International Journal of Communication 18(2024), 2685-2709 1932–8036/20240005

Online Disinformation in Brazil: A Typology of Discursive Action of Harmful Political Content on WhatsApp and Facebook

TATIANA DOURADO♦[1]
National Institute of Science and Technology for Digital Democracy, Brazil

VICTOR PIAIA
Getulio Vargas Foundation, Brazil

VIKTOR CHAGAS
DALBY DIENSTBACH
Fluminense Federal University, Brazil

MARCO AURELIO RUEDIGER
EURICO MATOS
Getulio Vargas Foundation, Brazil

JOÃO GUILHERME BASTOS DOS SANTOS
National Institute of Science and Technology for Digital Democracy, Brazil

This article investigates harmful political content in public WhatsApp and Facebook groups of the radical Right in Brazil. Considering harmful political content as that which generates direct damage to the quality, reasonableness, and plurality of public discussion, we

Tatiana Dourado: tatiana.dourado@inctdd.org
Victor Piaia: victor.piaia@fgv.br
Viktor Chagas: viktor@midia.uff.br
Dalby Dienstbach: dalbydienstbach@gmail.com
Marco Aurelio Ruediger: marco.ruediger@fgv.br
Eurico Matos: eurico.neto@fgv.br
João Guilherme Bastos dos Santos: santos.jgb@gmail.com
Date submitted: 2022-12-31

[1] The authors would like to thank Facebook Research for funding the project; the National Council for Scientific and Technological Development (Fellowship No. 306791/2021-8) and the Carlos Chagas Filho Foundation for Research Support of the State of Rio de Janeiro (Fellowships Nos. 259788 and 249104, respectively); colleagues at the National Institute of Science and Technology for Digital Democracy, Emerson Cervi and Samuel Barros, for their quantitative advice; and the anonymous reviewers for their invaluable feedback.

investigate the enunciative aspects of four specific types of discursive action (uncivil, conspiratorial, hateful, and dangerous) and the non-enunciative aspects used for harmful types of communication and interaction. The database consists of 3,503,540 messages propagated in 1,676 public groups during the electoral process. Through a quantitative approach to a sample of 2,201 unique messages, we found, among other things, that (1) harmful content was more present on Facebook than on WhatsApp; (2) messages about the elections were associated with uncivil speech; (3) uncivil speech was usually associated with dangerous speech and opposed to conspiratorial speech. The results allow for more nuanced reflections on the actions and strategy of the Far Right in the digital public debate.

Keywords: harmful political content, online hate speech, discursive action, radical Right, social media

The increasingly rapid exchange of political content on social media platforms has facilitated the cross-platform spread of harmful speech through unofficial and disinformation campaigns in electoral processes. Algorithm-driven social media, hyper-partisan media ecosystems, and user-generated content are aspects that have favored the digital articulation of the radical Right (Miller & Vaccari, 2020), structuring networks for the dissemination and amplification of information manipulation on different social media platforms (Bennett & Livingston, 2018). In particular, the Brazilian case is also marked by WhatsApp and its algorithm-free viral dynamics, shedding light on a complex combination of platforms and messaging applications.

In Brazil, the so-called New Right has been gaining strength online since 2008 (Rocha, 2021), with new collectives, YouTubers, WhatsApp groups, and influencers playing an important role in mobilizing individual views (Soares, Recuero, & Zago, 2018). New right-wing grassroots were key to the impeachment of President Dilma Rousseff in 2016 and the election of the Far Right candidate Jair Bolsonaro in 2018, with some of their emerging personalities jumping from social media to political representation. Bolsonaro's government relied on these (ultra)conservative societal movements, anchoring his discourse in religious elements (Rocha, 2021) and a systematic attack on the electoral system (Ruediger & Grassi, 2020). In this context, the public discussion has been characterized by levels of noxiousness as social interactions with radicalized repertories have increased.

This study seeks to provide a comprehensive understanding of harmful discursive action in digital political content that misinforms or is used for political influence. Although the large-scale spread of misinformation and disinformation on social media platforms has threatened the stability of several democracies around the world, platform standards tend not to address political content, allowing political and electoral informative deceptions to circulate more freely than other types of severe content, such as pornography, extremism, or other illegalities. It is therefore necessary to examine the characteristics of the discursive action of political content used by radical Right counter-publics to understand the dynamics of harmful speech during elections.

Additionally, there has been a debate about the impact of the platforms' infrastructures and functionalities on encouraging or restricting the circulation of different types of harmful discourse. Our methodology discusses a wider range of enunciative and non-enunciative harmful discourse types that can be mobilized by political activists in other different logics. In this sense, and considering the wide dissemination of WhatsApp in Brazil, we also invested in a comparative analysis between a more open and a more closed platform.

Therefore, the main objective of this study is to investigate the presence of harmful discursive action of political content spread in public groups on Facebook and WhatsApp during the Brazilian municipal elections of 2020. At the level of political content, this study assumes that a single digital political content can include one or more discursive actions in its enunciative and non-enunciative forms. Thus, this research systematized four types of discursive action—uncivil, conspiratorial, dangerous, and hateful—of political content, all of which are assumed to be harmful. Taking a quantitative approach, this study used 3,503,540 messages spread across 1,676 public Facebook and WhatsApp public groups during the 2020 Brazilian municipal elections, from a quantitative approach. As a result, this study contributes to a broader understanding of the implications of the discursive foundations of disinformation campaigns in the Brazilian political and electoral context.

The Discourse in Harmful Political Content Online

In digital media–based political activity, the harmful side of discourse can stem from attitude (behavior)—such as trolling, harassing, astroturfing, firehose, inauthentic accounts, mass messaging, and other methods of media manipulation (Bradshaw & Howard, 2017; Dourado, 2023)—and from content. In the case of the content itself, the topic, political actor, framing, and emotional appeal evoked matters for engaging polarized and, even more so, radicalized audiences. When becoming popular and viral, messages posted by ordinary, verified accounts can cause widespread distrust, severe public unrest, episodes of mass violence, and even democratic breakdowns.

The idea of harmful speech or content is a purposely broad theoretical category as it includes a diversity of occurrences of harm. Harmful speech "includes a variety of types of speech that cause different harms," ranging from offensive to extremist language and can be understood from speaker intent, intergroup dispute, and speech content (Faris, Ashar, Gasser, & Joo, 2016, pp. 5–6). Harmful forms of expression include hate, harassment, doxing, identity attacks, identity misrepresentation, insults, violence, self-inflicted harm, and ideological harm (Banko, Mackeen, & Ray, 2020). The latter can be understood as the "spread of beliefs that may lead to real world harm to society at large over time," and includes, in addition to extremism, terrorism, and organized crime as well as types of misinformation and disinformation capable of "leading to harmful global health or political events" (Banko et al., 2020, p. 135).

Therefore, discourses can have direct and indirect harmful effects, especially when integrated into the political context, when they evoke belief systems and inflame illiberal ideological tendencies (Benesch, 2020). While studies have been devoted to examining the type and severity of online harm (Jiang, Scheuerman, Fiesler, & Brubaker, 2021) and to building prevention protocols, mitigation mechanisms, and content moderation (Parekh, 2012), we seek to contribute to the field of political communication by calling

into question the harmfulness of misleading political content that is disseminated online either without propagandistic purpose (misinformation) or as part of interested action (disinformation), directing our gaze toward the enunciative and non-enunciative aspects of discursive action that affect online politics.

Types of Harmful Discursive Actions

From a functionalist perspective, the notion of discourse means language in use (or in concrete situations of use)—or, ultimately, "language in action" (Gee, 2017, p. 3). On that basis, we propose the concepts of (i) discursive action, referring to specific language-use situations; and (ii) types of discursive action, designating a collection of language-use situations with regular aspects in terms of content, style, form, and communicative purpose (Bhatia, 2004). It is through the identification of such communicative purposes that we seek to analyze our corpus, based on the following typology (Table 1).

Uncivil Speech

Rude messages about public life (Maisel & Parker, 2016), which even before the Internet were already amplified via comments on the radio, television, and political platforms, have become perennial in forums, websites, and blogs (Borah, 2013) and more pervasive with the widening use of social media. While civility is a social norm referring to respectful and courteous behavior as well as "a discourse that does not silence or derogate alternative views but instead evinces respect" (Jamieson, Volinsky, Weitz, & Kenski, 2017, p. 206), incivility means disrespectful, hostile, and inflammatory commentary, which has to do with tone and content (Kim, Guess, Nyhan, & Reifler, 2021). Thus, the problem of uncivil discourse does not lie in the negative approach of the message, which can be civil or uncivil (Brooks & Geer, 2007), but in the toxic verve of the content, which can inflate the perception of social, identity, and ideological segregation.

Conspiratorial Speech

Conspiracy theories presuppose the existence of secret plans elaborated by powerful people to manipulate public events as a kind of global governance that "inhibits rights, alters bedrock institutions, and commits large-scale fraud" (Uscinski, 2020, p. 22). In this sense, conspiracy theories are structured around logical explanations, have elements of evidence, and cannot be falsified (Hendricks & Vestergaard, 2019). Conspiracies circulate periodically in society through spontaneous and/or stimulated routes (via conspiracy entrepreneurs), are propagated by dispersed and chained information systems, are adherent among those "cognitively available," and are activated by emotions and antagonism among groups.

Hate Speech

Hate is a strong, progressive, and enduring feeling of dislike, antipathy, and contempt for something or someone (Cambridge Dictionary, n.d.). Hate encourages racist, discriminatory, and stigmatizing attitudes against those who are, or are perceived to be, different because of sexual orientation, gender identity, disability, or nationality (Yong, 2011). Hatred triumphed, in the form of anti-Semitism, in Nazi ideology, continues to operate from extremist movements scattered around the world today through their social media accounts (Lytvynenko, Silverman, & Boutilier, 2019). Hate speech "spreads, incites,

promotes or justifies racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance" (Council of Europe, 1997, p. 107).

Dangerous Speech

Based on a set of conditions, some discursive forms have the potential risk of catapulting social unrest, massive violence, and, at the extreme end, genocide, in the medium and long term. The designation of dangerous speech represents a type of inflammatory form of expression that encourages, directly or indirectly, acts of violence against those perceived as different by social groups or by governments, considering momentary circumstances and national historical processes (Benesch, 2020). While the use of hateful language is broader, guided by identity biases such as race or religion, and regulated by national laws, dangerous speech means "catalyzing or amplifying violence by one group against another" and is activated by the presence of variables such as (a) the speaker, (b) the audience, (c) the speech act itself, (d) the social and historical context, and (e) the mode of dissemination (Benesch, 2013, p. 1).

Table 1. A Systematized Typology of Harmful Discursive Actions Related to Disinformation Order.

Discursive Action	Adopted Definition
Uncivil speech	Rude, impolite, and negative messages, usually containing offensive language and swear words to present ideas, programs, and agendas, and to refer to actors, institutions, and public affairs in general.
Conspiratorial speech	Messages that affirm the existence of hidden organizations (with people, powerful groups, and the media) that plot against the population, commit various frauds, and manipulate the course of the country.
Hate speech	Messages that express stigma and mobilize discriminatory action against minorities and vulnerable groups, focusing on their social, ethnic, gender, racial, and other identities.
Dangerous speech	Messages that stimulate the individual to behave in a certain way and to tolerate violence against "outsiders," following a process of imaginary construction about "others" based on fear, threat, and sense of danger, according to the vulnerability of the political context.

Enunciative and Non-Enunciative Aspects

Discursive actions imply individual language uses, allowing them to be analyzed in terms of utterances. Such uses correspond to interactional events, in which language comes into play (Benveniste, 1989). The products of these uses are utterances—which, in this study, consist of WhatsApp messages and Facebook posts. Therefore, on the one hand, we approach the enunciative aspects belonging to the utterances themselves—in terms of both their textual (phrases and sentences) and co-textual (images, sounds, graphic elements) dimensions. The non-enunciative aspects, on the other hand, comprise contextual components that surround the enunciation act of the utterance, such as the sender's profile (author) and targeted audience (readers) and the norms, practices, and values related to the community, to mention a few.

In this study, the non-enunciative elements are attributive to messages and posts identified as having contextually antidemocratic bias. The contextually antidemocratic bias defined here is expressed by factors that concern the institutional conditions of democracy in a given context—in this specific case, the performance and political program of the Far Right leader and former president of Brazil, Jair Bolsonaro.

Platforms and Harmful Discourse

As previously mentioned, studies on the circulation and perception of harmful speech tend to associate its greater presence with platforms outside the mainstream, such as Gab, Parler, Gettr, 4chan, and Reddit (Zannettou et al., 2018). The literature points to three main factors for this greater presence: The first is the lack of regulation on the part of the platforms, in the name of freedom of expression (Bollinger & Stone, 2022); the second is related to the affordances of these platforms that allow, among other things, the use of anonymous profiles (Siegel, 2020); finally, these platforms tend to concentrate radicalized groups, creating what Cinelli and colleagues (2022) have called "echo platforms."

WhatsApp, however, has a distinct characteristic from the platforms usually studied in research in the Global North. Used by more than 96% of Brazilians with an Internet connection, it cannot be considered a dark platform (Statista, 2022). Associated with the user's cell phone number, it cannot be described as a platform that encourages anonymity. Despite this, its closed structure does not allow for the moderation of content, creating safe spaces for the circulation of all kinds of harmful discourse. In other words, messaging apps like WhatsApp have a hybrid structure compared with other more radicalized platforms, but their social capillarity can have a huge impact on the circulation of harmful discourse. At the same time, some research shows that the plurality of uses of WhatsApp, with the presence of groups of family and friends, can reduce the incentive to publish harmful political content (Santos, Freitas, Aldé, Santos, & Cunha, 2019).

WhatsApp and Facebook differ due to important elements, such as algorithmic regulation, the possibility of moderation, and the size and involvement of their communities, with WhatsApp presenting—in 2020—a more private group structure compared with the possibilities offered by Facebook. In this sense, we believe that comparing platforms allows us to delve even deeper into the research carried out so far. We can operationalize it in two initial research questions.

RQ1: What are the characteristics of harmful political content on WhatsApp and Facebook, considering distinct discourse action types and contextual factors?

RQ2: Considering the differences between open and closed platforms, is the circulation of harmful political content more frequent on WhatsApp or Facebook?

Considering the relationship between discursive types, their social uses, and the platforms, recent research has shown that uncivil treatment leads to increased enthusiasm about the discourse (Kosmidis & Theocharis, 2020) and that it remains more frequently on news sites than on platforms such as Facebook, especially when the subject involved engenders disagreement and polarization (Rossini & Maia, 2021). Adding to that, Kim and colleagues (2021) claim that on platforms under

algorithmic governance, these uncivil messages are more seen by others, which increases the visibility for toxicity. Following Rossini and Maia (2021), the threshold of what is democratically undesirable is crossed when incivility reflects intolerance.

Conspiracy thinking is a common element in different societies, which are marked by periods of greater and lesser adherence. Since the emergence of the Internet, conspiracy theories could be found on the Web, but access to them depended on a targeted and interested search. However, the contemporary dynamics of visibility of platforms change this, creating spaces for conspiracies to circulate on both dark platforms (Zeng & Schäfer, 2021) and mainstream platforms. In a study based on panel data from 17 countries, Theocharis and colleagues (2023) show that the symmetrical structure of Twitter constrains the circulation of conspiracies, while platforms such as YouTube, WhatsApp, Facebook, and Messenger might enable it.

Recognizing that online discriminatory actions are directly connected to different global contexts, Pohjonen (2019) advocates that the notion of online hate speech is better understood when seen through its performative function. Along with anonymity, the communities that do not require identity exposure and physical proximity are a delimiting feature of digital hate speech. This last characteristic also marks the profusion of dangerous speeches, especially in radicalized groups and spaces with low online visibility.

Types of Harmful Discursive Actions in Far Right Communication

The relationship between platforms and discourses, however, cannot be dissociated from the political contexts and repertoires of action that characterize the Far Right wave, which has gained social and institutional prominence in recent years (Bennett & Livingston, 2018). One of the main characteristics of these movements is, in addition to their radicalism, the demarcation of public discourses that go against universalist justifications and strain the logic of the democratic public sphere (Korstenbroek, 2022). This suggests that there are, therefore, different discursive logics that need to be balanced for the movements to survive and gain social and institutional space.

Despite being intimidating and confrontational, uncivil discourse can be reframed from the logics of polemic and humor, expanding the reach of messages through incivility (Nilsson, 2021). The use of uncivil language in social media comments and publications often affects the way users discuss real-world events and has "the potential to harmfully exacerbate group-based tensions" (Hiaeshutter-Rice & Hawkins, 2022, p. 11). During electoral processes, opponents and campaigns often engage in open attacks and counterattacks, using incivility as a political strategy (Brooks & Geer, 2007).

Hate speech and dangerous speech point to different targets and degrees of radicalization. However, both share a common characteristic: Due to their openly violent and confrontational nature, they tend not to be useful for expanding out-group messages—that is, they tend to be more effective for communication among those who are already radicalized or are in the process of radicalization (Ferguson, 2016). Despite this, hate speech and extremist content can be prioritized by social media recommendation systems, reaching an even wider audience than before (O'Callaghan, Greene, Conway, Carthy, & Cunningham, 2015).

Conspiratorial speeches, on the other hand, have a very distinct characteristic. They are not necessarily uncivil and can use sophisticated and complex arguments. One example is the way in which conspiracies about COVID-19 have been mobilized, with scientific and technical arguments (Oliveira, Wang, & Xu, 2022). Another important characteristic of conspiratorial discourse is that, unlike the hateful and the dangerous, it targets groups that are perceived as powerful. Conspiracy involves a minority group acting covertly to override the majority. In other words, compared with the other three types of discourse, the conspiratorial seeks to occupy a legitimate space in the public debate and can allow for a more perennial maintenance of the narrative lines linked to the political extremist (van Prooijen, Krouwel, & Pollet, 2015). Institutionally speaking, conspiracies can help expand the audience reached by the Far Right.

Finally, we understand that the institutional and social consolidation of the Far Right in Brazil involves the establishment of a communicative environment that can cause democratic damage even without the mobilization of specific discursive elements. The inclusion of a category referring to non-enunciative messages with contextually antidemocratic bias allows for a broader view of the expressions of this phenomenon in the digital public debate. In this sense, considering the Brazilian context and from a more exploratory point of view, we seek to answer two more research questions and test two hypotheses:

RQ3: Do the types of discourse action show any association with each other? If so, what kind of association?

RQ4: How do these associations differ across platforms (WhatsApp and Facebook)?

H1: Harmful political content is anchored more often in non-enunciative elements (contextual factors) than in enunciative elements (types of discursive action).

H2: Messages that mention the electoral dispute, and therefore are related to a polarized scenario in which candidates and the demographics more aligned to them are involved in contention, are associated with uncivil discourses.

Materials and Methods

From an exploratory-descriptive perspective, the empirical analysis of this study uses a quantitative approach to examine, specify, and compare the harmful political content, taking into account enunciative and non-enunciative aspects, in a more private medium (WhatsApp) and a more public medium (Facebook). While WhatsApp messages are protected by end-to-end encryption, are not mediated by relevance algorithms, and until recently were not parameterized by metrics (such as reactions, comments, and shares), Facebook provides some publicity of its metrics, allows some possibility of external monitoring through an application programming interface, and applies warnings to posts verified by fact-checking initiatives partners.

In this sense, as a methodological starting point, it is necessary to consider that the structure of the WhatsApp network of interconnected groups is shaped by social appropriation (it is common to have groups for

everyday matters), voluntary segmentation (people receive all the information that comes to a group; therefore, it is common to leave groups that repeatedly send content with which one strongly disagrees even if they are groups formed by users' family members), and social ubiquity (free data use packages make WhatsApp popular even among the population that otherwise does not have access to the Internet).

Data Collection

Given Brazil's recent political history of growing right-wing authoritarianism (Pinheiro-Machado & Scalco, 2020) and based on the understanding that WhatsApp is the main means for Brazilians to access the Internet and news consumption (Toff et al., 2021), WhatsApp was adopted as a reference in the process of data collection. Therefore, the first step was to select public groups representing the radical Right in Brazil and already monitored longitudinally from 513 groups of the messaging app.

For this, through an initial filter, all public WhatsApp groups that mentioned, in their titles, the terms "right," "conservatives," "patriotic," and symbols of the Brazilian radical Right were selected. These symbols include (a) praising patriotism and military intervention; (b) aversion to opponents based on ideologies, parties, politicians, and institutions; and (c) support for other right-wing national political figures (Appendix 1). After this step, we found 96 groups that were concentrated with hard-core supporters. This stage resulted in 766,865 messages spread in public groups on WhatsApp from September 27 to November 29, 2020.

From this first corpus, a lexical analysis was performed to elaborate linguistic structures (Ruediger, 2017) that could identify similar public groups on Facebook to guide data collection on this platform. We use the word2vec algorithm to associate groups based on the frequency of co-occurrence (and proximity) among words. Seven clusters of words were identified, four of which were discarded because they did not match with the scope of this project. The three selected thematic axes addressed "coronavirus and China," "Bolsonaro and institutions," and "criminalization of the left." Based on these themes, rules were developed to guide searches on Facebook using the free CrowdTangle tool.

This step found 1,994 public Facebook groups adhering to these thematic axes and with the potential for circulation of harmful political content in general. Regular expressions were then applied to this database to exclude public groups that might have been covered by the queries but do not correspond to accounts linked to the radical Right. This process resulted in 1,580 public Facebook groups, within which 2,736,675 posts were made during the analysis period. This period included the start of the official election propaganda authorized by the Electoral Tribunal, the day of voting (November 15) in the first round, and the day of the second round (November 29) of the presidential election. The election occurred under the presidency of Jair Bolsonaro (2018–2022), whose stance in favor of militarism and conservative policies contributed to the growth of radical Right groups that use a combination of social media and offline tactics for organization and influence (Rocha, 2021).

Sample and Data Representation

Five subsequent steps were carried out in the sample delimitation process (n = 2,201). First, we considered only unique messages, discarding repetitions present in the data collection. Second, we created an engagement ranking corresponding to the number of shares on Facebook and the number of forwarded

messages identified in the database for WhatsApp. Third, we separated messages with less than 70 characters to join them to possible previous messages sent in sequence, in cases where the time between each message did not exceed 20 seconds (50,467 on WhatsApp and 403,435 on Facebook). Fourth, we separated 20% of messages with the most engagement, considering all the remaining messages (10,093 on WhatsApp and 80,687 on Facebook). Fifth, we composed a random sample with a defined size so that there was a 95% degree of confidence and a 3% margin of error (966 on WhatsApp and 1,054 on Facebook).

Measures and Reliability

Three researchers manually coded the sample of 2,201 unique messages. Coding was based on a set of six different independent variables, with binary responses (1 = Yes; 2 = No). To measure the types of discursive action of harmful political content on WhatsApp and Facebook, the coders analyzed the content globally, which means considering the text of the post itself, the links, and the videos, where applicable. Once coded, the evaluations of these posts were transferred to all posts with the texts evaluated, considering their different repetitions in the corpus, enabling the crossing of textual analysis, engagement, and attributes obtained through human coding. To accelerate the codification stage, the variables were written in question form. The reliability test was performed in 10% of the analyzed sample, and Krippendorff's alpha was calculated (Table 2).

Table 2. Exploring Types of Discursive Action.

Variable	Question	Krippendorff's α
Incivility	Does the message . . . sound rude, offensive, crude, derogatory, or does it express disaffection or hostility?	0.858
Conspiracy	Does the message . . . insinuate, denounce, or affirm the existence of hidden entities, manipulation projects, and conspiratorial speculations?	0.743
Hate	Does the message . . . discriminate, stigmatize, or target the identity of vulnerable subjects, minorities, and groups?	1.000
Dangerous	Does the message . . . construct an idea of threat or sense of danger from "others" that can generate violence in the short, medium, or long term?	1.000
Contextually antidemocratic bias	Does the message . . . present aspects beyond discourse that contextually represent institutional insecurity, political antagonism, scientific denialism, censorship of contestation, censorship of free speech, contentions, discourse against pluralism, or uses of satire and irony to belittle political opponents?	0.803

In the case of the contextually antidemocratic bias, the coders were asked to identify markers that indicated contentious and antagonistic discourse, attacks on institutions, or any rhetorical appeal that signaled institutional insecurity, scientific denialism, censorship, or contestation of freedom of expression, discourses against political pluralism in general, or the discretionary and derogatory use of satire, irony, and debauchery to belittle political opponents. This variable can present contextual and para-discursive elements

since it demands a circumstantial understanding of the political conjuncture of a given region. In Brazil, for example, the evocation of an "anti-Petista" and anticommunist discourse has been associated with an antagonistic condition that fosters the suppression of the so-called progressive sphere in general.

Results and Discussion

This section presents and discusses the results of the quantitative analyses, starting with a simple statistical description and followed by binary logistic regression models. In RQ2, we asked whether harmful political content was more frequent on WhatsApp than on Facebook. The result of the analysis shows that harmful political content was identified more frequently on Facebook than on WhatsApp. Specifically, the analysis found that uncivil and conspiratorial speeches were the most frequent types of discursive action circulating in radical Right public groups. Hate and dangerous speeches appeared less frequently, with hate speech appearing twice as often as dangerous speech. In turn, harmful political content was identified through contextual factors, particularly contextually antidemocratic bias in this analysis, which was predominant compared with the four types of discursive action (Table 3).

Table 3. Absolute Number and Percentage of Speech Types by Platform.

	Facebook				WhatsApp			
	n	% Per Type of Speech	% On the Platform	Standardized Residuals	n	% Per Type of Speech	% On the Platform	Standardized Residuals
Uncivil speech	182	66.42	17.07	1.1381094	92	33.58	8.93	−1.4248965
Conspiratorial speech	141	54.44	13.23	−1.3616569	118	45.56	11.46	1.7047747
Hate speech	44	67.69	4.13	0.6852608	21	32.31	2.04	−0.8579366
Dangerous speech	19	63.33	1.78	0.1599819	11	36.67	1.07	−0.200295
Contextually antidemocratic bias	462	60.71	43.34	−0.1205774	299	39.29	29.03	0.1509611

Source: The authors.

A comparison of the platforms revealed that uncivil speech was more frequent on Facebook (std. res. = 1.13) than on WhatsApp (std. res. = −1.42), conspiratorial speech was a lot more frequent on WhatsApp (1.7) than on Facebook (std. res. = −1.36), and hate speech was reasonably more frequent on Facebook (std. res. = 0.68) than on WhatsApp (std. res. = −0.85). The platform analysis shows that, on Facebook, there was more uncivil speech, followed by conspiratorial speech, and less hateful and dangerous speech. On WhatsApp, conspiratorial speech stood out from the others, but there was also a prominent presence of uncivil speech, with less frequency of hateful and dangerous speech. On both platforms, harmful political content revealed by contextual factors was more frequent than the types of discursive action, which was more significant on Facebook than on WhatsApp (Table 4).

Table 4. Posts Mentioning the Elections per Type of Harmful Speech and per Platform.

	Facebook			WhatsApp		
	n	% In Relation to the Category n	Standardized Residuals	n	% In Relation to the Corpus	Standardized Residuals
Uncivil speech	41	22.53	0423	18	19.57	−0577
Conspiratorial speech	20	14.18	−0.58	15	12.71	0791
Hate speech	4	9.09	0.0491	2	9.52	−0067
Dangerous speech	4	21.05	0415	1	9.09	−0566
Contextually antidemocratic bias	65	14.07	−0.0862	36	12.04	0118

Note. CI = confidence interval. Source: The authors.

A small fraction of messages mentioned the 2020 municipal elections in Brazil, which demonstrates that the harmful content circulating in radical Right groups did not reflect official political campaigns or direct disputes between government plans. When it appeared, harmful political content identified through discursive type and contextually antidemocratic bias was more frequent on Facebook in all cases (Table 3).

Binary Logistic Regression Models

We analyzed binary logistic regression models to establish the dependence relationship between the discursive variables. Logistic regression is used to determine the probability of occurrence of a given event or class among a set of explanatory variables. Binary logistic regression refers to a dependent variable of categorical and binary character (true or false). Its objective is to present a model that enables the identification of factors that contribute to the occurrence of a given event, among a set of determinants selected a priori.

In this study, models were created considering each variable for the types of discursive action in relation to the full set of analyzed messages and according to each of the compared platforms (WhatsApp and Facebook). For instance, to determine which variables could explain uncivil speech, a binary logistic regression model was used that considered all other types of speech (conspiratorial speech, hate speech, dangerous speech, and harmful speech with contextually antidemocratic bias) in relation to the full set of messages. Then, these same variables were analyzed separately according to binary logistic regression models that considered only coded content from WhatsApp or only coded content from Facebook.

The same procedure was followed for all other variables, always assuming the variable to be analyzed as the dependent variable and all other discursive type variables as independent variables, which were then considered explanatory. The objective of this analysis was to understand how the discursive types were associated with each other, and how the incidence of a given type could determine the occurrence of another probabilistically.

The comprehensive analyses that sought to establish comparative patterns between the two platforms (WhatsApp and Facebook) will be presented in more detail in the text, while the results concerning the five main discursive types analyzed in relation to the full set of messages in the database for this study are summarized in Table 5.

The first observation arising from the binary logistic regression models is that all types of discursive action are significantly and reciprocally associated with contextually antidemocratic biased discourse. That is, all the other types of discursive action are explained by its presence, and vice versa.

The odds ratio (OR) values expressed in Table 5 demonstrate that the contextually antidemocratic bias can predict the incidence of other types of discursive action with reasonable accuracy, and it is also significantly explained by the presence of these same types. Due to this constancy, we will first emphasize the other results and then detail the discourse with contextually antidemocratic bias.

Table 5. Binary Logistic Regression Models.

	Predictors	Coef.	χ^2	*p* Value	OR	95% CI Lower	95% CI Upper
Uncivil speech	**Conspiratorial speech**	−.485	6317	.01196	.0616	.422	.899
	Hate speech	.469	2798	.09440	1598	.923	2767
	Dangerous speech	.966	6317	.01196	2627	1237	5580
	Contextually antidemocratic bias	1671	117663	< .001	5319	.053	.008
Nagelkerk's R^2	.140						
Conspiratorial speech	**Uncivil speech**	−.485	6334	.01185	.615	.421	.898
	Hate speech	−1135	8489	.00357	.321	.149	.689
	Dangerous speech	−1209	3789	.05160	.298	.088	1.00
	Contextually antidemocratic bias	3711	215429	< .001	40899	24917	.022
Nagelkerke's R^2	.368						
Hate speech	**Uncivil speech**	.450	2585	.10785	1569	0906	2716
	Conspiratorial speech	−1129	8389	.00378	.323	.150	.694
	Dangerous speech	.685	1737	.18746	1983	.716	5491
	Contextually antidemocratic bias	3775	39409	< .001	43599	13416	141692
Nagelkerke's R^2	.238						
Dangerous speech	**Uncivil speech**	.951	6109	.01345	2589	1217	5506
	Conspiratorial speech	−1196	3699	.05445	.302	.0893	1023
	Hate speech	.691	1762	.18435	1995	.719	5533
	Contextually antidemocratic bias	3134	17297	< .001	22976	5245	100647
Nagelkerke's R^2	.200						
Contextually antidemocratic bias	**Uncivil speech**	1673	116566	< .001	5327	3932	7218
	Conspiratorial speech	3723	216715	< .001	41420	25229	68003
	Hate speech	3779	39400	< .001	43769	13449	142435
	Dangerous speech	3146	17321	< .001	23239	23249	5283
Nagelkerke's R^2	.394						

Note. CI = confidence interval. Source: The authors.

First, uncivil speech was significantly predictable by the coincidence of dangerous speech (β = −.966, $p < .05$) and was opposed to conspiratorial speech (β = −.485, $p < .05$). That is, the more a discourse presented itself as conspiratorial, the less likely it was that it also constituted uncivil speech. Dangerous speech was approximately 2.6 times more likely to be classified as uncivil speech (OR = 2.627), that is, there was a predictive correlation between these two variables.

Conspiratorial speech, in turn, was significantly explained by the nonoccurrence of hate speech (β = −1.135, $p < .01$) and uncivil speech (β = −.485, $p < .05$), in that order. Simply put, the more a message had elements classifiable as hate speech or uncivil speech, the less likely it was to constitute conspiratorial speech. This result is consistent with the evaluation of conspiratorial speech as a speech that is, often, polite and does not necessarily contain any evidence of discriminatory rhetoric.

Correspondingly, hate speech was also significantly explained by the nonoccurrence of conspiratorial speech (β = −1,129, $p < .01$). This conclusion may seem counterintuitive at first, especially within the Brazilian context, since some messages were laden with content that was phobic about lesbians, gays, bisexuals, transgenders, and queers and simultaneously evoked conspiracies based on expressions such as "gender ideology," an idea according to which the progressive sphere is allegedly interested in "indoctrinating" children and "converting" them into homosexuals or transsexuals. However, the data seem to indicate that there was no convergence between these speech types in the period analyzed.

Dangerous speech, with messages that more openly incite violence, was significantly explained by uncivil speech (β = .951, $p < .01$). It was 2.58 times more likely that a speech classified as uncivil was also classified as dangerous (OR = 2.589). This association was also reciprocal, and it is possible to say that these types largely cooccur; that is, messages that contain offensive or rude language are more likely to incite violence, and the opposite is also true.

Last, harmful political content with contextually antidemocratic bias was also significantly predicted by uncivil speech (β = 1.673, $p < .01$), conspiratorial speech (β = 3.723, $p < .01$), hate speech (β = 3.779, $p < .01$), and dangerous speech (β = 3.146, $p < .01$). That is, the binary logistic regression model shows that all four discursive types helped explain the dependent variable. For instance, it was 43.7 times more likely that hate speech was also antidemocratic, 41.4 times more likely that conspiratorial speech was also antidemocratic, 23.2 times more likely that dangerous speech was also antidemocratic, 5.3 times more likely that uncivil speech was also antidemocratic (cf. OR). The model demonstrates how close these independent variables are to the dependent variable. However, as in the other models presented above, Nagelkerke's R^2 was relatively low (.394), which suggests that these variables were explanatory in only about 40% of the cases; that is, there was a reasonable number of arrangements in which these variables were not sufficient to explain the contextually antidemocratic biased discourse variable. The results suggest that, since the external characteristics of the groups are not being controlled, other elements may interfere in this situation. A subsequent analysis using post-estimation tests for omitted variable bias could perhaps help with this assessment. However, the variables selected here undoubtedly helped predict the dependent variable in a statistically significant way.

Regarding the differences between the platforms WhatsApp and Facebook, there were some small discrepancies between the data when considering the databases separately. For uncivil speech, for example, considering only the database of messages circulated in WhatsApp groups, dangerous speech (β = 1.389, OR = 4.008, $p < .05$) was the only one with explanatory potential, while considering only the database of messages circulated on Facebook, the absence of conspiratorial speech (β = −.658, OR = .518, $p < .01$) was the only one that presented a significant association. In summary, the association between uncivil speech and dangerous speech was more evident on WhatsApp than on Facebook, while the opposition between uncivil speech and conspiratorial speech was clearer on Facebook than on WhatsApp.

Something similar happened with conspiratorial speech. Considering only WhatsApp, there was a greater association between conspiratorial speech and the absence of hate speech (β = −2.387, OR = .092, $p < .05$) than there was on Facebook. On the other hand, uncivil speech (β = −.656, OR = .519, $p < .01$) was the only one significantly (and negatively) associated with conspiratorial speech, considering only the Facebook corpus.

Hate speech on WhatsApp was significantly and negatively associated with conspiratorial speech (β = −2.394, OR = .091, $p < .05$). On Facebook, there was no statistically significant predictive variable for this type of speech.

Binary logistic regression models also showed that dangerous speech was statistically associated with uncivil speech (β = −1.335, OR = 3.799, $p < .05$) on WhatsApp, but there was no statistically significant predictive variable on Facebook.

These differences suggest that there is a reciprocal and negative association between conspiratorial speech and hate speech that is more evident in the WhatsApp environment than on Facebook. On the other hand, conspiratorial speeches seem to be more associated with uncivil speech on Facebook than on WhatsApp.

Similarly, the reciprocal association between uncivil speech and dangerous speech is stronger in messages circulating in WhatsApp groups than in messages circulating on Facebook. This indicates that WhatsApp groups are potentially more radicalized and prone to inciting violence than Facebook groups, which could be a result of an environment with greater structural opacity and more timid content moderation (Chagas, 2023).

Last, we also sought to analyze how each of these discursive types affected messages that circulated in the observed period and made direct mention of the electoral race (Table 6). For that purpose, a last binary logistic regression model was developed using "messages that mention the electoral race" as a dependent variable and all the other five variables as independent. The result clearly demonstrated that messages about elections had a statistically significant relationship with uncivil speech; that is, among all discursive types, uncivil speech was the only one with an explanatory factor for this variable. The conclusion to be drawn from this analysis is that there is an association between messages about the elections that circulate in the WhatsApp and Facebook environments and messages that are offensive and rude.

Table 6. Binary Logistic Regression Model for Messages That Mention the Elections.

	Predictors	Coef.	χ^2	*p* Value	OR	95% CI Lower	95% CI Upper
Messages that mention the electoral race	**Uncivil speech**	.838	22681	.6902	2311	.106	3262
	Conspiratorial speech	.105	.226	< .001	1111	.678	1712
	Hate speech	.539	1453	.6348	.583	1637	1401
	Dangerous speech	.127	.061	.2281	1135	.720	3098
	Contextually antidemocratic bias	-.066	.159	.8047	.936	.243	1294
Nagelkerke's R^2	.0209						

Source: The authors.

The data presented in the analysis carried out through binary logistic regression models not only highlight striking relationships among the discursive types but also call attention to the fact that some explanatory reasons for the incidence of messages with contextually antidemocratic bias may not be restricted to enunciative-explicit discourses. Furthermore, there seems to be a coincidence between messages that mention the electoral race and a specific discursive type, uncivil speech. These observations support H1 and H2.

Conclusion

Our study aimed to investigate the presence of harmful discursive action of political content spread in public groups on Facebook and WhatsApp during the Brazilian municipal elections of 2020. The article contributes by addressing the problem of disinformation order from the perspective of enunciative and non-enunciative harmful types, which we understand to constitute an important part of the sense of harmfulness of typically digital communication such as fake news, memes, hyper-partisan clickbait, chains, and others.

First, the analysis highlights the existence of distinct types of discursive action and the associations among them. Uncivil speech stands out more on Facebook and conspiratorial speech stands out more on WhatsApp although they are the two most frequent types on both platforms. Hate speech and dangerous speech were more frequent on Facebook than on WhatsApp. Regarding the types of discourse action, the research identified that (a) uncivil speech is usually associated with dangerous speech and opposed to conspiratorial speech, (b) dangerous speech correlates with uncivil speech, (c) conspiratorial speech is not related to hate speech and uncivil speech, and last, (d) hate speech is not associated with conspiratorial speech.

The large presence of conspiratorial discourse and its non-association with types such as hate and uncivil suggests that its use seeks to occupy a legitimate place in public debate, escaping the typically aggressive rhetoric of Far Right supporters. Future studies could delve deeper into how this can contribute to the Far Right's institutional and electoral presence.

Regarding the platforms, our study shows that the association between uncivil and dangerous speech on WhatsApp is more evident than on Facebook, and the opposition between uncivil and conspiratorial speech on Facebook is more evident than it is on WhatsApp. In turn, we also show a reciprocal and negative association between conspiratorial speech and hate speech, which is more evident on WhatsApp than on Facebook, and a greater association between conspiratorial speech and uncivil speech on Facebook than on WhatsApp. Last, uncivil and dangerous speeches are more related to WhatsApp than to Facebook. Contextually antidemocratic bias, as a non-enunciative aspect, explains harmful content and is frequent on both platforms. Contextual antidemocratic bias explain harmful political content both linked to the type of discursive action and in isolation in an eminently non-enunciative character. Harmful political content that mentioned municipal elections hardly circulated in radical Right public groups, and, when it appeared, uncivil discourse was its only explanatory factor.

The present study focuses specifically on the association among different discursive types that constitute democratically harmful content. One limitation arising from this choice concerns the non-characterization of the groups that make up the research sample. Future studies can focus on understanding how the characteristics of the groups predict some kind of discourse. Understanding how the circulation of different types of socially harmful discourses is related to the interaction dynamics in these "micro" sociability environments is a promising path for future research.

These findings deserve attention because they shed light on the importance of considering conditions related to the political context (Salgado, 2019) to understand the harmfulness of online political content. Public groups are arenas that amplify aggressive and dirty political campaigns—during elections or unofficial campaigns—based on the dissemination of incivility, conspiracy, and, in many cases, hatred and danger. The fact that hateful and dangerous speeches appear less frequently does not mean that the type of harm is greater or lesser, which has not been gauged in this analysis. Subsequent research may move toward ascertaining correlations between types of discursive action and presumed public harm, as the quality of public discussion, the integrity of the public sphere, and democratic health are contemporary concerns of the first order. We emphasize that the research findings are limited to content circulated in radical Right public groups in a specific period. The development of research with a wider coded sample would be interesting to reinforce these findings.

References

Banko, M., MacKeen, B., & Ray, L. (2020). A unified taxonomy of harmful content. In Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran & Zeerak Waseem (Eds.). *Proceedings of the Fourth Workshop*

on Online Abuse and Harms (pp. 125–137). Kerrville, TX: Association for Computational Linguistics.

Benesch, S. (2013, February 23). *Dangerous speech: a practical guide*. Retrieved from https://dangerousspeech.org/guide/

Benesch, S. (2020). Proposals for improved regulation of harmful content online. In Y. Shany (Ed.), *Reducing online hate speech*: *Recommendations for social media companies and internet intermediaries* (pp. 247–306). Jerusalem: Israel Democracy Institute.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication, 33*(2), 122–139. doi:10.1177/0267323118760317

Benveniste, E. (1989). *Problemas de linguística geral II* [Problems in general linguistics]. Campinas, Brazil: Pontes.

Bhatia, V. (2004). *Worlds of written discourse: A genre-based view*. London, UK: Continuum. doi:10.1017/S027226310633029X

Bollinger, L. C., & Stone, G. R. (Eds.). (2022). *Social media, freedom of speech, and the future of our democracy* (online ed.). New York, NY: Oxford Academic. doi:10.1093/oso/9780197621080.001.0001

Borah, P. (2013). Interactions of news frames and incivility in the political blogosphere: Examining perceptual outcomes. *Political Communication, 30*(3), 456–473. doi:10.1080/10584609.2012.737426

Bradshaw, S., & Howard, P. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation. In Samuel Woolley & Philip N. Howard (Eds.). *Computational propaganda research project* (pp. 1–37). Oxford, UK: Oxford Internet Institute.

Brooks, D., & Geer, J. (2007). Beyond negativity: The effects of incivility on the electorate. *American Journal of Political Science, 51*(1), 1–16. doi:10.1111/j.1540-5907.2007.00233.x

Cambridge Dictionary. (n.d.). *Hate.* Retrieved from https://dictionary.cambridge.org/pt/dicionario/ingles/hate

Chagas, V. (2023). Far-right memespheres and platform affordances: The effects of environmental opacity on the spread of extremist memes on Twitter and WhatsApp. *Journal of Applied Communication Research, 51*(6), 702–719. doi:10.1080/00909882.2023.2290897

Cinelli, M., Etta, G., Avalle, M., Quattrociocchi, A., Di Marco, N., Valensise, C., . . . Quattrociocchi, W. (2022). Conspiracy theories and social media platforms. *Current Opinion in Psychology, October*(47), 101407. doi:10.1016/j.copsyc.2022.101407

Council of Europe. (1997). *Recommendation No. R (97) 20 of the committee of ministers to member states on "hate speech" adopted on October 30, 1997*. Retrieved from https://rm.coe.int/1680505d5b

Dourado, T. (2023). Who posts fake news? Authentic and inauthentic spreaders of fabricated news on Facebook and Twitter. *Journalism Practice, 17*(10), 2103–2122. doi:10.1080/17512786.2023.2176352

Faris, R., Ashar, A., Gasser, U., & Joo, D. (2016). Understanding harmful speech online. *Berkman Klein Center Research Publication, 2016–21*. doi:10.2139/ssrn.2882824

Ferguson, K. (2016). *Countering violent extremism through media and communication strategies: A review of the evidence*. Partnership for Conflict, Crime & Security Research. Retrieved from http://www.paccsresearch.org.uk/wp-content/uploads/2016/03/Countering-Violent-Extremism-Through-Media-and-Communication-Strategies-.pdf

Gee, P. (2017). *Introducing discourse analysis: From grammar to society*. London, UK: Routledge.

Hendricks, V. F., & Vestergaard, M. (2018). *Reality lost: Markets of attention, misinformation and manipulation*. Cham, Switzerland: Springer Open.

Hiaeshutter-Rice, D., & Hawkins, I. (2022). The language of extremism on social media: An examination of posts, comments, and themes on Reddit. *Frontiers in Political Science, 4*, 805008. doi:10.3389/fpos.2022.805008

Jamieson, K. H., Volinsky, A., Weitz, I., & Kenski, K. (2017). The political uses and abuses of civility and incivility. In K. H. Jamieson & K. Kenski (Eds.), *The Oxford handbook of political communication* (pp. 205–218). New York, NY: Oxford University Press.

Jiang, J. A., Scheuerman, M. K., Fiesler, C., & Brubaker, J. R. (2021). Understanding international perceptions of the severity of harmful content online. *PLOS ONE, 16*(11), 1–22. doi:10.1371/journal.pone.0256762

Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication, 71*(6), 922–946. doi:10.1093/joc/jqab034

Korstenbroek, T. (2022). Rethinking the public sphere in an age of radical-right populism: A case for building an empathetic public sphere. *Communication Theory, 32*(1), 68–87. doi:10.1093/ct/qtab005

Kosmidis, S., & Theocharis, Y. (2020). Can social media incivility induce enthusiasm? Evidence from survey experiments. *Public Opinion Quarterly, 84*(S1), 284–308. doi:10.1093/poq/nfaa014

Lytvynenko, J., Silverman, C., & Boutilier, A. (2019). *White nationalist groups banned by Facebook are still on the platform*. Retrieved from https://www.buzzfeednews.com/article/janelytvynenko/facebook-white-nationalist-ban-evaded

Maisel, L. S., & Parker, K. D. (2016). The negative consequences of uncivil political discourse. *Political Science and Politics, 45*(3), 405–411. doi:10.1017/S1049096512000467

Miller, M. L., & Vaccari, C. (2020). Digital threats to democracy: Comparative lessons and possible remedies. *International Journal of Press/Politics, 25*(3), 333–356. doi:10.1177/1940161220922323

Nilsson, P. (2021). "The new extreme right": Uncivility, irony, and displacement in the French re-information sphere. *Nordicom Review, 42*(s1), 89–102. doi:10.2478/nor-2021-0008

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (White) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review, 33*(4), 459–478. doi:10.1177/0894439314555329

Oliveira, T., Wang, Z., & Xu, J. (2022). Scientific disinformation in times of epistemic crisis: Circulation of conspiracy theories on social media platforms. *Online Media and Global Communication, 1*(1), 164–186. doi:10.1515/omgc-2022-0005

Parekh, B. (2012). Is there a case for banning hate speech? In M. Herz & P. Molnar (Eds.), *The content and context of hate speech: Rethinking regulation and responses* (pp. 37–56). Cambridge, UK: Cambridge University Press.

Pinheiro-Machado, R., & Scalco, L. M. (2020). From hope to hate: The rise of conservative subjectivity in Brazil. *HAU: Journal of Ethnographic Theory, 10*(1), 21–31. doi:10.1086/708627

Pohjonen, M. (2019). Extreme speech| A comparative approach to social media extreme speech: Online hate speech as media commentary. *International Journal of Communication, 13*, 3088–3103.

Rocha, C. (2021). From Orkut to Brasília: The origins of the new Brazilian right. In K. Hatzikidi & E. Dullo (Org.), *A horizon of (im)possibilities. A chronicle of Brazil's conservative turn* (1st ed., pp. 1–195). London, UK: University of London Press.

Rossini, P., & Maia, R. (2021). Characterizing disagreement in online political talk: Examining incivility and opinion expression on news websites and Facebook in Brazil. *Journal of Deliberative Democracy, 17*(1), 90–104. doi:10.16997/jdd.967

Ruediger, M. (Coord.). (2017). *Not so #simple: The challenge of monitoring public policies on social networks*. Rio de Janeiro, Brasil: FGV DAPP.

Ruediger, M., & Grassi, A. (Coord.). (2020). *Online disinformation and elections in Brazil: The circulation of links about mistrust in the Brazilian election system on Facebook and YouTube (2014–2020).* Rio de Janeiro, Brasil: FGV DAPP.

Salgado, S. (2019). Never say never . . . Or the value of context in political communication research. *Political Communication, 36*(4), 671–675. doi:10.1080/10584609.2019.1670902

Santos, J. B., Freitas, M., Aldé, A., Santos, K., & Cunha, V. (2019). WhatsApp, política mobile e desinformação: A hidra nas eleições presidenciais de 2018 [WhatsApp, mobile politics and misinformation: The hydra of Brazil's 2018 presidential election]. *Comunicação & Sociedade, 41*(2), 307–334. doi:10.15603/2175-7755/cs.v41n2p307-334

Siegel, A. (2020). Online hate speech. In N. Persily & J. Tucker (Eds.), *Social media and democracy* (pp. 56–88). Cambridge, UK: Cambridge University Press.

Soares, F. B., Recuero, R., & Zago, G. (2018). Influencers in polarized political networks on Twitter. In *Proceedings of the 9th International Conference on Social Media and Society* (pp. 168–177). New York, NY: Association for Computing Machinery. doi:10.1145/3217804.3217909

Statista. (2022). *Social media usage in Brazil*. Retrieved from https://www.statista.com/study/68696/social-media-usage-in-brazil/

Theocharis, Y.,Cardenal, A., Jin, S., Aalberg, T., Hopmann, D. N., Strömbäck, J., . . . Štětka, V. (2023). Does the platform matter? Social media and COVID-19 conspiracy theory beliefs in 17 countries. *New Media & Society, 25*(12), 3412–3437. doi:10.1177/14614448211045666

Toff, B., Badrinathan S., Mont'Alverne C., Ross Arguedas A., Fletcher R., & Nielsen R. (2021). *Overcoming indifference: What attitudes towards news tell us about building trust*. Oxford, NY: Reuters Institute for the Study of Journalism. Retrieved from https://reutersinstitute.politics.ox.ac.uk/depth-and-breadth-how-news-organisations-navigate-trade-offs-around-building-trust-news

Uscinski, J. (2020). *Conspiracy theories: A primer*. New York, NY: Rowman & Littlefield Publishers.

van Prooijen, J.-W., Krouwel, A. P. M., & Pollet, T. V. (2015). Political extremism predicts belief in conspiracy theories. *Social psychological and personality science, 6*(5), 570–578. doi:10.1177/1948550614567356

Yong, C. (2011). Does freedom of speech include hate speech? *Res Publica, 17*(4), 385–403. doi:10.1007/s11158-011-9158-y

Zannettou, S., Bradlyn, B., De Cristofaro, E., Sirivianos, M., Stringhini, G., Kwak, H., & Blackburn, J. (2018). What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of The Web Conference* (pp. 1007–1014). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. doi:10.1145/3184558.3191531

Zeng, J., & Schäfer, M. S. (2021). Conceptualizing "dark platforms". Covid-19-related conspiracy theories on 8kun and Gab. *Digital Journalism, 9*(9), 1321–1343. doi:10.1080/21670811.2021.1938165