Mapping Hate and Extremism: How Brazilian Reddit Communities Spread Toxic Discourse

Pedro Bento, Yan Aquino, Arthur Buzelin, Pedro Robles Dutenhefner, Arthur Chagas, Victoria Estanislau, Lucas Dayrell, Samira Malaquias, Caio Santana, Marcelo Sartori Locatelli, Gisele L. Pappa, Wagner Meira Jr., Virgilio Almeida

Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil

{pedro.bento, yanaquino, arthurbuzelin, arthurchagas, victoria.estanislau, lucasdayrell, samiramalaquias, caiosantana, locatellimarcelo, virgilio, glpappa, meira}@dcc.ufmg.br, pedroroblesduten@ufmg.br

Abstract

Online hate speech endangers minoritized groups and has been linked to offline harassment and violence (Müller and Schwarz 2020). Yet the mechanisms by which such toxic discourse spreads and evolves remain context-dependent. In this work, we focus on Brazilian Reddit communities to examine these dynamics at scale, focusing on both banned and non-banned subreddits. Leveraging over 20 million posts and comments, we employ two complementary approaches: Word Embedding Association Test (WEAT) to detect implicit semantic biases, and the Pysentimiento transformer model to classify hate speech in Portuguese-language content. Our results reveal that banned subreddits exhibit strong positive associations with hate-related terminology – particularly sexism – indicating a normalization of discriminatory language. In contrast, non-banned communities show broader but equally concerning patterns of prejudice across multiple hate categories, such as racism, body shaming, and ideological hostility. Additionally, we find evidence of a significant propagation effect: submissions containing hate speech substantially increase toxicity in subsequent comment threads. These findings highlight not only the diverse forms of extremism manifesting within Brazil's Reddit ecosystem, but also critical inconsistencies in content moderation practices. Our work contributes to a deeper, culturally contextualized understanding of online radicalization and offers scalable methodologies for monitoring and mitigating toxic discourse on digital platforms.

1 Introduction

The rise of online platforms has significantly changed how communities form and interact, allowing people from different locations and cultures to connect based on shared interests, ideas, and identities. These online spaces offer new opportunities for conversation, collaboration, and the growth of subcultures, but they also create new challenges for understanding how these digital environments influence collective behaviors and beliefs. Among these platforms, Reddit is particularly useful for studying community interactions. Its structure, consisting of thousands of subreddits – each acting as semi-independent discussion spaces – gives researchers a unique chance to see how language, norms, and ideas develop within separate online groups.

However, the same decentralized structure that makes Reddit a valuable environment for observing community-specific norms also facilitates the formation of harmful and ideologically extreme groups. The platform's model, where subreddits operate with relative autonomy, often under minimal moderation, can create fertile ground for the emergence and reinforcement of toxic subcultures. Within these spaces, repeated interactions and the continuous sharing of radical content can foster echo chambers (Efstratiou et al. 2023), creating ideal conditions for the emergence and growth of extremist fringe communities. Recently, researchers have begun paying closer attention to these groups to understand how online interactions contribute to radicalization and the spread of hateful content (Russo, Horta Ribeiro, and West 2024).

While these patterns have been well documented in regions like North America and Europe, a significant gap remains in investigations into how such phenomena manifest in other parts of the world. This geographical bias in the literature limits our understanding of platform-mediated radicalization within diverse and yet unexplored sociopolitical ecosystems, where distinct cultural and historical conditions may shape unique radicalization pathways. The absence of broader, cross-regional investigations restricts the development of comprehensive theoretical frameworks and evidence-based interventions for countering online extremism, undermining the effectiveness of platform governance and content moderation strategies on a global scale.

This paper narrows its focus to a specific and concerning subset of online extremism: Brazilian fringe communities active within Reddit. The urgent real-world significance of this study is underscored by recent cases in Brazil where fringe communities on platforms like Discord have demonstrated how social media interactions can escalate into serious criminal activities, specifically related to misogynistic behavior and violence against women, resulting in multiple arrests linked directly to online activity [1]. Despite these alarming developments, targeted research addressing Brazilian extremist communities, particularly on Reddit, remains limited. Reddit's Brazilian user base presents a unique context to explore the dynamics through which extremist narratives propagate and evolve within online communities. An in-depth investigation of these networks could shed light on

[1] https://g1.globo.com/fantastico/noticia/2023/06/25/rede-sem-lei-no-discord-criminosos-violentam-e-humilham-meninas-menores-de-idade.ghtml

underlying processes driving the dissemination and transformation of extremist content, ultimately informing broader efforts aimed at detecting, comprehending, and mitigating online radicalization.

Specifically, this research aims to: (1) uncover implicit linguistic biases and analyze the semantic associations linked to offensive and hate-related terms across different subreddit communities, (2) analyze the prevalence and characteristics of hate speech within banned versus non-banned subreddits, and (3) examine the propagation and amplification effects of hate speech in subreddit submissions and subsequent comments.

Our results reveal distinct patterns of semantic bias and toxic discourse propagation within these communities. To uncover these patterns, we conducted a semantic analysis using word embeddings to evaluate the implicit associations between subreddit language and hate-related terminology, focusing particularly on the proximity of offensive terms to positive or negative semantic anchors. This analysis revealed strong positive associations between banned subreddits and hate-related language, especially involving sexist content. For the detection and characterization of hate speech, we employed a transformer-based classifier to label individual messages, enabling us to compare the prevalence and intensity of toxic discourse across banned and non-banned communities. Furthermore, by analyzing both original submissions and their corresponding comment threads, we observed that posts containing hate speech significantly increase the likelihood of subsequent toxic interactions, producing cascading effects that amplify extremist narratives. These findings expose critical limitations in current content moderation strategies and highlight the need to address the diverse and evolving forms of online extremism in platform-specific and culturally contextualized ways.

2 Related Work

This section reviews related work across three key areas that inform our study: (1) the formation and dynamics of online extremist communities, (2) computational approaches for analyzing social media discourse, and (3) methods for detecting bias and hate speech in online environments.

2.1 Online Extremist Communities

A substantial body of work has examined the formation, dynamics, and evolution of online extremist communities. These studies explore how these communities use online platforms to propagate their ideologies, recruit new members, and coordinate activities (Winter et al. 2021; Youngblood 2020).

Research has shown that social media platforms (eg. Twitter, Telegram, Reddit) have been identified as fertile ground for extremist narratives due to their algorithm-driven content promotion and echo-chamber effects (Conway, Scrivens, and Macnair 2019; Ribeiro et al. 2020). Specifically, Reddit has been examined as a platform that facilitates extremist discourse (Gaudette et al. 2021). Studies have found that hate-driven and extremist subreddits create tightly-knit communities where radicalization processes unfold through repeated exposure to inflammatory content and social reinforcement (Schmitz et al. 2024; Gaudette, Scrivens, and Venkatesh 2020).

More specifically in the Brazilian Reddit context, a study conducted on internet memes in the r/brasilivre community (one of our leading study subjects) characterized the subreddit as a "far-right masculinist network that is divided into subgroups that hierarchize masculinities", diving into the concepts of how the "incel" terminology is applied in these contexts. These communities not only serve as spaces for ideological dissemination but also act as gateways to more extreme forms of engagement, including real-world mobilization (Mitts 2019).

Furthermore, studies have explored the characteristics and motivations of individuals who participate in these communities, as well as the potential for online radicalization (Schmitz et al. 2024). The growing presence of extremist communities online underscores the urgent need for platform regulation, automated content moderation, and counter-narrative strategies aimed at mitigating the spread of extremist ideologies in digital spaces (Scrivens, Amarasingam, and Hirschfield 2022).

2.2 Computational Analysis of Social Media

Computational methods have become increasingly important for analyzing social media data. Natural language processing (NLP) techniques, such as transformer-based models (e.g., BERT, RoBERTa), have revolutionized the analysis of online discourse by capturing nuanced semantic and syntactic patterns in user-generated text (Devlin et al. 2019; Liu et al. 2019).

Techniques like word embeddings and semantic analysis, have been widely used to understand the meaning and usage of language in online contexts (Matsui and Ferrara 2024; Çano and Morisio 2019; Bayrakdar et al. 2020).

Sentiment analysis has advanced beyond lexicon-based approaches, with multimodal frameworks integrating text, images, and metadata to detect hateful memes and contextual hostility (Kiela et al. 2021; Barbieri et al. 2020). Topic modeling has been used to identify dominant themes and topics of discussion (Zhu et al. 2024; Göçen, Ibrahim, and Khan 2024; Laureate, Buntine, and Linger 2023).

Neural architectures for topic modeling like BERTopic and Top2Vec identify latent themes in large-scale Reddit datasets, distinguishing between overt hate speech (e.g., racial slurs) and subtler ideological radicalization (Grootendorst 2022; Angelov 2020). These methods provide valuable tools for studying online communities and their discourse.

2.3 Bias and Hate Speech Detection

The detection and analysis of bias and hate speech in online communication is a critical area of research. Models such as Pysentimiento have been developed to identify and classify hate speech in Portuguese text posts (Pérez et al. 2023). The Word Embedding Association Test (WEAT) remains a valuable tool for measuring implicit biases embedded in word embeddings, offering insights into potential biases in language use (Caliskan, Bryson, and Narayanan 2017).

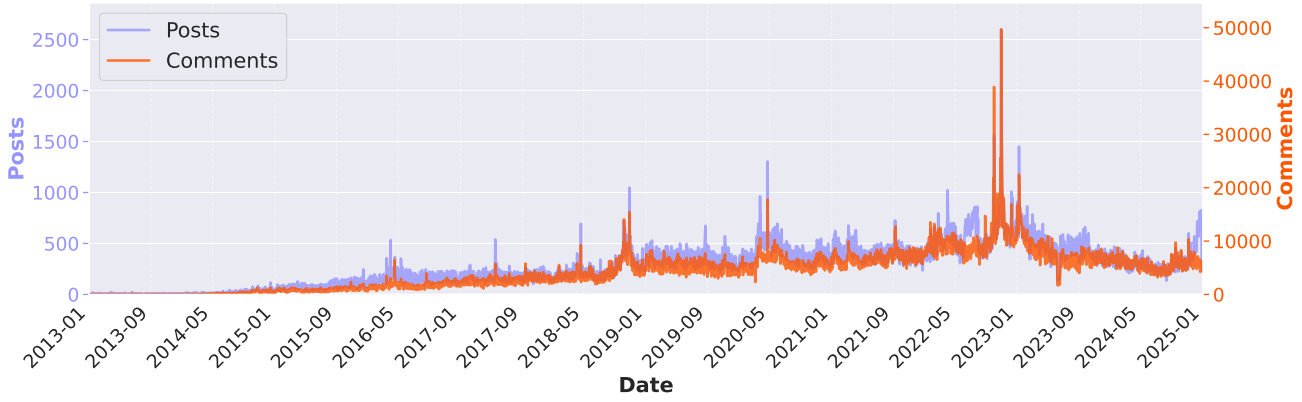

Figure 1: Monthly activity on Brazilian subreddits, showing the total number of posts and comments over time.

Recent research has increasingly focused on leveraging transformer-based models for improved hate speech detection in online platforms. These models have shown promising results in capturing contextual nuances crucial for identifying subtle forms of hate speech (Bender 2021; Farooqi, Ghosh, and Shah 2021). For instance, fine-tuned BERT models have outperformed traditional recurrent neural networks like SimpleRNN and LSTM in detecting hate speech in social media comments, achieving higher accuracy (Bender 2021).

Furthermore, several datasets specifically curated for hate speech detection on platforms like Reddit are available to facilitate researching this area. The ETHOS dataset, built from YouTube and Reddit comments, offers both binary and multi-label classification options for hate speech (Mollas et al. 2020). Similarly, the Ruddit dataset provides fine-grained, real-value offensiveness scores for English Reddit comments, allowing for a more nuanced analysis of online toxicity (Hada et al. 2021).

Research in this area is essential for understanding the harmful content present in online communities and developing strategies for mitigation.

3 Methodology

This section outlines the analytical approaches used to examine semantic associations and detect hate speech within Brazilian Reddit communities. We begin with data collection, describing the process of selecting relevant subreddits and assembling a large-scale dataset of posts and comments. We then detail our word embeddings analysis, where we apply the Word Embedding Association Test (WEAT) to uncover implicit semantic biases across communities and over time. Finally, we present our hate speech analysis, which leverages a transformer-based classifier to detect and categorize hate speech in both submissions and comments, enabling a comprehensive evaluation of toxic discourse and its propagation dynamics.

3.1 Data Collection

We collected Reddit data from an existing dataset available on Academic Torrents (Watchful1 2025), a historical aggregate of the top 40,000 subreddits from Reddit's history, retrieved via the Pushshift API (Baumgartner et al. 2020). From this dataset, we extracted data specifically from Brazilian Portuguese-speaking subreddits and user comments related to fringe communities. The selection of subreddits was guided by the goal of capturing a broad range of ideological positions and behavioral patterns associated with toxic discourse. To this end, we included subreddits representing opposite ends of the political spectrum that have been previously identified as polarized in the literature (Cunha and da Silva 2024; Vilaça 2022): *r/brasildob*, associated with far-left discourse, and *r/brasillivre*, aligned with far-right ideologies. In addition, we incorporated fringe subreddits that were permanently banned by the platform due to violations of Reddit's content policies (*r/brasil_liberal*, *r/brasil_drama*, *r/brasilimpo*, and *r/cornosbrasil*), which allow us to examine patterns of hate speech in communities that exceeded platform moderation thresholds.

We include the subreddit *r/brasil* as a baseline non-fringe community for comparative analysis. This subreddit was selected due to its large and diverse user base, its generalist nature, and its central position in the Brazilian Reddit ecosystem. Unlike the other subreddits analyzed – which are either ideologically polarized or have been banned for violating Reddit's content policies – *r/brasil* functions as a broad public forum for national discussions, encompassing a wide range of topics, from politics and news to entertainment and daily life. As such, it provides a representative linguistic and cultural baseline against which we can measure deviations in discourse patterns and the presence of hate-related content in fringe communities.

The final dataset includes the text of posts and comments, timestamps, and available user information. Figure 1 shows how posts and comments are distributed over time, while Table 1 summarizes the total volume of activity per subreddit.

Subreddit	Posts	Comments	Total
r/brasil	764,620	13,347,465	14,112,085
r/BrasildoB	59,632	430,166	489,798
r/brasil_drama	3,300	39,769	43,069
r/brasil_liberal	14,053	19,953	34,006
r/brasilimpo	22,739	5,460	28,199
r/brasilivre	326,826	5,154,242	5,481,068
r/cornosbrasil	17,133	38,138	55,271
Total	**1,236,041**	**19,058,641**	**20,294,682**

Table 1: Reddit Activity Metrics: Posts and Comments by Subreddit

3.2 Word Embeddings Analysis

One way to understand political discourse on Reddit is by examining implicit biases and the semantic contexts of key terms across different ideological communities. This can be effectively achieved through word embeddings, such as those generated by the Word2Vec model (Mikolov et al. 2013a,b). These embeddings map words into a high-dimensional vector space, where semantically similar terms are positioned closer together, capturing contextual relationships based on how language is used within and across subreddit communities.

To investigate these biases, we trained separate Word2Vec models for each selected subreddit (Mikolov et al. 2013b). This approach enables us to analyze how key terms are represented within each context, revealing shifts in their connotations and associations across different ideologies and fringe communities over time. However, segmenting the dataset limits the volume of data available for training each individual model, which may impact the quality of the resulting embeddings. To mitigate this issue, we initialized the training with a neutral *Word2Vec* model, pre-trained on a diverse corpus that includes Wikipedia articles, OpenCrawl data, and movie subtitles (Speer, Chin, and Havasi 2017).

To assess the semantic framing of hate-related terms (e.g., *feminista*, *militante*) within fringe subreddit communities, we apply the Word Embedding Association Test (WEAT) (Caliskan, Bryson, and Narayanan 2017). This method measures implicit biases in language by quantifying how strongly a target term t is associated with sets of positive (A_+, e.g., *bom* [good], *importante* [important]) and negative (A_-, e.g., *ruim* [bad] *terrível* [terrible]) reference words. The calculation is based on cosine similarity in the word embedding space. For a given target term t, the normalized association score $S(t, A_+, A_-)$ is defined as:

$$S(t, A_+, A_-) = \frac{\text{mean}(\text{sim}(t, A_+)) - \text{mean}(\text{sim}(t, A_-))}{\text{std}(\text{sim}(t, A_+ \cup A_-))}$$

where $\text{sim}(t, a)$ denotes the cosine similarity between the embedding of the target term t and an attribute word $a \in A_+ \cup A_-$. A positive value of $S(t, A_+, A_-)$ indicates a stronger semantic alignment of t with positive attributes, while a negative value suggests a closer association with negative terms. This methodology enables a systematic analysis of implicit linguistic biases embedded in the discourse of fringe subreddit communities.

To systematically explore specific hate speech categories within subreddit discussions, we utilized a predefined set of target concepts. The words were extracted from the Hatebase[2] dataset in Portuguese, a comprehensive lexicon of hate speech terms. Subsequently, they were manually classified into 12 concept classes: (1) *incel terminology*, (2) *relationship concepts*, (3) *negative emotions*, (4) *racial/ethnic slurs*, (5) *gender-based slurs*, (6) *LGBTQ+ slurs*, (7) *religious slurs*, (8) *disability and health-related slurs*, (9) *political slurs*, (10) *body and appearance-related slurs*, (11) *mental health-related terms*, and (12) *violence-related terms*.

This methodology follows a growing body of work that uses word embeddings and association tests to study online discourse and detect implicit biases in language (Ottoni et al. 2018; Buzelin et al. 2025).

3.3 Hate Speech Analysis

To systematically analyze hate speech within Brazilian subreddit communities, we utilized the transformer-based language model Pysentimiento (Pérez et al. 2023), which has been widely adopted in prior research for detecting hate speech in Portuguese-language content (do Santos et al. 2024; Buzelin et al. 2024a).

The model was initially pre-trained on millions of Portuguese-language tweets and subsequently fine-tuned on a curated dataset of 6,000 annotated tweets, labeled across six distinct hate speech categories: *sexism, racism, homophobia, ideology, body*, and *religion*.

Each Reddit post was classified into specific hate speech categories: *sexism, racism, homophobia, ideology, and body*, enabling detailed insights into prejudiced discourse within each community. We selected *Pysentimiento* over alternatives such as Perspective API (Lees et al. 2022) due to its superior performance on Portuguese-language datasets (Buzelin et al. 2024b).

To evaluate the effectiveness of the model, we randomly selected 100 messages, proportionally across communities, that had been classified by it, 50 labeled as hate speech and 50 as non-hate speech. Three undergraduate computer science students independently annotated these messages according to the guidelines outlined in the model's original paper. The inter-annotator agreement reached a Fleiss' kappa score of 0.87, and the model achieved an accuracy of 83%, indicating that it performs reliably and is suitable for our purposes.

We applied the classifier to the entire dataset of posts and comments from all selected subreddits. This application enabled us to quantify the prevalence of hate speech across different communities and interaction types (i.e., submissions versus replies), as well as to conduct comparative and temporal analyses of toxic discourse within and between fringe and non-fringe environments.

[2]https://hatebase.org/

4 Results

This section presents the results of our analysis and how they address the study's main objectives. We first examine semantic biases using word embeddings to reveal how hate-related terms are differently framed across communities (§4.1). We then analyze the distribution of hate speech categories in banned and non-banned subreddits (§4.2), followed by an investigation of how hate speech in submissions amplifies toxicity in subsequent comments (§4.3). These findings collectively highlight how fringe communities on Reddit express and reinforce hateful discourse.

4.1 Embedding Analysis

After training independent Word2Vec models for each selected subreddit, we applied the Word Embedding Association Test (WEAT) to quantify the implicit semantic associations of hate-related terms. The target terms were grouped into categories derived from the Hatebase 2.0 dataset, covering various forms of hate speech such as sexism, racism, homophobia, and political slurs. For each term category, we computed a WEAT score indicating its relative alignment with positive or negative reference terms.

Figure 2 presents the resulting association scores for each hate category across the subreddits. A positive score (closer to +1) indicates that the hate-related terms are semantically closer to positive attributes, while a negative score (closer to -1) indicates stronger proximity to negative attributes. A positive score would suggest that the term or slur tends to be seen more favorably in the community. This visualization allows for a comparative interpretation of how each community frames hate-related language in semantic space.

Baseline Community: *r/brasil* A few interesting observations emerge from this analysis. First, we notice that the hate-related terms appear significantly more negative in our baseline community, *r/brasil*. Across all categories, the average association scores in this subreddit fall below -0.5. This outcome is expected and encouraging. Hate-related terms associated with sensitive topics (e.g., homophobia, sexism) are strongly marked as negative, suggesting that users are aware of their harmful connotations and may be cautious about using them.

Extreme Fringe Communities With this in mind, we now turn our attention to the fringe communities. Two stand out immediately: *r/cornosbrasil* and *r/brasil_liberal*. The former is a sexually explicit subreddit centered on cuckold pornography and laced with misogynistic humor, while the latter focuses on libertarian and far-right political commentary in Brazil.

These distinct orientations help contextualize their WEAT positions: both communities are clearly located on the far right of Figure 2, indicating that, within them, hate-related terms carry notably positive connotations. This is extremely worrying, as it suggests a normalization, or even valorization, of discriminatory language. In these spaces, hate terms and other forms of hate speech are not only accepted but may even be embraced.

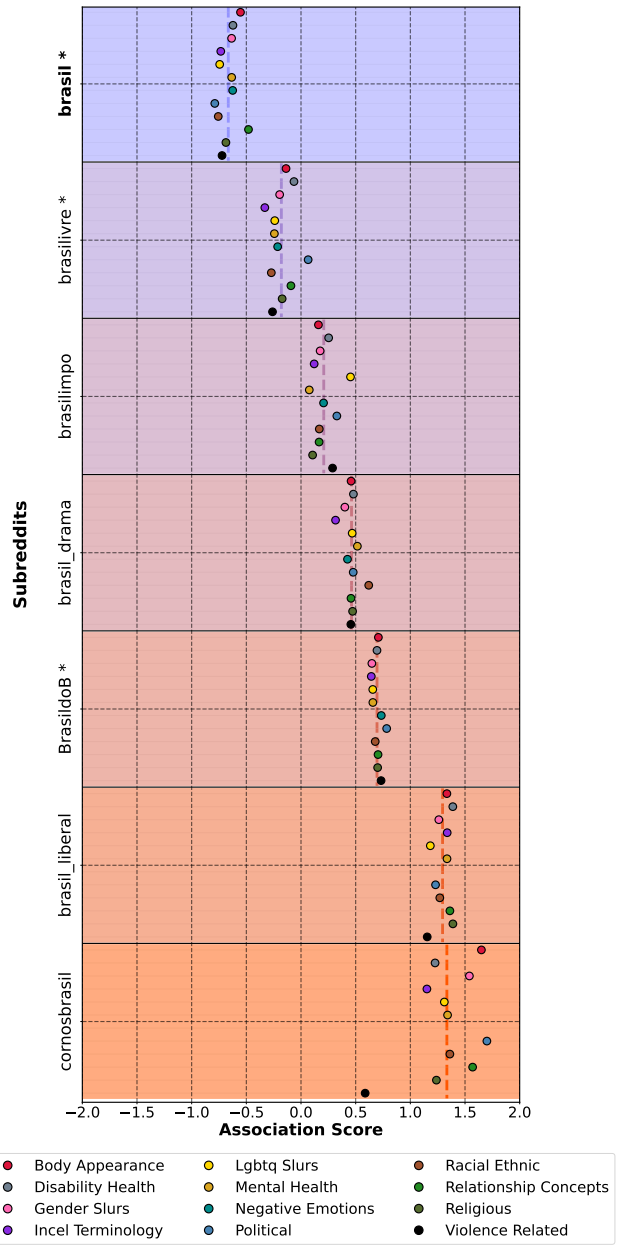

Figure 2: The graph presents the association scores for various categories across different subreddits. Each subreddit is represented by a separate row, and the dots indicate the association scores for each category. The vertical lines highlight distinct score averages for each subreddit, revealing how association varies between them. Subreddits marked with an asterisk (*) are non-banned communities.

These communities exhibited rhetoric laden with sexualized objectification and dehumanizing tropes, patterns commonly found in far-right and male-dominated online spaces (Weiss, Koban, and Matthes 2024). By intertwining misogynistic narratives with extremist ideologies, such platforms cultivate environments where harmful language is not only

normalized but actively encouraged.

Interestingly, despite *r/cornosbrasil*'s overt focus on sex and gender dynamics, the most extreme associations are not found in gendered slurs or relationship-related terms. Instead, the highest positive scores appear in categories related to body appearance and political offense, going beyond the community's pornographic facade and exposing a broader ideological hostility. The weaponization of physical appearance and political identity in this context underscores the community's embrace of hate not merely as a joke or fetish, but as a core part of its discourse.

It is also worth noting that, while these communities appear to be particularly harmful, the total number of posts within them is significantly lower compared to the other communities. This aligns with the idea that echo chambers (Cinelli et al. 2021) tend to attract and concentrate the most extreme voices in smaller, more insular spaces.

On the opposite end, we turn to *r/brasil_livre*. Among all the fringe subreddits, it has by far the highest number of posts. Supporting the trends observed earlier, this subreddit – while still positioned to the right of the baseline – shows that most hate-related terms, on average, carry a negative connotation. While this is not ideal, it still represents a notable improvement compared to more extreme communities. The only category with a positive association is political offenses, which aligns with the subreddit's focus on Brazilian far-right politics.

The remaining three communities show slightly positive values of association with hate terms. On average, their association scores range between +0.25 and +0.75. This indicates that, while they may not be as radicalized as *r/brasil_liberal* or *r/cornosbrasil*, they still exhibit a significantly favorable view of hate-related language. All hate categories in these communities lean toward the positive side of the spectrum, suggesting a consistent, although less extreme, normalization of harmful discourse.

Fringe vs. Baseline Comparison Figure 3 further illustrates how the different hate categories behave when comparing the average scores between the fringe communities and the baseline, *r/brasil*. It is immediately clear that, on average, all categories in the baseline fall on the negative side, whereas nearly all categories in the fringe communities lean toward the positive. Interestingly, political hate stands out as the most positively associated category among the fringe communities, while being the most negatively associated in the baseline. This contrast suggests a political nuance present in the fringe communities that is largely absent in *r/brasil*.

4.2 Hate Speech

From the WEAT analysis alone it is still unclear to which extent the different categories of hate speech appear on the studied communities. So, to systematically characterize the distribution of hate speech categories across banned and non-banned Brazilian Reddit communities, we analyzed the weighted average proportions of different categories, as illustrated in Figure 4. The weighted value for each category was computed as:

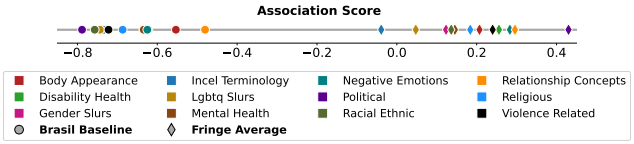

Figure 3: Association score comparison between the Brasil Baseline and Fringe communities across different categories. For visual clarity, the x-axis is restricted to the range [-0.8, 0.4].

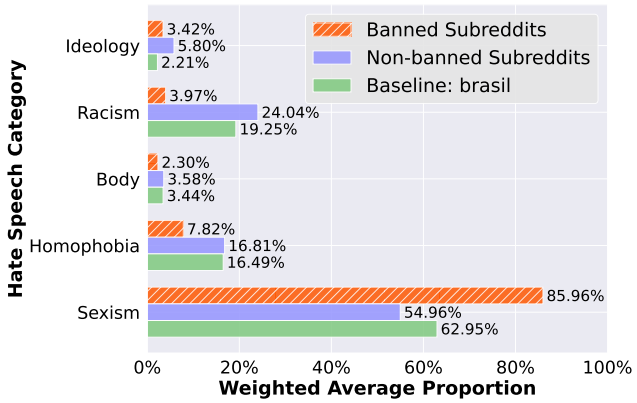

Figure 4: Bar plot presenting the comparison of the weighted average proportion of hate speech categories between banned and non-banned Subreddits. Banned subreddits include *r/brasil_drama*, *r/brasilimpo*, *r/brasil_liberal*, and *r/cornosbrasil*, while non-banned subreddits include *r/brasilivre* and *r/BrasildoB*. The categories are Sexism, Homophobia, Body, Racism, and Ideology, with banned subreddits showing higher proportions, especially in the Sexism category.

$$wv = \frac{(p_{\text{hate}} \cdot c_{\text{hate}}) + (p_{\text{non-hate}} \cdot c_{\text{non-hate}})}{c_{\text{total}}}$$

where p_{hate} is the proportion of hate content, c_{hate} is the count of hate instances, $p_{\text{non-hate}}$ is the proportion of non-hate content, $c_{\text{non-hate}}$ is the count of non-hate instances, and c_{total} is the total number of instances.

Figure 4 summarizes the average proportion of hate speech across five categories – Sexism, Homophobia, Body, Racism, and Ideology – for both banned and non-banned subreddits, alongside a baseline distribution from the broader *r/brasil* subreddit.

Our results reveal that sexism overwhelmingly dominates hate speech in banned communities, accounting for 86% of all hateful content. In contrast, non-banned subreddits consistently exhibit higher levels of hate speech across the remaining four categories – homophobia, body shaming, racism, and ideological attacks – when compared to their banned counterparts. Baseline averages tend to fall between the two groups across all categories.

This distribution raises critical questions about enforcement patterns on Reddit. The data suggest that moderation

may target communities dominated by sexist content while overlooking others with a broader and often higher overall presence of hate speech. In particular, non-banned subreddits exhibit more hate speech across most categories, yet appear to avoid the enforcement actions applied to communities with a narrower hate focus.

4.3 Hate Propagation

To further understand the dynamics of hate speech within the analyzed subreddits, we aimed to evaluate its propagation by comparing submissions and their corresponding comment threads. Specifically, we investigated whether submissions containing hate speech are associated with a higher proportion of hateful responses in the comments. This association may suggest that such submissions act as potential triggers for the amplification of toxic discourse. Among 443,683 total submissions, 13,153 (**2.96%**) were classified as containing hate speech.

Statistical analysis strongly supported this hypothesis, revealing that hate submissions had an average proportion of hate in comments of **0.1109**, compared to **0.0191** in non-hate submissions. This difference was statistically significant ($p < 0.001$), demonstrating that hate submissions contained approximately **5.81 times** more hateful comments. The 95% confidence interval for this difference was [0.0881, 0.0956], with a medium effect size (Fleiss' d = 0.68).

In Figure 5, we compare two groups of submissions: those classified as containing hate speech (*Hate*) and those without hate speech (*Non-Hate*). For each group, we evaluate the proportion of comments that also contain hate speech, allowing us to assess how the presence of hate in submissions influences the toxicity of ensuing discussions. The results provide compelling evidence of the amplification effect that hate-containing submissions have on comment toxicity.

In communities like *r/brasilivre* and *r/brasil_drama*, the contrast is particularly striking, with hate speech comments appearing **11.72** and **7.64** percentage points more frequently in submissions containing hate speech, respectively. This pattern remains consistent across all analyzed subreddits, with hate-containing submissions invariably generating a higher proportion of toxic comments compared to their non-hate counterparts.

Notably, r/cornosbrasil and r/brasil_liberal exhibited the smallest differences, with values of **0.69%** and **1.13%**, respectively. Despite having a higher percentage of hate speech in submissions containing hateful content, these communities demonstrated a broader distribution of toxicity. This suggests that harmful behavior was not confined to submissions with hate speech but permeated throughout the community, possibly contributing to their eventual ban.

Despite r/brasilivre and r/BrasildoB exhibiting some of the highest rates of hate speech propagation – with r/brasilivre showing the most extreme amplification effect – these communities remain active, non-banned subreddits. This discrepancy between content toxicity levels and moderation outcomes raises critical questions about the consistency and effectiveness of content policy enforcement within Brazilian subreddit ecosystems.

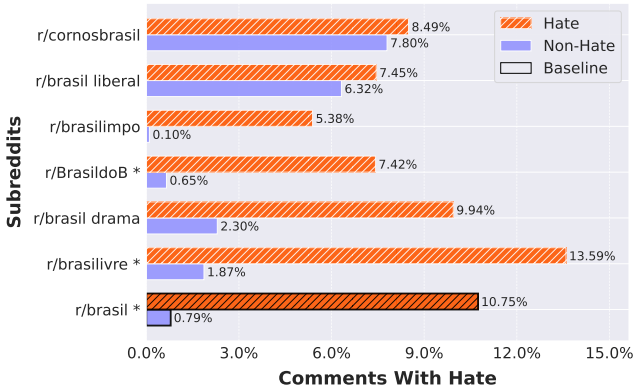

Figure 5: Percentage of comments containing hate speech in submissions with the presence or absence of hate speech. Subreddits marked with an asterisk (*) are non-banned communities.

5 Conclusion

This study sheds light on the dynamics of extremist discourse and hate speech in Brazilian Reddit communities by combining word embedding associations with transformer-based classification. Our findings reveal stark contrasts between banned, fringe, and mainstream subreddits – not only in the volume of toxic content but also in the semantic framing of hate-related terms. Banned communities consistently exhibited positive associations with discriminatory language, particularly related to sexism, suggesting a normalization of harmful rhetoric. Additionally, we identified a significant propagation effect: hate-laden submissions are strongly correlated with increased toxicity in comment threads, amplifying the reach and impact of harmful discourse.

Importantly, our results point to inconsistencies in Reddit's moderation practices. Some non-banned communities displayed hate speech levels and propagation dynamics similar to or greater than those of banned subreddits, raising questions about enforcement thresholds and possible biases in the content policy application. The lack of transparency regarding the full reasons for censorship is also a contentious issue, with some of the studied subreddits' admins alleging unfair treatment from reddit's staff [3].

By focusing on the Brazilian context, an understudied but increasingly relevant domain for online radicalization, this work contributes to the broader understanding of how digital platforms facilitate the spread of extremist narratives. Our methodology offers a replicable framework for identifying and monitoring harmful discourse in other linguistic and cultural contexts. Ultimately, we hope these insights can support more transparent, fair, and effective moderation strategies to foster healthier online ecosystems.

6 Limitations

This study focuses on a specific subset of Brazilian subreddits, primarily those that are ideologically polarized or

[3] https://www.reddit.com/r/brasilimpoalt/

have been banned for violating Reddit's content policies. While we intentionally included communities across the political spectrum and selected *r/brasil* as a generalist baseline, this sampling strategy may not capture the full diversity of Brazilian Reddit discourse. As such, our findings may not generalize to all user communities or thematic subreddits in the Brazilian context.

Additionally, our dataset relies on data collected via the Pushshift API (Baumgartner et al. 2020), which, while widely used in social computing research, has known limitations. Pushshift does not always provide real-time or complete coverage of Reddit activity, and its infrastructure has been subject to changes following Reddit's API restrictions in mid-2023. This may have affected the completeness of more recent posts and comments, particularly in low-traffic or newly created subreddits. Furthermore, metadata such as deleted or removed comments is inconsistently preserved, potentially underestimating moderation actions or the true extent of toxic discourse.

While the Word Embedding Association Test (WEAT) has been instrumental in exposing implicit biases within static embedding spaces (Caliskan, Bryson, and Narayanan 2017), its reliance on context-agnostic word vectors limits its sensitivity to nuanced usage in evolving online communities. Recent advances in contextualized language models—such as fine-tuning a transformer on subreddit-specific data and then querying prompts like "Is [term] good?"—offer more flexible and context-aware bias assessments. We leave the exploration of such methods, and their comparative evaluation against WEAT, as a direction for future work.

7 Ethical Considerations

Throughout our work we examine the differences between banned, mainstream, and fringe communities, finding differences that may have contributed to the eventual ban. Although these observed characteristics may be used for moderation, they must be approached with caution to avoid the risk of overgeneralization or reinforcing existing biases.

Importantly, all data used in this study were publicly available through Reddit and collected via the Pushshift API. At no point did our analysis involve private messages or restricted content. Furthermore, we did not perform any user-level tracking or profiling, ensuring that no individuals can be identified from the dataset. All results are presented in aggregate form, focusing solely on community-level trends. This approach adheres to ethical standards for research involving public online data and minimizes the risk of harm or privacy violations.

Acknowledgments

This work was partially funded by CNPq, CAPES, FAPEMIG, and IAIA - INCT on AI.

References

Angelov, D. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Barbieri, F.; Camacho-Collados, J.; Espinosa Anke, L.; and Neves, L. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. Online: Association for Computational Linguistics.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1): 830–839.

Bayrakdar, S.; Yucedag, I.; Simsek, M.; and Dogru, I. A. 2020. Semantic analysis on social networks: A survey. *International Journal of Communication Systems*, 33(11): e4424.

Bender, J. 2021. Detect hate speech in social media comments using deep learning. https://github.com/JensBender/hate-speech-detection.

Buzelin, A.; Aquino, Y.; Bento, P.; Dayrell, L.; Estanislau, V.; Malaquias, S.; Dutenhefner, P.; Porfírio, L.; Rigueira, P.; Grossi, C.; Evangelista, G.; Pappa, G.; and Jr., W. M. 2024a. Contextual BERT Model for Toxicity Detection in Messaging Platforms. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, 846–857. Porto Alegre, RS, Brasil: SBC.

Buzelin, A.; Aquino, Y.; Bento, P.; Malaquias, S.; Jr, W. M.; and Pappa, G. 2024b. A Change in Perspective: The Trade-Off Between Perspective API and Custom Models in Classifying Hate Speech in Portuguese. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 23–31. Porto Alegre, RS, Brasil: SBC.

Buzelin, A.; Dutenhefner, P. R.; Locatelli, M. S.; Malaquias, S.; Bento, P.; Aquino, Y.; Dayrell, L.; Estanislau, V.; Santana, C.; Alzamora, P.; Vasconcelos, M.; Jr., W. M.; and Almeida, V. 2025. Analyzing Political Discourse on Discord during the 2024 U.S. Presidential Election. arXiv:2502.03433.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Çano, E.; and Morisio, M. 2019. Word embeddings for sentiment analysis: a comprehensive empirical survey. *arXiv preprint arXiv:1902.00753*.

Cinelli, M.; De Francisci Morales, G.; Galeazzi, A.; Quattrociocchi, W.; and Starnini, M. 2021. The echo chamber effect on social media. *Proceedings of the national academy of sciences*, 118(9): e2023301118.

Conway, M.; Scrivens, R.; and Macnair, L. 2019. Right-Wing Extremists' Persistent Online Presence: History and Contemporary Trends. *International Centre for Counter-Terrorism*.

Cunha, G. F.; and da Silva, A. P. C. 2024. Caracterizando Polarização em Redes Sociais: Um Estudo de Caso das Discussões no Reddit sobre as Eleições Brasileiras de 2018 e 2022. In *Brazilian Symposium on Multimedia and the Web (WebMedia)*, 365–369. SBC.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for

Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

do Santos, V.; dos Santos, L.; Coelho, O.; de Araujo, R.; and de Oliveira, I. C. 2024. Mineração de Argumentos em Textos de Redes Sociais no Idioma Português. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 306–316. Porto Alegre, RS, Brasil: SBC.

Efstratiou, A.; Blackburn, J.; Caulfield, T.; Stringhini, G.; Zannettou, S.; and De Cristofaro, E. 2023. Non-polar Opposites: Analyzing the Relationship between Echo Chambers and Hostile Intergroup Interactions on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1): 197–208.

Farooqi, Z. M.; Ghosh, S.; and Shah, R. R. 2021. Leveraging transformers for hate speech detection in conversational code-mixed tweets. *arXiv preprint arXiv:2112.09986*.

Gaudette, T.; Scrivens, R.; Davies, G.; and Frank, R. 2021. Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society*, 23(12): 3491–3508.

Gaudette, T.; Scrivens, R.; and Venkatesh, V. 2020. The Role of the Internet in Facilitating Violent Extremism: Insights from Former Right-Wing Extremists. *Terrorism and Political Violence*.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Göçen, A.; Ibrahim, M.; and Khan, A. U. I. 2024. Public attitudes toward higher education using sentiment analysis and topic modeling. *Discover Artificial Intelligence*, 4.

Hada, R.; Sudhir, S.; Mishra, P.; Yannakoudakis, H.; Mohammad, S. M.; and Shutova, E. 2021. Ruddit: Norms of Offensiveness for English Reddit Comments. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2700–2717. Online: Association for Computational Linguistics.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Fitzpatrick, C. A.; Bull, P.; Lipstein, G.; Nelli, T.; Zhu, R.; Muennighoff, N.; Velioglu, R.; Rose, J.; Lippe, P.; Holla, N.; Chandra, S.; Rajamanickam, S.; Antoniou, G.; Shutova, E.; Yannakoudakis, H.; Sandulescu, V.; Ozertem, U.; Pantel, P.; Specia, L.; and Parikh, D. 2021. The Hateful Memes Challenge: Competition Report. In Escalante, H. J.; and Hofmann, K., eds., *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, 344–360. PMLR.

Laureate, C.; Buntine, W.; and Linger, H. 2023. A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, 56: 1–33.

Lees, A.; Tran, V. Q.; Tay, Y.; Sorensen, J.; Gupta, J.; Metzler, D.; and Vasserman, L. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 3197–3207. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393850.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Matsui, A.; and Ferrara, E. 2024. Word embedding for social sciences: an interdisciplinary survey. *PeerJ Computer Science*, 10: e2562.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Mitts, T. 2019. From Isolation to Radicalization: Anti-Muslim Hostility and Support for ISIS in the West. *American Political Science Review*, 113(1): 173–194.

Mollas, I.; Hutchinson, B.; Bade, N.; and Narayanan, A. 2020. ETHOS: An Online Hate Speech Detection Dataset. In *Proceedings of the Fourteenth International Conference on Language Resources and Evaluation (LREC 2020)*.

Müller, K.; and Schwarz, C. 2020. Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association*, 19(4): 2131–2167.

Ottoni, R.; Cunha, E.; Magno, G.; Bernardina, P.; Meira Jr., W.; and Almeida, V. 2018. Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, 323–332. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355636.

Pérez, J. M.; Rajngewerc, M.; Giudici, J. C.; Furman, D. A.; Luque, F.; Alemany, L. A.; and Martínez, M. V. 2023. pysentimiento: A Python Toolkit for Opinion Mining and Social NLP tasks. arXiv:2106.09462.

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V.; and Meira Jr, W. 2020. Auditing Radicalization Pathways on YouTube. *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, 131–141.

Russo, G.; Horta Ribeiro, M.; and West, R. 2024. Stranger Danger! Cross-Community Interactions with Fringe Users Increase the Growth of Fringe Communities on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1): 1342–1353.

Schmitz, M.; Muric, G.; Hickey, D.; and Burghardt, K. 2024. Do users adopt extremist beliefs from exposure to hate subreddits? *Social Network Analysis and Mining*, 14(1): 22.

Scrivens, R.; Amarasingam, A.; and Hirschfield, P. 2022. Combating Online Extremism: The Role of Counter-Narratives in Countering Violent Extremism. *Terrorism and Political Violence*.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Vilaça, G. 2022. VIRGIN X CHAD: Memes de Internet da machosfera brasileira no Reddit. In *Anais do 31º Encontro Anual da Compós*. Imperatriz - MA: Associação Nacional dos Programas de Pós-Graduação em Comunicação (Compós). Trabalho apresentado no GT Comunicação e Cibercultura, Universidade Federal do Maranhão, 06 a 10 de junho de 2022.

Watchful1. 2025. Subreddit comments/submissions 2005-06 to 2024-12.

Weiss, P.; Koban, K.; and Matthes, J. 2024. A narrow gateway from misogyny to the far right: empirical evidence for social media exposure effects. *Information, Communication & Society*, 0(0): 1–19.

Winter, C.; Neumann, P.; Meleagrou-Hitchens, A.; Ranstorp, M.; Vidino, L.; and Fürst, J. 2021. Online Extremism: Research Trends in Internet Activism, Radicalization, and Counter-Strategies. *International Journal of Conflict and Violence*, 14: 1–20.

Youngblood, M. 2020. Extremist ideology as a complex contagion: the spread of far-right radicalization in the United States between 2005 and 2017. *Humanities and Social Sciences Communications*, 7(1): 49.

Zhu, Y.; Zhou, R.; Chen, G.; and Zhang, B. 2024. Enhancing sentiment analysis of online comments: a novel approach integrating topic modeling and deep learning. *PeerJ Computer Science*, 10: e2542.

7.1 Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you run machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA