

Bubble reachers and uncivil discourse in polarized online public sphere

Jordan K Kobellarz¹, Milos Brocic^{2,✉}, Daniel Silver², Thiago H Silva^{1*},

¹ Informatics, Universidade Tecnológica Federal do Paraná, Curitiba, Brazil

² Sociology, University of Toronto, Toronto, Canada

✉Current Address: Sociology, McGill University, Montreal, Canada

* thiagoh@utfpr.edu.br (THS)

Abstract

Early optimism saw possibilities for social media to renew democratic discourse, marked by hopes for individuals from diverse backgrounds to find opportunities to learn from and interact with others different from themselves. This optimism quickly waned as social media seemed to breed ideological homophily marked by “filter bubbles” or “echo chambers.” A typical response to the sense of fragmentation has been to encourage exposure to more cross-partisan sources of information. But do outlets that reach across partisan lines in fact generate more civil discourse? And does the civility of discourse hosted by such outlets vary depending on the political context in which they operate? To answer these questions, we identified bubble reachers, users who distribute content that reaches other users with diverse political opinions in recent presidential elections in Brazil, where populism has deep roots in the political culture, and Canada, where the political culture is comparatively moderate. Given that background, this research studies unexplored properties of content shared by bubble reachers, specifically the quality of conversations and comments it generates. We examine how ideologically neutral bubble reachers differ from ideologically partisan accounts in the level of uncivil discourse they provoke, and explore how this varies in the context of the two countries considered. Our results suggest that while ideologically neutral bubble reachers support less uncivil discourse in Canada, the opposite relationship holds in Brazil. Even non-political content by ideologically neutral bubble reachers elicits a considerable amount of uncivil discourse in Brazil. This indicates that bubble reaching and incivility are moderated by the national political context. Our results complicate the simple hypothesis of a universal impact of neutral bubble reachers across contexts.

Introduction

Theorists have long recognized advances in communication technologies as double-edged swords. Over 2000 years ago, Plato worried about books replacing the experience of face-to-face conversation, even as he pioneered a novel genre of writing – the philosophical dialogue – designed to capture and disseminate that experience in a new medium [1, 2]. The printing press led to greater literacy, increased access to news and learning, and to the efflorescence of highly partisan pamphleteering marked by aggressive attacks on enemies with little regard for truth or civility.

Observers of social media have noticed similar tensions. Early optimism saw opportunities for a renewal of democratic discourse, with individuals from diverse

backgrounds finding possibilities to learn from and interact with others different from themselves [3, 4]. In this view, exposure to divergent views from members of outgroups would promote cross-cutting ties and thereby encourage mutual understanding, civility, and inclusivity. The Internet could be a school for deliberative democracy [5, 6].

This optimism quickly waned as online communication came to resemble the very fragmentation and balkanization that it had promised to overcome. Instead of cross-cutting ties, social media seemed to breed ideological homophily marked by “filter bubbles” or “echo chambers” [7, 8]. These, to some scholars, evoked the forms of interaction that arise when bridging ties are absent [9, 10]. By insulating partisans from countervailing views, existing viewpoints are intensified, and alternative ideas are delegitimized, potentially fraying the fabric of democratic society.

The metaphor of “echo chambers” [11] refers to epistemic environments in which individuals or groups interact with like-minded people or sources that reinforce their existing beliefs or opinions. In an echo chamber, information, ideas, and opinions are “echoed” and “amplified” within a closed network, leading to a reinforcement of existing viewpoints and a lack of exposure to diverse perspectives [11]. While the concept of “filter bubbles” refers to the algorithmic filtering that selectively curates content based on a user’s past behaviour, preferences, and demographic information [12]. In a filter bubble, users are presented with information that aligns with their interests and beliefs while filtering out opposing viewpoints or dissenting opinions [12].

A common response to the sense of fragmentation has been to encourage exposure to more cross-partisan sources of information [13, 14]. By offering citizens certified facts from neutral parties, platforms seek to reach across “bubbles” and “echo-chambers” to create a common basis for generating a meaningful conversation. However, recent research has largely popped the bubble on the ideological bubble perspective [15, 16]. In fact, much evidence indicates that online networks offer high exposure to diverse viewpoints and news sources and that users proactively seek out those with divergent views [17, 18]. They do so not necessarily to converse, however, but to engage the other in ways that do not moderate but reinforce prior commitments [19].

To examine these dynamics, we extend Kobellarz et al. [20], which investigated the role of central Twitter users, whom they characterized as “*bubble reachers*.” Bubble reachers are users who distribute content that reaches other users with diverse political opinions [20]. In order to identify these central users, the authors created a novel centrality metric called intergroup bridging [20]. By studying Twitter discourse during the 2018 Brazilian and 2019 Canadian elections, they identified bubble reachers such as @UOLNoticias in Brazil and @globalnews in Canada [20]. While such accounts disseminate content that engages ideologically diverse audiences, users nevertheless seem to respond to that content in strongly partisan ways: they share content that aligns with their own partisan orientation. This same dynamic held in both the Brazilian and Canadian contexts, despite the very different levels of political polarization in those countries.

The present article builds upon this research to study further unexplored properties of content shared by bubble reachers, specifically the quality of conversations and comments that it generates. Cross-cutting ties do not necessarily mitigate the tendency to maintain ideological groupings via selective reinforcement of existing views. Still, a question remains as to the quality of discussion that bubble reachers support. Do they offer spaces that mitigate rancorous discourse by providing a common starting point for conversation, even if that starting point is differently interpreted? Or do they provide the discursive equivalent of “combat zones,” gladiatorial arenas where partisans meet not to deliberate and converse, but to fight?

Bubble reachers may intervene in civil discourse in different ways: some bubble reachers may “open the hand” for broader conversations in a relatively neutral way,

while others “raise the fist” and offer not mutual understanding but mutual antagonism. But to what end? How bubble reachers navigate the civil sphere and the public reactions they elicit occur within a wider cultural context. Particularly relevant is the rise of populism and political polarization: relentless challenges against media institutions by populist leaders, epitomized by “fake news” narratives, have eroded trust in these institutions and credibility in their performances of neutrality [21]. In these contexts, a neutral bubble reacher may try to “open the hand” for broader conversation, but still receive the fist. As the room for neutrality narrows, even ostensibly non-political content may elicit antagonistic discourse as it becomes filtered through a partisan lens. Overall, whether bubble reachers support more or less hostile discursive spaces may not only vary by their styles of intervention, but also by the cultural context in which their performances are received.

Addressing these issues is the central goal of this paper: we examine how ideologically neutral bubble reachers differ from ideologically partisan accounts in the level of uncivil discourse they provoke, and explore how this varies in two contexts where populism is more and less prominent in political culture. We do that by examining the contexts of Brazil, where populism has deep roots in the political culture, and Canada, where the political culture is comparatively moderate. By analyzing the communication style provoked by neutral bubble reachers and ideologically partisan accounts in these distinct political cultures, this study marks a substantial step forward regarding past research.

The rest of this study is organized as follows. The section “Related Work and Hypotheses” presents a comprehensive literature review and related hypothesis about uncivil discourse and brokerage processes in online media, the relationship between the political context and neutrality, an overview of the political context of the two studied countries (Brazil and Canada), and methods for incivility detection in online discourse and its underlying challenges. The section “Materials and Methods” presents the data sources and methods we use to evaluate the hypotheses, including the measure used to identify bubble reachers, a presentation of the datasets, the text pre-processing method applied for toxicity inference from these datasets (using existing methods for inferring “toxicity” as a proxy for incivility), and methods for hypothesis testing. The “Results” section presents and discusses the results regarding our hypotheses. Finally, “Discussion and Conclusion” concludes the study and presents potential limitations and future work. Supporting information reports additional sensitivity analyses.

The pre-processed dataset and all the code necessary to replicate the statistical tests carried out in this research were made publicly available on the project website: <https://sites.google.com/view/onlinepolarization> [22].

Related Work and Hypotheses

Uncivil Discourse and Brokerage in Online Media

Scholars often evaluate discourse in the public sphere using the Habermasian normative ideal of deliberation based on reason-giving, reciprocity, and civility as a benchmark [23]. Applied to the online sphere, discourse arguably falls short of this standard: rather than a site for civil deliberation, social media resembles a rancorous town square, where uncivil discourse arguably holds sway [21]. Whereas civility involves mutual respect, incivility violates these norms [24, 25]. Uncivil discourse is not necessarily incompatible with deliberative democracy, to be sure [26]. Heated debates naturally involve aggressive and conflictual discourse, and studies find that this can make discussions more engaging while still appealing to deliberative persuasion [24, 27, 28]. Be that as it may, uncivil discourse is regarded as both symptomatic of polarization, and as a fuel

that exacerbates it further [29,30]. Research finds that uncivil discourse can elicit a feedback loop that provokes it in others, that alienates more civil participants from platforms, and distorts perceptions of political life altogether [31,32]. While it may not be inherently harmful to deliberative political talk, it is often deemed problematic to the extent that it contributes to polarization and intolerance [26].

Concerns over the proliferation of uncivil discourse have motivated inquiry into the conditions that underlie it. One approach emphasizes fragmentation in the media ecology [8,33,34]. Social media is putatively balkanized into “echo-chambers”, which amplify a set of in-group beliefs, discredit alternative viewpoints, and discourage charitable engagement with outsiders [35]. Recent work suggests fragmentation is not primarily about epistemic *closure* in so far as members are still exposed to outside information [15], but rather epistemic *discrediting* where outside information is readily dismissed [19]. According to Bright et al. [36], echo chambers thrive on steady exposure to low levels of oppositional views – they rely on conflict to enliven group identity, while providing cognitive tools to manage dissonance in ways that reinforce extreme belief. Fragmentation of this sort can set the backdrop for uncivil discourse. Segmented subcultures produce group identities with norms of civility that exclude outgroups, thereby licensing disparaging remarks against them as occasions to signal in-group affiliation [37]. While this can vary by communication norms in different communities [38] scholars also identify platform characteristics, whether user anonymity, homophily, or recommendation algorithms favouring intense emotional engagement, that undergird these dynamics, informing structural conditions that promote the expression of uncivil discourse [23,39,40].

But if uncivil discourse stems from fragmentation in the online public sphere, it also requires meeting points. “Bubble reachers” – central brokers that form bridges between opposing bubbles – can occasion the intersection of opposing partisans [20]. The term “bubble reachers” grows out of the computer science literature and refers to users on social media platforms who distribute content that reaches users with diverse political opinions, possibly bridging ideological divides and facilitating cross-partisan conversations [20]. They play a role in countering the phenomenon of “filter bubbles” [12] or “echo chambers,” [41] where individuals are exposed more frequently to information and perspectives that reinforce their prior beliefs. To understand the concept of bubble reachers in a broader network science context, we can draw parallels with the terms “brokerage” and “brokers.” In network science, brokerage refers to the role played by individuals or entities that connect otherwise disconnected groups or individuals within a network [42]. They occupy positions that span structural holes [42], which are gaps between distinct clusters or communities within a network. By spanning the structural holes, brokers create weak ties [43], acting as intermediaries and facilitating the flow of information, resources, or relationships between different parts of the network. Bubble reachers, in this sense, can be seen as performing a form of brokerage between partisan groups in a polarized context. They bridge ideological gaps and act as connectors between users with diverse political opinions, similar to how brokers bridge gaps between different groups or clusters in a network. Research suggests that, in practice, legacy media outlets are important bubble reachers in this regard [20,44,45]. For instance, Magin et al. [45] show that legacy media outlets help politically extreme individuals sustain engagement with the fundamental essence of public dialogue.

How bubble reachers relate to civil discourse is unclear and variable, however. By closing the distance between opposing groups, they open space for confrontation where greater incivility might be expected. At the same time, bubble reachers may also engage alternative perspectives out of a pluralist commitment to pursue neutral dissemination of information and uphold trust in a diverse public [46]. To the extent that bubble

reachers hold these commitments and are successful in their ambitions, exchange between opposing partisans may abide by norms of civility. It follows that neutral bubble reachers – those central nodes that avoid a partisan tilt - may, in some circumstances, temper the proliferation of uncivil discourse compared to those with a clear partisan valence. While this approach does not pre-determine what kind of media outlet emerges either as a bubble reacher in general or a neutral bubble reacher in particular, empirically neutral bubble reachers tend to be legacy media outlets [20].

These insights about the nature and types of bubble reachers inform the first hypothesis we examine in this paper:

H1: Neutral bubble reachers will tend to reduce uncivil discourse relative to partisan accounts.

Political Context and Performances of Neutrality

Scholarly focus on the role of platform characteristics and the structure of online networks tends to elide the role of differences in political culture. However, in the case of bubble reachers, there is good reason to believe their relation to uncivil discourse may vary across cultural contexts. Alexander [47] conceives of democratic life as consisting of symbolic performances that are interpreted as compelling or not based on their wider resonance with cultural structures. The appeal to principles of objectivity and fairness by neutral bubble reachers in their rendering of current affairs can likewise be interpreted as a performance of neutrality that the public assesses, and assigns varying degrees of credibility. Media brokers of this sort provide a crucial medium for how the public derives an understanding of world events and opinions on the state of affairs [48]. In contexts of consensus, the public assign credibility to the neutrality of bubble reachers, allowing them to provide what Talcott Parsons [49] referred to as a shared “definition of the situation” grounding conflict in common criteria for validity. When these claims lose their credibility, however, the civil sphere runs the risk of spiralling into crisis.

Against this backdrop, the rise of populism reflects a challenge to the social fabric: it is both a response to enervated mediating institutions and also a force working against the possibility of repair. Recent scholarship understands populism as a style or form of political communication which can be directed to many goals and ideological aims [50–56]. Although different in emphasis, these approaches agree that “the populist script” [54] tends to revolve around two core axes: a vertical one in which some “elites” are contrasted to “the people,” and a horizontal one, in which insiders are pitted against outsiders. Establishment media is often framed along the vertical axis as “elites” out of touch with the day-to-day concerns of “ordinary people”. Within this framework, universalistic ambitions of bubble reachers can falter. As Alexander [57] observes, a common trope in many populist scripts involves the critique of impartial expertise, in which experts’ claims to neutrality are reinterpreted as partisan defenses of “establishment” interests that run counter to those of the “common people”. Consistent with this, populist citizens are found to subscribe to a “false consensus”, assuming that their opinions are overwhelmingly shared by others, but unfairly maligned by hostile and biased mainstream media [58]. Performances of civility lose credibility in these contexts. Distinguished by its combative and morally-charged style, populist frames amplify authenticity over civility as a criterion for evaluation [59]. Careful and restrained forms of civil discourse fall short of these standards of expressivity, appearing weak and ambivalent by comparison [60].

Populism works in conjunction with polarization. As the legitimacy of mediating institutions recedes and neutral discursive spaces contract, the ambit of partisanship presumably expands – even non-political domains gradually become assimilated into partisan conflict. The “oil-spill” model of polarization, for instance, documents how

partisanship can permeate even areas of social life which are seemingly apolitical and inconsequential [61, 62]. This creates an additional reason why neutral bubble reachers may struggle to uphold spaces for civil discourse in a more polarized context. Significant scholarly literature suggests political discourse is especially rancorous. In contrast to news concerning science, technology, weather, art, and culture, Salminen et al. [63] find that topics with a political valence provoke more toxic comments. However, to the extent that polarization interfuses social life broadly, giving even non-political content a partisan valence, the distinction between the political and non-political may become blurred. As a result, in more polarized settings, uncivil discourse may be expected across different topics, jeopardizing neutral bubble reachers' ability to leverage non-political topics as meeting points for civility.

In these ways, contexts where political culture is characterized by populism and polarization, may affect uncivil discourse and the relation of neutral bubble reachers to it. Uncivil discourse may be expected to be more pronounced in these public spheres, and the role of neutral bubble reachers in curbing it may falter. Rather than tempering discourse, these actors may even elicit greater invective. Their claim to objectivity and fairness may be reframed as a veiled cover for the interests of the "establishment" and provoke greater scorn as a result [58]. Furthermore, as the space for neutrality contracts, uncivil discourse may be expected to enter even seemingly non-political domains. These considerations lead to three additional hypotheses:

H2: Uncivil discourse will be more pronounced in contexts that are more populist and politically polarized.

H3: Neutral bubble reachers will produce greater uncivil discourse than partisan accounts in contexts that are more populist and polarized.

H4: Uncivil discourse will be evident in both political and non-political topics in contexts that are more populist and polarized.

Populism and Polarization in Canada and Brazil

We examine the role of context by considering how bubble reachers operate differently in Canada and Brazil. Populism is not foreign to either context, but has deeper roots in Brazil. While populism has had a significant steady influence on Brazil's politics on both the left and the right since the early 20th century, its role in Canadian politics has been small by comparison. Lipset [64] famously enshrined the image of Canada as a country of political moderation when he contrasted the populist and anti-statist strains in American political culture to Canadian political culture, which was distinguished by its greater deference to elites and government authority. Some more recent scholarship disputes this characterization and suggests both populism and polarization are on the rise in Canada [65–67]. Nevertheless, studies still observe greater recent levels of affective polarization in Brazil as compared to Canada, and greater support for political leaders employing populism more generally [68–70]. These differences are especially pronounced in Canadian national electoral politics, where populism remains marginal compared to Brazil. To be sure, populist leaders have emerged in Canadian municipal politics [71]. But whereas Canadian populist parties like the People's Party of Canada have achieved little electoral success, both right-wing and left-wing populism have found success in Brazil at the national level. Jair Bolsonaro's election in 2018, for instance, was notably fuelled by an anti-establishment campaign which frequently targeted the media class as purveyors of "fake news." Scholars observed the important role of digital media in the culmination of these events. Digital media became a major conduit for anti-establishment news, and served to normalize Bolsonaro's populist discourse [72–74]. Communication campaigns mobilized citizens to engage in polarized debates, with platforms like Facebook and Twitter becoming the site of rancorous and even violent discourse.

Despite these differences, Canada and Brazil share important commonalities. Both are large federal systems and constitutional democracies, containing populations that are multicultural and racially diverse. Importantly, citizens in both countries are active on social media at roughly similar rates. Recent reports find that Facebook’s ad reach is equivalent to 50.5% and 53.5% of the total population, with 61.1% and 61.7% of the eligible population using Facebook in Brazil and Canada respectively. Given the different levels of populism and polarization in both countries, these structural similarities make Brazil and Canada appropriate cases for comparative analysis [75, 76]. They allow us to assess our hypotheses on how bubble reachers may operate differently in contexts where populism and polarization are more prominent and have deeper roots in political culture (Brazil), compared to contexts where political culture has historically been more moderate, and populism and polarization are less pronounced (Canada), while holding important structural variables constant.

Detecting Uncivil Discourse in Canada and Brazil

While the perception of incivility in discourse is to some degree a subjective matter, the research literature identifies some characteristic features that demarcate incivility, such as *ad hominem* attacks (direct attacks on the person instead of their arguments), vulgarity and exaggerations [77]. In short, incivility involves communication that violates norms of courtesy and respect in social interactions, especially in the political sphere, where discussion of ideas and opinions can be heated. In the online sphere, incivility theoretically overlaps with “toxic discourse”, which refers to disrespectful or irrational comments that may lead users to leave discussions [78]. Our empirical analysis detects incivility by engaging with work on toxic discourse.

To provide context, these are examples of web comments obtained from one of the datasets studied in this article, which were manually annotated by three independent evaluators regarding their toxicity [79], demonstrating uncivil discourse: “tirso, you drunk, I’m not just a man on the internet, come here or give me your address if you’re a man, and I’ll come and talk to you in person. My address is [...], (it’s very close to the current presidency of your luladrão boss) come here and talk this nonsense” and “KKKKKKK now DataFAKE does a minimally truthful survey, because it was already getting bad. kkkkkkkkk ridiculous! They tried to manipulate the Brazilian vote but this is over!! they just lost the rest of the credibility they had, if they had any!”. Meanwhile, these are examples of comments also manually annotated demonstrating civilized discourse [79]: “In the debate for governor of Rio held on Globo on the 25th, Eduardo Paes scored a great victory, he deserves to be elected.” and “Bolsonaro did not claim medical recommendation, the doctors themselves gave an interview recommending that he not go. Everyone saw it.”

Researchers have developed several methods to identify such comments, ranging from keyword-based proposals [80] to strategies that use machine learning [81–86]. Given the speed with which discussions are growing on the Web [87], different models for large-scale toxicity identification were created to solve specific challenges, such as for dealing with online comments [84, 86]. Among the challenges is the fact that this type of text contains varying degrees of subtlety inherent to the language, cultural aspects, specificities of context, presence of sarcasm, and use of figures of speech, which can make the actual toxicity of a comment ambiguous. Other challenges include the fact that online comments are usually short texts, often containing spelling errors that occur sparsely in the dataset [84], making it challenging to create models with a generalized capacity to identify uncivil instances. Likewise, machine learning-based approaches need considerable quantities of high-quality annotated data, which can present questions concerning model training [88]. Thus, we can find several studies that concentrate on studying and criticizing toxicity classification methods [88, 89]. While “toxicity” and

“incivility” may not perfectly overlap, the existing tools for measuring toxicity are highly developed and provide a useful proxy for online incivility. Future research may benefit from developing direct measurements of incivility purpose-built for that task.

Uncivil discourse is also sensitive to context. For example, expressing hostility about the indigenous community in Brazil will likely be different in the context of a pro-Bolsonaro discussion and a discussion in the context of Funai¹ supporters. Uncivil discourse also varies depending on target groups – for instance, expressions employed to express hate against a community in Brazil are distinct from expressions employed against Latin Americans in the United States. These challenges are even greater when it is necessary to classify and compare comments in different languages [84]. It is not uncommon for text analysis tools to work only for English or to have more resources available for this language, since it has been the subject of more training instances and research [90–92]. Given this situation, researchers often translate non-English content into English before performing automated text analysis tasks. Kobellarz et al. [79], however, found that while this approach can achieve good performance in some domains, such as sentiment analysis [91], performance for toxicity scores is poor. They, therefore, considered a translated version of a Brazilian Portuguese comments dataset to identify whether the original or the translated version would be more suitable for the study [79]. To infer toxicity scores in online comments, the authors used the Perspective API², a multilingual model widely employed for this task [63, 93–96]. They verified that the translation process artificially reduced the overall toxicity scores of the datasets by penalizing highly toxic comments in their original language [79]. The present study follows this guideline, keeping the datasets in their original language to infer toxicity.

Materials and Methods

To evaluate our hypotheses, we followed a method for comparing the toxicity scores in multiple comment datasets from distinct sources and languages. This involved a structured process of five sequential steps:

1. Apply a centrality metric to identify bubble reaching accounts.
2. Obtain the necessary data to study our hypothesis.
3. Perform essential standard text pre-processing steps.
4. Infer incivility levels from the studied comments.
5. Evaluate each hypothesis.

Fig 1 summarizes these steps, which are explained in the following sections.

Bubble Reacher Measure

The first crucial step to evaluating our hypotheses was to identify bubble reachers, central users who distribute content that reaches other users with diverse political opinions. To do so, we relied on the results of Kobellarz et al. [20], where the authors studied the role of these central users during the 2018 Brazilian and 2019 Canadian elections. For that, understanding the political context surrounding each election is fundamental for a comprehensive analysis. In the case of Brazil, during the election period, the polls³ revealed a notable degree of polarization among voters. The electoral

¹Brazilian governmental protection agency for Amerindian culture and their interests.

²<https://perspectiveapi.com>.

³<http://media.folha.uol.com.br/datafolha/2018/10/19/692a4086c399805ae503454cf8cd0d36IV.pdf>

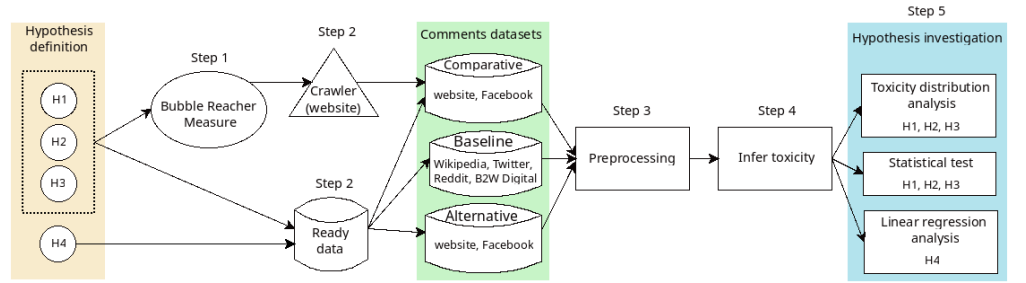


Fig 1. Overview of steps to answer our hypotheses.

contest revolved around two primary candidates: Jair Messias Bolsonaro, who symbolized the potential for a 15-year departure from the ruling Workers Party (PT), and Fernando Haddad, who represented the continuity of PT’s governance. This situation arose after a brief period during which a PT president elected was impeached and replaced by her vice-president. Ultimately, Bolsonaro emerged victorious with a majority of 55.13% of the valid votes, while Haddad secured 44.87% of the votes in his favor⁴. In Canada, Justin Trudeau represented the Liberals, which had previously held a parliamentary majority after unseating the Conservatives in 2015. In a close election, the Liberals won 157 (39.47%) seats in parliament, while the Conservatives, led by Andrew Scheer, won 121 (31.89%)⁵. The (left-wing) New Democratic Party continued to lose ground from its 2011 peak, especially in French Quebec, where the Liberals and the separatist Bloc Québécois subsequently gained ground. The 2019 election resulted in the Liberals forming a minority government, which has historically exhibited instability, since the prime minister relies on representatives of other parties to remain in power⁶. Despite Canada’s multi-party political system, national politics has predominantly revolved around the traditional left-right ideological differences since at least the 1980s [97].

To measure and find bubble reachers, Kobellarz et al. [20] developed a new centrality measure called intergroup bridging, which was inspired by the “bridgeness” algorithm [98], adapted from Brandes “faster algorithm” to calculate betweenness centrality [99]. This novel centrality measure demonstrated a significant improvement in identifying global bridges in a polarized network, highlighting users with a greater ability to distribute content to polarized bubbles with diverse political opinions.

Kobellarz et al. [20] used the intergroup bridging centrality metric to identify the most central users in *retweet* networks in the Canadian and Brazilian political situations. From the results, the 100 accounts with the highest intergroup bridging centrality value were selected. After that, the links from messages retweeted by other users from these bubble reaching accounts were extracted, as well as the respective domains and latent topics present in the contents pointed out by these links. In the last step, they identified the political polarity of the entities (contents, domains, and topics) using a metric that the authors called $RP(H)$ – relative polarity – which is a measure obtained from the weighted average of the polarity of the users who retweeted a certain entity. The $RP(H)$ metric, in short, expresses the political bias of an entity concerning the audience from which it generates engagement. The result of $RP(H)$ is a value in the continuous range $[-1.0; +1.0]$. Positive values represent a right-wing political orientation, negative values represent a left-wing orientation, and values close to 0.0 represent a neutral orientation.

⁴<https://politica.estadao.com.br/eleicoes/2018/cobertura-votacao-apuracao/segundo-turno>

⁵<https://newsinteractives.cbc.ca/elections/federal/2019/results>

⁶<https://web.archive.org/web/20130627154515/http://www2.parl.gc.ca/Parlinfo/compilations/parliament/DurationMinorityGovernment.aspx>

Based on $RP(H)$ value, entities were labelled according to their political orientation: the first third of values on the $RP(H)$ scale represents the left-wing (L) users, with $RP(H) \in [-1; -1/3[$, the second third, the neutral users (N), with $RP(H) \in [-1/3; 1/3]$ and the last third, the right-wing users (R), with $RP(H) \in]1/3; 1]$.

In the next section, we provide an explanation of how the intergroup bridging and $RP(H)$ metrics were used to create two of the datasets studied in this research.

Comments Datasets

Comparative Datasets (for H1, H2 and H3)

To assess the H1, H2, and H3 hypotheses, we first obtained representative comments made on content published by *neutral bubble reachers* and *partisan accounts*. We, therefore, generated four datasets, two to represent neutral bubble reachers comments, one in Brazilian Portuguese and the other in English, and two to represent partisan accounts comments, also one in Brazilian Portuguese and the other in English.

To capture neutral bubble reachers' comments, we relied on [20]. For the present study, we obtained comments made about news articles published by the top 3 bubble reaching domains with neutral $RP(H)$ in each country. To simplify the understanding of which dataset we refer to in the next sections, we've labelled Brazil's dataset NEUTRAL_REACHER_pt and Canada's dataset NEUTRAL_REACHER_en – the prefix NEUTRAL_REACHER means “neutral bubble reacher” to facilitate the identification of these specific cases on further analysis. For each domain in these datasets, we created a web scraper to capture all comments from news articles published during the electoral period in each country.

Considering that all comments in NEUTRAL_REACHER_pt and NEUTRAL_REACHER_en were made on content produced by neutral bubble reachers, we also collected comments made on contents produced by partisan accounts to enable us to compare these distinct cases. Following the recommendations of [79] we kept the comments in their original language, even in comparisons between different languages [79]. We, therefore, collected two additional datasets to be compared to NEUTRAL_REACHER_pt and NEUTRAL_REACHER_en, one in Portuguese and the other in English, respectively, containing comments on content distributed by partisan accounts. To ensure a representative sample from these accounts, we identified users among the top 100 bubble reachers in the Canadian and Brazilian datasets [20] with partisan $RP(H)$ scores ($RP(H) < -0.333$ or $RP(H) > +0.333$) – those who act as bridges between bubbles but are more likely to generate engagement on one side of the political spectrum, given by their $RP(H)$ value.

Unlike what was possible to do with neutral bubble reachers to collect comments on news articles on their respective websites, it was not possible to replicate this same step for the case of partisan accounts, since among the top 100 bubble reachers with partisan $RP(H)$, there were few linked to sites with comments sections and none with a considerable number of comments. Therefore, alternative sources of comment data on partisan news media pages on Facebook in Portuguese and English were obtained to compare with the NEUTRAL_REACHER_pt and NEUTRAL_REACHER_en datasets, respectively.

For the English case, we cross-referenced the top 100 bubble reachers with partisan $RP(H)$ in Canada with a Facebook dataset containing comments on 20,000 posts made by American news organizations and personalities [100]. This dataset has been labelled as PARTISAN_REACHER_en. The American political culture is recognized for its polarization [101, 102], and some evidence shows that it is higher than the Canadian scenario [69]. Despite differences in political culture, Canadians also engage with these sources, albeit to varying degrees. Thus, we selected this dataset as an approximation

for partisan accounts for the Canadian situation.

However, for the Portuguese case, it was not possible to cross-reference this same Facebook dataset with any bubble reacher with partisan $RP(H)$ like what was possible for `PARTISAN_REACHER_en`, so we obtained another Facebook dataset containing comments on Brazilian Conservative pages from 2012-01-01 to 2018-12-31, which can be considered partisan representative accounts (mostly right-wing) [103]. This dataset has been labeled as `PARTISAN_pt`. Note the lack of the “reacher” in the name of this dataset: it was removed intentionally, because it does not include comments from cross-referenced bubble reaching accounts, in contrast to what was possible with other datasets.

In summary, here are the datasets introduced in this section:

NEUTRAL_REACHER_pt: this dataset comprises comments captured from news media websites representing these top three neutral bubble reaching domains in the Brazilian election setting: `noticias.uol.com.br`, `g1.globo.com`, and `extra.globo.com`.

NEUTRAL_REACHER_en: this dataset comprises comments captured from news media websites representing these top three neutral bubble reachers in the Canadian election setting: `cbc.ca`, `globalnews.ca`, and `theglobeandmail.com`.

PARTISAN_pt: This dataset comprises comments captured from Brazilian conservative Facebook pages (mostly right-wing), including: `padrepaulo`, `flaviobolsonaro`, `CampanhadoArmamento`, `carvalho.olavo`, `OPesadelodeQualquerPolitico2.0` and `MisesBrasil`. It is worth mentioning that comments in this dataset are not linked to the specific electoral contexts in Brazil or Canada, like the ones obtained to represent neutral bubble reachers. Also, these pages can’t be considered bubble reaching sources, like what was possible with other datasets with “reacher” in their name.

PARTISAN_REACHER_en: this dataset comprises comments captured from Facebook news pages representing the partisan accounts (left-wing or lean toward the left according to the All Sides Media Bias Report⁷): `bloomberg-politics`, `huffington-post`, `ny-times`. It is also important to note that comments in this dataset are not linked to national elections in Brazil or Canada.

We call this set of datasets “Comparative datasets.”

Reference Datasets

As our hypotheses concern comparisons regarding incivility, we need a method to identify and compare the incivility of comments in the neutral bubble reachers (`NEUTRAL_REACHER_pt` and `NEUTRAL_REACHER_en`) and partisan accounts (`PARTISAN_pt` and `PARTISAN_REACHER_en` datasets). To do so, we utilize four reference datasets; two of them composed of highly uncivil comments, one in English (`UNCIVIL_en`) and another in Brazilian Portuguese (`UNCIVIL_pt`), and the other two containing more civil comments, also one in English (`CIVIL_en`) and another in Brazilian Portuguese (`CIVIL_pt`). These reference datasets are presented below.

UNCIVIL_en⁸: contains highly uncivil United States English comments (not translated) taken from an open dataset with human-labelled Wikipedia comments

⁷The tool assigns a rating of Left, Lean Left, Center, Lean Right, or Right to each media source. The ratings reflect the average view of people across the political spectrum, obtained by exploring a scientific, multipartisan analysis. More details about the method can be found in: <https://www.all-sides.com/media-bias/media-bias-rating-methods> – The All Sides Media Bias report was obtained on Dec 12, 2022

⁸Disclaimer: This file includes words or language that is considered profane, vulgar or offensive by some readers. Due to the topic studied in this article, quoting offensive language is academically justified but we nor PLOS in no way endorse the use of these words or the content of the quotes. Likewise, the quotes do not represent the opinions of us or that of PLOS, and we condemn online harassment and offensive language.

according to different categories of toxic behaviour⁹, including “toxic”, “severe_toxic”, “obscene”, “threat”, “insult”, and “identity_hate”. The variable “toxic”, representing the degree of toxicity of the comment, was used to select 5,000 comments with the highest value for this metric and compose the UNCIVIL_en dataset. Examples of comments can be found on the project’s website¹⁰.

UNCIVIL_pt⁸: contains highly uncivil comments in Brazilian Portuguese obtained from tweets manually annotated according to different toxicity categories in a publicly available dataset [104]. The available categories were “non-toxic”, “LGBTQ+ phobia”, “obscene”, “insult”, “racism”, “misogyny”, and “xenophobia” [104]. To select a representative sample, the number of toxicity categories linked to each comment was counted, except for the “non-toxic” category. This count was used to select 5,000 comments with the highest amount of toxic categories to compose the UNCIVIL_pt dataset. Examples of this dataset can be found on the project’s website¹⁰.

CIVIL_en: contains more civil comments in United States English (not translated) obtained through Reddit’s public API, a network of communities where people with common interests interact in a forum-like system. To collect the data, the most popular 100 posts from the communities (*subreddits*) “AskHistorians”, “changemyview”, “COVID19”, “everythingScience”, and “science” were selected. These communities were chosen because they contain potentially more civil discussions. Indeed, inspection of a sample of responses showed a high level of civility and earnestness. These features are likely due to the fact that stricter rules for posting were explicitly informed and seemed to be followed by their participants. Thus, this dataset was composed mostly of constructive comments. In addition to the text of the comments, other attributes were obtained, among them the “score” of the comment, which is a metric calculated by subtracting the negative votes from the positive votes that a given comment received. This metric was used to select 5,000 comments with the highest score to compose the CIVIL_en dataset. Some comments examples in this dataset are: “According to a paper published in IEEE Transactions on Computational Social Systems by researchers at The University of Notre Dame, some 73 percent of posts on Reddit are voted on by users that haven’t actually clicked through to view the content being rated. Hopefully, this information allows 3 out of 4 people to not have to read through the article.” and “I have lived with the prospect of an early death for the last 49 years. I’m not afraid of death, but I’m in no hurry to die. I have so much I want to do first.”.

CIVIL_pt: composed of more civil comments in Brazilian Portuguese obtained from a database of product reviews in a famous e-commerce business, B2W Digital, responsible for the *americanas.com* website, whose data were obtained between January and May 2018 [105]. This dataset was selected considering that among positive and lengthy product reviews, there would be a less hostile and uncivil discussion. Therefore, during data cleaning, evaluations with a score lower than 5 and that contained less than 20 unique characters were eliminated. This is important because the incidence of texts with repeated words was identified in cases where the evaluator only filled in the text field with no intention of making a careful assessment. After cleaning, the longest 5,000 evaluations were selected to compose the CIVIL_pt dataset. Some comments examples (translated to English) on this dataset are: “I had researched the product previously and it met my expectations, despite the design flaws that I was already aware of. Furthermore, the product arrived very quickly and in perfect condition.” and “The table is light and small, as the name suggests. It has an excellent texture and the pen feels comfortable in the hand. The settings are in a very intuitive interface and overall allow us to have excellent adjustments for use. Excellent for beginners who want to

⁹<https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data>.

¹⁰ <https://sites.google.com/view/onlinepolarization>

experiment with digital art.”

We call this set of datasets “Reference datasets.”

Alternative datasets (for H4 and sensitivity analysis)

In politically polarized and populist environments like Brazil, even neutral content may acquire a political connotation. This raises the question of whether ideologically neutral bubble reachers behave differently in such contexts. H4 delves into this possibility by examining how the incivility of comments varies between political and non-political content in contexts that are populist and polarized, like in Brazil. To analyze this, we gathered discussions from ideologically neutral bubble reachers that could be distinguished based on whether they specifically addressed political topics or not. For this purpose, comments were obtained from the Globo G1 [106] and the New York Times¹¹ news websites. Both websites organize comments by topic, enabling us to identify which comments explicitly referenced political or non-political topics. This categorization allows us to examine incivility in political and non-political content in the Brazilian populist and polarized context using Globo G1 data, and compare it with the American context using New York Times data. These datasets, labelled as G1_SITE_pt and NYT_SITE_en, respectively, are presented below.

G1_SITE_pt: composed of comments written in Brazilian Portuguese on news articles published on the Globo G1 (g1.globo.com) website between March 28, 2020, and November 11, 2020 [106]. It comprises a total of 1,059,672 comments spanning 18,014 news articles [106]. Considering that g1.globo.com domain is also present in NEUTRAL_REACHER_pt dataset, but contains comments in distinct time periods, it also enables an investigation into potential variations in toxicity levels between comments made in a political context, as captured in NEUTRAL_REACHER_pt, and those made in other context, as captured in G1_SITE_pt.

NYT_SITE_en: composed of comments obtained from the New York Times website during March 2018¹¹, from which were selected 5,000 random comments from political sections and 5,000 random comments from non-political sections.

Additional data was collected to address limitations regarding comparative datasets, which are presented in S1 Appendix. We call this set of datasets “Alternative datasets.”

S2 Appendix presents a summary of all datasets and their corresponding sources for Comparative (Table 6), Reference (Table 7) and Alternative (Table 8) groups for quick reference.

Note on using Facebook as a data source

It is important to consider the validity of using Facebook data, considering that neutral bubble reachers were identified from Twitter data [107]. First, the neutral bubble reachers that emerged from the Twitter analysis were major legacy media outlets. That these outlets were central in reaching across ideological bubbles is consistent with other work likewise finding evidence that legacy outlets have the greatest audience overlap in consumption of digital news [44]. This makes us confident that the neutral bubble reachers identified using Twitter data should characterize these outlets more broadly, including on Facebook. Second, incorporating the Facebook dataset allows for greater information diversity in the analysis of comments. Relying exclusively on Twitter as a data source would lead to results that are skewed toward Twitter platform practices and the political representativeness of its users. Research by Barbera and Rivero has shown that Twitter users who engage in political discussions tend to be male, reside in urban areas, and possess extreme ideological preferences [108]. Moreover, Twitter’s character

¹¹<https://www.kaggle.com/datasets/aashita/nyt-comments>:<https://www.kaggle.com/datasets/aashita/nyt-comments>

limit constraints discourse complexity inhibits reflexivity and tends to encourage uncivil behaviour due to the impersonal nature of the platform [109]. By including data from Facebook, we can broaden our understanding of the phenomenon across multiple platforms. Furthermore, previous studies on polarization and hate speech have heavily relied on Twitter data alone [41, 110, 111]. Our research aims to expand the scope by extrapolating our findings to other platforms, enabling a more comprehensive examination of the phenomenon.

Pre-processing

The pre-processing step was performed to maintain maximum integrity in the datasets presented in the previous sections, only removing noisy instances that could negatively impact performance. S3 Appendix presents pre-processing details.

Inferring Toxicity as a Proxy for Incivility

A widely used model for identifying toxicity in academia and industry is the Perspective API [78], which is a multilingual tool available for free through a Google initiative called Perspective API². This model can be accessed through a public API, which allows users to identify different toxicity types in online comments. Given that this is a widely adopted tool by major news outlets for moderating comments on their portals, in addition to the fact that it is openly available, having support and good performance on different languages [79], Perspective API was chosen to be applied to this research. Previous studies indicate that Perspective API satisfactorily captures the toxicity of social media content [94–96]. For instance, Rajadesingan et al. [94] show that its performance in identifying toxicity is comparable to toxicity labelled by humans. Some studies suggest that Perspective API has the potential for racial bias against speech performed by African Americans [112], but this possible bias should not compromise our analyses. A recent study systematically tested for adversarial examples and other recognized vulnerabilities in state-of-the-art multilingual models for toxicity detection and found that the latest production model of the Perspective API, the one used in this study, outperforms strong baselines [90]. Considering the widespread adoption of this tool and its state-of-the-art performance, we deem this tool suitable for use in this research.

All comments presented in Section Comments Datasets were pre-processed by the Perspective API, which assigns a continuous score between 0 and 1 to comments according to their toxicity. A higher score indicates a greater likelihood that a reader will perceive the comment as containing the given attribute, e.g., toxicity [78]. For example, as presented in the API documentation, a comment like “You are an idiot” may receive a probability score of 0.8 for the toxicity attribute, indicating that 8 out of 10 people would perceive that comment as toxic [78]. With this, it is possible to use this score, for example, to moderate toxic comments with a certain score [78]. Perspective architecture is composed of multilingual BERT-based models trained on data from online forums that are distilled into single-language Convolutional Neural Networks (CNNs) for each language that they support – distillation ensures the models can be served and produce scores within a reasonable amount of time [78]. Perspective has so-called production attributes, tested in various domains and trained on significant amounts of human-annotated comments. These attributes are available in English, Portuguese, and many other languages. Also, it contains experimental attributes – English only – that are not recommended for professional use at this time [78]. In this work, we focus only on the toxicity attribute, the main feature among the production attributes. A comment with a high toxicity score is described in terms that resonate with the notion of incivility: “a rude, disrespectful or irrational comment that is likely

to cause people to leave a discussion.” To allow a fair comparison between datasets, toxicity scores were normalized using a min-max strategy for each dataset separately. Following a guideline presented in [79], we considered the identification of toxicity in comments in their original language. As noted above, while we believe that these measures of toxicity provide valuable proxies for incivility in the present study, future research may enrich this work by seeking to measure incivility directly.

Tables 9 and 10 were added to S2 Appendix for quick reference of summarized statistics of datasets and sources, respectively. These tables encompass comment counts before and after pre-processing, along with mean, median, and standard deviation for toxicity scores. Table 10 shows that the impact of cleaning noisy instances was greater for PARTISAN_pt sources. Despite this, the remaining dataset size is still large enough to consider in the analyses. Also, it is important to note that “extra.globo.com” source has the least amount of comments. Due to the sample’s limited size, it has been excluded from subsequent analyses.

Methods for Hypothesis Investigation

These are the three main methods used in this study:

- **Toxicity score distribution analysis:** We explore the distribution of toxicity scores on the different datasets and sources, separately. For that, we apply box plot visualizations.
- **Statistical test:** We conducted a statistical test to examine if there is a significant difference between toxicity scores across the evaluated datasets. Before defining an appropriate statistical test for this purpose, we check whether the data follows a normal distribution. To do this, we applied the D’Agostino K-squared test [113] and relied on Q-Q plots and histograms for each dataset to visually identify normal curve patterns (plots were not included for brevity). If data were non-normal, we apply the Kolmogorov-Smirnov Goodness of Fit test [114] pairwise between datasets. For this test, the hypotheses were H_0 : datasets originate from the same distribution, and H_1 : datasets do not originate from the same distribution.
- **Linear regression analysis:** We performed a linear regression analysis considering toxicity as a dependent variable and as the independent variable, the one we want to study in relation to toxicity.

Table 1 summarizes all datasets and methods used to study the hypotheses, while Table 2 cross-classifies the comparative datasets. Table 2 helps to organize the datasets and analysis by sorting them between media partisanship (neutral vs. partisan) and national polarization (high vs. low). The cells show datasets within each combination (e.g. highly partisan sources in a highly partisan context versus neutral sources in a less polarized context). As we were not able to clearly measure the combination of partisan sources with a less polarized national context due to the data limitations noted on Section Comparative Datasets (for H1, H2 and H3), we consider this cell to be an approximation (see also the note to Table 2).

Results

In this section, we present the results regarding our hypotheses.

Table 1. Summary of all datasets and methods used to evaluate hypotheses.

Hypothesis	Datasets and Sources	Methods
H1: Neutral bubble reachers will tend to reduce uncivil discourse, relative to partisan accounts.	Comparative NEUTRAL_REACHER_pt PARTISAN_pt NEUTRAL_REACHER_en PARTISAN_REACHER_en Reference CIVIL_pt CIVIL_en UNCIVIL_pt UNCIVIL_en Alternative G1_SITE_pt NYT_SITE_en	Toxicity distribution analysis between datasets and respective sources, separately. Statistical test: Kolmogorov-Smirnov Goodness of Fit test [99] pairwise between datasets.
H2: Uncivil discourse will be more pronounced in contexts that are more populist and politically polarized.		
H3: Neutral bubble reachers will produce greater uncivil discourse than partisan accounts in contexts that are populist and polarized.		
H4: Uncivil discourse will be evident in both political and non-political topics in contexts that are populist and polarized.		Linear regression analysis to verify the relationship between the source type (political or non-political) and toxicity scores. For this test, toxicity scores were considered as the dependent variable, and source type (political or non-political) the independent variable.

Table 2. Cross-reference of comparative datasets applied for H1, H2 and H3 hypotheses testing.

		National polarization	
		High	Low
Media type	Neutral Bubble Reacher	NEUTRAL_REACHER_pt ($N = 123,212$) (Brazil): - extra.globo.com ($N = 48$ - removed from analysis due it's size) - g1.globo.com ($N = 55,057$) - noticias.uol.com.br ($N = 68,155$)	NEUTRAL_REACHER_en ($N = 113,114$) (Canada): - globalnews.ca ($N = 2,038$) - cbc.ca ($N = 97,932$) - theglobeandmail.com ($N = 13,144$)
	Partisan	PARTISAN_pt ($N = 24,741$) (Brazil): - CampanhadoArmamento ($N = 4,199$) - MisesBrasil ($N = 3,508$) - OPesadelodeQualquerPolitico2.0 ($N = 4,306$) - Carvalho.olavo ($N = 4,218$) - Flaviobolsonaro ($N = 4,307$) - padrepaulo ($N = 4,203$)	— Approximation — PARTISAN_REACHER_en* ($N = 28,070$) (Canada-USA): - huffington_post ($N = 13,013$) - ny-times ($N = 14,169$) - bloomberg_politics ($N = 888$)

*This dataset was obtained by cross-referencing the top 100 biased bubble reachers in Canada (measured by their $RP(H)$) with a Facebook dataset containing comments on 20,000 posts made by American news organizations and personalities [100]. See Comparative Datasets (for H1, H2 and H3) section for a complete reference. Note that this dataset represents an approximation for the Canadian scenario.

Incivility and neutral bubble reachers

Fig 2 shows a box plot to help understand the difference in incivility (as measured by the toxicity scores) regarding the comparative and reference datasets. On this figure, it is possible to note that the reference pairs UNCIVIL_pt and UNCIVIL_en are similar to each other, as well as the pairs CIVIL_pt and CIVIL_en, which indicates that these pairs share characteristics in common, despite linguistic and data source differences. As expected, the pairs UNCIVIL_pt and UNCIVIL_en proved to be uncivil, while CIVIL_pt and CIVIL_en proved to be the most civil, as also expected. These results indicate that the reference pairs present the desired characteristics for comparison with the other datasets. Furthermore, the slight difference observed within the reference pairs suggests that language might exert a minimal influence on the toxicity scores detected by the Perspective API in these extreme cases [79].

Fig 2 also highlights the difference between neutral bubble reachers and partisan accounts. The most surprising finding here is that, contrary to H1, NEUTRAL_REACHER_pt (neutral bubble reachers in Brazil) have more uncivil comments (according to their median toxicity scores) compared to PARTISAN_pt (partisan accounts in Brazil). By contrast, NEUTRAL_REACHER_en (neutral bubble reachers in Canada) fits the expectations of H1: the neutral bubble reachers inspire less uncivil comments than their partisan counterparts in PARTISAN_REACHER_en (an

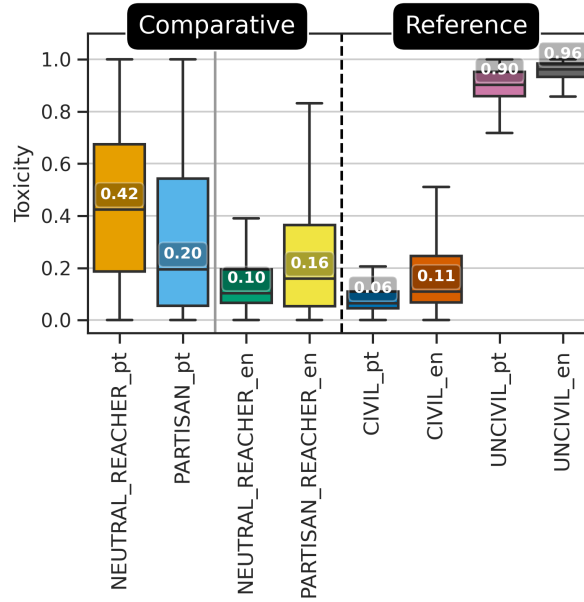


Fig 2. Box plots showing the distribution of incivility (measured by toxicity scores) for the comparative and reference datasets.

approximation for partisan accounts related to Canada). This contrast between the more polarized Brazilian context and the less polarized Canadian one suggests support for H3: neutral bubble reachers in the polarized context generate more uncivil discourse than partisan sources, whereas in the Canadian context, the opposite is the case¹². Just as intriguing is that the median toxicity score on the Canadian dataset (NEUTRAL_REACHER.en) is close to the median for CIVIL.en and CIVIL_pt references, while all these three datasets have, at least, 3.8 times lower toxicity median score compared to NEUTRAL_REACHER_pt, and, at least, 1.8 times lower toxicity median score compared to PARTISAN_pt. These results lend support to H2: uncivil discourse, in general, is lower in the less polarized and populist Canadian context relative to the more polarized Brazilian setting.

To verify whether differences between median toxicity scores observed in Fig 2 were statistically significant, we applied the tests explained in Section Methods for Hypothesis Investigation. First, Comparative datasets (NEUTRAL_REACHER_pt, NEUTRAL_REACHER.en, PARTISAN_pt, and PARTISAN_REACHER.en) were subjected to a normal test for the toxicity scores applying D’Agostino K-squared test [113]. Table 3 presents these test results, showing that none of the datasets followed a normal distribution ($p < .001$). We also relied on Q-Q plot and histogram visualizations for each dataset, confirming this result (plots were not included for brevity). Considering that the data were not following a normal distribution, we applied the Kolmogorov-Smirnov Goodness of Fit test [114] pairwise between datasets to verify whether the samples originate from the same distribution. Results for this test were included in a separate spreadsheet for consultation¹³, which shows that none of the comparative datasets originated from the same distribution ($p < .05$), meaning that the

¹²Regarding this result, it is important to recall that PARTISAN_pt is mostly composed by comments made on content published by extreme right Facebook pages which probably do not promote an encounter between distinct political views like what could be happening on NEUTRAL_REACHER_pt, which is composed mostly by comments on news accessed by users on distinct positions on the political spectrum.

¹³Statistical tests results available at: https://drive.google.com/drive/folders/1iAhm9lapHA0y2qtTB8Bn_4sJqVPg7rks?usp=drive_link

observed differences on box plots are statistically significant.

Table 3. D’Agostino K-squared (k^2) tests for Comparative datasets.

Dataset	k^2
NEUTRAL_REACHER_pt	471793.46***
NEUTRAL_REACHER_en	45481.83***
PARTISAN_pt	4763.45***
PARTISAN_REACHER_en	4285.82***

Note: k^2 represent the D’Agostino K-squared [113] statistic. All statistics are significant at $p < .001$ (***) level.

These results suggest that, in general, there does not appear to be a simple and direct relationship between being a neutral bubble reacher and reducing uncivil discourse across all situations. Rather, the relationship between “reaching the bubble” and incivility appears to depend on the broader national context in which neutral and partisan accounts operate. Table 4 summarizes these results along the same lines as Table 2, organizing them by media source partisanship and national polarization.

Table 4. Summary of results for H1, H2 and H3. Plus (+) sign indicates the incivility level, ranging from + + + + + (highest) to + (lowest).

Media type	Neutral Bubble Reacher	National polarization	
		High	Low
		HIGHEST INCIVILITY (+ + + + +) Neutral bubble reachers elicited the highest incivility compared to all other cases in a highly polarized situation (Brazil).	LOWEST INCIVILITY (+) Neutral bubble teachers elicited the lowest incivility in a less polarized situation (Canada).
	Partisan	MODERATE TO HIGH INCIVILITY (+ + + +) Partisans also elicited considerable incivility, but slightly lower when compared to neutral bubble reachers in a highly polarized situation (Brazil).	LOW TO MODERATE INCIVILITY (+ + +) Partisans elicited some incivility, which was slightly higher than bubble reachers in a less polarized situation (Canada-USA).

We conducted a number of additional analyses to increase our confidence in these results. These are reported in S4 Appendix and S5 Appendix. It is possible that specific sources influence the results. Therefore, in S4 Appendix we decompose the analysis by all individual sources. We found that only in the Canadian context were there more prominent differences across sources, with globalnews.ca showing a higher toxicity score than other Canadian sources. Possible reasons for that difference include the fact that, relative to globalnews.ca, the other organizations in the Canadian dataset are more traditional and established sources, whereas Global News emerged more recently and grew out of talk radio. It is also possible that the difference arises from the fact that globalnews.ca integrates its discussion board automatically with Facebook, and Facebook comments could tend to be less civil than those on the internal news organization systems, which are often moderated and require users to register. For this reason, we evaluate this possibility in S5 Appendix and find little evidence that uncivil comments are differentially related to Facebook than elsewhere, giving us greater confidence in our results. This analysis also indicates that in the case of globalnews.ca, other factors could be responsible for greater incivility. This would be a topic worth pursuing in an analysis more specifically focused on the Canadian media ecosystem.

The politicization of non-political content in a polarized context

The preceding analysis suggests ideologically neutral bubble reachers operate differently in Canada and Brazil. One possibility is that the heightened polarization in the Brazilian context makes any performance of ideological neutrality more suspect, resulting in political contention seeping into even seemingly non-political discussions. H4 explores this possibility by examining how incivility levels vary by political vs

non-political content shared by ideologically neutral bubble reachers. Scholars find that political content elicits more uncivil discourse. However, evidence of minimal toxicity score discrepancy between political and non-political content would suggest this distinction is blurred in a highly polarized context. If polarization closes the room for neutrality, this could explain why ideologically neutral bubble reachers are unable to temper uncivil discourse in Brazil.

To examine H4, we investigated comments on the G1_SITE_pt, collected from the portal at g1.globo.com. This dataset is composed of comments extracted from the same website as g1.globo.com source included in the NEUTRAL_REACHER_pt dataset, but at distinct moments. This analysis was conducted to identify whether there are differences between comments made about politics in a polarized situation, represented by the g1.globo.com source from NEUTRAL_REACHER_pt dataset, and comments made about distinct topics and situations, represented by the G1_SITE_pt dataset and underlying sources (each of which representing news columns on G1_SITE_pt). Fig 3 shows the distribution of toxicity scores for the G1 sources.

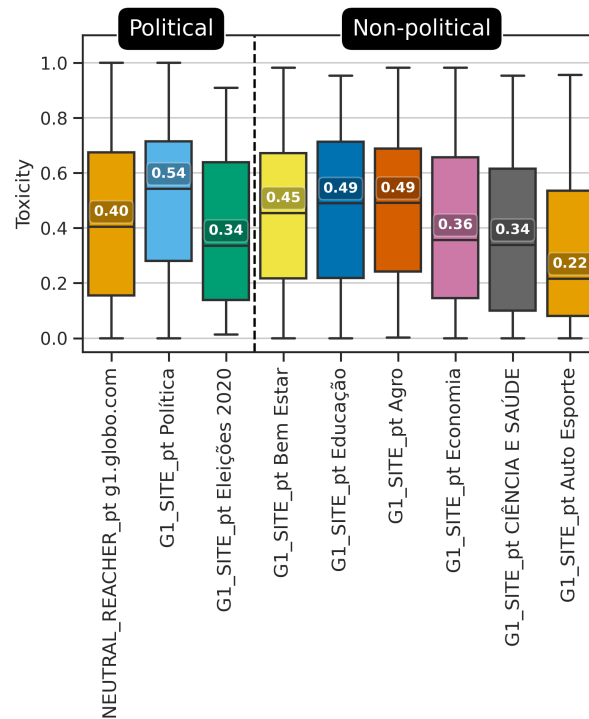


Fig 3. Box plots showing the distribution of toxicity scores for the G1 sources.

These box plots present some initially surprising results. Consistent with previous literature, political content generates high toxicity scores. G1_SITE_pt “Política” - Political, translated to English - is among the most uncivil columns in G1_SITE_pt. However, in line with H4, the differences in its toxicity scores compared to some of the other subjects are minimal. Consider “Bem estar”, “Educação”, and “Agro” – Wellness, Education and Agriculture, translated to English. These columns show a high toxicity score, which is surprising given that these columns appear to contain relatively neutral topics. To try to understand these cases, we analyzed samples of comments in each of these G1_SITE_pt columns. Here are some examples for the Wellness column with the highest toxicity score (*toxicity* = 0.99): “Globo vai se fu** seus M****”, “Vcs bolsonaristas vai pagar caro por cada morte de um brasileiro. Eu juro que vão.

Fanáticos de m****” e “Bolsonaro, faz um favor para esse pobre e f***** o país: SE MATA CARA !!!”. And some examples with the lowest toxicity score (*toxicity* = 0.00): “Ciência prova eficácia da hidroxicloroquina no tratamento inicial de COVID-19 - O tratamento com hid (...)”, “Pesquisadores analisaram dados de satélites comerciais ao redor de cinco hospitais de Wuhan, comparan (...)” e “E CONHECEREIS A VERDADE, E A VERDADE VOS LIBERTARÁ”. By analyzing these comments, as well as Education and Agriculture comment samples (not included for brevity), it was possible to understand why there is high incivility on seemingly neutral subjects. These cases usually contain comments discussing the political situation related to health (in the Wellness column), Education or Agriculture issues. Thus, this finding suggests that H4 holds, that even apparently neutral subjects are assimilated into political conflict in a highly polarized context (see [115] for a recent study finding that Political posts from influential users on Twitter engage users more than non-political posts).

To verify if the differences between toxicity scores observed in Fig 3 were statistically significant, we applied the tests explained in Section Methods for Hypothesis Investigation – the same one applied for datasets in Section Incivility and neutral bubble reachers, but changing the tested variables to be the G1 sources toxicity scores presented in this figure. These G1 sources were subjected to a normal test applying D’Agostino K-squared test [113]. Table 5 presents these test results, showing that none of the Facebook sources followed a normal distribution ($p < .001$). We also relied on Q-Q plots and histogram visualizations for each G1 source, confirming this result (plots were not included for brevity). Considering that the data were not following a normal distribution, we applied the Kolmogorov-Smirnov Goodness of Fit test [114] pairwise between G1 sources to verify whether the samples originate from the same distribution. Results for this test were included in a separate spreadsheet for consultation ¹³, which shows that none of the G1 sources originated from the same distribution ($p < .05$), except for two cases: the pairs “Eleições 2020” and “Economia”, as well as, “Eleições 2020” and “CIÊNCIA E SAÚDE”. These cases, as can be seen in the box plots in Fig 3 have similar distributions and medians, so this was expected and does not impact the conclusion. Therefore, the observed differences in box plots are statistically significant, except for these specific cases.

Table 5. D’Agostino K-squared (k^2) tests for G1 sources.

Dataset	Source	k^2
NEUTRAL_REACHER_pt	g1.globo.com	982481.041***
G1_SITE_pt	Política	235072.0699***
	Eleições 2020	2163.3942***
	Bem Estar	137482.9471***
	Educação	118091.6763***
	Agro	2445.8281***
	Economia	115085.351***
	CIÊNCIA E SAÚDE	3860.0542***
	Auto Esporte	252.9476***

Note: k^2 represent the D’Agostino K-squared [113] statistic. All statistics are significant at $p < .001$ (***) level.

To study H4 more formally, we examined the influence of political topics on the toxicity score of comments with regression analyses regarding political and non-political comments in the G1_SITE_pt and NYT_SITE_en datasets (see Section Methods for Hypothesis Investigation). Like Brazil, the American context from which NYT_SITE_en was collected is regarded as more polarized. In both cases, the linear regression coefficients were statistically significant ($p < .001$), being $r = 0.116$ *** for G1 and $r = 0.093$ *** for New York Times. Still, while statistically significant, the influence was weak for both cases, showing that political topics’ influence on incivility is minimal.

Overall, this result indicates that political topics seem to elicit slightly more uncivil comments compared to non-political topics.

Discussion and Conclusion

Bubble reachers manage to reach ideologically diverse users by distributing information across politically distinct homophilic groups, but those who are exposed to such content tend to share what is aligned with their own political orientation [20]. The present study extended these findings, moving beyond the political valence of content that users share to the discursive style of that content. In particular, we examined a series of hypotheses that stem from asking whether ideologically neutral bubble reaching accounts create venues where users engage in more civil discourse, as compared to partisan accounts.

Our results complicated the simple hypothesis of a universal impact of neutral bubble reachers across contexts. Instead, we found evidence that the relationship between bubble reaching and incivility is moderated by the national political culture. In the Canadian case, we find that ideologically neutral bubble reachers tend to elicit more civil discourse. Compared to more ideologically partisan outlets, neutral bubble reachers carve out discursive spaces where uncivil commentary is less pronounced. This is consistent with the stated goals of these outlets, where aspirations towards objectivity and fairness serve as a foundation for more civil discourse. However, incivility was more pronounced in the Brazilian context among the ideologically neutral bubble reachers.

By exploring further hypotheses, we found evidence as to the source of these differences. It appears that the capacity of neutral bubble reachers to operate as mediating institutions may be undermined in highly-polarized contexts where populist discourse is prominent. Bubble reachers may present themselves as neutral arbiters of impartial expertise, but this performance may backfire under these conditions [57]. As discussed in the literature review, populists often undermine trust in mediating institutions by arguing those institutions and the elites who staff them advance the particular interests of the status quo. Polarization, more generally, raises the stakes of politics, creating a climate where passionate partisanship is incentivized while remaining neutral or unbiased becomes treated with suspicion. This context works against the performances of ideologically neutral bubble reachers, constraining the discursive space for civility.

The greater prominence of populism and polarization in the Brazilian context may explain the discrepancies we observe. Consistent with this depiction, our analysis finds greater levels of incivility in the Brazilian context across both neutral and partisan accounts, suggesting a more inflamed discursive space in general. We also find that while explicitly political content tends to elicit somewhat more uncivil discourse, the discrepancy is lower than expected: even ostensibly non-political content gets pulled into the ambit of partisan politics. Collectively, this may explain the role of ideologically neutral bubble reachers in Brazil – populism and polarization challenge these outlets’ credibility, turning discursive spaces into meeting points for rancorous conflict rather than mediation, and shrinking the room for civility. Under these circumstances, ideologically neutral bubble reachers may “open the hand” for broader discussion, but still receive the fist.

Our findings support and extend research on the contextual nature of incivility. Scholars argue that incivility is inherently sensitive to cultural context – prevailing interaction norms dictate whether an expression is perceived to be disrespectful or not [116]. But even those who breach norms of civility do so in ways that are attuned to contextual factors at different levels [117]. While several studies point to variation in topical context of news stories as an important factor [24, 118], the role of national context remains unclear. One of the few comparative studies, in fact, finds that while

levels of uncivil conflict vary by national context, the relationship between incivility and political engagement is not mediated by national context [119]. In contrast, we find that national context mediates where incivility is more likely to be expressed - whether in relation to neutral bubble reaching outlets, or across topical issues. While Otto et al.'s [119] comparative analysis focused on three European countries, our findings suggest that when examining countries with more dissimilar political cultures, the patterning of incivility may differ. Further systematic comparative analysis exploring how incivility is mediated in different national contexts would be a worthwhile direction for scholars to pursue.

That the neutral bubble reachers detected by our analysis corresponded with legacy media outlets also makes our findings relevant to related scholarship. Research suggests that hostile media perceptions - the tendency of partisans to view media coverage as biased against them even if coverage is even-handed - contribute to incivility in discourse [58, 120]. Scholars highlight cognitive, emotional, and behavioural dimensions of this process: motivated reasoning distorts how partisans receive information from outlets, contributing to perceptions of bias that incite feelings of media indignation, and increase willingness to engage in uncivil discourse in turn [121]. Our findings are consistent with this work, showing how greater incivility can ensue as legacy media outlets become meeting-points for partisan conflict. We extend this work, however, by underscoring the importance of cultural context in cueing this package of cognitive, emotional, and behavioural processes. The underlying mechanisms are not entirely clear. For instance, a populist context may frame the reputation of legacy media outlets in ways that disproportionately attract antagonistic partisans while alienating moderates, heightening incivility as a result of altering the composition of commenters [122]. Alternatively, cultural norms in a populist context may heighten media indignation towards these outlets across the board, increasing the expression of incivility for both partisans and moderates. Future research can work towards clarifying the mechanisms whereby cultural context mediates the relation between legacy media and incivility.

This study represents a significant advancement from previous research by analyzing the communication style provoked by neutral bubble reachers. The findings from this study could help researchers, platform designers, and policymakers seeking to promote healthier online interactions. Recognizing the contextual factors that impact the effectiveness of neutral bubble reachers can inform the development of strategies to mitigate incivility and foster more constructive dialogue in online spaces. Moreover, understanding the interplay between polarization, populism, and the performance of mediating institutions can contribute to efforts to address the broader challenges associated with political polarization and the erosion of civility in public discourse.

While our study is highly suggestive of these relationships, some possible limitations should be highlighted. One was regarding the `PARTISAN_REACHER_en` dataset, which was used to approximate Canada's less polarized political culture. This approximation was achieved by cross-referencing the top 100 bubble reachers with partisan $RP(H)$ on Canada with a Facebook dataset containing comments on 20,000 posts made by American news organizations and personalities [100] – see Comparative Datasets (for H1, H2 and H3) section for a complete reference.

The fact that the comparative, reference and alternative datasets were obtained from different sources makes comparisons less robust, since the dynamics of interactions on each platform can influence the degree of incivility. For example, `NEUTRAL_REACHER_pt` and `NEUTRAL_REACHER_en` comments were extracted from content produced on distinct news websites, while `PARTISAN_pt` and `PARTISAN_REACHER_en` comments were extracted from content on Facebook Pages in distinct situations and dates. In this sense, each source (news website) has distinct moderation mechanisms that may or may not allow uncivil behaviour in comments, thus

making some news outlets less uncivil by default. These differences make it difficult to conduct a perfectly fair comparison between them. In this same direction, cultural, contextual and linguistic differences may also directly affect the toxicity we observe in comparing datasets in distinct languages and situations. Additionally, it is noteworthy that `PARTISAN_pt` exclusively comprises right-wing pages, whereas `PARTISAN_REACHER_en` includes left-wing or left-leaning pages, as indicated by the All Sides Media Bias Report. This distinction is pertinent, as varying political perspectives may prompt disparate online behaviours in terms of information consumption [110, 123] (e.g. liberals being more likely than conservatives to engage in cross-ideological dissemination) [41]). Considering these aspects, the differences that can be observed in datasets and sources and conclusions that were drawn from these results need to be considered with caution. We tried to address some of these limitations by including additional analyses in S5 Appendix and S6 Appendix.

Regarding differences between datasets and sources, it is worth noting that the standard deviation for some datasets, such as, for example, `NEUTRAL_REACHER_pt`, `PARTISAN_REACHER_en` and `PARTISAN_pt` datasets is high (see Table 9 from S2 Appendix). This can potentially limit conclusions about these cases, even when there are statistically significant differences between them. We also recognize that the pre-processing step, although not aggressive, may have eliminated some representative comments in some datasets, such as `PARTISAN_pt`. Despite this, the number of comments in our analysis was high in all datasets and sources, which reduces the chance that the pre-processing step greatly influenced the results. Another limitation refers to the polarization characteristics in the `NEUTRAL_REACHER_pt` and `NEUTRAL_REACHER_en` datasets. In this sense, several comments presented subtle uncivil characteristics, mainly with sarcasm related to the political context itself. These subtleties could be too complex to be captured by a generic model, perhaps even by humans without detailed local context-sensitivity [79].

Our study opens up several possibilities for future work. For instance, while we analyzed political polarization in two different contexts, it is essential to investigate more contexts, including similar contexts in other countries and new contexts. Extending our analysis beyond Canada and Brazil could provide more robust support for the observed role of political culture in mediating how bubble reachers operate. Further, examining the hypotheses in different periods for the same country is also an important future step; this enables the understanding of, for example, how changes in political culture affect the results. For instance, we explain higher incivility levels in Brazil due to greater polarization and populism in their political culture – evidence of longitudinal growth in incivility corresponding with polarization would be in line with this explanation and could parse out the time-ordering between these two. Likewise, a longitudinal analysis could examine how the diffusion of populist framing in public culture affects whether citizens adopt more toxic styles in their expression [124]. Another motivation for examining our hypothesis in a different period is because COVID-19 has been associated with an increase in affective polarization, exacerbated by factors such as the politicization of public health measures and information, leading to heightened ideological divides [125]. Although most of our data do not cover this period (only `G1_SITE_pt` and possibly `UNCIVIL_pt`, which does not provide their collection date), polarization could have changed after this period. Additionally, given that the neutral bubble reacher accounts examined were comprised of legacy news outlets, it is essential to ponder whether similar conclusions can be extended to other account types, such as bubble reaching accounts entirely unrelated to politics. These expanded possibilities contribute to the generalizability of the findings across diverse political cultures and various types of bubble reaching accounts.

Supporting information

S1 Appendix. Alternative datasets used in sensitivity analysis.

This appendix describes the alternative datasets applied in sensitivity analysis presented in S5 Appendix and S6 Appendix.

To address limitations related to the comparative datasets, particularly the exclusive collection of data for partisan accounts from the Facebook platform (for `PARTISAN_pt` and `PARTISAN_REACHER_en`), additional data was gathered. Specifically, we investigated to what extent the Facebook platform could influence the toxicity levels of comments to ensure a fair comparison between the comparative datasets. These datasets are summarized below.

FACEBOOK_NEUTRAL_en: obtained by selecting the top 3 Facebook pages by the number of comments from central leaning news media sources according to the All Sides Media Bias Report⁷. Selected pages were: `bbc`, `the_hill`, `abc_news`.

FACEBOOK_PARTISAN_en: obtained by selecting the top 3 Facebook pages by the number of comments from left and top 3 from right-leaning news media sources according to the All Sides Media Bias Report⁷. Selected pages were: `cnn`, `msnbc`, `raw_story` with left-leaning and `fox_news`, `breitbart`, `the_blaze` with right-leaning.

FACEBOOK_PERSON_en: obtained by selecting the top 3 Facebook pages from influential people (not news media) with the most comments in the dataset. Selected pages were: `megyn_kelly`, `bill_mahar` and `rachel_maddow`;

FACEBOOK_OTHER_en: obtained by randomly selecting 4 Facebook pages that weren't selected for `FACEBOOK_NEUTRAL_en`, `FACEBOOK_PARTISAN_en` or `FACEBOOK_PERSON_en`. Selected pages were: `los_angeles_times`, `mother_jones`, `npr`, and `yahoo_news`.

In addition to the alternative Facebook datasets, we also obtained comments from the Yahoo News site, which was labelled `YAHOO_SITE_en`. This dataset was used alongside the New York Times (`NYT_SITE_pt`) presented in the methodology section to conduct additional sensitivity analyses (described in S6 Appendix).

YAHOO_SITE_en: composed of 10,000 random comments from Yahoo News website comments [126]. This dataset was selected to compare with comments obtained from `yahoo_news` Facebook page contained in `FACEBOOK_OTHER_en` to understand the influence of the Facebook platform by comparing the toxicity on comments made on Facebook and on the news website.

S2 Appendix. Datasets and associated sources summary.

This appendix serves as a quick reference for datasets and associated sources. Tables 6, 7 and 8 present a summary of datasets and associated sources data for the Comparative, Reference and Alternative groups, respectively. Tables 9 and 10 provide summarized statistics for datasets and associated sources, respectively, including the initial comments count before and after pre-processing (Original and Final size columns, respectively), the percentage of comments that were removed after pre-processing (% column), and the mean, median and standard deviation values for the toxicity variable (Mean, Median and σ columns, respectively).

S3 Appendix. Comments pre-preprocessing.

The pre-processing step was performed to maintain maximum integrity in the datasets presented in Section Materials and Methods, only removing noisy instances that could negatively impact performance during toxicity identification. Escape sequences (such as line breaks), Markdown tags, and links have been removed. Sentences repeated in more than one comment, such as "Reply to @" and "REPORT COMMENT," which are clearly not part of the construction of a comment, were also identified and removed.

Since online comments tend to be noisy, usually containing spelling errors, it was necessary to apply an autocorrection step. For this task, an algorithm called *SymSpell*

Table 6. Summary of datasets composing the Comparative group.

Dataset	Dataset summary	Source	Source summary
NEUTRAL_REACHER_pt	Comments captured from Brazilian news media websites during the 2018 presidential election representing the top three neutral bubble reaching domains in the Brazilian election setting: noticias.uol.com.br, g1.globo.com, and extra.globo.com [20].	extra.globo.com	Comments extracted on political topics during the 2018 Brazilian Presidential Elections from the news website extra.globo.com.
		g1.globo.com	Comments extracted on political topics during the 2018 Brazilian Presidential Elections from the news website g1.globo.com.
		noticias.uol.com.br	Comments extracted on political topics during the 2018 Brazilian Presidential Elections from the news website noticias.uol.com.br.
NEUTRAL_REACHER_en	Comments captured from Canadian news media websites during the 2019 federal elections representing the top three neutral bubble reachers in the Canadian election setting: cbc.ca, globalnews.ca, and theglobeandmail.com	globalnews.ca	Comments extracted on political topics during the 2019 Canadian Federal Elections from the news website globalnews.ca.
		cbc.ca	Comments extracted on political topics during the 2019 Canadian Federal Elections from the news website cbc.ca.
		theglobeandmail.com	Comments extracted on political topics during the 2019 Canadian Federal Elections from the news website theglobeandmail.com.
PARTISAN_pt	Comments captured between 2012-01-01 and 2018-12-31 from Brazilian conservative Facebook pages: padrepaolo, flaviobolsonaro, CampanhadoArmamento, carvalho.olavo, DireitaConservadoraOficial, OPesadelodeQualquerPolitico2.0, and MisesBrasil [103].	CampanhadoArmamento	The Facebook page Campanha do Armamento is associated with the Instituto Defesa, a non-profit organization dedicated to advocating for and preserving the right to access firearms and self-defense in Brazil.
		MisesBrasil	The Facebook page Mises Brasil is associated with the Ludwig von Mises Brazil Institute (IMB), a think tank dedicated to producing and disseminating economic and social science studies that promote the principles of free markets and a free society.
		OPesadelodeQualquerPolitico2.0	The Facebook page O Pesadelo de Qualquer Político 2.0 was a conservative page used to diffuse misinformation about politics. It is not available any more on the platform.
		Carvalho.olavo	This Facebook page is dedicated to disseminating the ideas and perspectives of Olavo de Carvalho, a Brazilian philosopher and political commentator known for his conservative and controversial views.
		Flaviobolsonaro	This Facebook page is associated with Flávio Bolsonaro, a right-wing Brazilian politician and senator, and serves as a platform for promoting his political agenda, sharing news and updates, and engaging with his supporters and followers.
		padrepaolo	This Facebook page is dedicated to promoting the teachings and activities of Father Paulo, a religious figure who shares spiritual guidance, discusses social issues, and engages with his followers through inspirational content and community outreach initiatives.
PARTISAN_REACHER_en	Comments captured from Facebook news pages representing the partisan accounts: bloomberg-politics, huffington-post, and ny-times [100].	huffington-post	The Facebook page Huffington Post is a platform for sharing news, opinion pieces, and engaging content covering a wide range of topics, including politics, culture, and current events.
		ny-times	The Facebook page New York Times provides comprehensive coverage of national and international news, investigative journalism, and in-depth analysis on various subjects, serving as a reliable source of information.
		bloomberg-politics	The Facebook page Bloomberg Politics offers insightful news and analysis on politics, policy, and global economic issues, providing readers with a deeper understanding of the intersection between business and politics.

Table 7. Summary of datasets composing the Reference group.

Dataset	Dataset summary
UNCIVIL_en	Highly uncivil comments in English obtained from an open dataset with human-labeled Wikipedia comments according to different categories of toxic behavior ¹⁴ .
UNCIVIL_pt	Highly uncivil comments in Brazilian Portuguese obtained from tweets manually annotated according to different toxicity categories [104].
CIVIL_en	More civil comments in English obtained through Reddit’s public API from communities known for moderated discussions.
CIVIL_pt	More civil comments in Brazilian Portuguese obtained from a database of product reviews on a famous e-commerce website (B2W Digital) [105].

¹⁶ was applied, which depends on the creation of a dictionary with the correct number of words that will be used to replace those whose spelling does not conform to the dictionary standard. Thus, a dictionary was created for each dataset, including the words that appeared at least 5 times in the respective dataset, given that more frequent tokens tend to be the correct version. In contrast, rare tokens tend to be spelling errors. Then, into this initial dictionary, we concatenated a standard dictionary of the language

¹⁶<https://symspellpy.readthedocs.io>.

Table 8. Summary of datasets composing the Alternative group.

Dataset	Dataset summary	Source	Source summary
FACEBOOK_NEUTRAL_en	Top 3 Facebook pages [100] with comments from central leaning news media sources according to the All Sides Media Bias Report ⁷ : bbc, abc_news, and the.hill.	bbc	The Facebook page BBC provides global news coverage across various topics including politics, current affairs, culture, and entertainment, with a focus on objective reporting and analysis.
		abc_news	The Facebook page ABC News delivers breaking news, features, and in-depth stories covering a wide range of topics, including politics, business, technology, and popular culture, with an emphasis on American perspectives.
		the.hill	The Facebook page The Hill offers political news, analysis, and opinion pieces covering US politics, Congress, the White House, and policy debates, catering to a wide audience including policymakers and political enthusiasts.
FACEBOOK_PARTISAN_en	Top 3 Facebook pages [100] with comments from left-leaning (cnn, msnbc, raw_story) and right-leaning (fox_news, breitbart, the.blaze) news media sources.	cnn	The Facebook page CNN delivers comprehensive news coverage on various topics including politics, business, world affairs, and entertainment, with a reputation for breaking news stories and in-depth reporting.
		msnbc	The Facebook page MSNBC offers liberal-leaning news, analysis, and opinion pieces on politics, social issues, and culture, catering to progressive viewers and providing coverage on a wide range of topics.
		raw_story	The Facebook page Raw Story features progressive news and investigative reporting, covering politics, social justice, science, and culture, with a focus on critical analysis and alternative perspectives.
		fox_news	The Facebook page Fox News provides conservative-leaning news and opinion pieces on current events, politics, business, and culture, with a focus on American perspectives and a strong following among conservative viewers.
		breitbart	The Facebook page Breitbart offers conservative news, opinion pieces, and analysis on politics, world events, culture, and technology, catering to a right-leaning audience and promoting conservative viewpoints.
		the.blaze	The Facebook page The Blaze provides conservative news, commentary, and analysis on politics, culture, and current events, offering a platform for conservative voices and catering to an audience seeking right-leaning perspectives.
FACEBOOK_PERSON_en	Top 3 Facebook pages [100] with comments from left-leaning (cnn, msnbc, raw_story) and right-leaning (fox_news, breitbart, the.blaze) news media sources.	person_rachel_maddow	Rachel Maddow's Facebook page features content from her television show, offering progressive commentary, analysis, and investigative reporting on politics, current events, and social issues from a liberal perspective.
		person_megyn_kelly	Megyn Kelly's Facebook page features content from her media ventures, offering news, interviews, and commentary on politics, culture, and current events, with a focus on delivering diverse perspectives and engaging discussions.
		person_bill_mahar	Bill Mahar's Facebook page features content from his shows, providing progressive commentary, satire, and political humor, with a focus on social and political issues, current events, and interviews with notable figures.
FACEBOOK_OTHER_en	Randomly selected 4 Facebook pages [100] not included in FACEBOOK_NEUTRAL_en, FACEBOOK_PARTISAN_en, or FACEBOOK_PERSON_en: los.angeles.times, mother.jones, npr, and yahoo.news.	npr	The Facebook page NPR delivers news, analysis, and storytelling on a wide range of topics including politics, culture, science, and arts, with a reputation for in-depth reporting and high-quality journalism.
		yahoo.news	The Facebook page Yahoo News offers a mix of news articles, videos, and features covering politics, current events, lifestyle, and entertainment, catering to a broad audience with diverse interests.
		los.angeles.times	The Facebook page Los Angeles Times provides comprehensive news coverage on local, national, and international stories, including politics, entertainment, sports, and culture, with a focus on the Los Angeles area.
		mother.jones	The Facebook page Mother Jones offers progressive news, investigative reporting, and commentary on politics, social justice, and environmental issues, with a reputation for in-depth reporting and inquisitive journalism.
NYT_SITE_en	Comments from the New York Times website [19].	Politics	5,000 random comments from politics related sections.
YAHOO_SITE_en	10,000 random comments from Yahoo News website [126].	Non-politics	5,000 random comments from non-politics related sections.
G1_SITE_pt	Comments obtained from the g1.globo.com domain during different contexts other than the 2018 Brazilian presidential elections [106].	Economia	This column covers the latest updates and analyses on financial markets, business news, economic policies, and trends affecting Brazil and the global economy.
		Bem Estar	This column provides information and tips on nutrition, fitness, mental health, and lifestyle choices for a balanced and healthy life.
		Educação	This column explores educational topics, including school policies, educational reforms, advancements in teaching methods, and initiatives aimed at improving the education system in Brazil.
		CIENCIA E SAÚDE	This column covers groundbreaking research, medical advancements, public health issues, and scientific discoveries in various fields to keep readers informed about the latest developments.
		Agro	This column delves into the agricultural sector, discussing farming practices, rural policies, market trends, innovations in agriculture, and the impact of agriculture on the economy and society.
		Auto Esporte	Focusing on the automobile industry, this column provides news, reviews, and insights on cars, motorcycles, and related technologies, keeping readers updated on the latest trends in the automotive world.
		Eleições 2020	This column covered the 2020 elections in Brazil.
		Política	Covering domestic and international politics, providing insights into the political landscape in Brazil and beyond.

of each dataset, Brazilian Portuguese for datasets with _pt prefix and United States English for datasets with _en prefix. These dictionaries were obtained from the OpenOffice repository¹⁷. The resulting dictionary is then processed by *SymSpell*, which generates permutations of the words through the character deletion procedure, resulting in a dictionary of permutations with words that would potentially be misspellings. After this procedure, the comments' autocorrection was performed. In this step, *SymSpell* was configured to autocorrect only words with a maximum edit distance of 2 characters (the number of characters needed to turn a misspelled word into the correct word). It was observed that the increase to a maximum edit of 3 or more characters generated incorrect respellings. Short words with wrong spelling were more likely to be replaced by words with correct spelling but unrelated to the corrected word. It is important to mention that the autocorrection step does not eliminate the possibility of interference from adversarial attacks [127], but it can reduce their incidence.

¹⁷<https://www.openoffice.org/lingucomponent/dictionary.html>.

Table 9. Datasets size before and after pre-processing and statistics.

Dataset	Size			Toxicity		
	Original	Final	%	Mean	Median	σ
NEUTRAL_REACHER_pt	128,898	122,836	95.30	0.44	0.42	0.27
PARTISAN_pt	30,000	24,741	82.47	0.32	0.20	0.30
NEUTRAL_REACHER_en	115,779	113,114	97.70	0.17	0.10	0.17
PARTISAN_REACHER_en	31,075	28,070	90.33	0.24	0.16	0.23
CIVIL_pt	5,000	4,989	99.78	0.12	0.06	0.15
CIVIL_en	5,000	4,976	99.52	0.19	0.11	0.19
UNCIVIL_pt	5,000	4,718	94.36	0.87	0.90	0.14
UNCIVIL_en	5,000	4,927	98.54	0.95	0.96	0.06
FACEBOOK_NEUTRAL_en	60,155	57,839	96.15	0.23	0.14	0.23
FACEBOOK_PARTISAN_en	121,928	118,910	97.52	0.27	0.20	0.24
FACEBOOK_PERSON_en	70,717	70,290	99.40	0.30	0.24	0.24
FACEBOOK_OTHER_en	60,337	56,404	93.48	0.22	0.13	0.23
G1_SITE_pt	342,405	328,026	95.80	0.47	0.50	0.26
NYT_SITE_en	256,507	255,012	99.42	0.19	0.15	0.16
YAHOO_SITE_en	10,000	9,026	90.26	0.17	0.08	0.20

Considering that even with the autocorrection procedure, the comments could still present noise from rare or unusual words, a pre-processing step was added to remove comments with many poorly recognized instances. For this, the comments were processed by a tool widely applied in the literature to recognize linguistic, psychological, and social characteristics called LIWC [128]. This tool was chosen because it has good coverage for several dictionaries of different languages and does not require complex steps in the pre-processing of the text to be analyzed. For datasets in English, the official internal LIWC dictionary was used, and an unofficial dictionary [129]¹⁸ for datasets in Brazilian Portuguese. After processing the datasets with LIWC, the “Dic” attribute was used to keep only comments with at least 50% of the words recognized by the respective LIWC dictionary – this attribute counts the percentage of words identified in the respective dictionary. Finally, considering that short comments can influence toxicity results, all comments containing less than 10 words were removed using LIWC’s WC (Word Count) parameter, which counts the number of words identified in a text.

S4 Appendix. Sensitivity Analysis - Toxicity by Sources.

Our central hypotheses concern whether neutral bubble reachers tend to reduce uncivil discourse. We found that they do in Canada but not in Brazil. To probe this result further, we examined specific accounts to understand if their distinctive histories and audiences might impact results, as well as whether the commenting systems and moderation policies they use would perhaps generate more heated discussions. For this purpose, we conducted the same analysis for the datasets in Incivility and neutral bubble reachers section. S4 Fig shows the toxicity score box plots grouped by source for this analysis.

This figure reveals valuable insights that were not possible with Fig 2. Noticias.uol.com.br and g1.globo.com showed a similar toxicity level for the Brazilian neutral bubble reachers. This suggests that the toxicity level was likely not a direct product of those platforms, but more probably related to the general political situation of the country.

Regarding the Canadian neutral bubble reachers, both theglobeandmail.com and cbc.ca exhibited very low toxicity scores, while, compared to them, globalnews.ca shows a 1.6 times higher toxicity median score. This finding is interesting for at least two reasons. For one, it suggests that if globalnews.ca had produced toxicity scores similar to the other Canadian news sources, the effect of neutral bubble reachers would have been even stronger. Second, it raises questions about why globalnews.ca differs from the

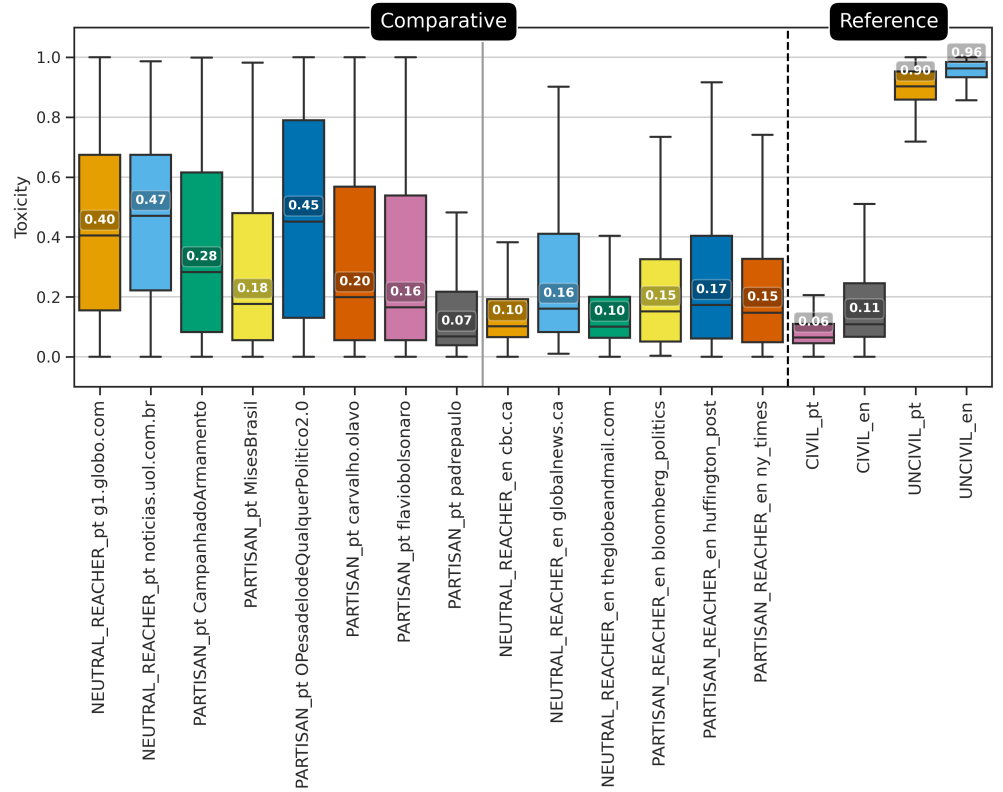
¹⁸<http://143.107.183.175:21380/portlex/index.php/pt/projects/liwc>.

Table 10. Sources size before and after pre-processing and statistics.

Dataset	Source	Size			Toxicity		
		Original	Final	%	Mean	Median	σ
NEUTRAL_REACHER_pt	extra.globo.com	51	49	96.08	0.59	0.67	0.25
	noticias.uol.com.br	69,814	67,779	97.09	0.45	0.47	0.25
	g1.globo.com	59,033	55,057	93.26	0.42	0.40	0.27
PARTISAN_pt	CampanhadoArmamento	5,000	4,199	83.98	0.36	0.28	0.29
	MisesBrasil	5,000	3,508	70.16	0.29	0.18	0.27
	OPesadelodeQualquerPolitico2.0	5,000	4,306	86.12	0.46	0.45	0.32
	carvalho.olavo	5,000	4,218	84.36	0.32	0.20	0.30
	flaviobolsonaro	5,000	4,307	86.14	0.30	0.16	0.28
	padrepaulo	5,000	4,203	84.06	0.16	0.07	0.20
NEUTRAL_REACHER_en	theglobeandmail.com	13,328	13,144	98.62	0.17	0.10	0.17
	cbc.ca	100,358	97,932	97.58	0.17	0.10	0.16
	globalnews.ca	2,093	2,038	97.37	0.28	0.16	0.25
PARTISAN_REACHER_en	bloomberg_politics	961	888	92.40	0.23	0.15	0.22
	ny_times	15,754	14,169	89.94	0.22	0.15	0.21
	huffington_post	14,360	13,013	90.62	0.26	0.17	0.24
FACEBOOK_NEUTRAL_en	the.hill	20,490	20,070	97.95	0.28	0.20	0.25
	abc.news	16,598	15,663	94.37	0.22	0.11	0.22
	bbc	23,067	22,106	95.83	0.20	0.11	0.19
FACEBOOK_PARTISAN_en	the.blaze	15,112	14,714	97.37	0.25	0.18	0.23
	raw_story	20,076	19,703	98.14	0.35	0.31	0.27
	breitbart	20,218	19,175	94.84	0.28	0.18	0.27
	msnbc	21,207	20,917	98.63	0.27	0.21	0.22
	cnn	22,325	21,618	96.83	0.24	0.17	0.21
	fox.news	22,990	22,783	99.10	0.26	0.20	0.22
FACEBOOK_PERSON_en	person_rachel_maddow	24,123	24,092	99.87	0.32	0.28	0.22
	person_megyn_kelly	22,747	22,488	98.86	0.24	0.17	0.21
	person_bill_mahar	23,847	23,710	99.43	0.34	0.29	0.25
FACEBOOK_OTHER_en	los_angeles_times	11,465	10,234	89.26	0.24	0.16	0.24
	yahoo_news	14,421	13,668	94.78	0.23	0.16	0.21
	npr	16,959	15,396	90.78	0.15	0.07	0.18
	mother_jones	17,492	17,106	97.79	0.27	0.20	0.24
G1_SITE_pt	Economia	34,518	33,437	96.87	0.40	0.36	0.26
	Bem Estar	65,606	63,011	96.04	0.45	0.45	0.26
	Eleições 2020	725	702	96.83	0.39	0.34	0.26
	Educação	5,227	5,086	97.30	0.47	0.49	0.26
	CIÊNCIA E SAÚDE	2,492	2,410	96.71	0.38	0.34	0.27
	Agro	2,024	1,962	96.94	0.46	0.49	0.25
	Auto Esporte	1,747	1,683	96.34	0.30	0.22	0.26
	Política	230,066	219,735	95.51	0.49	0.54	0.25
NYT_SITE_en	Politics	57,934	57,502	99.25	0.22	0.18	0.17
	Non-politics	198,573	197,510	99.46	0.19	0.14	0.15
YAHOO_SITE_en	–	10,000	9,026	90.26	0.17	0.08	0.19

others. There are a number of possible explanations for this difference. For example, the CBC is a government news source, and The Globe and Mail is one of Canada’s oldest national news organizations with a strong reputation for being moderate. Global News, by contrast, is a newer source that grew out of talk radio, a legacy that might carry forward into it, hosting more lively and heated conversations. Another possibility concerns the commenting systems used by the sites. CBC requires users to sign into their site to comment and use a verified email address and does not permit pseudonyms. The Globe and Mail requires users to have a verified subscription to comment. By contrast, Global News’ discussion board is automatically integrated with Facebook. This difference might account for the greater toxicity of Global News comments, in that Facebook conversations may tend to be more heated and less cautionary, and also are not confined to subscribers. This observation motivates additional sensitivity analysis of the impact of Facebook on toxicity levels presented on S5 Appendix and S6 Appendix.

To verify if the differences between toxicity scores observed in S4 Fig were



S4 Fig. Box plots showing the distribution of toxicity scores grouped by source for the comparative and reference comments datasets.

statistically significant, we applied the tests explained in Section Methods for Hypothesis Investigation (the same one applied for datasets in Section Incivility and neutral bubble reachers, but changing the tested variables to be the Comparative group sources toxicity scores presented on Fig S4 Fig. These sources were subjected to a normal test applying D’Agostino K-squared test [113]. Table 11 presents these test results, showing that none of the datasets followed a normal distribution ($p < .001$). We also relied on Q-Q plots and histogram visualizations for each source, confirming this result (plots were not included for brevity). Considering that the data were not following a normal distribution, we applied the Kolmogorov-Smirnov Goodness of Fit test [114] pairwise between sources to verify whether the samples originate from the same distribution. Results for this test were included in a separate spreadsheet for consultation¹³, which shows that none of the Comparative sources originated from the same distribution ($p < .05$), except for the case of the pair of sources ny_times and bloomberg_politics, which the test showed that were originated from the same distribution. Therefore, the observed differences in box plots are statistically significant, except for this specific case.

S5 Appendix. Sensitivity Analysis - Incivility on Facebook by Different Accounts.

Overall, our analysis suggests that neutral bubble reachers have a variable effect on uncivil discourse. Some evidence suggests that in more highly polarized contexts, their performances of neutrality lose credibility, thereby eliciting greater toxicity. The strength of our results, however, relies on controlling for unaccounted sources of bias.

Table 11. D’Agostino K-squared (k^2) tests for Comparative sources.

Dataset	Source	k^2
NEUTRAL_REACHER_pt	g1.globo.com	982.481.04***
	noticias.uol.com.br	92,809.12***
PARTISAN_pt	flaviobolsonaro	559.56***
	MisesBrasil	409.62***
	OPesadelodeQualquerPolitico2.0	27,061.59***
	padrepaulo	1,299.81***
	carvalho.olavo	1,017.08***
	CampanhadoArmamento	2,932.33***
NEUTRAL_REACHER_en	cbc.ca	40,039.28***
	theglobeandmail.com	5,002.15***
	globalnews.ca	312.46***
PARTISAN_REACHER_en	huffington.post	1,626.50***
	ny_times	2,478.50***
	bloomberg_politics	174.65***

Note: k^2 represent the D’Agostino K-squared [113] statistic. All statistics are significant at $p < .001$ (***) level.

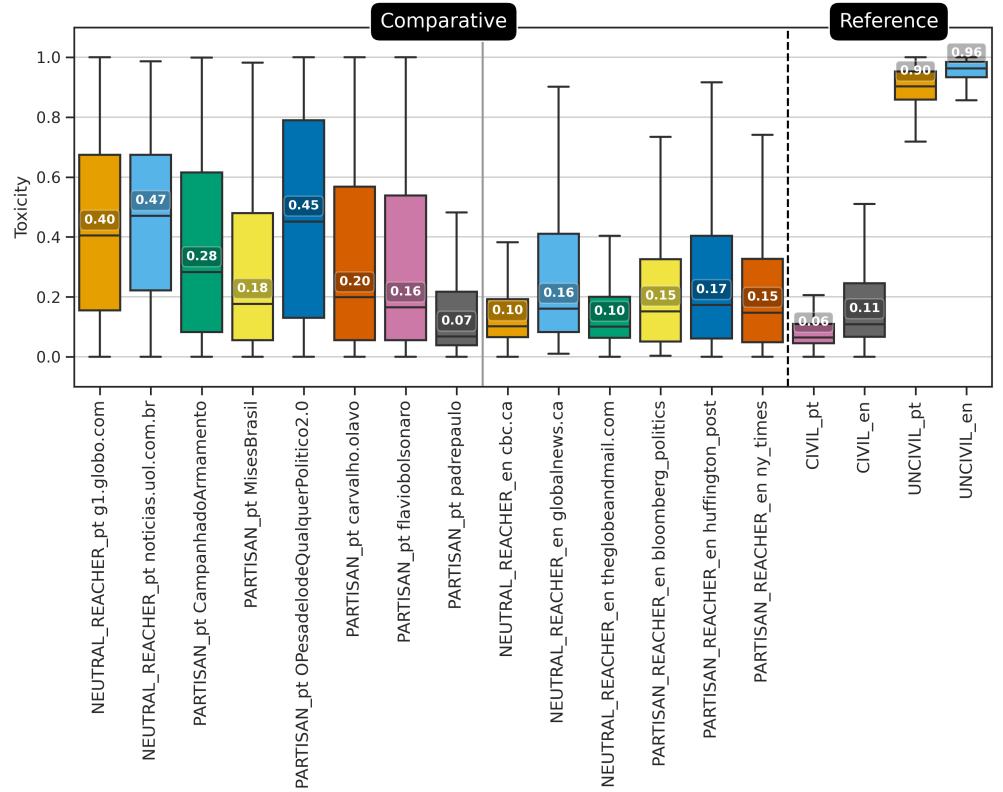
One limitation in the analysis was that data on ideologically partisan accounts were exclusively collected from Facebook, raising the possibility that the results may be unique to the affordances of this platform.

We therefore compare the datasets with comments obtained from Facebook pages to identify whether Facebook accounts for different toxicity scores over and above bubble reaching. This is important for determining if the toxicity levels identified in the PARTISAN_pt and PARTISAN_REACHER_en datasets is potentially a product of partisan accounts, or if it is more likely a product of Facebook’s platform own users’ normal behaviour. In the first case, the toxicity scores should be different between the datasets, while in the second case, the toxicity scores should be similar.

S5 Fig show the toxicity box plot to compare comments on distinct Facebook pages. The vertical gray line in this figure separates the Brazilian Portuguese (left) from the English pages (right).

The differences between pages in both languages are clearly noticeable, suggesting that toxicity scores are not products of the Facebook platform alone. On the PARTISAN_pt dataset, for example, there are cases with relatively low toxicity, like padrepaulo’s page, but also cases with the highest toxicity among other pages, like in OPesadelodeQualquerPolitico2.0 and Campanha do Armamento. When it comes to personal pages from famous personalities, like rachel_maddow (political news commentator) and bill_mahar (also a political news commentator, but in a satirical tone), the toxicity scores are slightly higher compared to pages from news media outlets on FACEBOOK_OTHER_en and PARTISAN_REACHER_en. It is worth noting that PARTISAN_REACHER_en sources, in general, had median values close to the sources in FACEBOOK_NEUTRAL_en, except for the the_hill Facebook page. Also, the sources in FACEBOOK_PARTISAN_en and FACEBOOK_PERSON_en have median toxicity scores higher than the sources in PARTISAN_REACHER_en and FB_CENTER_en, except for the the_hill Facebook page. Overall, variation in toxicity across these pages suggests that toxicity is not a constant feature of the Facebook platform, and the selection of sources from this platform should not systematically bias our earlier analysis.

To verify if the differences between toxicity scores observed in S5 Fig were statistically significant, we applied the tests explained in Section Methods for Hypothesis Investigation (the same one applied for datasets in Section Incivility and neutral bubble reachers, but changing the tested variables to be the Facebook sources toxicity scores presented in this figure). These Facebook sources were subjected to a normal test applying D’Agostino K-squared test [113]. Table 12 presents these test



S5 Fig. Box plots showing the distribution of toxicity scores for the Facebook Pages. The vertical gray line in this figure separates the Brazilian Portuguese (left) from the English pages (right).

results, showing that none of the Facebook sources followed a normal distribution ($p < .001$). We also relied on Q-Q plots and histogram visualizations for each Facebook source, confirming this result (plots were not included for brevity). Considering that the data were not following a normal distribution, we applied the Kolmogorov-Smirnov Goodness of Fit test [114] pairwise between Facebook sources to verify whether the samples originate from the same distribution. Results for this test were included in a separate spreadsheet for consultation ¹³, which shows that none of the Facebook sources originated from the same distribution ($p < .05$). Therefore, the observed differences in box plots are statistically significant.

S6 Appendix. Sensitivity Analysis - Facebook Influence on Incivility.

To ensure that our previous results are not biased by platform selection, we consider comments made in each news media on their Facebook account or website. In the ideal case, we could compare comments on the same articles on different platforms, but this data was unavailable. Therefore, we compare toxicity scores for a set of articles on the news organization's own website to another set of articles on Facebook. We were able to make this analysis for Yahoo News and the New York Times. To achieve this, we sourced the comments from each of these media outlets based on their origin (website or Facebook page). Concerning Yahoo News, the website comments were sourced from the YAHOO_SITE.en dataset, while comments from its corresponding Facebook page were extracted from the yahoo_news page within the FACEBOOK_OTHER.en dataset. In

Table 12. D’Agostino K-squared (k^2) tests for Facebook sources.

Dataset	Source	k^2
PARTISAN_pt	CampanhadoArmamento	2932.33***
	MisesBrasil	409.62***
	OPesadelodeQualquerPolitico2.0	27,061.59***
	carvalho.olavo	1,017.08***
	flaviobolsonaro	559.56***
	padrepaulo	1,299.81***
PARTISAN_REACHER_en	bloomberg_politics	174.65***
	huffington_post	1,626.50***
	ny_times	2,478.50***
FACEBOOK_NEUTRAL_en	abc_news	2,280.71***
	bbc	3,797.29***
	the_hill	2,072.24***
FACEBOOK_PARTISAN_en	breitbart	2,114.90***
	cnn	2,332.63***
	fox_news	2,048.13***
	msnbc	1,845.90***
	raw_story	3,887.21***
	the_blaze	1,464.03***
FACEBOOK_PERSON_en	person_bill_mahar	3,225.41***
	person_megyn_kelly	2,687.70***
	person_rachel_maddow	1,880.39***
FACEBOOK_OTHER_en	los_angeles_times	1,614.26***
	mother_jones	1,650.20***
	npr	4,354.89***
	yahoo_news	1,469.69***

Note: k^2 represent the D’Agostino K-squared [113] statistic. All statistics are significant at $p < .001$ (***) level.

the case of the New York Times, website comments were extracted from the NYT_SITE_en dataset, and comments from its Facebook page were extracted from the ny_times page within the PARTISAN_REACHER_en dataset.

A linear regression analysis was performed, considering toxicity scores as the dependent variable and the comment’s origin (website or Facebook page) as an independent variable. The results revealed statistically significant values for Pearson’s coefficients in both cases ($p < .001$), with $r = 0.144$ for Yahoo News and $r = 0.044$ for the New York Times. This result indicates that comments on Facebook had slightly higher toxicity scores than on the website for Yahoo News; in the case of the New York Times, there is no correlation. Thus, we have a suggestion that Facebook’s influence on toxicity results is perhaps more minimal than one might expect. The fact that we have different articles of the same news media on different platforms could help explain this minimum influence observed for Yahoo News.

Although this new analysis helps in understanding the influence of the platform where the comments were made, it holds a limitation due to the unavailability of data for directly comparing comments on the same articles across different platforms. Consequently, the comparison made between toxicity levels on these news organizations’ websites and their corresponding Facebook page might not fully capture the complete impact of platform-specific differences on toxicity. This is a topic worth exploring in future work.

Acknowledgments

We want to thank the authors who made their data available, enabling some analysis of this study. We also want to thank all the anonymous reviewers for their valuable feedback. This research was partially supported by the SocialNet project (process 2023/00148-0 of São Paulo State Research Support Foundation - FAPESP) and by the

References

1. Sharp K. *Socrates and the Second Person: The Craft of Platonic Dialogue*. University of Chicago, Committee on Social Thought; 2006.
2. Hackforth R. *Plato: Phaedrus*. 119. Cambridge University Press; 1972.
3. Shirky C. *Here Comes Everybody: The Power of Organizing Without Organizations*. Reprint edition ed. New York: Penguin Books; 2009.
4. Halpern D, Gibbs J. Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior*. 2013;29(3):1159–1168.
5. Dahlberg L. The Internet and Democratic Discourse: Exploring The Prospects of Online Deliberative Forums Extending the Public Sphere. *Information, Communication & Society*. 2001;4(4):615–633. doi:10.1080/13691180110097030.
6. Price V. Citizens Deliberating Online: Theory and Some Evidence. In: *Online Deliberation: Design, Research, and Practice*. Chicago, IL: University of Chicago Press; 2009. p. 37–58.
7. Pariser E. *The Filter Bubble: What the Internet is Hiding from You*. Penguin Books; 2012.
8. Sunstein CR. *#Republic: Divided Democracy in the Age of Social Media*. Illustrated edition ed. Princeton ; Oxford: Princeton University Press; 2017.
9. Cota W, Ferreira SC, Pastor-Satorras R, Starnini M. Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Science*. 2019;8(1):1–13. doi:10.1140/epjds/s13688-019-0213-9.
10. Garimella K, Morales GDF, Gionis A, Mathioudakis M. Reducing Controversy by Connecting Opposing Views; 2018. Available from: <http://arxiv.org/abs/1611.00172>.
11. Sunstein CR. *Echo Chambers: Bush V. Gore, Impeachment, and Beyond*. Princeton University Press; 2001. Available from: <https://books.google.com.br/books?id=sEgHAAAACAAJ>.
12. Pariser E. *The Filter Bubble: What The Internet Is Hiding From You*. UK: Penguin Books Limited; 2011.
13. Liu X, He D, Wu D. Breaking Social Media Bubbles for Information Globalization: A Cross-Cultural and Cross-Language User-Centered Sense-Making Approach. *Data and Information Management*. 2020;4(4):297–305. doi:10.2478/dim-2020-0020.
14. Spohr D. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*. 2017;34(3):150–160. doi:10.1177/0266382117722446.
15. Möller J. *Filter bubbles and digital echo chambers*. London: Routledge; 2021. Available from: <https://dare.uva.nl/search?identifier=0ae00d60-d14e-480f-9c43-16e04ccfeabd>.

16. Dubois E, Blank G. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society*. 2018;21(5):729–745. doi:10.1080/1369118X.2018.1428656.
17. De Francisci Morales G, Monti C, Starnini M. No echo in the chambers of political interactions on Reddit. *Scientific Reports*. 2021;11.
18. Dylko I, Dolgov I, Hoffman W, Eckhart N, Molina M, Aaziz O. Impact of Customizability Technology on Political Polarization. *Journal of Information Technology & Politics*. 2018;15(1):19–33.
19. Bail CA, Argyle LP, Brown TW, Bumpus JP, Chen H, Hunzaker MBF, et al. Exposure to opposing views on social media can increase political polarization. *PNAS*. 2018;115(37).
20. Kobellarz JK, Bročić M, Graeml AR, Silver D, Silva TH. Reaching the bubble may not be enough: news media role in online political polarization. *EPJ Data Science*. 2022;11(1):1–28. doi:10.1140/epjds/s13688-022-00357-3.
21. Gurri M. The revolt of the public and the crisis of authority in the new millennium. Second edition. ed. San Francisco, CA: Stripe Press; 2018.
22. Kobellarz JK, Brocic M, Silver D, Silva T. Bubble reachers and uncivil discourse in polarized online public sphere comments dataset; 2023. Available from: <https://doi.org/10.5281/zenodo.10443022>.
23. Maia RCM, Rezende TAS. Respect and Disrespect in Deliberation across the Networked Media Environment: Examining Multiple Paths of Political Talk. *Journal of Computer-Mediated Communication*. 2016;21(2):121–139. doi:10.1111/jcc4.12155.
24. Coe K, Kenski K, Rains SA. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication*. 2014;64(4):658–679. doi:10.1111/jcom.12104.
25. Muddiman A, Stroud NJ. News Values, Cognitive Biases, and Partisan Incivility in Comment Sections. *Journal of Communication*. 2017;67(4):586–609. doi:10.1111/jcom.12312.
26. Rossini P. Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk. *Communication Research*. 2022;49(3):399–425. doi:10.1177/0093650220921314.
27. Borah P. Does It Matter Where You Read the News Story? Interaction of Incivility and News Frames in the Political Blogosphere. *Communication Research*. 2014;41(6):809–827. doi:10.1177/0093650212449353.
28. Brooks DJ, Geer JG. Beyond Negativity: The Effects of Incivility on the Electorate. *American Journal of Political Science*. 2007;51(1):1–16. doi:10.1111/j.1540-5907.2007.00233.x.
29. Suhay E, Bello-Pardo E, Maurer B. The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments. *The International Journal of Press/Politics*. 2018;23(1):95–115. doi:10.1177/1940161217740697.
30. Anderson AA, Brossard D, Scheufele DA, Xenos MA, Ladwig P. The “Nasty Effect:” Online Incivility and Risk Perceptions of Emerging Technologies. *Journal of Computer-Mediated Communication*. 2014;19(3):373–387. doi:10.1111/jcc4.12009.

31. Kim JW, Guess A, Nyhan B, Reifler J. The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication*. 2021;71(6):922–946. doi:10.1093/joc/jqab034.
32. Yang J, Rojas H, Wojcieszak M, Aalberg T, Coen S, Curran J, et al. Why Are “Others” So Polarized? Perceived Political Polarization and Media Use in 10 Countries. *Journal of Computer-Mediated Communication*. 2016;21(5):349–367. doi:10.1111/jcc4.12166.
33. Garrett RK. Echo chambers online?: Politically motivated selective exposure among Internet news users1. *Journal of Computer-Mediated Communication*. 2009;14(2):265–285. doi:10.1111/j.1083-6101.2009.01440.x.
34. Iyengar S, Hahn KS. Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication*. 2009;59(1):19–39. doi:10.1111/j.1460-2466.2008.01402.x.
35. Nguyen CT. ECHO CHAMBERS AND EPISTEMIC BUBBLES. *Episteme*. 2020;17(2):141–161. doi:10.1017/epi.2018.32.
36. Bright J, Marchal N, Ganesh B, Rudinac S. Echo chambers exist!(But they’re full of opposing views). *arXiv preprint arXiv:200111461*. 2020;.
37. Rajadesingan A, Budak C, Resnick P. Political Discussion is Abundant in Non-political Subreddits (and Less Toxic). *Proceedings of the International AAAI Conference on Web and Social Media*. 2021;15:525–536. doi:10.1609/icwsm.v15i1.18081.
38. Gibson A. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society*. 2019;5(1):2056305119832588. doi:10.1177/2056305119832588.
39. Santana AD. Virtuous or Vitriolic. *Journalism Practice*. 2014;8(1):18–33. doi:10.1080/17512786.2013.813194.
40. Rowe I. Civility 2.0: a comparative analysis of incivility in online political discussion. *Information, Communication & Society*. 2015;18(2):121–138. doi:10.1080/1369118X.2014.940365.
41. Barberá P, Jost JT, Nagler J, Tucker JA, Bonneau R. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*. 2015;26(10):1531–1542.
42. Burt RS. *Structural holes: The social structure of competition*. Harvard university press; 1995.
43. Granovetter MS. The strength of weak ties. In: *Social networks*. Elsevier; 1977. p. 347–367.
44. Mukerjee S, Majó-Vázquez S, González-Bailón S. Networks of audience overlap in the consumption of digital news. *Journal of Communication*. 2018;68(1):26–50.
45. Magin M, Geiß S, Stark B, Jürgens P. Common Core in Danger? Personalized Information and the Fragmentation of the Public Agenda. *The International Journal of Press/Politics*. 2022;27(4):887–909. doi:10.1177/19401612211026595.

46. Christians CG, Glasser T, McQuail D, Nordenstreng K, White RA. Normative Theories of the Media: Journalism in Democratic Societies. University of Illinois Press; 2010.
47. Alexander JC. Civil sphere. S.l.: Oxford University Press; 2006. Available from: <http://myaccess.library.utoronto.ca/login?url=http://books.scholarsportal.info/viewdoc.html?id=/ebooks/ebooks0/oxford/2009-11-30/3/0195162501>.
48. Lippmann W. Public Opinion. First edition. ed. London: Taylor and Francis; 2017.
49. Parsons T, Platt GM. The American University:. Cambridge, MA: Harvard University Press; 1973.
50. Laclau E. On populist reason. Verso; 2005.
51. da Silva FC, Vieira MB. Populism and the politics of redemption. Thesis Eleven. 2018;149(1):10–30. doi:10.1177/0725513618813374.
52. Mudde C, Kaltwasser CR. Populism: A very short introduction. Oxford University Press; 2017.
53. Jansen RS. Situated political innovation: explaining the historical emergence of new modes of political practice. Theory and Society. 2016;45:319–360.
54. Brubaker R. Why populism? Theory and society. 2017;46:357–385.
55. Ostiguy P. 73Populism: A Socio-Cultural Approach. In: The Oxford Handbook of Populism. Oxford University Press; 2017. Available from: <https://doi.org/10.1093/oxfordhb/9780198803560.013.3>.
56. Silver D, Taylor Z, Calderón-Figueroa F. Populism in the city: The case of Ford Nation. International Journal of Politics, Culture, and Society. 2020;33:1–21.
57. Alexander JC, Kivisto P, Sciortino G. Populism in the Civil Sphere. John Wiley & Sons; 2020.
58. Schulz A, Wirth W, Müller P. We are the people and you are fake news: A social identity approach to populist citizens' false consensus and hostile media perceptions. Communication Research. 2020;47:201–226. doi:10.1177/0093650218794854.
59. Mudde C, editor. The Populist Radical Right: A Reader. 1st ed. London ; New York: Routledge; 2016.
60. Hahl O, Kim M, Sivan EWZ. The authentic appeal of the lying demagogue: Proclaiming the deeper truth about political illegitimacy. American Sociological Review. 2018;83:1–33. doi:10.1177/0003122417749632.
61. DellaPosta D. Pluralistic Collapse: The “Oil Spill” Model of Mass Opinion Polarization. Amer Soc Rev. 2020;85(3):507–536.
62. Rawlings C, Childress C. The Polarization of Popular Culture: Tracing the Size, Shape, and Depth of the Oil Spill. SocArXiv. 2022;.
63. Salminen J, Sengün S, Corporan J, Jung Sg, Jansen BJ. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. PLOS ONE. 2020;15(2):1–24. doi:10.1371/journal.pone.0228723.

64. Lipset SM. *Continental Divide: The Values and Institutions of the United States and Canada*. 1st ed. New York: Routledge; 1990.
65. Dalton RJ. *An Undivided Continent ? Re-examining the Political Culture of Canadians and Americans*; 2016. Available from: <https://sites.socsci.uci.edu/~rdalton/archive/canada2016.pdf>.
66. Garzia D, Ferreira da Silva F, Maye S. Affective Polarization in Comparative and Longitudinal Perspective. *Public Opinion Quarterly*. 2023; p. nfad004. doi:10.1093/poq/nfad004.
67. Thomas DD, Sands C. *Canada and the United States: Differences That Count, Fifth Edition (Paperback)*. 5th ed. University of Toronto Press; 2023. Available from: <https://www.libertybaybooks.com/book/9781487544225>.
68. Orhan YE. The relationship between affective polarization and democratic backsliding: comparative evidence. *Democratization*. 2022;29(4):714–735. doi:10.1080/13510347.2021.2008912.
69. Van Der Veen O. *Political Polarisation Compared: Creating a Comprehensive Index of Political Polarisation [PhD Thesis]*. Doctoral dissertation, Central European University; 2021.
70. Phillips D, Burke J, Lewis P. How Brazil and South Africa became the world's most populist countries. *The Guardian*. 2019;.
71. Silver D, Taylor Z, Calderón-Figueroa F. Populism in the City: the Case of Ford Nation. *International Journal of Politics, Culture, and Society*. 2020;33(1):1–21. doi:10.1007/s10767-018-9310-1.
72. Araújo B, Prior H. Framing Political Populism: The Role of Media in Framing the Election of Jair Bolsonaro. *Journalism Practice*. 2021;15(2):226–242. doi:10.1080/17512786.2019.1709881.
73. Dourado T, Salgado S. Disinformation in the Brazilian pre-election context: probing the content, spread and implications of fake news about Lula da Silva. *The Communication Review*. 2021;24(4):297–319. doi:10.1080/10714421.2021.1981705.
74. Vale HFd. *Brazil's Digital Politics and the Crisis of Democracy (2013-2018)*. IntechOpen; 2022. Available from: <https://www.intechopen.com/online-first/83475>.
75. Kemp S. *Digital 2023: BRAZIL*; 2023. Available from: <https://datareportal.com/reports/digital-2023-brazil>.
76. Kemp S. *Digital 2023: CANADA*; 2023. Available from: <https://datareportal.com/reports/digital-2023-canada>.
77. Kenski K, Coe K, Rains SA. Perceptions of uncivil discourse online: An examination of types and predictors. *Communication Research*. 2020;47(6):795–814.
78. Jigsaw G. *Perspective API*; 2022. Available from: <https://perspectiveapi.com>.
79. Kobellarz JK, Silva TH. Should we translate? evaluating toxicity in online comments when translating from portuguese to english. In: *Brazilian Symposium on Multimedia and Web*; 2022. p. 89–98.

80. Hatebase. Hatebase; 2022. Available from: <https://hatebase.org>.
81. Warner W, Hirschberg J. Detecting Hate Speech on the World Wide Web. In: Proceedings of the Second Workshop on Language in Social Media. Montréal, Canada: Association for Computational Linguistics; 2012. p. 19–26.
82. Yin D, Xue Z, Hong L, Davison BD, Kontostathis A, Edwards L. Detection of harassment on web 2.0. In: Proceedings of the Content Analysis in the WEB. vol. 2. Madrid, Spain; 2009. p. 1–7.
83. Park JH, Fung P. One-step and Two-step Classification for Abusive Language Detection on Twitter. In: Proceedings of the First Workshop on Abusive Language Online. Vancouver, BC, Canada: Association for Computational Linguistics; 2017. p. 41–45. Available from: <https://aclanthology.org/W17-3006>.
84. Srivastava S, Khurana P, Tewari V. Identifying Aggression and Toxicity in Comments using Capsule Network. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). Santa Fe, New Mexico, USA: Association for Computational Linguistics; 2018. p. 98–105.
85. Georgakopoulos SV, Tasoulis SK, Vrahatis AG, Plagianakos VP. Convolutional Neural Networks for Toxic Comment Classification. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence. SETN '18. New York, NY, USA: Association for Computing Machinery; 2018.
86. Almerexhi H, Kwak H, Salminen J, Jansen BJ. In: Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions. New York, NY, USA: Association for Computing Machinery; 2020. p. 3033–3040.
87. Kumar S, Hamilton WL, Leskovec J, Jurafsky D. Community Interaction and Conflict on the Web. In: Proceedings of the 2018 World Wide Web Conference. WWW '18. Republic and Canton of Geneva, CHE; 2018. p. 933–943.
88. Fortuna P, Soler J, Wanner L. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2020. p. 6786–6794.
89. Jain E, Brown S, Chen J, Neaton E, Baidas M, Dong Z, et al. Adversarial Text Generation for Google's Perspective API. 2018 International Conference on Computational Science and Computational Intelligence (CSCI). 2018; p. 1136–1141.
90. Lees A, Tran VQ, Tay Y, Sorensen J, Gupta J, Metzler D, et al. A new generation of perspective api: Efficient multilingual character-level transformers. arXiv preprint arXiv:2202.11176. 2022;.
91. Araújo M, Pereira A, Benevenuto F. A comparative study of machine translation for multilingual sentence-level sentiment analysis. Information Sciences. 2020;512:1078–1102.
92. Pereira DA. A survey of sentiment analysis in the Portuguese language. Artificial Intelligence Review. 2021;54(2):1087–1115.

93. Rieder B, Skop Y. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society*. 2021;8(2).
94. Rajadesingan A, Resnick P, Budak C. Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*. 2020;14(1):557–568. doi:10.1609/icwsm.v14i1.7323.
95. Zannettou S, Elsherief M, Belding E, Nilizadeh S, Stringhini G. Measuring and Characterizing Hate Speech on News Websites. In: *Proceedings of the 12th ACM Conference on Web Science*. WebSci '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 125–134. Available from: <https://doi.org/10.1145/3394231.3397902>.
96. Hua Y, Ristenpart T, Naaman M. Towards measuring adversarial twitter interactions against candidates in the us midterm elections. In: *Proceedings of the international AAAI conference on web and social media*. vol. 14; 2020. p. 272–282.
97. Cochrane C. *Left and right: The small world of political ideas*. vol. 3. Canada: McGill-Queen's Press-MQUP; 2015.
98. Jensen P, Morini M, Karsai M, Venturini T, Vespignani A, Jacomy M, et al. Detecting global bridges in networks. *Journal of Complex Networks*. 2016;4(3):319–329.
99. Brandes U. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*. 2001;25(2):163–177.
100. Bencina J. Facebook News Dataset; 2017. Available from: <https://github.com/jbencina/facebook-news>.
101. Doherty C. 7 things to know about polarization in America; 2014. Available from: <https://www.pewresearch.org/short-reads/2014/06/12/7-things-to-know-about-polarization-in-america>.
102. Fiorina MP, Abrams SJ. Political Polarization in the American Public. *Annual Revof Pol Sci*. 2008;11(1):563–588.
103. Lerner C. Brazilian Facebook Public Pages - Posts and Comments: a research on conservative mentality; 2019. Available from: <https://doi.org/10.5281/zenodo.3226970>.
104. Leite JA, Silva DF, Bontcheva K, Scarton C. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis; 2020. Available from: <https://arxiv.org/abs/2010.04543>.
105. Real L, Oshiro M, Mafra A. B2W-Reviews01-An open product reviews corpus. In: *the Proceedings of the XII Symposium in Information and Human Language Technology*; 2019. p. 200–208.
106. Velho EG, Montardo SP. O que os textos dizem sobre Jair Bolsonaro? Desenvolvimento de uma rede de coocorrência lexical dos comentários de um portal de notÁcias. In: *ABCIBER XIII-Simpósio Nacional da ABCIBER 2020*; 2021.

107. Kobellarz JK, Bročić M, Graeml AR, Silver D, Silva TH. Reaching the bubble may not be enough: news media role in online political polarization. *EPJ Data Science*. 2022;11(1):47.
108. Barberá P, Rivero G. Understanding the political representativeness of Twitter users. *Social Science Computer Review*. 2015;33(6):712–729.
109. Ott BL. The age of Twitter: Donald J. Trump and the politics of debasement. *Critical studies in media communication*. 2017;34(1):59–68.
110. Conover M, Ratkiewicz J, Francisco M. Political polarization on twitter. In: *Proc. of ICWSM*. Barcelona, Spain; 2011. p. 89–96.
111. Morales AJ, Borondo J, Losada JC, Benito RM. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos*. 2015;25(3).
112. Sap M, Card D, Gabriel S, Choi Y, Smith NA. The risk of racial bias in hate speech detection. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*; 2019. p. 1668–1678.
113. d’Agostino RB. An omnibus test of normality for moderate and large size samples. *Biometrika*. 1971;58(2):341–348.
114. Massey Jr FJ. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*. 1951;46(253):68–78.
115. Oliveira LSd, Amaral MS, Vaz-de Melo PO. Long-term Characterization of Political Communications on Social Media. In: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*; 2021. p. 95–102.
116. Herbst S. *Rude democracy: civility and incivility in American politics*. Philadelphia: Temple University Press; 2010. Available from: <http://site.ebrary.com/id/10408649>.
117. Ziegele M, Jost P, Springer N, Wright S. Online user comments across news and other content formats: Multidisciplinary perspectives, new directions. *Studies in Communication and Media*. 2017;6(4):315–332. doi:10.5771/2192-4007-2017-4-315.
118. Salminen J, Sengün S, Corporan J, Jung Sg, Jansen BJ. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLOS ONE*. 2020;15(2):e0228723. doi:10.1371/journal.pone.0228723.
119. Otto LP, Lecheler S, Schuck ART. Is Context the Key? The (Non-)Differential Effects of Mediated Incivility in Three European Countries. *Political Communication*. 2020;37(1):88–107. doi:10.1080/10584609.2019.1663324.
120. Post S. Incivility in Controversies: The Influence of Presumed Media Influence and Perceived Media Hostility on the Antagonists in the German Conflict Over Aircraft Noise. *Communication Research*. 2017;44(8):1149–1175. doi:10.1177/0093650215600491.
121. Post S. Polarizing Communication as Media Effects on Antagonists. *Understanding Communication in Conflicts in Digital Media Societies*. *Communication Theory*. 2019;29(2):213–235. doi:10.1093/ct/qty022.
122. Prior M. *Media and Political Polarization*; 2013.

123. Kobellarz JK, Graeml A, Reddy M, Silva TH. Parrot Talk: Retweeting Among Twitter Users During the 2018 Brazilian Presidential Election. In: Proc. of ACM Webmedia. Rio de Janeiro: SBC; 2019. p. 221–228.
124. Lizardo O. Improving Cultural Analysis: Considering Personal Culture in its Declarative and Nondeclarative Modes. *American Sociological Review*. 2017;82(1):88–115. doi:10.1177/0003122416675175.
125. Druckman JN, Klar S, Krupnikov Y, Levendusky M, Ryan JB. Affective polarization, local contexts and public opinion in America. *Nature human behaviour*. 2021;5(1):28–38.
126. Yang Z, Xu C, Wu W, Li Z. Read, Attend and Comment: A Deep Architecture for Automatic News Comment Generation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 5076–5088.
127. Hosseini H, Kannan S, Zhang B, Poovendran R. Deceiving google’s perspective api built for detecting toxic comments. *arXiv preprint arXiv:170208138*. 2017;.
128. Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates. 2001;71.
129. Balage Filho PP, Pardo TAS, Aluisio SM. An evaluation of the Brazilian Portuguese LIWC Dictionary for sentiment analysis. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology. Fortaleza, CE, Brazil: Sociedade Brasileira de Computação; 2013. p. 215–219.