

Analyzing the Role of Negativity in Brazilian Reddit Discussions

Jonas Elias Santos Kretli
jonas.kretli@aluno.ufop.edu.br
Department of Computing and Systems, UFOP
Brazil

Josemar Alves Caetano
josemar@ufop.edu.br
Department of Computing and Systems, UFOP
Brazil

ABSTRACT

Online discussions are increasingly organized as discussion cascades, where users interact through sequences of replies that can grow into large, complex conversational structures. While prior work has studied sentiment, toxicity, and user behavior in online communities, less is known about how negativity relates simultaneously to discussion structure, user interaction, and language. We investigate the role of negativity in Brazilian Reddit discussion cascades, analyzing over one million comments from 20 subreddits. Using a transformer-based sentiment classifier, we group cascades and users into quartiles by the proportion of negative comments, and examine structural, interactional, and linguistic patterns across negativity levels. Negativity is associated with multiple dimensions of discussion dynamics, but its effects are consistently non-monotonic: cascades grow larger, deeper, and more viral as negativity rises up to moderately high levels, then contract sharply at the highest levels. Homophily follows the same pattern, peaking among moderately negative users, who occupy a central position in the interaction network, while the most negative users show low homophily and interact mainly with moderately negative peers. Linguistically, negative discussions show a steady rise in anger-related and conflict-oriented language, while thematic diversity peaks at moderate negativity and declines at both extremes. These findings indicate that negativity is not a simple linear driver of engagement: moderate negativity sustains the largest, most cohesive, and most diverse discussions, whereas extreme negativity is reflected mainly in hostile language rather than structural or social amplification. This distinction has direct implications for content moderation and community health monitoring, as platforms that focus exclusively on the most extreme content may overlook the moderately negative, high-engagement core that sustains much of the conversational activity.

KEYWORDS

Negativity, Reddit, Sentiment Analysis, Discussion Cascades, Homophily, Thematic Diversity

1 INTRODUCTION

Online discussions increasingly unfold through threaded interaction structures, in which users respond to one another and collectively build discussion cascades rather than posting isolated messages [6]. Such cascades are particularly prominent on Reddit, where nested comment threads enable conversations to evolve through sequences of replies that vary considerably in size, depth,

and participation [23]. These interaction structures provide a valuable opportunity to investigate how discussions grow, how users interact, and how emotional expressions emerge throughout online conversations [20, 23].

Previous research has examined several aspects of online discussions, including information diffusion, political polarization, toxicity, sentiment, and user behavior across social media platforms [21, 29]. However, many studies focus either on the textual content of individual messages or on aggregate community-level properties. As a result, less attention has been devoted to understanding how emotional characteristics are associated with the structural organization of discussions and with the interaction patterns established among users. This gap is particularly relevant because negative interactions are often associated with polarization, toxicity, conflict escalation, and reduced deliberative quality in online environments [9, 30].

Understanding negativity in Reddit discussions requires going beyond the detection of negative comments in isolation. A discussion may contain negative content without necessarily becoming large, deep, or socially cohesive. Conversely, some negative discussions may attract sustained participation, generate long reply chains, and concentrate interactions among users with similar emotional profiles. Thus, negativity should be examined not only as a textual attribute, but also as a factor potentially associated with the growth of discussions, the organization of user interactions, and the linguistic and thematic characteristics of online participation [25, 32, 35]. This perspective is especially important for Reddit, where community norms, nested replies, and topic-centered participation jointly shape how conversations develop [23, 37].

In this paper, we investigate the role of negativity in Brazilian Reddit discussions. Rather than focusing on predefined subreddit categories, we analyze how discussions differ according to their level of negativity. To this end, we model each discussion thread as a set of cascades of comments and replies and classify cascades according to the proportion of negative comments they contain. We also characterize users according to the proportion of negative comments they produce. This approach allows us to examine whether increasingly negative discussions and users exhibit distinct structural, social, linguistic, and thematic patterns.

From a structural perspective, we characterize cascades using metrics such as size, depth, breadth, and structural virality [13]. From a social perspective, we analyze user interaction patterns through reply networks and homophily measures [4, 5, 25]. Finally, we investigate linguistic and thematic differences across negativity levels using sentiment-aware lexical analysis and topic modeling techniques [14, 27]. Together, these analyses allow us to examine negativity as a multidimensional phenomenon that affects not only what users say, but also how discussions grow and how participants interact.

More specifically, we address the following research questions:

RQ1: How does negativity relate to the structural properties of discussion cascades?

RQ2: How does negativity influence user interaction patterns within discussion cascades?

RQ3: How does negativity shape the linguistic and thematic profiles of discussion cascades and users?

Our study makes three main contributions, corresponding to RQ1–RQ3: a structural characterization of cascades across negativity levels, an analysis of homophily and reply patterns among users with different negativity profiles, and a linguistic and thematic characterization based on lexical analysis and topic modeling. Together, these contributions offer a comprehensive view of how negativity manifests in online discussions, revealing differences not only in cascade growth and user interactions, but also in the language and themes that characterize such discussions.

This work offers practical implications for content moderation and community health monitoring: rather than focusing exclusively on the most overtly negative content, platforms may benefit from treating negativity as a continuous signal, as moderately negative users tend to occupy the structural core of contentious discussions and sustain much of the conversational activity. Moderation tools that incorporate this graduated view may enable earlier and more targeted interventions before discussions escalate to their most extreme forms.

The remainder of this paper is organized as follows. Section 2 reviews related work on discussion cascades, sentiment and toxicity in online communities, homophily, and linguistic characterization of online discussions. Section 3 describes the dataset, cascade construction, sentiment analysis, negativity-based grouping, and the methods used to analyze structure, interactions, and language. Section 4 presents the results for the three research questions. Section 5 discusses the main findings, limitations, and directions for future work. Finally, Section 6 discusses the ethical aspects related to data collection and use in this work.

2 RELATED WORK

This work intersects four research strands: structural analyses of online discussion cascades, homophily and interaction patterns, sentiment and toxicity in Brazilian online communities, and linguistic and thematic characterization of online discussions. Together, these strands motivate the need for a joint, negativity-centered analysis of Reddit discussions.

Structural Properties of Cascades and Negativity: Reddit’s nested reply architecture has been widely used to study how discussions grow through metrics such as cascade size, depth, breadth, and structural virality [13, 23]. Prior work shows that toxic or conflictual conversations often develop into larger, wider, and deeper reply trees than non-toxic ones, suggesting an association between negative content and structural amplification [32]. Related evidence from intercommunity conflict and cross-platform cascade comparisons also indicates that emotional or adversarial content can shape how discussions spread [6, 20, 37]. In the Brazilian Reddit context, toxic discussion trees have been found to attract more replies and extend across more interactions than non-toxic ones, while hate-laden

submissions can amplify toxicity in subsequent comments [3, 28]. These studies rely mostly on binary toxic/non-toxic distinctions and suggest a broadly monotonic relationship between negativity and cascade growth. Our results complicate this expectation: structural amplification peaks at moderate-to-high negativity and contracts at the most extreme levels, indicating that this relationship is not strictly monotonic.

Homophily and Interaction Patterns: Homophily is a central concept in social network analysis and has been widely used to study political alignment, ideological segregation, echo chambers, and online polarization [5, 9, 30, 36]. However, recent work cautions against assuming that charged environments always produce self-contained homophilic clusters. For example, Reddit users in news discussions may engage with ideologically opposed users while still exhibiting other forms of segregation [25]. Other studies have extended homophily analysis to behavioral and sentiment-based profiles [4, 36], but less is known about how homophily varies across different degrees of negativity. We find that homophily peaks among moderately negative users rather than the most negative ones, indicating that negativity-based clustering is not a simple function of how extreme a user’s profile is.

Sentiment, Toxicity, and Negativity in Brazilian Online Communities: Recent work has expanded computational resources for studying sentiment and toxicity in Brazilian Portuguese, including annotated datasets, and contextual approaches for toxicity detection [21, 24, 31]. Studies on Brazilian Reddit further show that Portuguese BERT-based models can achieve reasonable agreement with human annotations, although irony, sarcasm, regional slang, and ambiguous comments remain persistent challenges [29]. Prior work also indicates that translating Portuguese content into English before toxicity detection can reduce classification reliability [18], reinforcing the need for native-language inference. At the same time, analyses of Brazilian Reddit communities reveal that toxicity and hate speech are not limited to obviously problematic spaces, making it important to characterize negativity beyond banned/non-banned or toxic/non-toxic labels [3]. We adopt this approach, using sentiment classification to define negativity levels and examining how these levels relate to structural, interactional, and linguistic properties.

Linguistic and Thematic Characterization: Lexical and psycholinguistic frameworks such as LIWC have long been used to characterize affective, cognitive, and social dimensions of online language [7, 27]. Topic modeling methods, including BERTopic, combined with entropy-based measures, have also been used to analyze thematic diversity and concentration in large-scale textual corpora [8, 14]. These approaches have been combined with emotion analysis to study discourse dynamics and emotional intensity during conflict escalation [1]. Since topic modeling pipelines designed for English do not necessarily transfer well to Portuguese, prior work recommends language-specific preprocessing and Portuguese embeddings [12, 34]. We extend this line of research by jointly examining LIWC-based lexical markers and topic entropy across negativity levels in Brazilian Reddit discussions.

Positioning of This Work: Prior research provides strong, largely independent evidence that negativity is associated with structural

amplification [32], that interaction patterns among emotionally or ideologically similar users are more complex than simple homophily accounts suggest [25], and that Brazilian Reddit text poses specific challenges for sentiment and toxicity analysis [18, 21, 29, 35]. However, these dimensions have not been examined jointly, nor has prior work tested whether structural amplification holds uniformly across the full range of negativity.

Moreover, prior work on toxicity in Brazilian Reddit communities has relied on binary toxic/non-toxic or banned/non-banned distinctions [3, 28], rather than treating negativity as a continuous dimension capable of revealing non-monotonic patterns. Our work addresses this gap by treating negativity as a continuous organizing dimension and jointly analyzing its structural, interactional, and linguistic correlates in Brazilian Reddit discussions. As shown in Section 4, this joint perspective reveals that negativity's relationship with discussion dynamics is consistently non-monotonic across all three dimensions: a pattern that isolated analyses of structure, interaction, or language alone would not have surfaced.

3 METHODOLOGY

In this Section, we first present the data collection process and the resulting corpus. We then describe how discussion cascades and user negativity profiles are constructed, followed by the analyses performed to address each research question.

3.1 Dataset

Our study is based on a large-scale dataset collected from Reddit, one of the largest online discussion platforms [26]. The dataset comprises 1,071,383 comments associated with 9,959 posts written by 118,330 unique users from 20 Brazilian communities¹. The collected communities cover a diverse range of topics, including sociopolitical and news discussions, humor and memes, labor and opinion-based forums, as well as entertainment and hobby-oriented communities. This stratification ensures observed effects can generalize across discourse types.

Data were collected from publicly available Reddit content published between May 11, 2025 and May 21, 2026. To ensure access to complete discussion structures and associated metadata, we directly retrieved publicly accessible Reddit data and reconstructed the corresponding discussion threads. The collection process included an additional retrieval stage to recover comments located in deeper levels of discussion trees, reducing the likelihood of missing interactions in highly nested conversations.

3.2 Discussion Cascade Construction

Reddit discussions are organized as hierarchical reply structures, where users can respond either to the original post or to previous comments. Following prior work on online discussion cascades [13, 23], we model each discussion thread as a collection of discussion cascades.

Each first-level comment in response to the original submission is considered the root of a discussion cascade, with subsequent replies

recursively connected to their parent comments, forming an n -ary tree that captures the conversational structure of the discussion.

This representation enables the computation of structural properties commonly used to characterize online discussions, including cascade size, depth, breadth, and structural virality. Moreover, it preserves the reply relationships among users, allowing the analysis of interaction patterns and conversational dynamics throughout discussions.

To reduce the loss of contextual information, multimedia content was enriched before analysis [11, 16, 33]. When available, images were converted into textual descriptions using a vision-language model, YouTube links were replaced by their corresponding video titles, and external hyperlinks were preserved. This procedure allows sentiment and linguistic analyses to incorporate contextual information that would otherwise be unavailable in text-only representations. In total, we identified 71,281 cascades originating from the 9959 posts in the dataset.

3.3 Sentiment Analysis

To investigate the role of negativity in Reddit discussions, each comment in the dataset was assigned a sentiment label using XLM-RoBERTa (XLM-R), a multilingual transformer model for natural language processing tasks [10]. We employ the Portuguese sentiment classification model *cardiffnlp/xlm-roberta-base-tweet-sentiment-pt*, available through the HuggingFace Transformers library². This model fine-tunes the XLM-R architecture on a multilingual Twitter sentiment benchmark covering eight languages, including Portuguese [2], making it suitable for the informal, user-generated linguistic register found in Reddit discussions. Following the inference process, comments were classified into one of three categories: *Positive (POS)*, *Neutral (NEU)*, or *Negative (NEG)*.

We note that this classifier was fine-tuned on Twitter data rather than Reddit-specific Brazilian Portuguese text; as discussed in Section 2, some misclassification on ambiguous, sarcastic, or slang-heavy comments is therefore expected [29], a limitation we revisit in Section 5.

The sentiment labels obtained in this stage serve as the basis for all subsequent analyses. At the cascade level, they are used to compute the negativity ratio of each discussion cascade, defined as the proportion of comments classified as negative. At the user level, they are used to estimate individual negativity profiles based on the proportion of negative comments produced by each participant.

Because Reddit discussions frequently rely on images, videos, and external links as contextual anchors, sentiment inference was performed on the enriched textual representations described in Section 3.2. This procedure allows sentiment classification to incorporate contextual information associated with multimedia content, reducing the likelihood of misinterpreting comments whose meaning depends on external context.

Among the 1,071,383 comments analyzed, 578,032 (53.95%) were classified as negative, 385,797 (36.01%) as neutral, and 107,554 (10.04%) as positive.

3.4 Negativity-Based Grouping

We define negativity at both the cascade and user levels.

¹*r/brasil*, *r/brasillivre*, *r/brasildob*, *r/debatesbr*, *r/noticiasbr*, *r/botcodoreddit*, *r/faralimabets*, *r/memesbr*, *r/shitpostbr*, *r/antitrampo*, *r/opiniaoiimpopular*, *r/opiniaoburra*, *r/filosofia*, *r/infernosocial*, *r/futebol*, *r/gamesecultura*, *r/videogamesbrasil*, *r/carros*, *r/computadores*, and *r/soapaulo*

²<https://huggingface.co/cardiffnlp/xlm-roberta-base-tweet-sentiment-pt>

For a discussion cascade (c), the negativity ratio is defined as:

$$Neg(c) = \frac{N_{NEG}(c)}{N(c)}$$

where $N_{NEG}(c)$ denotes the number of comments classified as negative and $N(c)$ denotes the total number of comments in the cascade.

Similarly, for a user (u), the negativity ratio is defined as:

$$Neg(u) = \frac{N_{NEG}(u)}{N(u)}$$

where $N_{NEG}(u)$ represents the number of negative comments produced by the user and $N(u)$ corresponds to the total number of comments authored by that user.

Using these measures, cascades and users are independently ranked according to their negativity ratios and partitioned into four quartiles. For cascades, Q1 contains the least negative discussions, whereas Q4 contains the most negative ones. Analogously, users are grouped into four negativity profiles, ranging from UQ1 (least negative users) to UQ4 (most negative users).

Due to ties in negativity ratios, the resulting groups contain different numbers of observations. Table 1 summarizes the negativity-based partitioning adopted throughout the paper, including the ranges associated with each quartile and the corresponding numbers of cascades and users. We denote cascade quartiles as Q1–Q4 and user quartiles as UQ1–UQ4. In both cases, the first quartile corresponds to the least negative observations, whereas the fourth quartile corresponds to the most negative ones.

Table 1: Negativity-based quartiles used throughout analyses.

Quartile	Negativity Range	# Cascades	# Users
Q1 / UQ1	0–25%	12,953	36,926
Q2 / UQ2	25–50%	21,570	26,742
Q3 / UQ3	50–75%	22,384	20,333
Q4 / UQ4	75–100%	14,374	34,329

The cascade quartiles are used throughout RQ1 and RQ3 to compare structural, linguistic, and thematic characteristics across different levels of negativity. The user quartiles are employed in RQ2 to investigate interaction patterns and homophily among users with different negativity profiles.

3.5 Structural Analysis of Cascades (RQ1)

To address RQ1, we investigate whether discussion cascades with different levels of negativity exhibit distinct structural properties. Following previous work on information diffusion and online discussion cascades [13, 23], we characterize each cascade using four complementary metrics: size, depth, breadth, and structural virality.

Size corresponds to the total number of comments contained in a cascade. Depth is defined as the length of the longest reply chain, measured as the maximum distance between the root comment and any descendant comment. Breadth represents the maximum number of comments observed at the same depth level. Finally, structural virality captures the overall shape of the cascade and is

computed using the Wiener Index, corresponding to the average shortest-path distance between all pairs of nodes in the cascade [13].

To compare cascades across negativity levels, we analyze the distributions of these metrics for each negativity quartile (Q1–Q4) using complementary cumulative distribution functions (CCDFs). This approach enables the comparison of highly skewed distributions commonly observed in online discussions and allows us to assess whether increasingly negative discussions tend to generate larger, deeper, broader, or more viral conversational structures.

3.6 User Interaction Analysis (RQ2)

To address RQ2, we investigate whether users with different negativity profiles exhibit distinct interaction patterns. For this purpose, we construct a user interaction network from the discussion cascades, where each node represents a user and a directed edge $u_i \rightarrow u_j$ indicates that user u_i replied to a comment authored by user u_j .

Users are grouped into four negativity quartiles (UQ1–UQ4) according to the proportion of negative comments they produce, as described in Section 3.4. These groups range from the least negative users (UQ1) to the most negative users (UQ4).

To quantify the tendency of users to interact with others exhibiting similar negativity levels, we measure homophily [17]. For a user (u), homophily is defined as the fraction of neighbors belonging to the same negativity quartile:

$$H(u) = \frac{|v \in N(u) : T(v) = T(u)|}{|N(u)|}$$

where $N(u)$ denotes the set of users directly connected to u and $T(u)$ represents the negativity quartile assigned to user (u). Higher values of $H(u)$ indicate stronger interaction preferences toward users with similar negativity profiles.

In addition, we analyze the interaction matrix between user quartiles. Specifically, we compute the proportion of interactions originating from users in quartile UQ_i and directed toward users in quartile UQ_j . This matrix provides a global view of interaction preferences and allows us to identify whether interactions are concentrated within the same negativity group or distributed across different user profiles.

3.7 Linguistic and Thematic Analysis (RQ3)

To address RQ3, we investigate whether discussions and users with different levels of negativity exhibit distinct linguistic and thematic characteristics. While RQ1 focuses on cascade structure and RQ2 on interaction patterns, this analysis examines how negativity is reflected in the language employed by participants and in the diversity of topics discussed.

First, we employ the Linguistic Inquiry and Word Count (LIWC) framework [7, 27] to characterize linguistic patterns associated with different negativity levels. LIWC maps words to psychologically and socially meaningful categories, enabling the analysis of dimensions such as affective processes, social references, cognitive processes, and conflict-related language. We compare the prevalence of these categories across negativity quartiles to identify linguistic markers associated with increasingly negative discussions and users.

Finally, we investigate thematic characteristics using BERTopic [14]. Given that our datasets are in Portuguese, we employ the BERTimbau model [34]. The model is applied to the entire corpus to identify latent discussion topics and assign topic distributions to cascades and users. Based on these distributions, we compute thematic diversity using the normalized entropy of the topic distribution [8]:

$$H = - \sum_{k=1}^N p(k) \log p(k),$$

where $p(k)$ denotes the proportion of messages assigned to topic k and N is the total number of topics. Entropy is normalized to the interval $([0,1])$, where values close to 0 indicate discussions concentrated on a small number of topics, whereas values close to 1 indicate a more diverse distribution of attention across topics.

4 RESULTS

As reported in Section 3, comments classified as negative (53.95%) dominate the dataset, motivating our investigation of how different levels of negativity relate to discussion dynamics. The remainder of this section addresses RQ1 (structural properties), RQ2 (interaction patterns), and RQ3 (linguistic and thematic characteristics), in turn.

4.1 Negativity and Structural Properties of Discussion Cascades (RQ1)

To answer RQ1, we compare the distributions of four structural properties across the cascade negativity quartiles (Q1–Q4): structural virality, maximum depth, maximum breadth, and cascade size. To assess statistical significance, we first apply a Kruskal–Wallis test [19] to each metric, followed by pairwise Kolmogorov–Smirnov (KS) tests [22] between negativity quartiles.

Figure 1 shows the distributions of these four metrics across negativity quartiles. Q1 exhibits the steepest CCDF decay in all cases, indicating that low-negativity discussions are substantially less likely to develop large, deep, broad, or structurally viral conversational structures.

The strongest differences emerge for structural virality (Figure 1a) and cascade size (Figure 1d). Cascades in Q3 exhibit the highest average structural virality ($\mu = 1.24$) and the largest average size ($\mu = 10.03$), whereas Q1 presents considerably lower values ($\mu = 0.95$ and $\mu = 5.76$, respectively). Similar patterns are observed for maximum depth (Figure 1c) and maximum breadth (Figure 1b), where Q3 again dominates the remaining groups.

These visual differences are statistically supported: a Kruskal–Wallis test indicates significant differences among quartiles for all structural metrics ($p < 10^{-250}$). Pairwise KS comparisons involving Q1 and Q3 yield the largest effect sizes overall, with $D = 0.2365$ for structural virality, $D = 0.2445$ for cascade size, $D = 0.1922$ for maximum depth, and $D = 0.1737$ for maximum breadth. Because our central claim of non-monotonicity hinges specifically on whether the decline from the peak (Q3) to the most negative quartile (Q4) reflects a genuine effect rather than sampling variation, we test this pairwise comparison directly: the contraction from Q3 to Q4 is itself statistically significant across all four metrics ($D = 0.1161$, $D = 0.1334$, $D = 0.0728$, and $D = 0.1068$, respectively; all $p < 10^{-40}$), confirming that the observed non-monotonicity is not an artifact of sampling noise. Applying a Bonferroni correction [15] for the 24

pairwise comparisons ($\alpha \approx 0.002$), all reported differences remain significant except for the Q2–Q4 comparison for maximum depth ($D = 0.0171$, $p = 0.013$), which does not affect our central Q1–Q3–Q4 comparisons. We additionally verified that cascade size shows a negligible correlation with the negativity ratio used to define quartiles (Spearman’s $\rho = 0.0437$), and that user activity volume shows an equally negligible correlation with user-level negativity ratio (Spearman’s $\rho = 0.0536$); both correlations reach statistical significance only due to the large sample size ($p < 0.001$ in both cases, with exact values of 1.68×10^{-31} and 4.85×10^{-76} , respectively). This rules out the possibility that the observed non-monotonic pattern, as well as the user-level patterns reported in RQ2 and RQ3, are driven by a mechanical relationship between sample or activity size and proportion-based negativity scores. These results suggest that the transition from low-negativity discussions to moderately high-negativity discussions is associated with substantial changes in cascade structure.

4.2 Negativity and User Interaction Patterns (RQ2)

To address RQ2, we analyze homophily, i.e., the tendency of users to interact with others in the same negativity quartile, and the interaction matrix between quartiles, which identifies preferential reply patterns across negativity levels.

Figure 2 presents the average homophily index for each user quartile. Users in UQ1 exhibit very low homophily ($\mu = 0.11$), indicating that interactions among the least negative users are relatively uncommon. Homophily increases considerably for UQ2 ($\mu = 0.37$) and reaches its highest value in UQ3 ($\mu = 0.54$), suggesting that moderately negative users tend to preferentially interact with others exhibiting similar negativity levels. Surprisingly, the most negative users (UQ4) exhibit a much lower homophily index ($\mu = 0.13$), comparable to that observed for UQ1. This pattern is statistically significant: a Kruskal–Wallis test across the four user quartiles confirms a significant global effect ($H = 20438.95$, $p < 10^{-300}$). Because the central pattern here is the sharp homophily peak at UQ3, we directly test the two comparisons that bracket it: the rise from UQ1 to UQ3 ($D = 0.7092$, $p < 10^{-300}$) and the drop from UQ3 to UQ4 ($D = 0.6844$, $p < 10^{-300}$) are both extremely large effects, confirming that the peak at UQ3 reflects a substantial, non-incidental difference in homophily.

To better understand this pattern, Figure 3 shows the response proportion matrix between user quartiles. Each row represents the distribution of replies generated by users in a given quartile, while each column indicates the quartile of the replied-to user.

First, interactions involving UQ1 are relatively dispersed. Users in this group direct only 12.1% of their replies to other UQ1 users, while most interactions are directed toward UQ2 (41.0%) and UQ3 (40.8%). Similarly, users in UQ2 preferentially interact with UQ3 users (48.0%) rather than with members of their own quartile (36.9%).

Second, UQ3 appears as the central interaction group in the network. Regardless of the source quartile, the largest share of replies is consistently directed toward UQ3 users. For example, 56.6% of replies produced by UQ3 users target other UQ3 users, while 58.2% of replies generated by UQ4 users are also directed

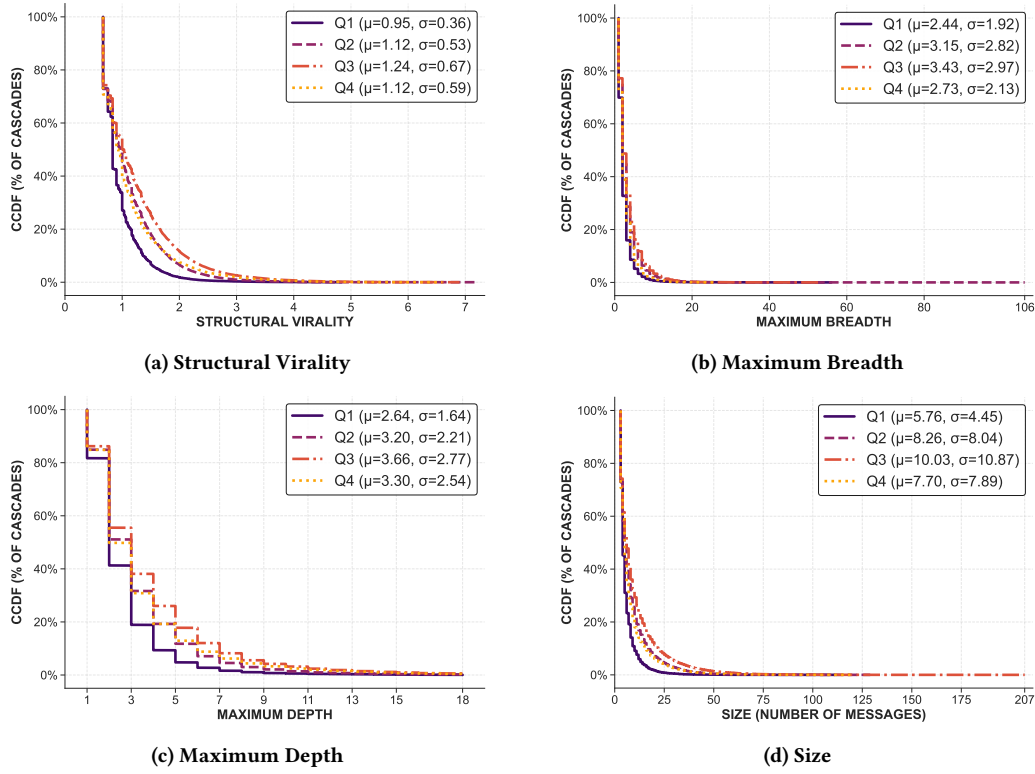


Figure 1: Structural properties of cascades across negativity quartiles.

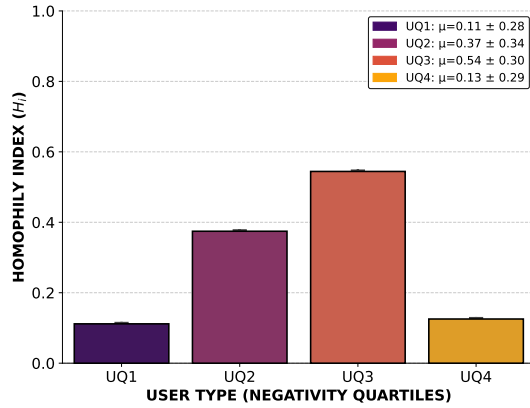
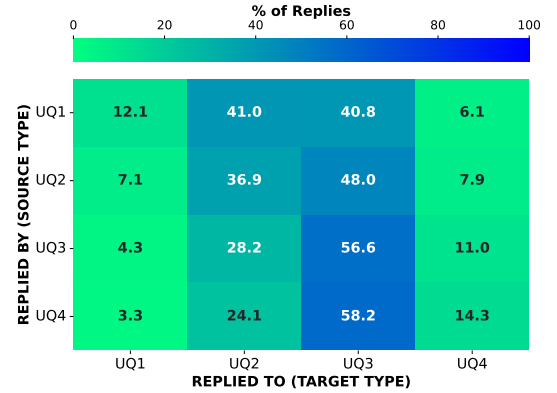
Figure 2: Avg. Homophily Index (H_i) and standard deviation.

Figure 3: Response proportion matrix (replies).

toward UQ3. Even UQ1 and UQ2 users predominantly interact with UQ3.

Finally, interactions involving UQ4 are comparatively rare. Despite representing the most negative users, UQ4 receives only a small fraction of replies from all groups. UQ4 users direct only 14.3% of their replies to other UQ4 users, while most of their interactions target UQ3. This finding explains the low homophily observed for UQ4 in Figure 2. Rather than forming highly self-contained interaction groups, the most negative users appear to engage primarily

with moderately negative users. This interaction asymmetry is itself statistically significant: a Kruskal–Wallis test on the distribution of reply targets across source quartiles confirms a significant global effect ($H = 13267.58$, $p < 10^{-300}$). The comparison that most directly speaks to this central-group pattern is UQ3 versus UQ4: their target distributions are significantly different ($D = 0.0624$, $p = 7.99 \times 10^{-128}$) but far more similar to each other than either is to UQ1's (UQ1 vs UQ3: $D = 0.2074$, $p < 10^{-300}$), confirming that UQ3 and UQ4 source users converge on a shared, UQ3-centered

reply pattern that sets them apart from the more dispersed behavior of UQ1.

4.3 Negativity, Language, and Thematic Diversity (RQ3)

To address RQ3, we measure thematic diversity using topic entropy derived from BERTopic, and examine the prevalence of linguistic categories associated with emotions, social behavior, and psychological processes using LIWC.

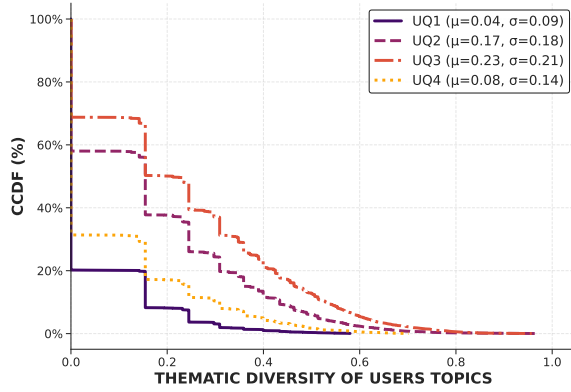


Figure 4: Thematic diversity (normalized Shannon entropy) of topics discussed by users.

Figure 4 presents the CCDF of user-level thematic diversity. Users in UQ3 exhibit the highest thematic diversity ($\mu = 0.23$), followed by UQ2 ($\mu = 0.17$), whereas UQ1 ($\mu = 0.04$) and UQ4 ($\mu = 0.08$) display substantially lower entropy values. These results indicate that moderately negative users tend to participate in a broader range of topics, while both the least negative and the most negative users concentrate their activity on a more limited set of themes. This pattern is statistically significant: a Kruskal–Wallis test across the four user quartiles confirms a significant global effect ($H = 1941.21$, $p < 10^{-300}$). The comparisons that most directly test the peak-at-UQ3 pattern, UQ1 vs UQ3 and UQ3 vs UQ4, both show large effects ($D = 0.4625$ and $D = 0.3779$, respectively; $p < 10^{-170}$ in both cases), confirming that the rise toward and fall away from the UQ3 peak are themselves strong, non-incidental effects.

A robustness check at the cascade level reveals the same qualitative pattern (Q1 $\mu = 0.16$; Q2 $\mu = 0.21$; Q3 $\mu = 0.23$; Q4 $\mu = 0.17$; Kruskal–Wallis $H = 879.78$, $p = 2.16 \times 10^{-190}$). The peak-focused comparisons, Q1 vs Q3 ($D = 0.1424$, $p = 7.85 \times 10^{-118}$) and Q3 vs Q4 ($D = 0.1191$, $p = 2.85 \times 10^{-86}$), confirm this pattern but with effect sizes notably smaller than at the user level, suggesting that the moderate-negativity peak in thematic diversity, while present at the cascade level, is primarily a user-level phenomenon.

This pattern mirrors the interaction dynamics reported in RQ2: the same group with the highest homophily (UQ3) also shows the greatest thematic diversity, suggesting a central role in discussions, while the most negative users (UQ4) are comparatively more specialized, concentrating on a smaller set of recurring themes.

Figure 5 reveals substantial linguistic differences across cascade negativity quartiles: the frequency of words related to anger grows from approximately 0.4% in Q1 to more than 1.2% in Q4, while swear words more than double between the least and most negative quartiles, with a similar increasing trend for risk-related language.

Other categories follow similar trends: references to power rise steadily from Q1 to Q4, suggesting greater association with dominance, authority, or control, while certainty-related terms become progressively more frequent, indicating stronger assertions and less linguistic uncertainty.

In contrast, netspeak is the only category exhibiting a consistent decline, from most frequent in Q1 to least frequent in Q4, suggesting that highly negative discussions rely less on casual online language and more on emotionally charged, conflict-oriented communication. All linguistic differences described above are statistically robust: Kruskal–Wallis tests across the four negativity quartiles are significant for every category (all $p < 10^{-30}$), and pairwise KS tests between Q1 and Q4 show effect sizes ranging from large (anger: $D = 0.4816$; swear: $D = 0.3448$; risk: $D = 0.3421$) to moderate (certain: $D = 0.2646$; power: $D = 0.2624$; netspeak: $D = 0.1243$), all significant at $p < 10^{-90}$.

5 DISCUSSION, CONCLUSION AND FUTURE WORK

Our results show that negativity is associated with multiple dimensions of Reddit discussions, but, consistently across all three research questions, this relationship is non-monotonic rather than linear. From a structural perspective, cascade size, depth, breadth, and structural virality all increase as negativity rises from the least negative discussions (Q1) up to moderately high levels (Q3), after which the most negative cascades (Q4) exhibit a measurable contraction across every metric. This pattern complicates the expectation, raised by prior structural analyses of toxic conversations, that negativity and structural complexity should scale together monotonically [32]: in our data, the most negative discussions are not the largest or most structurally developed ones — moderately negative discussions are. This suggests that moderate negativity may stimulate disagreement, emotional engagement, and sustained participation, while the most negative discussions become comparatively more restricted in their interactional scope.

At the user level, we observed an analogous pattern. Homophily peaks among moderately negative users (UQ3), who also occupy a central position in the interaction matrix, attracting replies from users across all other negativity profiles. In contrast, the most negative users (UQ4) display homophily comparable to the least negative users and do not form a strongly self-contained group, interacting predominantly with moderately negative peers instead. This finding echoes recent evidence that homophily should not be assumed as a default outcome of emotionally or ideologically charged interaction [25]: in our case, it is moderate, not extreme, negativity that drives the strongest clustering. This suggests that contentious discussions are not sustained only by the most negative participants; instead, moderately negative users appear to play an important role in connecting different user profiles and maintaining conversational activity.

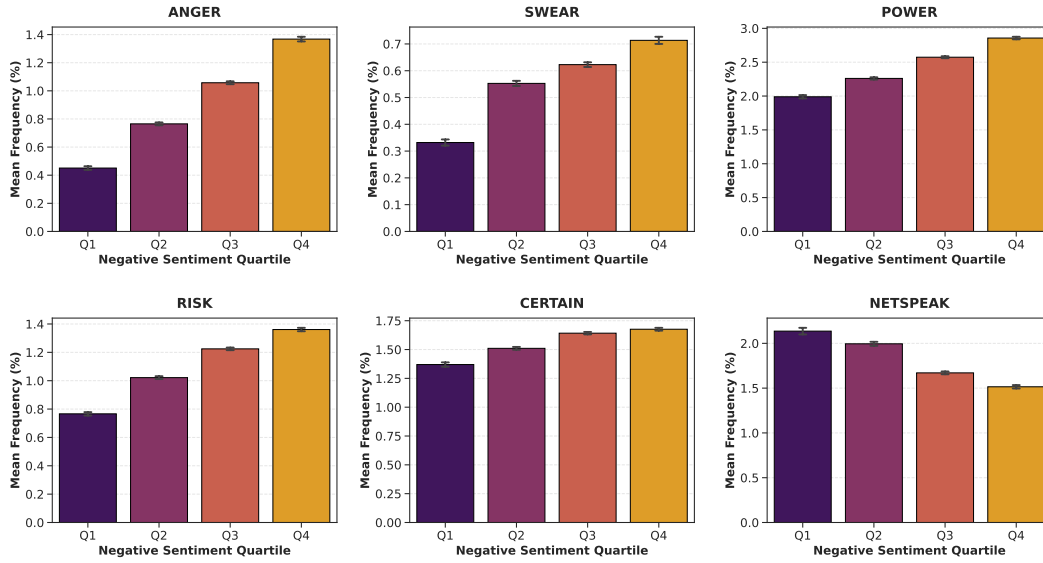


Figure 5: Average number of words associated with emotional categories in LIWC.

The linguistic and thematic analyses reveal a partially different pattern, which helps explain the apparent contrast between structural contraction and emotional intensity at the highest negativity levels. Cascades with increasingly negative content contain more words associated with anger, swearing, power, risk, and certainty, while relying less on informal netspeak. These results indicate that negativity is accompanied by a shift toward more aggressive, assertive, and conflict-oriented language, refining prior accounts that link negativity and out-group hostility directly to engagement and virality [30]: such effects appear to hold for language intensity, but not uniformly for structural or social amplification. At the same time, user-level topic entropy shows that moderately negative users discuss a broader range of topics than both low-negativity and high-negativity users. Thus, the most negative users are not necessarily the most thematically diverse; rather, moderate negativity is associated with broader participation across topics.

These findings suggest that negativity helps explain how Reddit discussions grow, how users interact, and how language changes across conversational contexts. More importantly, they show that negativity should not be treated as a simple linear phenomenon. Moderate negativity appears to function as a structural and social core of contentious discussions, associated with larger cascades, stronger interactional cohesion, and broader thematic participation. Extreme negativity, in contrast, is more clearly reflected in hostile and assertive language, but not necessarily in greater structural expansion or social cohesion.

These results show that negative discussions are not only defined by the presence of negative content; they also differ in terms of structure, interaction, and language, and these differences do not move together in the same direction. Therefore, analyses based only on isolated comments or aggregate negativity levels may overlook important social dynamics. In particular, focusing exclusively on

the most negative users may miss the interactional core formed by moderately negative users, who appear to sustain much of the conversational activity; a consideration that may be relevant for content moderation strategies that prioritize the most extreme cases while overlooking moderately negative, high-engagement discussions.

Limitations. Our findings should be interpreted as associations rather than causal effects. Although negativity is consistently related to structural, social, and linguistic differences, our analyses do not determine whether negativity causes cascades to grow or whether larger discussions become more negative over time. A second limitation concerns the sentiment classifier, fine-tuned on multilingual Twitter rather than Reddit-specific Brazilian Portuguese text [2]; as discussed in Sections 2 and 3, this likely introduced some misclassification on ironic, sarcastic, or slang-heavy comments [29], possibly attenuating quartile boundaries near the median negativity range.

Future Work. Future work could address the causality limitation identified above by examining the temporal evolution of negativity within cascades, identifying when negative comments emerge, whether they trigger subsequent replies, and how negativity increases or dissipates throughout a discussion. The classifier limitation could be addressed by fine-tuning or validating sentiment models directly on annotated Brazilian Reddit data, following the methodology proposed in [29], or by incorporating human-in-the-loop validation for cascades near quartile boundaries. Finally, future studies could investigate whether the non-monotonic pattern observed here generalizes to other languages, platforms, and cultural contexts, or whether it is specific to the structural and normative characteristics of Brazilian Reddit communities.

6 ETHICAL CONSIDERATIONS

This work complies with CNS Resolution No. 510/2016, as it did not involve direct contact with human subjects, consisting instead of the collection and analysis of publicly available discussion data from Reddit communities.

The dataset used in this study was collected from publicly accessible Reddit content and comprises comments, reply structures, and author identifiers from 20 Brazilian subreddits published between May 2025 and May 2026.

Considering the nature of the data, measures were adopted to protect personal data, in compliance with the Brazilian General Data Protection Law (LGPD, Law No. 13.709/2018). Author-identifying information associated with comments (e.g., usernames) was anonymized prior to analysis, and no such information is disclosed in this article. User-level analyses, including the construction of interaction networks and the computation of negativity and homophily profiles, were conducted exclusively at the aggregate level; no individual user can be identified from the results reported here.

The use of generative artificial intelligence technologies in this work was limited to spell-checking and grammar review of the text and was not employed in the generation of results or the writing of any section.

REFERENCES

- [1] Despoina Antonakaki and Sotiris Ioannidis. 2026. Cross-Platform Digital Discourse Analysis of the Israel-Hamas Conflict: Sentiment, Topics, and Event Dynamics. arXiv:2601.02367 doi:10.48550/ARXIV.2601.02367
- [2] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France.
- [3] 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., and Virgilio Almeida. 2025. Mapping Hate and Extremism: How Brazilian Reddit Communities Spread Toxic Discourse. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM '25)*. AAAI. doi:10.36190/2025.18
- [4] Carmela Bernardo, Dora Ricci, and Francesco Vasca. 2026. Homophily Detection in Opinion Dynamics from Reddit Discussions. In *Advances in Social Networks Analysis and Mining*, I-Hsien Ting, Panagiotis Karampelas, Min-Yuh Day, and Reda Alhajj (Eds.). Springer Nature Switzerland, Cham, 251–262.
- [5] Josemar A. Caetano, Helder S. Lima, Mateus F. Santos, and Humberto T. Marques-Neto. 2018. Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election. *Journal of Internet Services and Applications* 9, 1 (2018), 18. doi:10.1186/s13174-018-0089-0
- [6] Josemar Alves Caetano, Gabriel Magno, Marcos Gonçalves, Jussara Almeida, Humberto T. Marques-Neto, and Virgilio Almeida. 2019. Characterizing Attention Cascades in WhatsApp Groups. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 27–36. doi:10.1145/3292522.3326018
- [7] Flavio Carvalho, Rafael Rodrigues, Gabriel Santos, Pedro Cruz, Lilian Ferrari, and Gustavo Guedes. 2019. Avaliação da versão em português do LIWC Lexicon 2015 com análise de sentimentos em redes sociais. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining* (Belém). SBC, Porto Alegre, RS, Brasil, 24–34. doi:10.5753/brasm.2019.6545
- [8] Baitong Chen, Satoshi Tsutsui, Ying Ding, and Feicheng Ma. 2017. Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics* 11, 4 (2017), 1175–1189. doi:10.1016/j.joi.2017.10.003
- [9] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118. doi:10.1073/pnas.2023301118
- [10] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Jiatao Ji, Guillaume Agrawal, Francisco Guzman, Rahul Goyal, Ritesh Kumar, Mylène Ott, et al. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).
- [11] Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *Comput. Surveys* 55, 13s (2023), 1–38.
- [12] Antônio Pereira De Souza Júnior, Pablo Cecilio, Felipe Viegas, Washington Cunha, Elisa Tuler De Albergaria, and Leonardo Chaves Dutra Da Rocha. 2022. Evaluating Topic Modeling Pre-processing Pipelines for Portuguese Texts. In *Proceedings of the Brazilian Symposium on Multimedia and the Web (Curitiba, Brazil) (WebMedia '22)*. Association for Computing Machinery, New York, NY, USA, 191–201. doi:10.1145/3539637.3557052
- [13] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J. Watts. 2016. The Structural Virality of Online Diffusion. *Management Science* (2016). doi:10.1287/mnsc.2015.2158
- [14] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. doi:10.48550/ARXIV.2203.05794
- [15] Winston Haynes. 2013. *Bonferroni Correction*. Springer New York, New York, NY, 154–154. doi:10.1007/978-1-4419-9863-7_1213
- [16] Jack Hessel, Lillian Lee, and David Mimno. 2017. Cats and Captions vs. Creators and the Clock: Comparing Multimodal Content to Context in Predicting Relative Popularity. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, 927–936. doi:10.1145/3038912.3052684
- [17] Abbas K. Rizi, Riccardo Michielan, Clara Stegehuis, and Mikko Kivelä. 2025. Homophily within and across groups. *Nature Communications* 16, 1 (Dec. 2025). doi:10.1038/s41467-025-66415-2
- [18] Jordan K. Kobellarz and Thiago H. Silva. 2022. Should We Translate? Evaluating Toxicity in Online Comments when Translating from Portuguese to English. In *Proceedings of the Brazilian Symposium on Multimedia and the Web (Curitiba, Brazil) (WebMedia '22)*. Association for Computing Machinery, New York, NY, USA, 89–98. doi:10.1145/3539637.3556892
- [19] William H. Kruskal and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621. doi:10.1080/01621459.1952.10483441
- [20] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. ACM Press, 933–943. doi:10.1145/3178876.3186141
- [21] Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2024. Toxic Content Detection in online social networks: a new dataset from Brazilian Reddit Communities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, Pablo Gamallo, Daniela Claro, António Teixeira, Livy Real, Marcos Garcia, Hugo Gonçalo Oliveira, and Raquel Amaro (Eds.). Association for Computational Linguistics, Santiago de Compostela, Galicia/Spain, 472–482. <https://aclanthology.org/2024.propor-1.48/>
- [22] Frank J. Massey. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Amer. Statist. Assoc.* 46, 253 (1951), 68–78. <http://www.jstor.org/stable/2280095>
- [23] Alexey N. Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. 2019. The Anatomy of Reddit: An Overview of Academic Research. In *Dynamics On and Of Complex Networks III*. Springer International Publishing, 183–204. doi:10.1007/978-3-030-14683-2_9
- [24] Gabriel Melo and Flavio Figueredo. 2025. What makes a sentence toxic? A critical analysis of specialist models for toxicity detection. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web (Rio de Janeiro/RJ)*. SBC, Porto Alegre, RS, Brasil, 376–384. doi:10.5753/webmedia.2025.15684
- [25] Corrado Monti, Jacopo D'Ignazi, Michele Starnini, and Gianmarco De Francisci Morales. 2023. Evidence of Demographic rather than Ideological Segregation in News Discussion on Reddit. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*. Association for Computing Machinery, New York, NY, USA, 2777–2786. doi:10.1145/3543507.3583468
- [26] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David AL Levy, and Rasmus Kleis Nielsen. 2025. Reuters Institute Digital News Report 2025. Retrieved June 27, 2026 from <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025>
- [27] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate G. Blackburn. 2015. The Development and Psychometric Properties of LIWC2015. <https://api.semanticscholar.org/CorpusID:151038946>
- [28] Giovana Piorino, Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2025. Toxicidade e Gatilhos: Um Estudo de Caso em Comunidades do Reddit no Brasil. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web (Rio de Janeiro/RJ)*. SBC, Porto Alegre, RS, Brasil, 575–579. doi:10.5753/webmedia.2025.15984
- [29] Giovana Piorino, Vitor Moreira, Luiz Henrique Quevedo Lima, Ana Clara Souza Pagano, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2025. Sentiment Analysis of Shared Content in Brazilian Reddit Communities. *Journal on Interactive Systems* 16, 1 (Aug. 2025), 666–686. doi:10.5753/jis.2025.5564
- [30] Steve Rathje, Jay J. Van Bavel, and Sander van der Linden. 2021. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences* 118, 26 (2021), e2024292118. doi:10.1073/pnas.2024292118
- [31] Francisco Ricarte Neto, Rafael Anchieta, Raimundo Moura, Pedro Santos Neto, and Andre Santana. 2025. Investigating Contextual Word Embeddings in Semi-Supervised Learning for Toxic Comment Detection. In *Proceedings of the 31st*

- Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 330–339. doi:10.5753/webmedia.2025.15994
- [32] Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1086–1097. doi:10.1145/3442381.3449861
- [33] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A Survey of Multimodal Sentiment Analysis. *Image and Vision Computing* 65 (2017), 3–14.
- [34] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Intelligent Systems*, Ricardo Cerri and Ronaldo C. Prati (Eds.). Springer International Publishing, Cham, 403–417.
- [35] Fernanda Luiza Tobias and Ana Paula Couto da Silva. 2025. Linguistic Analysis of Human and Language Model Comments on Posts in Brazilian Reddit Communities. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 521–525. doi:10.5753/webmedia.2025.15908
- [36] Lanqin Yuan, Philipp J. Schneider, and Marian-Andrei Rizoiu. 2025. Behavioral Homophily in Social Media via Inverse Reinforcement Learning: A Reddit Case Study. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*. Association for Computing Machinery, New York, NY, USA, 576–589. doi:10.1145/3696410.3714618
- [37] Yiming Zhu, Gareth Tyson, and Pan Hui. 2026. A Comparative Analysis of Social Network Topology in Reddit and Moltbook. arXiv:2602.13920 [cs.SI] <https://arxiv.org/abs/2602.13920>