on Twitter and Facebook / Coordination Marco Aurélio Ruediger, Amaro Grassi. – Rio de Janeiro : FGV DAPP, 2021.
1 recurso online (60 p.) : PDF

Inclui bibliografia.
ISBN: 978-65-86845-14-3

1. Mídia digital - Censura. 2. Políticas públicas. 3. Discurso de ódio na Internet. 4. Liberdade de expressão. 5. Redes sociais on-line. 6. Discriminação. I. Ruediger, Marco Aurelio, 1959- . II. Grassi, Amaro. III. Fundação Getulio Vargas. Diretoria de Análise de Políticas Públicas.

CDD – 302.23

How to cite

RUEDIGER, M. A.; GRASSI, A. (Coord.). **Hate speech in digital environments : definitions, specificities and context of online discrimination in Brazil on Twitter and Facebook**. Policy paper. Rio de Janeiro: FGV DAPP, 2021.

EDITORIAL STAFF

Founded in 1944, Fundação Getulio Vargas was born from the goal to promote socioeconomic development in Brazil through the formation of qualified administrators in the public and private sectors. Over time, FGV has expanded its activities into other fields of knowledge, such as Social Sciences, Law, Economics, History, Applied Mathematics and International Relations, becoming a reference in quality and excellence with its ten schools.

Edifício Luiz Simões Lopes (Headquarters)
Praia de Botafogo 190, Rio de Janeiro
RJ - CEP 22250-900
P.O. Box 62.591 Zip Code 22257-970
Tel (21) 3799-5498 | www.fgv.br

Director
Marco Aurelio Ruediger
PhD in Sociology and Director of Public Policy Analysis of Fundação Getulio Vargas (FGV/DAPP). His main fields of interest are political sociology, communication, social networks, and technological innovation and its impacts on democracy. He is a consulting member of the D4D initiatives of the National Democratic Institute and Social Science One. He is currently leading the creation of the School of Communication, Media and Information of Fudação Getulio Vargas and of the project "Digitalization and Democracy in Brazil", carried out with the support of the Ministry of Foreign Affairs of Germany and of the German Embassy in Brasília.
E-mail: marco.ruediger@fgv.br

FGV DAPP
(21) 3799-6208
www.dapp.fgv.br | dapp@fgv.br

Research Coordination
Marco Aurelio Ruediger
Amaro Grassi

Researchers
Dalby Hubert
Danielle Sanches
Eurico O. Matos Neto
Luiza C. Santos
Lucas Roberto da Silva
Polyana Sampaio Barbosa
Renata Tomaz

Technical Review
Renata Tomaz

Graphic Project
Luis Gomes
Daniel Cunha

CONTENTS

EXECUTIVE SUMMARY

Online hate speech practices are a challenge to democracies, particularly because they touch on issues related to one of its pillars, namely freedom of speech. Seeking to understand the dynamics of negotiation between the safety of protected groups and the insurance of freedom of speech, this study investigated the public debate on Twitter and Facebook about hate speech and censorship. For that purpose, we will first present an overview of the definitions and legislations informing the understanding about hate speech, discussing the specificities of the online context and the forms of definition and self-regulation proposed by the platforms. Afterwards, based on a data collection on social networks, specifically Twitter and Facebook, we will analyze the public debate about hate speech and censorship, especially considering the peaks of discussion about these topics and events associated with them. Lastly, we will identify the major actors and interactions happening at the highest peak of debate about hate speech and censorship, using graphs that present a structural analysis of social networks. The study used bibliographic and documental sources as well as data collected on Facebook and Twitter, which were then analyzed both qualitatively and quantitatively.

Keywords: hate speech; digital platforms; censorship; freedom of speech; social networks.

SUMMARY OF RESULTS

- ➔ Our theoretical review combined with the analysis of the platforms revealed that, although there are guidelines for sanctions of accounts that publish offensive content, it is very difficult and complex to detect this type of post;
- ➔ The research focused on collecting data about hate speech and censorship, gathering around 11 million posts on Twitter and 34 thousand posts on Facebook between November 2020 and January 2021;

- ➔ Regarding the debate about hate speech on Twitter and Facebook, the posts focused on topics such as misogyny and racism. The peak of posts about the topic on both platforms happened on November 20, 2020, Black Awareness Day. The number of posts increased due to the discussion about structural racism, with the news that a black customer had been beaten to death in the previous day at a Carrefour supermarket in Porto Alegre (RS);
- ➔ The discourse about censorship and freedom of speech increased due to the sanctions imposed by the platforms for accounts such as that of the former president of the United States, Donald Trump, whose posts questioning the results of the presidential elections were associated with the storming of the United States capitol by extremist groups. The suspension of accounts mobilized the networks to discuss the approach taken by the platforms regarding the censorship of opinions expressed by users, both on Twitter and on Facebook.

PRESENTATION

The maintenance of democratic regimes relies, among other factors, on a healthy balance between freedom of speech and a fight against intolerance, especially intolerance targeting minority groups. An equally important factor in addition to each individual's possibility to express freely is the maintenance of a democratic environment in which discourses attacking diversity are adequately contained, such as hate speeches or violent speeches. The possibilities for public expression generated by social media bring new challenges to this already controversial equation.

Different lines of thought in law understand the role of the State in maintaining this environment in different forms: countries such as the United States, for instance, protect freedom of speech above all else; other countries, such as Germany, have more strict limitations about what can be said. In the online context, it is also necessary to include the positions of the digital platforms where these discourses circulate, which are usually opposed to state regulations and defend self-regulation.

Considering the need for an ecosystem where freedom of speech and security can coexist in all groups, the issue of hate speeches is critical when talking about digital democracy. This issue has been increasingly approached both in larger studies, such as those investigating disinformation (BARBOSA et al., 2020), and in more specific mappings (NERIS et al., 2020), which analyze its constitutive aspects. In this study, we will present an overview of the definitions and legislations regarding hate speech, discussing the specificities of the online context and the forms of definition and self-regulation proposed by the platforms. Based on a data collection on social networks, specifically Twitter and Facebook, we will analyze the public debate about hate speech and censorship, taking into account especially the peaks of mentions about these topics and events associated with them. We will also analyzed the major actors and interactions mobilized from the moments of high activity in the debate about hate speech and censorship, using graphs that present a structural analysis of social networks.

This is the third policy paper launched through the project Digitalization and Democracy in Brazil, a partnership between the Department of Public Policy Analysis of Fundação Getulio Vargas (FGV DAPP) and the German Embassy in Brazil. The project will run until 2022 and take advantage of FGV DAPP's experience with research applied to social networks and expertise in public policy analysis developed over the past decade.

The goal is to develop strategies to fight and understand the new and imminent challenges faced by the Brazilian democracy, which is now represented by an established scenario of digital extremism and its consequences for society. To that end, the project focuses first and foremost on the strengthening of democratic values and on the development of the necessary mechanisms and knowledge to promote a digital public debate that is consistent and fruitful.

METHODOLOGY

This study used bibliographic and documentary sources, as well as data collected in platforms of digital social networks – specifically Facebook and Twitter. The bibliographic research informs the first and second parts of this document, presenting the state of the art in the discussion around hate speech in digital environments. The documentary research informs the first part of the study, focusing on the legal guidelines in force in Brazil, and the second part of the study, focusing specifically on the positions of the platforms regarding the hate speech that happens in their network. To map the concepts of hate speech and the behavioral guidelines established by the platforms, we studied the Terms of Use and the Community Guidelines on Twitter, Facebook, Instagram and YouTube. Both documents are openly available for consultation in specific tabs in each platform's website, and are periodically updated. Additional links found in these websites were also studied whenever deemed relevant. The registers found in this study refer to the Terms of Use and Community Guidelines in force during February 2021 in their respective platforms.

Regarding the analysis of data in social networks that completes this study, our empirical approach included automatic methods to collect and classify posts in social network platforms. First, aiming to collect data from Twitter, we developed a list of terms that would be able to deal with the thematic scope that is the focus of this study – that is, hate speech and, ultimately, freedom of speech. Part of this collection was carried out in the Trendsmap application, which allows the real-time monitoring of posts with a high level of engagement on Twitter, both in terms of shares and in terms of comments and replies. In addition, this list of terms was used to filter posts made during the analysis period – between November 16, 2020 and February 6, 2021 –, which were stored in a PostgreSQL database.

Based on the initial data collection on Twitter, we classified the data obtained using linguistic rules developed about the topic (RUEDIGER, 2017). This stage included the development of search syntaxes driven by the two central topics in this study, namely “hate speech” and “freedom of speech” (and “censorship”). Search syntaxes are a set of linguistic values articulated through boolean (or logical) operators based on semantic and pragmatic elements with which a given topic is usually discussed. These search syntaxes were executed through a textual search on the database. In turn, to collect and classify the data on Facebook, we adapted the search syntaxes developed for the Twitter data classification – for the same two topics, that is, “hate speech” and “freedom of speech” – in order to execute them on Crowd Tangle, a post monitoring application on Facebook associated with the platform itself.

Although the methodology we adopted relies on the construction of structured linguistic rules, the target topic of our research (the debate about hate speech) is extremely complex. For that reason, we sometimes found posts using a more simplified vocabulary, but with different meanings. This challenge led us to establish some filters after collecting the data, focusing on the peaks of mentions to analyze the level of engagement in the debate. Therefore, the qualitative analysis of these moments when the discussion intensifies has become a methodological strategy to deal with the limitations of the results of data collection based on the development of search syntaxes.

THEORETICAL CONCEPTS AND DEFINITIONS

1. What is hate speech?

It is extremely difficult to arrive at a definition for hate speech, either online or offline. This topic has been widely debated by academics in different fields of knowledge, jurists and legislators, with no consensus. It can be said that hate speech is motivated by prejudice, offenses and aggressive discourses against a person or group due to their characteristics (COHEN-ALMAGOR, 2011; FARIS et al. 2016). However, Sellars (2016) explains that, even though there is extensive literature about the causes and effects of this type of discourse, there is still a literature gap about the definition and systematization of the term.

A large variety of topics may fall under the definition of hate speech, for example, slander and insults that are easily identifiable. However, language has many nuances, and content may not always be necessarily considered hate speech by the speaker or target recipient. Another point we must highlight is the use of code to express hatred, especially in online communities, which makes it even harder to detect this discourse. The literature about this topic reveals the existence of so-called violent discourses, which incite violence against vulnerable groups in the offline environment (SIEGEL, 2020).

Therefore, the existing definition of hate speech can be extremely general, dealing with a variety of discourses targeted at a given group or individual based on their physical characteristics or gestures that are different from established standards (PAREKH et al., 2012). On the other side we have definitions that result in prejudice. The more specific definitions of hate speech state that it is directly associated with inciting mass violence or physical harm against an external group (BENESCH, 2013). This duality of concept (general definition and specific definition) reflects the difficulty of obtaining an adequate definition for the plurality of phenomena that may be considered hate speech.

Despite the inherent complexity of the concept and the challenges imposed by this issue, in this study we adopted the definition proposed by the "Guide for analyzing hate speech", according to which hate speeches "are manifestations that negatively evaluate a vulnerable group, or an individual that is a member of a vulnerable group, in order to establish that they are less worthy of rights, opportunities or resources than other groups and individuals that are members of other groups, which consequently legitimizes the practice of discrimination or violence" (LUCCAS; GOMES; SALVADOR, 2020, p. 4). According to Nandi (2018), the groups targeted by hate speech are commonly members of minorities who are socially and economically vulnerable, such as black people, indigenous people, homosexuals, women and religious minorities. Still according to the author, the circulation of hate speech is related to intimidating the victim or victims to keep them from exercising a right; for this reason, it is directed towards individuals or groups that become a highlight in academic debates, judicial decisions or legislative debates. According to Nandi, hate speech always has a political bias, since the attack has the function of delegitimizing pleas.

An important point to be made in this debate are the specificities of online hate speech, because it involves instigators, different types of targets, and a variety of motives and tactics. In addition, there is the issue of the aggressors having their own followers, increasing their reach and their ability to target certain individuals or groups. This is a point of concern for academics in the debate about hate speech, because an utterance that incites violence spoken by a single individual is different when used as a type of "hate campaign" coordinated by a digital crowd (SELLARS, 2016).

In this sense, the discussions are always accompanied by the relationship between freedom of speech and hate speeches. Although this debate belongs to the field of Law, it can be said that the freedom of speech provided for in international treaties and in the national legislation has been increasingly interpreted, by different areas, as a non-absolute right that is limited by other guarantees, with responsibilities and restrictions in terms of individual freedom (LUNA; SANTOS, 2014). However, there is rarely a direct connection between freedom of speech and violence, since, according to Antoine Buyse (2014), it is very difficult to agree on the legal and social limits of freedom of speech.

Therefore, fear speech is different from hate speech. More than directing hateful words towards individuals belonging to certain groups, fear speech is usually targeted at members of the social group of the speakers themselves. The objective is to create an internal feeling of fear against individuals outside of that community. Through this mechanism, fear legitimizes violence. Therefore, the emergence of violent discourses is enabled by the feeling of fear spread inside social groups. Antoine Buyse, a professor of Human Rights at Utrecht University, explains that, in contexts in which there are strict laws against hate speech, individuals tend to avoid direct manifestations inciting violence against certain social groups, and start adopting more indirect ways to promote hostility against certain communities.

It is interesting to point out that the fear caused by violent discourse in minority groups or targeted groups is often worse than violence itself (such as armed or physical violence). Fear speech is a resource used as preventive violence, establishing positions among those who adopt this discourse and those who are targeted by it.

One important aspect about the issue of fear speeches is the impact they have on a certain group or on an entire population. The messages we see daily in media channels (television, printed or digital) about criminal activity in Brazilian cities or attacks orchestrated by terrorist groups in Europe cause the discourse about exterminating these groups – the idea coming from the common belief that "a good criminal is a dead criminal" – to be adopted by the population, surpassing the notions of justice and imprisonment. According to Anna Simons and John Mueller (2001), this is due to the incorporation of fear speech from one group to another, which can lead to extreme reactions and cause violence to be accepted for the sake of the well-being of all[1].

Another significant conceptual distinction is the understanding of what hate speech is and what extremist speech is. According to a report published by UNESCO (ALAVA; FRAU-MEIGS;

[1] The example given deals with issues that affect the whole society. However, any group that is persecuted, oppressed or affected by the actions of another group can use fear speech or incorporate it, accepting physical and armed violence as a resource.

HASSAN, 2017), the use of extreme violence is more commonly detected on the internet, as it is a vector that spreads the radicalization of extremist discourses and ideologies. However, according to that same report, despite allowing the proliferation of these discourses, there is no causality link between radical discourses on the internet and real acts of extreme violence. This debate about to what extent public or private messages in the virtual environment are transferred to daily social practices of certain groups invites a reflection on whether social media companies are responsible for fighting these discourses, moderating the content of publications and demanding the creation of laws allowing them to regulate the online actions of their users, or a joint strategy is required to regulate this type of content.

Lastly, hate speech is not a homogeneous phenomenon. In fact, discriminatory discourses and practices against minorities are directly related to the social and political contexts in which they are expressed. The different forms through which hate speech is manifested in different contexts is a factor that makes the task of formulating strategies to fight the dissemination of hatred in online or offline environments even harder. Therefore, it is more adequate to approach the phenomena as “hate speeches”, in the plural, since this plurality and connection to the context are defining aspects of its practice in social contexts.

2. Legal approaches and classifications of hate speech in the world and in Brazil

The increasingly complex notion of hate speeches, including its use in the plural, is highly dependent on the efforts made in the legal field. International treaties, particularly those created after the two great world wars, have been feeding discussions and debates that demand that States create legal mechanisms to restrain and punish actions motivated by the desire to disqualify, silence, ostracize and dehumanize vulnerable groups. Although international treaties do not provide definitions of hate speech, they establish important parameters to identify it and, consequently, fight its direct and indirect effects.

These letters, pacts and declarations are based on the principle of human dignity, which was strongly evoked after the “untold sorrow” of war, according to the preamble to the Charter of the United Nations, the document that founded the United Nations (UN, 1945). The text defends the “respect for human rights and fundamental freedoms for all without distinction as to race, sex, language, or religion”. This idea that the human condition is sufficient to ensure the equality of rights and the enjoyment of freedoms provided a base for the legal understanding of the fight against hate speeches. The Declaration of Human Rights (UN, 1948) expands the fight against discriminations in the context of “political or other opinion, national or social origin, property, birth or other status” (Article 2) and advances actions “against any incitement to such discrimination” (Article 7) or attacks against “honor and reputation” (Article 12). In this sense, it also fights these discourses.

In the second half of the 20th century, the opposition to the crimes of prejudice and to incitation mechanisms intensified. The International Convention on the Elimination of All Forms of Racial Discrimination (BRAZIL, 1969), adopted by the UN in 1965, introduced the term “racial hatred” into the debate, which quickly expanded into the national and religious contexts in the International Covenant on Civil and Political Rights (BRAZIL, 1992a), adopted by the UN in 1966, and in the American Convention on Human Rights (BRAZIL, 1992b), adopted by the Organization of American States in 1969. The three documents defended that racial, national and religious hatred should become a legally punishable crime. In 1992, Brazil enacted both the International Covenant and the American Convention, granting them the status of laws in the country.

The Declaration and the Programme of Action of the third World Conference against Racism, Racial Discrimination, Xenophobia and Related Intolerance (2001) demanded that countries adopt less punitive and move preventive and protective measures. It recognized the need for the participation of multiple actors in the fight against “hatred”, and listed its different forms of manifestation as well as its main targets. It highlighted the role of the different types of media, especially digital, in the dissemination of ideas that incite hatred towards vulnerable groups, on one hand, and in the promotion of actions to fight them, on the other hand. In addition, Article 147 of the Programme of Action demanded that

countries, while guaranteeing the right to freedom of opinion and expression, also adopt and apply “to the extent possible, appropriate legislation for prosecuting those responsible for incitement to racial hatred or violence through the new information and communications technologies, including the Internet”.

Freedom of expression is recognized by the international treaties and by national legislation, such as the Brazilian Constitution of 1988, as a fundamental right to establish a complete democratic regime. However, the increase in the possibilities to spread hate speeches in the digital environment has driven discussions in the first decades of the 21st century about the role of the State in ensuring it, but also limiting it when it harms principles such as human dignity. Those who oppose this type of intervention believe that interfering with the free manifestation of ideas, including hate speech, violates this freedom (DWORKIN, 2006; 2009), which weakens the pillars of democracy. In turn, those who defend that the State should impose limits to the freedom of expression argue that this is necessary to fight intolerance (WALDRON, 2012), which is an obstacle to establishing human dignity and, consequently, the democratic spirit itself. The former position is usually associated with American theorists, who are usually based on the principle of neutrality of the State. European countries, which experienced two great wars and the holocaust, tend to have a more strict position regarding the limits of freedom (BRUGGER, 2007; CONSANI, 2015).

Therefore, it is not simple to provide legal norms to deal with the social dynamics involved in hate speeches and the contexts from which they emerge. The so-called soft laws, also known as soft norms, provide support to overcome the limitations found in the international treaties. They are provisions that seek to resolve impasses in the context of international law, such as in cases when freedom of expression is evoked in order to protect hate speech. That is the case of the Special Rapporteur for Freedom of Expression, created by the Organization of American States to provide recommendations based on the provisions of the American Convention on Human Rights.

Brazilian legislation is more aligned with the European concept, prioritizing equality of rights and human dignity over the right to freedom of expression. In Brazil, discrimination by race, color, ethnicity, religion and national origin is a crime under Law 7.716 (BRAZIL, 1989), which also makes the dissemination of Nazism punishable. Bill No. 7582/2014, which criminalizes hatred and intolerance, expands the scope of the target groups to include social class and origin, migrant, refugee or internally displaced status, sexual orientation, gender identity and expression, age, homelessness and disability. Until these criteria are included in legislation, cases of hate speeches involving these groups are usually ajudged following the principles of the Federal Constitution (BRAZIL, 1988), such as human dignity (Article 1) and equality before the law (Article 5), which includes gender equality and not being subjected to inhuman or degrading treatment (Article 5).

It is important to mention that the Penal Code (BRAZIL, 1940) criminalizes slander, aggravated by Law 10.741 (BRAZIL, 2003) when including “elements referring to race, color, ethnicity, religion, origin, elderly status or disability”. Another important legal provision to fight hate speeches in Brazil is Law 12.735 (BRAZIL, 2012). It expands the legislation that criminalizes prejudice and discrimination by including “actions carried out through the use of electronic, digital or similar systems”, indicating an effort to make the current legislation applicable to digital contexts.

In the context of the digital environment, the most important Brazilian legal provision is the Civil Rights Framework for the Internet, which came into force with the approval of Law 12.965 (BRAZIL, 2014). Although the text states in Article 2 that “The matter of internet use in Brazil is based on the respect for freedom of expression”, the single paragraph in Article 3 clarifies that “The principles expressed in this Law do not exclude other provisions of the national legal system related to this matter, or international treaties of which the Federative Republic of Brazil is a part”. In this sense, Brazil demonstrates a wider social and legal understanding that the guarantee of freedom of expression is not a basis for free hate speech.

More recently, the Bill 2.630 of July 3, 2020 is pending in the Congress. It institutes the Brazilian Law on Freedom, Responsibility and Transparency on the Internet – the so-called "Fake News Law". The bill was authored by Senator Alessandro Vieira (Cidadania-SE) and approved by the Senate in 2020, and is currently pending at the Chamber of Deputies. Some of the changes proposed in the original text are related to the topic of hate speech. One of them demands that social media platforms disclosure publicly available trimestral reports on the accounts and contents removed, and the permission to exclude content without notifying the user in cases of incitation of violence and immediate damage that is difficult to repair[2]. The proposal has been criticized, especially regarding the establishment of vigilance mechanisms and the potential effects for the internet ecosystem, as discussed by Coletivo Intervozes and Revista Piauí.

Therefore, the legal approaches and classifications mentioned above are examples of important understandings in the fight against the production and dissemination of hate speeches in network interactions in the Brazilian context. The idea that freedom of speech is not an absolute right prevails in the current and proposed legal mechanisms; therefore, it must be guaranteed in line with fundamental rights and based on constitutional principles such as human dignity.

3. Hate speech in digital environments

The issue of hate speech precedes the internet. However, the proliferation and intense use of online social media brings specific issues into its discussion and the fight against it. Until recently, online hate speech practices were considered occasional activities. However, in the past years, their prominence and presence in mainstream spaces of the

[2] So far, in Brazil, the Civil Rights Framework for the Internet (BRAZIL, 2014) allows the removal of content by the platforms whenever a publication violates the community guidelines, which are signed by the user upon account creation. Anything not included in the guidelines cannot be removed, even with a report, except under a court order. In Germany, for example, a controversial law that is specific to the administration of illegal content in the online context, the Network Enforcement Act, provides that it is the duty of digital platforms to remove hate speech content within 24 hours, risking expensive fines (SILVA et al., 2019).

internet[3] have made this topic increasingly visible. The growing offline consequences of coordinated online hate speech actions have been intensifying this situation, as they realize attacks (SIEGEL, 2020) such as the Suzano school shooting in the metropolitan region of São Paulo.[4]. In this section, we will debate how hate speech gains new nuances when manifested in digital environments. We will start by indicating some of the characteristics of digital environments that may contribute to the practice and circulation of hate speeches, considering the specificities of social media and the forms of human organization that happen inside them. The structure of the social networks makes it difficult to restrict the propagation of these narratives, creating a challenge for both national legislation and the self-regulations of the platforms, which we will approach in the next section.

To adequately assess the first issue, we must first understand that online and offline hate speech are different from one another. Systematic literature reviews indicate that some characteristics of digital media are important components of their proliferation, namely: anonymity, invisibility, the creation of communities around affinities (including hatred) with no geographic barriers, low investment of time and money to propagate this type of discourse, and the instantaneousness enabled by this type of media (BROWN, 2018).

The combination of these factors intensifies online discrimination, offering new possibilities for communities driven by hatred, and partially explains the presence of these discourses in this context. The anonymity of users, even if partial (except when considering the deep web), is particularly important in removing barriers of accountability for acts of hatred in the online context and in decreasing the possibilities for reactions or physical

[3] We define mainstream digital platforms as those used the most by Brazilians, and not those belonging to particular niches of interest or even the platforms used the most by the hate groups to organize themselves (such as the chans). According to a research by We are Social with Hootsuit, the ten social networks used the most by Brazilians in 2020, in descending order of user numbers, are Facebook, Whatsapp, YouTube, Instagram, Facebook Messenger, LinkedIn, Pinterest, Twitter, TikTok and Snapchat.

[4] A school shooting that happened on March 13, 2019 at the Professor Raul Brasil State School in the city of Suzano, São Paulo. Two shooters, Guilherme Monteiro e Luiz Henrique de Castro, who were former students at the school, killed five students, two employees, and Monteiro's uncle. Afterwards, one of them killed the other and then committed suicide. They were both users of a chan – a type of anonymous forum –, where they discussed and received guidance for the attack. The involvement of hate communities present in chans and their impact on real situations was widely discussed by the Brazilian media.

confrontations between aggressors and victims. In addition, the invisibility created by the lack of a visual presence of the aggressor and victim makes executing the attacks easier, since their effects on the victim are not visible to the attacker. Therefore, online hate speech, with its partial invisibility and anonymity, *may seem* less real, with fewer implications than it actually has (BROWN, 2018).

The role of the internet in creating communities based on affinities, cultivating feelings of belonging around characteristics or interests shared by individuals of different geographic origins, also has implications for the proliferation of hate speech. The ease of access to digital communication resources, which saves time and money when organizing and spreading hate speeches, is also a significant issue (BROWN, 2018). Hate groups use the internet systematically to recruit and expand their collaborators, cultivating communities and values that enable the strengthening of identities in the digital context (WEAVER, 2013). Inside these spaces, it is possible to feed the community feeling of those involved, promoting closer ties and identities and leading to a potential increase in the reach of hate groups from the engagement of their members (BOWMAN-GRIEVE, 2009). This engagement could be seen on social networks in the previously mentioned hate campaigns, for example, with the coordination of efforts by members to demonstrate regarding a common topic or to attack a given actor on the network, simultaneously.

Lastly, the possibility to act according to one's impulses also creates a type of proliferation of hate speech that happens incidentally, with no time for careful thinking about the action. This time for careful thinking usually happens in traditional media outlets due to the costs of transaction in the production and the circulation of these types of discourse, such as producing, printing and publishing a pamphlet. Today, it is easy to pick up one's phone, open the Twitter app, write a hate message and press send. Therefore, the way the internet works encourages this type of spontaneous hate speech (BROWN, 2018).

Another characteristic of digital platforms that may contribute to the circulation of hate speeches is its operating logic based on algorithms, in which the management of the visibility of content is directly related to the user's preferences (GILLESPIE, 2018) – this is

what some authors call “echo-chambers”[5]. Since the timelines prioritize content that agrees with the user’s opinion, algorithms can also produce a distorted perception about social issues and induce polarization through the creation of ideological bubbles. The way the algorithms work to build these customized content visibility regimes is fully disclosed by the digital platforms, constituting opaque systems for most users (JURNO; D’ANDRÉA, 2017).

According to this argument, by aggregating people with similar thoughts, positions and opinions, the algorithmic logic of social networks can induce users into having a biased view about the environment of opinions predominating in a society. If what I can see through my social networks is that most people accept (and often reward) certain types of social behavior, I am more likely to imagine that that behavior can be considered a socially adequate standard. Inside this logic, there is a risk that hate speeches can be interpreted as socially acceptable by users in a given community when they are exposed to this type of content in their timelines. Therefore, people may tend to feel comfortable posting hate speeches when they see that this type of behavior can lead to social rewards inside the networks (likes, shares, interaction, etc.).

Unlike the content control that happens in mass communication media, where checking happens before publishing, user discourses in digital platforms are checked only after their publication, and usually after other users report them. This also happens due to specificities of the digital context: the volume of massive content produced by thousands of users everyday makes previous checking an impediment. In addition, the notion of freedom of speech remains as a major value on the digital platforms, as we will see in the next section. Therefore, hate speech in mainstream online platforms is fought only after it has been circulated. In turn, in the offline context, hate speech tends to stay outside the circle of large media outlets, circulating only in a marginal way.

[5] An opposed perspective is composed by authors who defend the ideas of selective exposure and incidental exposure, in which users of social media are exposed to content that opposes their points of view. See more about this debate in (2015).

Another important difference between the propagation in digital platforms and in traditional media companies is that, in digital platforms, the companies do not produce the content circulating in the environment, while traditional media companies do. For that reason, it is not possible to simply transpose or compare the ways hate speech operates and is controlled and fought in the two contexts. Consequently, the responsibilities are also not the same. However, even though digital platforms do not produce discursive content, they create the environment where these messages are propagated, and obtain financial gains with the interaction and attention of users because of them. Therefore, digital platforms are also not like the usual public spaces, since they are companies, or private spaces that are partially responsible for what happens inside them. They are not content producers in the traditional sense of the media, but they are not completely public discussion spaces either, and this makes the solutions to fight hate speech (and other topics) more complex.

Previous studies have shown that the proportion between hate speech content in mainstream platforms is low when compared to the total amount of content circulating inside them. However, the strategies used to spread this discourse generate visibility, which means a large part of individuals interacting online has been exposed to this type of content. Communities that organize attacks based on hate speech mostly target public people, such as artists, politicians and journalists, and they act in a coordinated way, expanding their reach (SIEGEL, 2020).

Today, however, systematic literature reviews indicate that most studies about the topic of online hate speech focus on communities organized inside niche platforms. For that reason, our knowledge about actors that proliferate hate speech in an informal way, particularly in widely used platforms, is still limited. Aiming to contribute to filling this gap, this report will present a temporal study in the Brazilian context about hate speech in the platforms Twitter and Facebook. According to data from a survey carried out by We are social and Hootsuit in 2020, Facebook has 130 million Brazilian users, and Twitter has 16.6 million. Together with YouTube and Instagram, they are the major networks where ideas are openly expressed in the public debate inside the Brazilian digital context.

It is important to establish the difference between the definition of hate speech and the categorization of the types of hate speech. The definition is closer to a concept; therefore, it is abstract and can encompass several specific contexts. The classification, on the other hand, is the process of determining what is and what is not hate speech based on real events, supported by the conceptual definition (be it legal, academic, or operational inside the platforms). Therefore, while a definition of hate speech may seem conceptually clear, it can become obscure in concrete situations that are based on context, uses of language, behaviors in certain cultures, and even specific linguistic appropriations and meanings inside communities.

This situational dynamics causes a lot more practical challenges in detecting and fighting hate speech in the platforms than the conceptualization. In digital platforms such as Twitter and Facebook, some subjects are priority targets of hate speech due to specific characteristics such as sexual orientation, social class, gender, ethnicity and physical features (SILVA et al., 2016). Some of these characteristics are understood by the platforms as requiring protection, in line with legal frameworks, even though these characteristics are not always the same.

4. Guidelines of digital platforms about hate speech

As discussed in the previous section, online hate speech has its own specificities. In addition to the particularities coming from the communication media itself, the interactions in each platform are also shaped by their Terms of Use. The Terms of Use of an online service are documents that explain the way platforms work and the rules users are subjected to while using them. When we use a service, we are subjected to and agree with its terms of use – even though most users do not read this service contract[6]. Social media platforms usually indicate in their Terms of Use that users must follow the Community

[6] Consent in digital environments is based on the principle of "Notice and Choice", which means the services notify their users about potential risks involved in using the platform, and users indicate their approval and consent by clicking on "I accept". To learn more about this principle and its implications on the debate about privacy in digital environments, see Sloan and Warner (2014).

Guidelines, and that they may be penalized if they do not. Therefore, we will describe, compare and analyze the community guidelines of the most used digital platforms in Brazil, which include the definitions established by the platforms for hate speech.

The Terms of Use and Community Guidelines are sets of rules that establish the behavior expected from users, and change from platform to platform. Those are the documents where we can find what a social network understands as hate speech, dangerous or violent discourses, as well as their tolerance regarding these issues and the way the negotiate users' expectations of freedom of speech and their safety, especially those who are members of protected categories. Although some platforms base themselves on scientific research to propose measures regarding this issue, such as Facebook and Twitter, there is no standardized or common understanding among all platforms.

Despite the specific efforts to create automated ways to detect hate speech on social networks, the primary way this works in the platforms analyzed is through reports made by other users. This means any user can report content from a publication that violates an aspect of the Community Guidelines, and it will be analyzed inside its context by the moderators. Since 2013, Twitter, Facebook and YouTube have been signatories of the agreement to fight hate speech led by the Anti-Defamation League, a non-profit organization from the United States. Based on the document Best Practices for Responding to Cyberhate, the three digital platforms are committed to some practices, including a commitment to analyzing hate speech reports and stories in a timely manner, offering users a clear explanation of how they moderate content, applying the provided sanctions in a consistent and fair manner, and offering simple ways to report hate content (SILVA et al., 2019).

This agreement is one of the drivers of the changes and improvements implemented by digital platforms since 2015 in the fight against hate speech. A few of the changes made over time by some digital platforms include providing clearer terms of use and community guidelines, implementing content moderation reports, and developing automated and proactive techniques to detect hate speech. Other factors also contributed to these

improvements, such as new laws about online hate content (such as in the case of Germany), data leaks that became public, terrorist attacks motivated by online groups, and governmental interference (SILVA et al., 2019).

We will now present a summary of the current Community Guidelines on Facebook, Twitter, Instagram and YouTube, to understand the level of conceptual detail and the ways to deal with hate speech and its variables in the online context. Despite the efforts in legislation both nationally and internationally, and despite the Community Guidelines and Terms of Use in the platforms, the relationships between the freedom of speech of individuals and hate speeches are not always clear and can have different understandings, as discussed previously. The balance in this relationship is one of the major issues addressed by the platforms when dealing with this subject: how to guarantee the safety of protected categories while interfering as little as possible with users' freedom of speech.

In addition, the difficulties imposed by the inherent context of hate speech are an issue raised by Twitter, Facebook, Instagram and YouTube in their pages: words historically associated with degrading meanings may acquire new meanings in certain social groups; new forms of verbal incitation or aggression may appear, as well as deciphered forms. Therefore, culture is an important element to correctly detect these types of discourse.

Overview

Facebook, Twitter, Instagram and YouTube are among the ten major digital platforms used by Brazilians to produce and consume content. By investigating their Terms of Use and Community Guidelines, we can see that all four of them have in their guidelines some level of restriction or repudiation of hate speech. In some cases, there are also guidelines against violent, extremist or dangerous discourses. However, the level of detail, conceptualization of terms and exemplification of guidelines varies from platform to platform.

Instagram is the platform with the lowest level of detail regarding these aspects, presenting more succinct and generic Community Guidelines. However, it informs users that they are also subjected to the Facebook policies, since the platforms belong to the same group. The lack of specificity is an indication of an even bigger difficulty to moderate, which is delicate due to the challenges of context and freedom of speech.

Another aspect shared by the platforms analyzed is the mention to protected characteristics or categories and specific groups, which may or may not be listed by each platform. In Instagram, they are not listed. Facebook, Twitter and YouTube provide a list of the categories or groups considered protected, often indicating age, gender, sexual orientation, ethnicity, race, religion, and immigrant status. However, some of those categories are not a consensus, appearing in only one or two of the platforms. Facebook and Twitter include physical characteristics or diseases; Twitter includes marginalized and historically underrepresented communities; YouTube includes social class and war veterans. The platforms agree on three of the protected categories, which represent groups that are more likely to be targeted by hate speech according to the literature: sexual orientation, gender and ethnicity (SILVA et al., 2016). However, two of the characteristics considered to be vulnerable by the literature are present in the guidelines of only one of the four platforms: physical features and social class (SILVA et al., 2016).

Additionally, of the four platforms analyzed, only YouTube indicates in its Community Guidelines that, in addition to reports of inadequate content made by users, they also have their own monitoring through an automated signaling system to detect content that violates the community guidelines. It is possible that this higher concern on YouTube about verifying the content posted is related to video monetization, since publications with offensive content are not eligible for monetization in that platform. The other platforms inform only that they analyze content signaled by users inside a context, and do not indicate other initiatives to control and monitor content in their Community Guidelines. However, both Facebook and Twitter use automated mechanisms to detect content that violates their community guidelines (SILVA et al., 2019).

We will now look into the Community Guidelines for Facebook and Twitter in terms of hate speech and related topics, since these are the platforms from which we collected data to inform the analyses found in this study.

Facebook

Facebook mentions in their Community Standards that their "commitment to expression is paramount", signaling that they are in line with the American perspective of preserving freedom of speech. The argument of freedom of speech is used to justify their standards regarding hate speech "to ensure that everyone's voice is valued", including the voices of the so-called protected categories.

Hate speech is defined as "a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability". Attack is defined as "violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation". Therefore, the platform is able to provide a clear understanding of gate speech, even though, in practice, the evaluation of this understanding by moderators relies on contextual or even personal factors. By indicating the protected characteristics, the platform also indicates the groups more susceptible to this type of attack and their focus of attention when moderating.

The guidelines indicate three different levels of hate speech: level 1 includes violent discourse, degrading comparisons, mockery and negationism of historical situations regarding protected categories; level 2 includes statements of inferiority and belittling declarations regarding protected categories; and level 3 includes actions of segregation or encouraging the exclusion of protected categories. In a separate tab, Facebook also indicates that they do not tolerate the exaltation and celebration of violence and explicit content in any format.

Twitter

Facebook mentions in their Community Standards that their "commitment to expression is paramount", signaling that they are in line with the American perspective of preserving freedom of speech. The argument of freedom of speech is used to justify their standards regarding hate speech "to ensure that everyone's voice is valued", including the voices of the so-called protected categories.

Hate speech is defined as "a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability". Attack is defined as "violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation". Therefore, the platform is able to provide a clear understanding of gate speech, even though, in practice, the evaluation of this understanding by moderators relies on contextual or even personal factors. By indicating the protected characteristics, the platform also indicates the groups more susceptible to this type of attack and their focus of attention when moderating.

The guidelines indicate three different levels of hate speech: level 1 includes violent discourse, degrading comparisons, mockery and negationism of historical situations regarding protected categories; level 2 includes statements of inferiority and belittling declarations regarding protected categories; and level 3 includes actions of segregation or encouraging the exclusion of protected categories. In a separate tab, Facebook also indicates that they do not tolerate the exaltation and celebration of violence and explicit content in any format.

Twitter discusses prohibited behaviors when using their platform in their "Rules and Policies" section, indicating three categories that are relevant for this study: The

“glorification of violence policy”, the “hateful conduct policy”, and the “violent threats policy”. Of the four social networks, Twitter offers the most detailed document regarding what is and is not allowed, and how it understands the topic.

The Glorification of violence policy clarifies that users “can’t glorify, celebrate, praise or condone violent crimes, violent events where people were targeted because of their membership in a protected group, or the perpetrators of such acts”, followed by a definition of what they understand as glorification and a list of examples of these actions. This category in the platform is directly associated with what we discussed regarding dangerous discourses.

The Hateful conduct policy focuses on groups that are more susceptible to online harassment, especially those who have intersectional characteristics: “women, people of color, lesbian, gay, bisexual, transgender, homosexuals, intersex, asexual individuals, marginalized and historically underrepresented communities”. The policy establishes that users “may not promote violence against or directly attack or threaten other people” based on protected characteristics. In addition, users are not allowed to post imagens propagating hatred, such as images historically associated with hate groups, or content referring to violent attacks against protected groups.

The Violent threats policy explains that Twitter users “may not threaten violence against an individual or a group of people”, especially in the form of a declaration of intention, including sexual violence. According to the documents, the platform evaluates some of these expressions, including variations of “I’ll kill you”, in a contextual manner, in order to exclude hyperbolic speech.

Public interest

The four platforms admit that, in special circumstances of public interest or awareness raising and education about the topic, they may allow content that does not comply with the community standards regarding hate speech to remain visible, even if they have received reports. Examples of this are posts aiming to raise public awareness regarding an issue and to encourage public discussion about the topic, provided that the user's intention of doing so is clear.

Sanctions for guideline violations

None of the four platforms provides clear and transparent details about the proceedings involved in the sanctions applied to users who publish hate speech or similar content. The platforms are vague when explaining what happens to those who violate the terms of use: Instagram is the platform showing the least details about sanctions and proceedings, and YouTube tends to provide more details on that regard. This transparency is, of course, partial, since the platform has unlimited access to the data informing their decisions and reports.

Facebook claims to analyze the severity of the violation and the user's history in the platform to make a decision, which can range from notification to restrictions of use and the complete deactivation of the profile. The user's intention is part of the evaluation both for removing the content and for the sanction applied. In case of real risk of physical harm of public safety, the authorities are notified.

Twitter claims to take into account only the user's history in the platform, indicating that the first violation causes the removal of the content and potentially a temporary restriction

of use of the account. Recurring violations cause the permanent suspension of the profile, but Twitter does not provide the number of repeated violations before this happens.

Instagram explains that the penalty for violating their guidelines is the removal of the content and the eventual cancelation of the user's account, without specifying any criteria used to inform this decision. YouTube explains that the first violation of the Community Guidelines causes a simple signaling, with no penalty. Further violations lead to the removal of the content (a video or a comment) in addition to a notification, which also limits the user's actions in the platform for a period of one or two weeks. After three notifications within a 90-day period, the channel is shut down and all its videos are deleted.

Although the platform does not explain what they take into account when analyzing reports, the details of their proceedings indicate that the criteria for applying penalties are the severity and frequency of violations. YouTube also maintains a transparency report that is easy to access about their sanctions regarding violations of community guidelines. According to the report, between October 2020 and December 2020, there were 2,055,515 channels removed. The second main reason indicated for removing a channel was "hateful or abusive" content, with 8.3% of the total. The first reason, with more than 75% of the total removals, was spam, misleading or scams. The table below demonstrates the main findings in our comparative investigation of the community guidelines for Facebook, Twitter, Instagram and YouTube.

Table 1 – Comparison of Community Guidelines on Facebook, Twitter, Instagram and YouTube

Comparison – Community Guidelines				
	Facebook	**Twitter**	**Instagram**	**YouTube**
Guidelines on hate speech	Yes	Yes	Yes	Yes

Definition of hate speech	"[...]a direct attack on people based on what we call protected characteristics."	"[...]promote violence against or directly attack or threaten other people" based on protected characteristics.	Not provided.	"[...] content promoting violence or hatred against individuals or groups based on any of the [protected] attributes".
Protected categories and groups	Race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.	Categories: race, ethnicity, national origin, sexual orientation, sex, gender identity, religion, age, and serious disease or disability. Groups: "women, people of color, lesbian, gay, bisexual, transgender, homosexuals, intersex, asexual individuals, marginalized and historically underrepresented communities".	Does not define protected categories or groups, but explains that "it's never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases".	Age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, veteran status.
Criteria for evaluating reports	Severity, intention, and user's history in the platform.	User's history in the platform.	Not specified.	Severity and frequency of violations.
Sanctions for violations	Ranges from notification to restrictions of use of the platform and the complete deactivation of the profile.	Removal of the content and potentially a temporary restriction of use of the account. Suspension of the account only after persistent violations.	Removal of the content and the eventual removal of the user from the network.	First violation: signaling with no penalty. For further violations, removal of the content (a video or a comment) in addition to a notification and limitation of the user's actions in the platform for a period. After three notifications within 90-days, the channel is shut down and the content is deleted.

Public interest	Content that violates the guidelines may be kept visible for the purposes of awareness raising, helping fight the issue or educating about the topic.	Content that violates the guidelines may be kept visible for the purposes of awareness raising, helping fight the issue or educating about the topic.	Content that violates the guidelines may be kept visible for the purposes of awareness raising, helping fight the issue or educating about the topic.	Content that violates the guidelines may be kept visible for the purposes of awareness raising, helping fight the issue or educating about the topic.

Source: Community Guidelines of each platform | Elaborated by: FGV DAPP

ANALYSIS OF THE DEBATE ABOUT HATE SPEECHES AND CENSORSHIP ON TWITTER AND FACEBOOK

In this section, we will present the analysis of the debate about hate speech on the networks. We collected posts from Twitter and Facebook because these two networks represent a significant number of users in Brazil, and have similar aspects regarding the characteristics allowed for posts and the forms of engagement.

The data collected on Twitter and Facebook were categorized into two major groups: the **debate about hate speech** and the **debate about censorship** in each platform, separately. The peak of the debate about hate speech both on Twitter and on Facebook happened on the same day, November 20, 2020. Regarding the debate about hate speech, we found that users in the social networks analyzed tend to understand the topic as the publication of offenses, both online and offline, against a social minority group. For that reason, debates about racism and misogyny are constantly detected. It is interesting to highlight that, based on the data collected, the debate about this topic on the networks is always driven by situations that happened in the offline environment, allowing the identification of the victim and/or the aggressor.

The data collected indicates that the debate about hate speech is constantly accompanied by a discussion about censorship or about attacks against freedom of expression, in line with the findings of the literature on this topic. There is a recurring idea among the comments and publications analyzed that blocking accounts due to posts considered offensive is a violation of the individual freedom to freely express an opinion. In this mapping, we found that this debate is centered around the sanctions imposed to accounts of users in the social networks, since the focus is on the measures taken by the platforms when “taking down” posts, suspending or blocking profiles, such as the case of the former president of the United States, Donald Trump.

The posts on social networks are concentrated around this discussion regarding hate speech and freedom of expression based on the data collected about this topic. Based on that data, we chose to analyze the peak of mentions, meaning the dates during the period analyzed when the volume of posts about the topics was the highest. We conducted a qualitative and quantitative analysis of the data. First, we mapped all the publications about the topic, and then we focused on the peak of publications to understand the engagement, that is, finding which post, news story or user contributed for the debate to grow.

Our analysis of the engagement focused on the reach of each post; therefore, we observed all the possibilities a publication has to involve users on Facebook and Twitter, beyond the so-called “vanity metrics” defined by Rogers (2019), that is, those that simply yield likes and followers for a certain profile. Instead, we adopted a critical analysis posture of social networks, going beyond the simple counting of metrics in publications or profiles (SANCHES, 2020). From this perspective, we developed a qualitative analysis based on cases identified through the quantitative collection of interactions, such as likes, shares and comments of a given account, in order to understand the reach of the post and the importance of the author or publication in the discussions in the network. We will now discuss the findings of this analysis in the debates that happened on Twitter and Facebook, starting with an investigation of the peaks of posts about hate speech and then about censorship on these networks.

Debate about hate speech and censorship on Twitter

Between November 16, 2020 and February 6, 2021, hate speech was the topic of approximately 11.6 million posts on Twitter. The debate was driven by the celebration of Black Consciousness Day, the assassination of a black customer by a security guard at a Carrefour supermarket in Porto Alegre (RS), the storming of the United States Capitol, and the ensuing suspension of the accounts of the former president of the United States, Donald Trump, on social networks. It became evident that the two topics driving the debate on Twitter were hate speech (the debate about structural racism) and freedom of speech (the debate about censorship). As indicated by the literature, the debate about online offenses is always accompanied by questions about the violation of individual freedoms and the censorship imposed by the platforms against opinions expressed by users.

In the period analyzed, other news stories also contributed significantly to this debate. Regarding hate speech, the harassment episode of the congresswoman Isa Penna (PSOL-SP) was permeated by a discussion about the vulnerability of women and mysogyny in different spaces[7]. Regarding freedom of expression, there was a publication by the Ministry of Health about early treatment for Covid-19, which was considered by Twitter to be misleading.

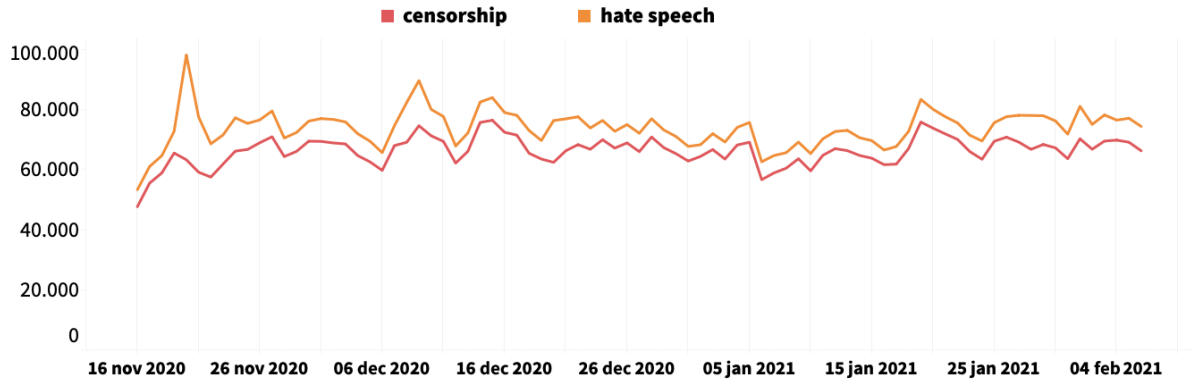
As indicated previously, the data collection about online hate speech or offenses is always accompanied by posts about the users' freedom to express freely. During the period analyzed, there was an intense debate about the suspension of accounts or the removal of content in social network websites, such as Facebook, Instagram or Twitter itself, due to violations of the platforms' guidelines. Some posts coming from profiles of politicians or bloggers with partisan, ideological or identity alignments questioned the decisions made

[7] During a section of the State Legislature of São Paulo on December 16, 2020, the congresswoman Isa Penna (PSOL-SP) was sexually harassed by the congressman Fernando Cury (Cidadania-SP). The harassment consisted of forced physical contact with an intimate part of the congresswoman's body.

by the websites, claiming that the content that was blocked did not violate the community standards or that the reasons for the suspension were not rational, and that those users were actually being targeted by defamatory campaigns started by opposing groups. The suspensions are eventually compared to other cases in which users indicate posts that also contained sensitive content, such as offenses or threats against certain politicians or partisan groups, or the dissemination of fake news or information, but those posts were not removed by the platforms or caused the authors' accounts to be suspended.

The following figure shows the evolution of the debate on Twitter. Based on the peaks indicated in it, notably on November 20 (about hate speech) and on December 15, 2020 and January 19, 2021 (about freedom of speech), we will conduct a qualitative analysis based on the engagement around the topics. We will take into account the number of shares, likes and followers in each account in order to detect the publications that drove the debate.

Figure 1 – Volume of Interactions in Twitter Categories
Period: From November 16, 2020 to February 6, 2021



Source: Twitter | Elaborated by: FGV DAPP

Regarding the debate about hate speech, there was a focus on the topic of structural racism in the Brazilian society. This discussion was driven by two events that fed off each other: the Black Consciousness Day, celebrated on November 20, and the assassination of

João Alberto Freitas by security guards at a Carrefour supermarket in Rio Grande do Sul. The discussion was centered around whether João Alberto's death was a racial crime or a consequence of the guards truculence and aggressiveness. The engagement in this debate happened due to the publication of the event by different news channels, with high repercussion due to a demonstration against the assassination held at the Avenia Paulista, in São Paulo, also on November 20, 2020. Posts made by the channels Mídia Ninja, Quebrando o Tabu and other accounts about the protest further increased the volume of debate about racism on the networks.

Regarding the discussions about censorship and freedom of speech inside the platforms, there was a significant amount of tweets about the topic on December 15, 2020, with around 76,467 posts. We found two posts that drove the engagement in this debate: one was made about President Jair Bolsonaro and the other was made by the congresswoman Carla Zambelli (PSL-SP). The congresswoman made an explicit reference to Bill 2630/2020, proposed by the congressman Alessandro Vieira (Cidadania), which was discussed in section 2 of this report. According to the publication, the Federal Senate had an open public consultation in which the "Yes" vote was winning, while the congresswoman and some of her allies wanted their followers to vote "No". The bill provides, among other measures, that platforms regulate and establish sanctions against false and offensive posts. The proposal was considered an act of censorship against individual freedom of speech. Driven by these posts, other publications mentioned different moments when Twitter prohibited liking or commenting on some contents considered a violation of the platform's Terms of Use or Community Guidelines.

A clear example of this debate was the large mobilization about Twitter's decision to block the official account of Donald Trump, which happened during the first weeks of January 2021, after the former American president posted a video in which, considering the comments related to that content, he incited his supporters to invade the United States Congress to protest the result of the American election in 2020, in which his opponent, Joe Biden, won. The decision made by the platform was based on the allegation that the video could incite violent acts from the demonstrators. Afterwards, other social network

websites, such as Facebook, Instagram and Snapchat, also announced the suspension of Trump's accounts in their respective platforms.

The suspension of the former American president's accounts in different websites led to a discussion about impasses between hate speech and freedom of speech on social networks. Some profiles who opposed the protests at the Capitol in Washington, D.C. approved Twitter's decision, denouncing the "message of hate" propagated by Trump and stating the freedom of speech – a cause defended by supporters of the former president – must have limits. It is important to highlight that several media outlets that covered the event showed that many of the demonstrators who invaded the American Congress wore clothes or bore symbols recognized as representations of white supremacy, which corroborates those who defend limits for free opinion.

However, the predominant opinion in the debate about this episode was that Twitter's decision was arbitrary. This opinion was supported mostly by profiles of conservative politicians, bloggers and influencers. With the argument that Trump was being persecuted, this discourse interpreted the platform's measure as a strategy of left-wing groups to try and take over power in the United States. The posts argued that the website's decision was incoherent, since accounts belonging to other rulers who are "dictators" were kept active on the social networks, such as the profiles of Nicolás Maduro, president of Venezela, and Hassan Rouhani, president of Iran. This part of the debate also argued that comments criticizing Trump's message were hypocritical, stating that, even though they claim to be defenders of democracy, they were celebrating an act of censorship against the former president. Some users, motivated by a publication made by the influencer Allan dos Santos, classified the measure as absurd, dictatorial and silencing voices, which intensified the debate.

Another point that motivated the debate about censorship and freedom of speech was the classification of a post made by Brazil's Ministry of Health about early treatment against Covid-19 as misleading. The engagement was based on a publication made by the portal Brasil 247 on the investigation initiated by Goiás State Prosecution Office about the case.

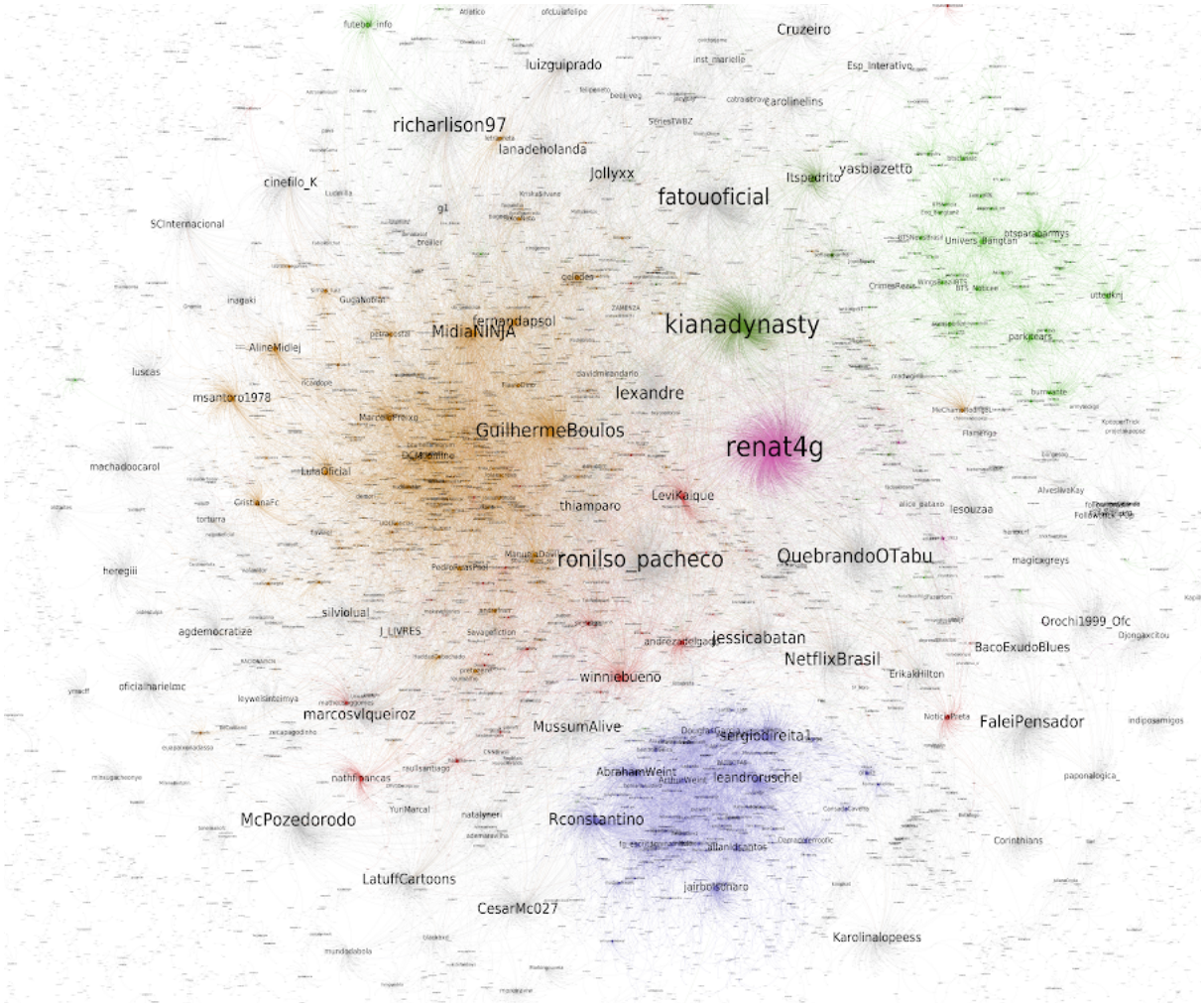
Although the debate at this point was permeated by posts about Covid-19 and the severity of the disease, it was also marked by the platform's interference with an official governmental account, arguing about the risks of that attitude in terms of freedom of speech.

In the following section, we will understand how these profiles interacted and who was at the center of the debate. However, what we can conclude from the facts presented is that the debate about censorship was present during the whole analysis period and driven by relevant national political figures.

Analysis of Twitter actors

On November 20, 2020, the assassination of João Alberto Silveira de Freitas at a supermarket in Porto Alegre, on the previous day, was the highlight in the debate about hate speech on Twitter. The circulation of individual opinions and interpretations about the case allowed the identification of specific groups in that context.

Figure 2 – Map of interactions in the debate about hate speech on Twitter
Period: November 20, 2020.



Source: Twitter | Elaborated by: FGV DAPP

Orange – 11.21% of the interactions

Group composed of left-wing politicians and media channels lamenting the assassination of João Alberto and insisting on the racist nature of the crime. Some posts also called for the mobilization of social movements to organize protests against the episode.

Blue – 6.63% of the interactions

Group influenced by right-wing politicians, bloggers and celebrities of conservative orientation, who argued that it was precipitate to classify the assassination of João Alberto as a racial crime. The posts shared an alleged declaration made by a chief of police involved in the investigation, who dismissed indications of racism in that case. Some profiles also recovered a video from 2012 in which, during an interview, the American actor Morgan Freeman questioned the existence of the Black Consciousness Month, arguing that there should be a White Consciousness Month. Outside of the scope of João Alberto's case, there were also publications about legal proceedings initiated by the former athlete Ana Paula Henkel due to offenses and attacks she suffered on social networks, as well as the circulation of the right of reply granted to the influencer Olavo de Carvalho after declarations made by the journalist Thaís Oyama about him.

Light Green – 4.98% of the interactions

This group was formed by entertainment channels that, in the context of the assassination of João Alberto, published other cases of racism involving black professionals and influencers in Brazil. The highlights were the death of a supermarket employee in Recife (PE) in August 2020, whose body was kept hidden under umbrellas while the shop operated normally, the case of the youtuber @badgallore, who was humiliated at a bank location after being questioned about the high amount of money in her account, and the case of a manager of a department store in Minas Gerais, in which a customer doubted his position at the company due to him being black.

Red – 3.17% of the interactions

Group composed of black activists who reacted to the assassination of João Alberto and denounced the relativization of racism and victim blaming in cases of violence. Contesting the alleged declaration made by a chief of police responsible for the investigation, who supposedly said there were no evidence of racism in the case, the posts questioned the Brazilian judiciary branch, comparing the statement with the "involuntary rape"

judgement in the trial of the sexual harassment suffered by the blogged Mariana Ferrer in September 2020.

Purple – 2.69% of the interactions

Group mobilized by profiles of black activists and influencers, demonstrating their revolt and indignation with the fact that João Alberto was beaten to death, especially with the crime happening close to Black Consciousness Day in Brazil. Arguing that there would be no reason to celebrate on that date, many profiles published data about the reality of the black population in the country in order to demonstrate that the crime was not an isolated case, but an example of the situation of black people in Brazil.

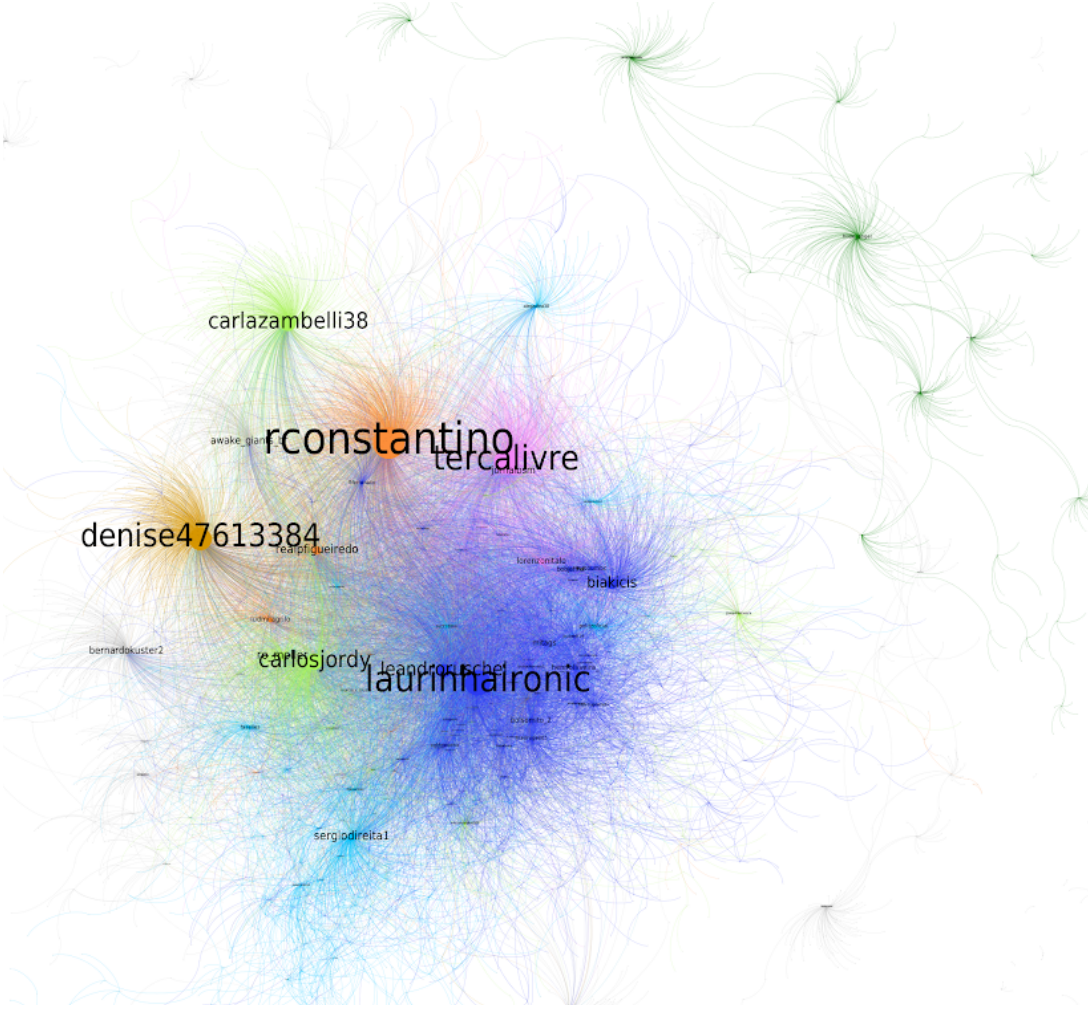
Dark Green – 2.50% of the interactions

Group composed of profiles of social influencers who attacked the supermarket chain in which the assassination of João Alberto happened, highlighting that it was not the first time that company was a location of atrocities, and insisting that the security guards responsible for the crime should be exposed, not the victim.

Regarding freedom of speech – and, ultimately, censorship –, the debate on Twitter focused on different episodes in order to question the decisions taken by several social networks in an attempt to moderate the content circulating in their platforms.

Figure 3 – Map of interactions in the debate about censorship on Twitter

Period: January 14 to 15, 2021



Source: Twitter | Elaborated by: FGV DAPP

Dark Blue – 39.23% of the interactions

Group mobilized by right-wing politicians and comic profiles aligned with the right, who called for their followers to join the campaign #silenceday, in protest against the supposed violation of freedom of speech promoted by social networks such as Facebook, Twitter, Instagram and YouTube. In response to moderation measures taken against content published mostly by conservative profiles, the campaign instructed users not to post, like or share anything in the platforms for a day.

Light Green – 11.17% of the interactions

During the health crisis that hit Manaus (AM) in January due to the Covid-19 pandemic, this group composed of right-wing politicians, bloggers and digital influencers of conservative orientation questioned the attempt made by traditional media outlets and left-wing politicians to blame the federal government for the situation. Accusing the Brazilian media of spreading fake news, the profiles argued that the governor of the state of Manaus should be blamed for the collapse of the health system in the city.

Pink – 9.03% of the interactions

Centered around alternative media channels aligned with the right wing, this group criticized decisions made by some social networks, such as Twitter and Facebook, to moderate official channels due to content published by the then American president Donald Trump regarding the 2929 presidential elections in the United States. Arguing that the decisions made by the platforms were an example of censorship, the profiles also mobilized the campaign #silenceday.

Orange – 8.89% of the interactions

Calling for profiles to join the campaign #silenceday, this group was based on profiles of conservative journalists and attacked legal actions attempting to prevent the judge Ludmila Lins Grilo to publish, on her social network profiles, messages supporting more flexible social distancing measures and mask use in the context of the Covid-19 pandemic. This part of the debate claimed that those measures were in violation of the magistrate's freedom of speech.

Dark Green – 3.11% of the interactions

This group included profiles of journalists, political commentators and digital influencers, who mocked the campaign #silenceday and questioned the allegation that the moderation of posts and channels used by the former American president Donald Trump on social

networks were a violation of the principle of freedom of speech. Criticizing declarations made by politicians associated with the federal government in which they downplayed the Covid-19 pandemic or opposed the recommendations of health authorities, the posts demanded the social network platforms to act by suspending accounts and removing content.

Light Blue – 12.34% of the interactions

Group mobilized by profiles of right-wing politicians and conservative information channels and digital influencers who refused to join the campaign #silenceday, arguing that the initiative did little to guarantee that freedom of speech would be respected. Instead, these profiles argued that users should make an effort to post on their social networks and question the discourse of the traditional media, which blamed the federal government for the health crisis in Manaus. In the context of this mobilization, the posts insisted that the governor of Amazonas should be blamed for the collapse of the health system in the region.

Yellow – 4.16% of the interactions

Group composed of profiles of ordinary users commenting on the news that the Russian government allegedly criticized the decision made by social networks to suspend posts and accounts of the former American president Donald Trump, and that Vladimir Putin was supposedly willing to fight the platforms' attempts to violate freedom of speech in his country. The group also shared the information that, due to the alleged censorship of Trump, websites such as Twitter and Facebook were suffering financial losses.

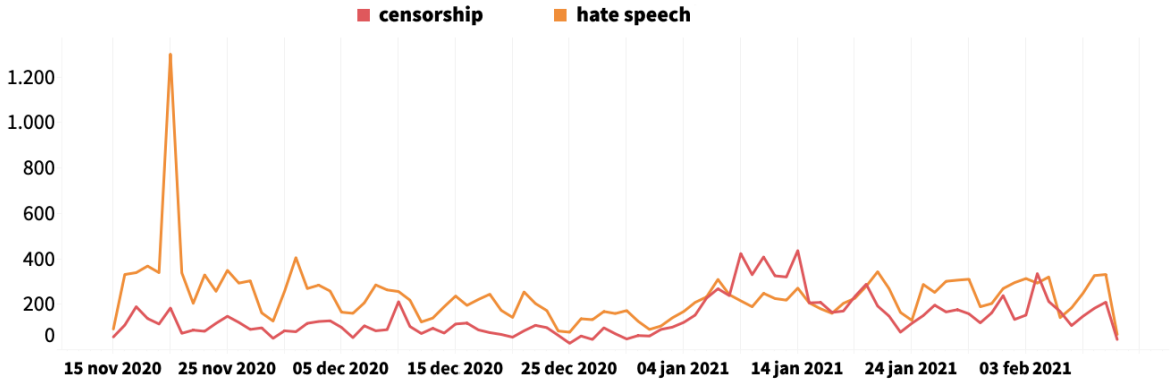
Debate about hate speech and censorship on Facebook

This section will present the findings of the analysis conducted on Facebook between November 15, 2020 and February 11, 2021. We collected 21,149 posts about hate speech

and 12,773 posts about the topics of freedom of speech or censorship. The evolution of both debates was observed in parallel, considering the direct relationship between the two topics which we mentioned previously. Thinking about this articulation is important for this study because there is a growing demand for mechanisms to curtail hate speech and, similarly, impose limits to freedom of speech.

Figure 3 – Evolution of the debate about hate speech and censorship on Facebook

Period: From Sunday, November 15, 2020 to Thursday, February 11, 2021



Source: Facebook | Elaborated by FGV DAPP

Starting with the posts related to hate speech, the corpus indicated a few specific debates that stood out in the period analyzed. The topic was a frequent subject in some profiles, who argued about the need to fight discriminatory discourses on the network, such as a commitment made by a city councilor elected for Jaru (RO). There was also a proposal made by a church in Osasco (SP), who was willing to discuss the hate culture in one of its masses. Hate speeches were also mentioned in pages that consider its debate a way to prioritize social relationships that are marked by offenses before others. Specifically, the data revealed the engagement in discussions related to racism, misogyny and homophobia.

Since the sample included the two rounds of voting during the Brazilian municipal elections, the publications were affected by that context, especially in November in

December. For example, there were mentions of the first black woman to be ellected for the city council of Joinvile (SC), who was the targeted of racial abuse and received death threats on the networks. There were also posts criticizing a comic video made by the channel *Porta dos Fundos*, referring to the most voted city councilor of Curitiba (PR), attributing a sexual connotation to her successful result.

Similar to the Twitter analysis, the highest peak of debate about the topic happened on November 20, Black Consciousness Day, when there were 1,301 posts about hate speech in the platform, more than five times the daily average of 237,6 publications, in the period analyzed. The date alone already drives content production about the need to fight racism, especially coming from public figures. However, the death of João Alberto, the black consumer who was beaten the day before the holiday at the car park of a Carrefour supermarket in Porto Alegre (RS), increased the levels of engagement.

Throughout the day, the publications focused, on one side, on discussing the case as an example of structural racism, and on the other side, on arguing against the idea that the crime was motivated by racism. Although most of the posts about the episode associated it with the holiday as an evidence of the low level of understanding about equality of rights in the country, the publication that produced the most engagement was made by the profile of President Jair Bolsonaro. In the message, which had 41,787 shares, he stated that hate is produced by those who want to divide the Brazilian people through resentment, making the population vulnerable to controlling mechanisms.

This approach was chosen by those arguing that mentioning racism in Brazil is a way to generate discord among Brazilians and feed a left-wing project for chaos. However, the data collection identified posts made by non-conservative profiles, to whom treating the case of João Alberto as racism makes the class struggle invisible and reduces the possibilities to fight inequality. One side believes that debate about racism is harmful because it corrodes social relationships, weakening an alleged national unity. The other side believes the debate is an oversimplification, since it reduces the possibilities of resistance in favor of a niche type of activism that cannot change the system.

Of the ten most shared publications on that day, seven dealt specifically with the assassination of João Alberto and its direct association with racism. The arguments to justify this association included statistic data revealing that black Brazilians are more frequently the target of assassinations than white Brazilians. Almost every post had suggestions about how Carrefour could and should be legally punished, but also about how people could demonstrate their indignation by boycotting the supermarket. Notes of repudiation were criticized and considered palliative, and some users remembered other episodes involving the company in cases of racism and death. In general, there were recurring calls for people to get involved in the fight against racism, both individually and collectively.

Six publications were always present when we changed the engagement criteria (number of likes, shares or comments), even though they appeared in different positions in the ranking. Of those, four were made by public figures: in addition to Jair Bolsonaro, the major influencers of the debate on the peak day were the artists Nando Reis (Mais Brasil) and Maria Rita, and the priest Reginaldo Manzotti. When observed separately, the different engagement criteria enabled the identification of other profiles of public figures. The second post with the most likes in the share, for example, was made by the Brazilian football player Fred, of Manchester United. The publication made by the culinary blogger Rita lobo appeared among the ten with the most engagement when the criterion was the number of comments. Except for the president, all the publications mentioned the Black Consciousness Day as a day to reflect about the suffering caused by racial prejudice.

Therefore, in the period analyzed, the debate about hate speech was driven by the Black Consciousness Day, November 20, and was combined with posts about the assassination of João Alberto in Rio Grande do Sul. In general, the authors of the publications understood that the homicide was a crime motivated by racial prejudice, not only because the victim was black, but also because of the way he was beaten, with no right to defend himself. As shown by the large literature on the topic, the basis for defining hate speech is the idea of superiority, according to which certain groups do not deserve to be treated

equally. The social vulnerability faced by this group, which is materialized in specific characteristics, makes them the target of attacks and hate speeches.

Regarding the debate about censorship no Facebook, in addition to posts discussing the topic in specific contexts such as religion and art, two other types of publications made in the analyzed period stood out: posts accusing the platforms, especially Twitter and Facebook, of violating freedom of speech on the network; and posts celebrating notable legal decisions characterized by terms such as “the victory of freedom of speech”.

The posts accusing the platforms included publications defending organized actions by progressive groups of left-wing orientation, participating in communication media, platforms, public institutions, and especially the judiciary branch, allegedly aiming to dismantle right-wing conservative movements. One of the most commented topics in this part of the discussion was the impossibility to share a link to a news story published by the New York Post, due to it allegedly containing information that was not checked. The text made accusations against the son of then democrat candidate for the US presidential elections, Joe Biden. The posts analyzed accused Facebook and Twitter of conducting a “pro-democrat edition” of its content, and, by doing that, censoring anything outside that criteria.

A second group of publications that stood out mentioned legal actions motivated by accusations of hate speech and disinformation, among other subjects. This group includes a series of posts celebrating a favorable decision by a sports court regarding a “#ForaBolsonaro” (out Bolsonaro) cry made by the volleyball player Carol Solbergduring an interview. There was also engagement in publications celebrating the fact that the Supreme Court did not consider illicit the posts made by the congressman Eduardo Bolsonaro (PSL-SP) against the Homeless Workers Movement (MTST).

The posts about censorship reached a daily average of 145.5 publications, a lower volume than that of hate speech. However, on January 9, 11 and 14, the volume of posts in the debate about censorship was higher than that of hate speech. The change was caused by

the news that the then president of the United States, Donald Trump, had his accounts cancelled permanently by Twitter, and indefinitely by Facebook and Instagram. On those three days, the debate about censorship saw a peak of publications, registering 423, 408 and 436 posts, respectively.

Three movements marked the discussions about censorship. The first one was related to the suspension of Trump's accounts in those platforms and to the ban of the app Parler, in which the former president kept an account, from the Google, Apple and Amazon app stores. The measures were considered an act of censorship against the conservative movement in general, and an attack against a legitimate representative of the far right. The second movement was a strong engagement in publications promoting other platforms: Parler itself, before the ban, and also Telegram and Paatria. On January 14, there was repercussion of the news from January 9 and 11, and the posts were also motivated by the #silenceDay, whose goal was avoid posting, liking, marking or interacting in any way on Twitter, Facebook, Instagram, YouTube and WhatsApp on January 15, to protest against the actions of the platforms.

In this group of posts, the one with the most engagement was published by the congresswoman Carla Zambeli (PSL-SP), in which she celebrated the news that Trump would create a new platform "with other companies to fight censorship against the right wing". The publication obtained 29,524 shares and 13,825 comments. In addition to Zambelo, other profiles of public figures were identified in the sample among the authors of the publications generating the most engagement in the debate, especially President Jair Bolsonaro, the congresswomen Bia Kicis (PSL-DF) and Caroline de Toni (PSL-SC), and the congressmen Filipe Barros (PSL-PR) and Carlos Jordy (PSL-RJ). The profiles República de Curitiba and Jovem Pan News also made posts that drove the discussions. Unlike the debate about hate speech, the topic of censorship was marked by more homogeneity of its authors, showing that the debate was articulated by right-wing conservative groups in the period analyzed.

The analysis of the debates about hate speech and censorship on Facebook demonstrates the important role of the public agendas and news media in driving political debates. The combination of a holiday that encompasses a series of actions and celebrations organized by governments and the society with the tragic death of João Alberto, and the international repercussion of the cancellation of Trump's accounts, motivated the publications made by users on the network. Both debates increased in volume through the accounts of public figures in politics, music, sports, religion and the internet, in the debate about hate speech, and particular in politics in the debate about censorship. This demonstrates the relevance of the involvement of important names with large networks to increase the volume of discussion and reach different levels of society. It is also an evidence of the responsibility public authorities have in directing the course of debate.

CONCLUSIONS

Although the offenses, threats and discrimination that are part of hate speeches have subjective effects, considering the great psychological suffering they produce, their investigation cannot dismiss their political impacts. The silencing and invisibility of vulnerable groups weaken their demands and pleas, removing them from democratic decision-making processes. Therefore, identifying and fighting hate speeches is important for democracy.

This study presented a literature review about the topic, aiming to demonstrate the major understanding about what hate speeches are in the academy and in the legal field, their specificities in the digital environment, and how they are defined by digital platforms in their self-regulation protocols. The bibliographic and documentary investigation revealed that while it is possible to conceptualized gate speeches in theory, it is very complex to define them in concrete situations. The challenge becomes harder in online pages with memes, videos, organized communities, lexical variations and other aspects.

The material demonstrated that, in addition the adoption of international treaties to fight the different forms of discrimination, it is necessary that countries formulate legislation combining their social, historical and especially cultural contexts with the specificities of the online universe. In that sense, the challenge was not easier for the platforms. When explaining their perception of the different manifestations of hate speeches in their Terms of Use and Community Guidelines, the platforms left a series of gaps about how to approach and moderate the authors of that content. One hypothesis on that regard is that global platforms need to know the sociocultural particularities which their users belong to in order to develop viable and more productive solutions when fighting hate speeches.

The empirical analysis in this study consisted in mapping the debate of hate speeches and its relationship with freedom of speech on Twitter and Facebook between November 2020 and February 2021. Even though they are both digital social networks with a significant number of users, including in the Brazilian context, we understand that they have limitations in terms of representativeness, both regarding the amount of digital interactions and the diversity of authors they enable. This analysis offered an incomplete view of the complex ecosystem involved in the creation, proliferation and maintenance of online hate speeches.

Regarding our methodological concerns, one challenge related to the development of search syntaxes – used to collect and classify the data to be analyzed – was the presence of subtleties in the conceptualization and semantic-pragmatic scope of the topic targeted by the study, that is, “hate speech”. Considering that it is very difficult to conclusively and indisputably determine what actually is hate speech, or even offensive speech, and especially considering how this discourse happens in the public debate on social networks, finding linguistic and discursive strategies to analyze this phenomenon proved to be an undeniably flawed task.

Based on the data collection and analysis we conducted, we found that the volume of debate about hate speech and censorship increased during a holiday and based on the repercussion of national and international news. This leads us to believe that the existence

of a public agenda and the participation of the news media contributed to the intensification of ongoing discussions in the platforms we investigated. In addition to these elements, the posts about the topic made by public figures who have a large network of followers produced a high volume of interactions, which demonstrates the politically strategical role of public figures in the production and reach of the debate.

Considering the evolution of the online debate in the period analyzed, this study indicated the existence of tensions related to the structural and personal understandings of discrimination. This means that, for some of the authors, this debate is critical to the development of democracy and must be present in the public arena, while for other authors, this discourse destabilizes the political field without producing effective change. Lastly, the sanctions applied by the platforms interfered with the volume of publications and with the participation of certain actors, intensifying the debate about freedom of speech and leading to a discussion that does not question their legitimacy, but their limits.

By identifying the arguments and authors that compose this debate, this study lays the groundwork for new investigations about the dissemination of hate speeches on the networks. Among them are the appearance of new platforms in response to the sanctions, the expansion of the lexical diversity associated with the different possibilities of language, the relationship between the self-regulatory procedures of platforms and the sociocultural contexts in which they are applied, and the strategies of social actors to disarticulate online discrimination.

REFERENCES

ALAVA, S.; FRAU-MEIGS, D.; HASSAN, G. **Youth and violent extremism on social media:** mapping the research. Paris: Organização das Nações Unidas para a Educação, a Ciência e a Cultura, 2017.

BARBERÁ, P. How social media reduces mass political polarization: evidence from Germany, Spain, and the US. In: AMERICAN POLITICAL SCIENCE ASSOCIATION ANNUAL MEETING, 2015. **Proceedings**. American Political Science Association, São Francisco, 2015. Available at http://pablobarbera.com/static/barbera_polarization_APSA.pdf. Accessed on: February 26, 2021.

BARBOSA, B.; MARTINS, H.; VALENTE, J. **Fake news:** como as plataformas enfrentam a desinformação. São Paulo: Intervozes, 2020.

BENESCH, S. Dangerous speech: A proposal to prevent group violence. **Dangerous Speech Project proposal paper**. February 23, 2013. Available at: https://dangerousspeech.org/wp-content/uploads/2018/01/Dangerous-Speech-Guidelines-2013.pdf. Accessed on: March 01, 2021.

BOWMAN-GRIEVE, L. Exploring Stormfront: a virtual community of the radical right. **Studies in Conflict and Terrorism**, v. 11, n. 31, p. 989-1007, 2009.

BRASIL. Decreto-lei nº 2.848, de 7 de dezembro de 1940. Código Penal. Available at: http://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm. Accessed on: February 26, 2021.

BRASIL. Decreto nº 65.810, de 8 de dezembro de 1969. Promulga a Convenção Internacional sobre a Eliminação de todas as Formas de Discriminação Racial. Available at: http://www.planalto.gov.br/ccivil_03/decreto/1950-1969/D65810.html. Accessed on: February 26, 2021.

BRASIL. Constituição da República Federativa do Brasil. Brasília, 1988. Available at: http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm. Accessed on: February 25, 2021.

BRASIL. BRASIL. Law 7,716, Thursday, January 5, 1989. Define os crimes resultantes de preconceito de raça ou de cor. Available at: http://www.planalto.gov.br/ccivil_03/leis/l7716.htm. Accessed on: February 26, 2021.

BRASIL. Decreto nº 592, de 6 de julho de 1992a. Promulga Atos Internacionais. Pacto Internacional sobre Direitos Civis e Políticos. Available at: http://www.planalto.gov.br/ccivil_03/decreto/1990-1994/d0592.htm. Accessed on: February 26, 2021.

BRASIL. Decreto nº 678, de 6 de novembro de 1992b. Promulga a Convenção Americana sobre Direitos Humanos (Pacto de São José da Costa Rica), de 22 de novembro de 1969. Available at: http://www.planalto.gov.br/ccivil_03/decreto/d0678.htm. Accessed on: February 26, 2021.

BRASIL. Declaração e programa de ação [da III Conferência Mundial de Combate ao Racismo, Discriminação Racial, Discriminação Racial, Xenofobia e Intolerância Correlata]. Brasília, DF: Ministério da Cultura, 2001. Available at: http://www.unfpa.org.br/Arquivos/declaracao_durban.pdf. Accessed on: February 26, 2021.

BRASIL. BRASIL. Law 10,741, Wednesday, October 1, 2003. Dispõe sobre o Estatuto do Idoso e dá outras providências. Available at: http://www.planalto.gov.br/ccivil_03/leis/2003/l10.741.htm. Accessed on: February 26, 2021.

BRASIL. BRASIL. Law 12,735, Friday, November 30, 2012. Altera o Decreto-Lei nº 2.848, de 7 de dezembro de 1940 – Código Penal, o Decreto-Lei nº 1.001, de 21 de outubro de 1969 – Código Penal Militar, e a Lei nº 7.716, de 5 de janeiro de 1989, para tipificar condutas realizadas mediante uso de sistema eletrônico, digital ou similares, que sejam praticadas contra sistemas informatizados e similares; e dá outras providências. Available at: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12735.htm. Accessed on: February 26, 2021.

BRASIL. BRASIL. Law 12,965, Wednesday, April 23, 2014. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Available at: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Accessed on: February 26, 2021.

BROWN, A. What is so special about online (as compared to offline) hate speech? **Ethnicities**, v. 18, n. 3, p. 297-326, 2018.

BRUGGER, W. Proibição ou proteção do discurso do ódio? algumas observações sobre o direito alemão e o americano. **Revista de Direito Público**, n. 15, p. 117-136, 2007.

BUYSE, A. Words of violence: "fear speech," or how violent conflict escalation relates to the freedom of expression. **Human Rights Quarterly**, v. 36, n. 4, p. 779-797, 2014.

COHEN-ALMAGOR, R. Fighting hate and bigotry on the Internet. **Policy and Internet**, v. 3, n. 3, 2011. pp.: 1-26.

CONSANI, C. F. Democracia e os discursos de ódio religioso: o debate entre Dworkin e Waldron sobre os limites da tolerância. **Etic@,** v. 14, n. 2, p. 174-197, 2015.

DWORKIN, R. **O direito da liberdade**. São Paulo: Martins Fontes, 2006.

DWORKIN, R. Foreword to Extreme Speech and Democracy. In: HARE, I.; WEINSTEIN, J. (eds). **Extreme Speech and Democracy**. New York: Oxford University Press, 2009, p. v-ix;

FARIS, R.; ASHAR,, A.; GASSER, U.; JOO, D. Understanding harmful speech online. **Berkman Klein Center Research Publication**, n. 2016-21, 2016. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract%5Fid=2882824. Accessed on: March 01, 2021.

FERNANDEZ-MATAMORO, A.; FARKAS, J. Racism, hate speech, and social media: a systematic review and critique. **Television & New Media,** v. 22, n. 2, p. 205-224, 2021.

FORTUNA, P.; SOLER-COMPANY, J.; WANNER, L. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In: CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 12, 2020. **Proceedings**. European Language Resources Association, Marseille, 2020. p. 6786-6794.

FUNDAÇÃO GETULIO VARGAS. Diretoria de Análise de Políticas Públicas. **Monitor de temas**. Rio de Janeiro: FGV DAPP, 2015. Available at: http://bibliotecadigital.fgv.br/dspace/handle/10438/15262. Accessed on: February 26, 2021.

GILLESPIE, T. A relevância dos algoritmos. **Parágrafo**, v. 6, n. 1, p. 95-121, 2018.

JURNO, A. C.; D'ANDRÉA, C. (In)visibilidade algorítmica no "feed de notícias" do Facebook. **Revista Contemporânea**, v. 15, n. 2, p. 463-484, 2017.

LUCCAS, V. N.; GOMES, F. V.; SALVADOR, J. P. F.. **Guia de análise de discurso de ódio**. Rio de Janeiro: Fundação Getulio Vargas, 2020. Available at: https://www.conib.org.br/wp-content/uploads/2019/11/Guia-de-An%C3%A1lise-de-Discurso-de-%C3%93dio.pdf. Accessed on: February 26, 2021.

LUNA, N e SANTOS, G. Liberdade de expressão e discurso do ódio no Brasil. **Revista Direito e Liberdade**, v. 16, n. 3, 2014, p. 227–255. Available at: http://ww2.esmarn.tjrn.jus.br/revistas/index.php/revista_direito_e_liberdade/article/view/780. Accessed on: March 01, 2021.

NANDI, J. **O combate ao discurso de ódio nas redes sociais**. Trabalho de Conclusão de Curso (graduação). Universidade Federal de Santa Catarina (UFSC), Campus Araranguá, Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2018. Available at: https://repositorio.ufsc.br/handle/123456789/187510. Accessed on: March 01, 2021.

NERIS, N. (coord); VALENTE, M.; CRUZ, F.; OLIVA, T. **Outras vozes:** gênero, raça, classe e sexualidade nas eleições de 2018. InternetLab, São Paulo, 2019. Available at: https://www.internetlab.org.br/wp-content/uploads/2019/10/OutrasVozes_2018.pdf. Accessed on 3/10/2021.

ONU (ORGANIZAÇÃO DAS NAÇÕES UNIDAS). **Carta das Nações Unidas**. São Francisco: ONU, 1945. Available at: https://www.un.org/en/charter-united-nations/index.html. Accessed on: February 26, 2021.

ONU (ORGANIZAÇÃO DAS NAÇÕES UNIDAS). **Declaração Universal dos Direitos Humanos**. Paris: ONU, 1948. Available at: https://www.un.org/en/universal-declaration-human-rights/. Accessed on: February 26, 2021.

PAREKH, B. Is there a case for banning hate speech? In: HERZ, M.; MOLNAR, P. (eds.). **The Content and Context of Hate Speech:** Rethinking Regulation and Responses. Cambridge: Cambridge University Press, 2012, p. 37-56.

PEARSON, J.; CHILD, J.; DEWITT, L.; KAHL, D. Facing the fear: an analysis of speech-anxiety content in public-speaking textbooks. **Communication Research Reports**, v. 24, p. 159-168, 2007.

ROGERS, R. Engajados de outra maneira: As Mídias Sociais - Das Métricas de Vaidade à Análise Crítica. In: OMENA, J. J. (ed.). **Métodos Digitais:** Teoria e Prática. Lisboa: Editora da Universidade Nova de Lisboa, 2019. p, 73-96.

RUEDIGER, M. (coord.). **Nem tão #simples assim:** o desafio de monitorar políticas públicas nas redes sociais. Rio de Janeiro: FGV DAPP, 2017. Available at: http://hdl.handle.net/10438/19436. Accessed on 3/1/2021.

SANCHES, D. Construindo grafos de conhecimento (knowledge graphs) para análise de redes sociais: uma análise sobre discurso de ódio em Portugal. In: WORKSHOP ON MEDIA, INFORMATION AND DATA SCIENCE, 1, 2020. **Anais.** Universidade Federal de Goiás, Goiânia, 2020.

SELLARS, Andrew. Defining Hate Speech. **Berkman Klein Center Research Publication**, n. 2016-20, Boston Univ. School of Law, Public Law Research Paper, p. 16-48, 2016. Available at: https://ssrn.com/abstract=2882244 or http://dx.doi.org/10.2139/ssrn.2882244. Accessed on: March 01, 2021.

SHEPHERD, A. Extremism, free speech and the rule of law: evaluating the compliance of legislation restricting extremist expressions with article 19 ICCPR. **Utrecht Journal of International and European Law**, v. 33, p. 62-83, 2017.

SIEGEL, A. Online hate speech. In: PERSILY, N.; TUCKER, J. (orgs.). **Social media and democracy**. Cambridge: Cambridge University Press, 2020. p.: 56-88.

SILVA, L. R.; BOTELHO-FRANCISCO, R. E.; OLIVEIRA, A. A.; PONTES, V. R. A gestão do discurso de ódio nas plataformas de redes sociais digitais: um comparativo entre Facebook, Twitter e Youtube. **Revista Ibero-americana de Ciência da Informação**, v. 12, n. 2, p. 470-492, 2019.

SILVIA, L.; MONDAL, M.; CORREA, D.; BENEVENUTO, F.; WEBER, I. Analyzing the targets of hate in online social media. In: **Proceedings of the Tenth International AAAI conference on Web and Social Media**, 2016. Available at: https://arxiv.org/abs/1603.07709v1.

SIMONS, A.; MUELLER, John. The Dynamics of Internal Conflict. **International Security**, n. 24, v. 4, p. 187-192, 2001.

SLOAN, R. H.; WARNER, R.. Beyond notice and choice: privacy, norms, and consent. **Journal of High Technology Law**, v. 14, p. 370, 2014.

WALDRON, J. **The harm in hate speech**. Cambridge: Harvard University Press, 2012.

WEAVER, S. A rhetorical discourse analysis of online anti-Muslim and anti-Semitic jokes. **Ethnic and Radical Studies**, v. 3, n. 36, p. 483-499, 2013.

WEINSTEIN, J.; HARE, I. General introduction: free speech, democracy, and the suppression of extreme speech past and present. In: HARE, I.; WEINSTEIN, J. (eds.). **Extreme speech and democracy**. Nova York: Oxford University Press, 2009, p. v-ix.