The Hidden Rise of Toxicity: How Retweets Obscure Increasing Hostility in Brazilian Politics

Soroush Karimi
University of Exeter
Exeter, United Kingdom
sk931@exeter.ac.uk

Marcos Oliveira
University of Exeter
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
m.a.oliveira@vu.nl

Diogo Pacheco
University of Exeter
Exeter, United Kingdom
d.pacheco@exeter.ac.uk

Abstract

Toxicity on social media is often assessed using aggregate trends, yet different forms of engagement may shape these patterns in distinct ways. We analyse approximately 100 million Twitter posts collected during the 2018 Brazilian presidential election and the following year to examine how engagement type, automation, and user activity relate to online toxicity. Using a fine-tuned toxicity model, we find that 12% of posts are classified as toxic at a 0.5 threshold (6% at 0.8). Although aggregate toxicity appears relatively stable over time, this masks diverging engagement dynamics. Retweets, which account for about 60% of posts, are consistently less toxic (relative risk $RR \approx 0.5$) and exhibit a slight downward trend. In contrast, replies are twice as likely to be toxic ($RR \approx 2$) and increase significantly over time. Replies to toxic parent posts are themselves more likely to be toxic ($RR \approx 1.6$), indicating conversational propagation of hostility. Contrary to the common assumption that automated accounts primarily amplify content through retweeting, we find that higher automation levels are associated with fewer retweets and more replies and original tweets. Consistent with the higher toxicity of replies, automated accounts are disproportionately more likely to post toxic replies. By contrast, highly active accounts favour retweeting behaviour and, due to this amplification pattern, are not the primary drivers of toxicity. These findings demonstrate that rising hostility is concentrated in conversational exchanges and partially obscured by large volumes of low-toxicity amplification, challenging simplified assumptions about automation and online toxicity.

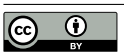
CCS Concepts

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → *Lexical semantics*.

Keywords

Online toxicity, Social bots, User activity, Political discourse, User engagement

ACM Reference Format:
Soroush Karimi, Marcos Oliveira, and Diogo Pacheco. 2026. The Hidden Rise of Toxicity: How Retweets Obscure Increasing Hostility in Brazilian Politics. In *18th ACM Web Science Conference (WebSci '26), May 26–29, 2026, Braunschweig, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3795766.3799769



WebSci '26, Braunschweig, Germany

ACM ISBN 979-8-4007-2504-3/2026/05
https://doi.org/10.1145/3795766.3799769

Introduction

Social media platforms create conditions for individuals to express their thoughts more openly than they might in real-world situations. On many of these platforms, individuals can remain anonymous and share their preferred content without fear of judgment or repercussions. Although this phenomenon may encourage more open expression, it also contributes to what is known as "toxic disinhibition," which fosters the spread of hostility and hate speech [1]. Though online, toxic discourse is tied to real-world conditions: it can reflect the state of real-world tensions and social issues [2, 3] and it can provoke real-world harm against its victims [4], even when based on fabricated claims [5]. Yet the state of online toxicity, such as insults and hate speech, is difficult to track, as large volumes of activity can obscure underlying trends. In this work, we investigate how online toxicity evolves, how repost volume can hide toxicity trends, and how activity levels and bot-like accounts influence the toxic landscape.

Online toxicity has been studied from multiple perspectives. In terms of user attributes, research shows that verified users are less likely to engage in toxic replies [6], while toxic language does not necessarily discourage participation in long conversations [7]. Another important aspect is the relationship between online toxicity and real-world events: offline incidents can trigger broad cascades of online hate speech, such as the murder of George Floyd on May 25, 2020 [2]. Brazil provides a particularly salient example of this dynamic, where political polarisation intensified in the aftermath of the 2015 elections and the impeachment of President Dilma Rousseff [8]. This polarised atmosphere worsened during the 2018 presidential cycle, as incidents of physical violence intensified disputes between conservative groups, which are frequently characterized by their opposition to the Workers' Party (PT), and liberal left-wing groups. In polarised environments, automated accounts, or social bots, also shape online discourse by amplifying specific narratives and spreading disinformation [9, 10]. Some studies suggest that bots exhibit higher levels of toxicity than human users [11], while others report they may be less toxic [12]. In this political context, the roles of bots, the roles of bots [13], hate speech trends [14], and the network of anonymous social media accounts that coordinated Jair Bolsonaro's aggressive hate speech efforts, which magnified his and his family's sexist and misogynistic remarks [15], are particularly evident. However, limited research has systematically compared how toxicity manifests across different

interaction mechanisms, such as replies versus retweets, or disentangled the specific influence of automated behaviour from high posting frequency.

Despite significant progress in analysing online toxicity, important aspects such as engagement dynamics and account activity levels remain poorly understood. Similarly, existing studies do not fully clarify the role of bot-like accounts in shaping toxic interactions. To the best of our knowledge, the study of how repost volume might mask underlying trends in original toxic content, particularly in political contexts, has not yet been explored.

Here, we investigate how user behaviour and engagement types relate to the spread of toxic content on Twitter throughout the 2018 Brazilian presidential campaign and the subsequent months, providing a more nuanced view of how toxic content propagates and evolves. We compare the behaviour of bot-like and human-like accounts, as well as different levels of user engagement. We address three key research questions: (1) How does online toxicity evolve in original tweets compared with different engagement types (i.e., retweets, replies, and quotes)? (2) How does the rate of toxic content differ between posts shared by bot-like and human-like accounts? (3) Does the frequency of account activity correlate with the toxicity of the posts they share?

Our findings reveal that the trends of toxicity across different engagement types do not follow the original tweets in terms of trend and level. Specifically, retweets are less toxic than the rest of the post types, suggesting that more toxic content tends to receive less engagement of this kind. Conversely, replies to toxic posts are more likely to be toxic than replies to non-toxic posts. The analysis shows that bot-like accounts exhibit a disproportionately higher likelihood of producing toxic replies. In contrast, increased activity levels are associated with a shift toward more retweets and fewer replies, resulting in a lower likelihood of toxicity in replies but a higher likelihood in original tweets.

Related Work

Online toxicity trend: Research on online discourse among engaged partisans on Reddit highlights their disproportionate contribution to toxic content. In contrast, Saveski et al. [16] showed that on Twitter, toxicity is distributed among numerous low to moderately toxic users rather than concentrated among a small number of highly toxic users. Mamakos and Finkel [17] found that political discourse on social media is intensely toxic because it attracts dispositionally uncivil individuals who display toxic behaviour even in non-political contexts. Törnberg and Chueri [18] revealed a marked increase in toxic discourse among political elites over a five-year period. However, not all studies find evidence of a growing trend. Szabó et al. [19], in their dictionary-based quantitative content analysis of 17.6 million online comments in Hungary, reported that the frequency of incivility remained steady over the three-year period examined. Similarly, Goovaerts and Turkenburg [20], analysing televised election debates in Belgium over three decades, found that levels of incivility fluctuated but did not follow a consistent upward trajectory. Sun et al. [21] also noted that when considering Reddit as a whole, the overall proportion of incivility has remained relatively constant across time.

Bot-like accounts and online toxicity: Xu et al. [22] analysed human–bot engagement on Twitter and Reddit and showed that automated accounts are significant spreaders of toxic content, often acting more aggressively and exerting greater influence than human users. Similarly, Uyheng and Carley [23], in their study of Twitter conversations in the United States and the Philippines during the COVID-19 pandemic, revealed that bots play a complex role in amplifying online hate speech. While the dynamics varied between the two countries, their findings establish a clear link between bot activity and higher levels of hate, particularly within dense and isolated online communities. In another study, Stella et al. [11] show that bots operating from the sidelines target key human users. These bots specifically flood some groups with violent posts, increasing their exposure to negative, inflammatory information and exacerbating online conflict. To compare the bot-like and human-like accounts, Rossetti and Zaman defined toxic users whose average toxicity score was greater than or equal to 0.9 [12], and reported that bots may display lower levels of toxicity than human users. In our research, we considered different thresholds for defining the toxic account, and the result was not robust, so we needed to use another measurement. Instead of analysing individual accounts, we grouped accounts into bot-like and human-like categories and compared the probability of sharing toxic posts between these groups. This comparison is quantified as the relative risk of sharing toxic posts.

Activity level and online toxicity: Blumer and Kleinberg [24] showed that the relationship between activity and toxicity is not universal but platform-dependent: on Reddit, the most active users tend to produce more toxic content in aggregate, whereas on Wikipedia, the least active users are more likely to display toxic behaviour. from another perspective, Jiang et al. [25] showed communities characterised by users with diverse interests tend to have a lower level of toxicity.

Methods

Dataset

We analyse a one-year slice of a dataset containing tweets from August 30, 2018, to August 30, 2019, comprising around 100 million posts from about 5 million unique user accounts [26]. The data were gathered by tracking official accounts and hashtags associated with presidential candidates in the 2018 election cycle, with supplementary data derived from governmental accounts and hashtags (see Table 1). We use this dataset to investigate the spread of aggressive communicative behaviours during pivotal moments in Brazil's political context.

Toxicity Measurement

To assess toxicity in our dataset, we utilised a publicly available fine-tuned model developed by Leite et al. [27], built on the distilbert-base-multilingual-cased architecture, a lighter and more computationally efficient version of BERT. Leite et al. fine-tuned this model for binary classification of toxic language on the ToLD-Br dataset (Toxic Language Dataset for Brazilian Portuguese), a corpus containing 21,000 manually annotated tweets. In this framework, the notion of toxicity includes comments containing insults and obscene language, alongside hate speech. We use this model to

Table 1: List of the hashtags and keywords used for data gathering

Name	Account	Hashtag
Álvaro Dias	@alvarodias_	#AlvaroDias19
Cabo Daciolo	@CaboDaciolo	#Daciolo51
Ciro Gomes	@cirogomes	#Ciro12
José Maria Eymael	@Eymaeloficial	#Eymael27
Fernando Haddad	@Haddad_Fernando	#Haddad13
Geraldo Alckmin	@geraldoalckmin	#Alckmin45
Guilherme Boulos	@GuilhermeBoulos	#Boulos50
Henrique Meirelles	@meirelles	#Meirelles15
Jair Bolsonaro	@jairbolsonaro	#Bolsonaro17
João Amoêdo	@joaoamoedonovo	#Amoedo30
João Goulart Filho	@joaogoulart54	#Goulart54
Luiz Inácio Lula da Silva	@LulaOficial	#Lula13
Marina Silva	@MarinaSilva	#Marina18
Vera Lúcia	@verapstu	#Vera16
Superior Electoral Court	@TSEjusbr	#Eleições2018

generate a probability score for each tweet, which serves as a continuous measure of its likelihood of containing toxic content. We note that their fine-tuning methodology demonstrates strong performance, achieving a macro F1-score of up to 76% on the toxic language detection task.

To avoid recalculating the toxicity score for retweets, which contain the same content as the original tweets, we compute the toxicity score for the remaining records and then apply those scores to the retweets using their linked IDs. While the toxicity of a retweet is the same as the toxicity of the parent tweet, the cumulative toxicity of the retweet category can diverge significantly from that of original tweets, based on which posts with what toxicity scores are retweeted more. Furthermore, to analyse the toxicity of the posts to which replies were directed, we linked each reply in the dataset to its corresponding parent tweet.

Using Relative Risk to Highlight Toxicity Bias Across Categories

To compare the behaviour of different categories (e.g., bot-like vs. human-like accounts), we must decide how to define a toxic account. This required specifying a threshold for the least number of toxic posts that have been shared by an account. However, depending on the chosen threshold, either bot-like or human-like accounts can appear more toxic, making comparisons unstable.

To address this issue, we avoid comparing toxicity at the account level. Instead, we compare the probability of toxic posts within each category. For that, we use relative risk analysis to assess differences across categories [28].

Relative risk (RR) is the ratio of the probability of an event occurring in an exposed group to the probability of the same event in an unexposed group. The formula for relative risk is:

$$RR = \frac{a/(a+b)}{c/(c+d)},$$

where the variables correspond to: a is the number of individuals in the exposed group who experience the outcome, b is the number of individuals in the exposed group who do not experience the outcome, c is the number of individuals in the unexposed group who experience the outcome, and d is the number of individuals in the unexposed group who do not experience the outcome. This compares the cumulative incidence of the outcome in the exposed group to that in the unexposed group.

In our analysis, the "exposure" corresponds to membership in a specific category (e.g., a post type, activity level, or bot-score group), while the "outcome" corresponds to whether the post is classified as toxic. Accordingly:

- a: number of toxic posts within a given category,
- b: number of non-toxic posts within that category,
- c: number of toxic posts outside that category,
- d: number of non-toxic posts outside that category.

This formulation allows comparison of the likelihood of toxicity within each category relative to all others combined.

On Twitter, posts fall into four categories: tweets (original posts), retweets (shared posts), replies (responses to others), and quotes (shared posts with your own comment). We apply relative risk analysis to compare different post categories, activity-level groups, and bot-score groups. We retain the natural distribution of samples to represent the actual composition and ongoing dynamics within each category.

Bot-Score Calculation

To identify potentially automated accounts, we use the bot detection framework introduced by Yang et al. [29]. This method calculates a bot-score based on user profile metadata embedded within each tweet. According to Yang et al., the model achieves high internal validity (AUC = 0.98 in cross-validation) and robust generalization to unseen bot classes (AUC up to 0.99, F1 = 0.77 in cross-domain tests) through strategic training data selection. Using this framework allows for the evaluation of a user's profile as it existed at the moment the tweet was generated. The framework produces a score ranging from 0 to 1, where higher values indicate a higher likelihood that an account exhibits bot-like behaviour. For each

account, we used the average bot-score across their tweets to identify potentially automated accounts. To compare accounts in terms of bot-likeness or human-likeness, we classified them into three groups based on their average bot-score.

This categorisation facilitates the comparison of toxicity across user groups. Accounts in [0, 0.33) are those whose posts, on average, received a low bot-score and are more likely to be human. On the other hand, the accounts within [0.66, 1] are the accounts that are mostly recognised as bots. The rest of the accounts are in between [0.33,0.66).

User Activity Measurement

We use the total number of posts shared by each account (including all types of posts) to determine the activity level. This activity count is limited to posts captured through our data collection process, which relied on the political keywords and hashtags listed in Table 1. For our analysis, we consider users with at least 20 posts over the entire period. We group accounts into three categories based on their total number of posts over the study period:

- Low activity: 20 to 100 posts
- Medium activity: 100 to 1,000 posts
- High activity: more than 1,000 posts

Results

The results come in five parts. First, we examine the distribution of toxicity scores and the proportion of posts classified as toxic under different threshold definitions. Second, we analyse temporal trends in the proportion of toxic posts and how these trends vary across post types (i.e., tweets, retweets, quotes, and replies). Third, we characterize the share of posts in each category and compare the relative risk of toxicity across engagement types. Fourth, we compare the cumulative behaviour of bot-like and human-like accounts. Finally, we conduct a similar analysis across accounts grouped by activity level.

Most posts are non-toxic, but a small fraction is highly likely to be toxic

To assess the toxicity, we use the probability scores produced by the fine-tuned model described in the Methods section (Toxicity Measurement). The toxicity scores range from 0 (highly likely non-toxic) to 1 (highly likely toxic) [27].

We find that most posts have low toxicity scores, with only a small number scoring close to 1, and a roughly flat distribution in the middle (see Figure 1). By using a 0.5 threshold, we classify about 12% of posts as toxic, whereas a stricter 0.8 threshold yields only 6%. In the following analyses, we primarily use the 0.5 threshold, and we also use the 0.8 threshold to assess the robustness of the findings. Table 2 presents sample toxicity probability distributions derived from the dataset.

Aggregate toxicity is almost stable overall, but mainly because of the impact of the retweets

Online toxicity varies with day-to-day politics. To track these fluctuations, we measure toxicity on a daily basis to monitor general trends and variations in the online political environment over the

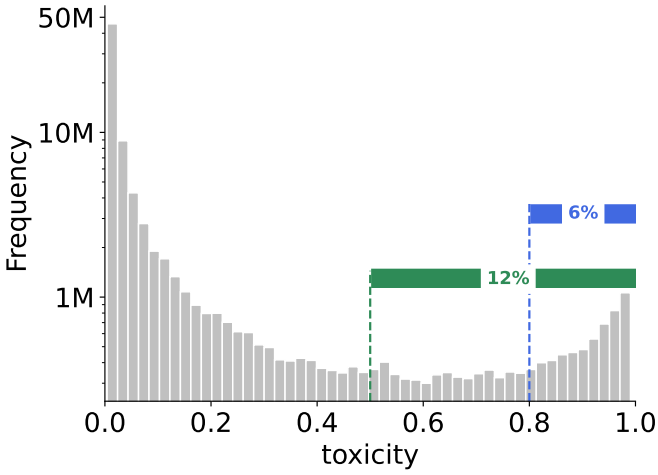


Figure 1: Most posts have low toxicity scores, while a smaller peak near 1 represents posts that are highly likely to be toxic. Toxicity ranges from 0 to 1, where higher values indicate a greater likelihood of the post being toxic. The fraction of posts classified as toxic decreases as the toxicity threshold increases. For instance, when the threshold is 0.5, 12% of posts are labelled toxic, whereas at 0.8, only 6% are classified as toxic.

one-year period, which includes the first and second rounds of the presidential election (Figure 2). Although they are mostly distributed around 12%, there are specific days that the percentage of toxicity reaches 20%. A linear regression model reveals a small but statistically significant upward trend in the percentage of toxic posts. Over the period, the percentage of toxic posts increased by approximately 1.34 units (coef. = 0.0038, p-value = 0.003).

However, when we examine post types separately, different patterns emerge (Figure 3). Our results reveal that replies show the strongest rise in toxicity, with a marked upward trend (coef. = 0.013, $p < 0.001$) and a 10-day moving average that remains higher than for any other post type. We also find that retweets move in the opposite direction: their toxicity declines slightly over time (coef. = -0.0042, p = 0.012) and stays consistently below that of original tweets, suggesting a preference for resharing less toxic content. We note that the gap between tweet and retweet toxicity is small at first but widens over time, especially after the second round of elections, indicating a growing divergence in how users tweet versus retweet toxic content. We further show that original tweets become more toxic over time, increasing by about 3.74 units (coef. = 0.0106, $p < 0.001$), roughly 2.8 times the aggregate trend. Quoted posts follow a similar significant upward trend (coef. = 0.0079, p-value < 0.001).

Toxicity varies across post types, with replies being more likely to be toxic than all other types

Next, we characterise how toxicity is distributed across post types. When we consider all posts, regardless of toxicity, most of the activity arises from engagement with existing content rather than from new posts. Only about 4% of posts are original tweets; the rest

Table 2: A sample set of tweets is categorized into five levels based on their toxicity score, with English translations.

Probability Level	Sample (Portuguese)	English Translation
0 - 0.2	@GuilhermeBoulos Em 2015 o Museu Nacional fechou as portas por atrasos no repasse de verbas federais. Tu lembra, quem era presidente?	In 2015 the National Museum closed its doors due to delays in federal funding. Do you remember who was president?
0.2 - 0.4	Até o Ciro Gomes de vez em quando não fala merda https://t.co/fVaeChul7Z	Even Ciro Gomes doesn't talk nonsense every now and then.
0.4 - 0.6	@geraldoalckmin Vc foi um péssimo governador e não será presidente!!!	You were a terrible governor and you won't be president!!!
0.6 - 0.8	@anaamelialemos @geraldoalckmin Coitados!!!	Poor things!!!
0.8 - 1	O @geraldoalckmin ataca Bolsonaro porque falou para uma idiota que ela era idiota. Respondeu ao ataque da Maria do Rosário como deveria. O Geraldo ataca, mas se olhar no espelho verá que só um é corrupto, Ladrão de merenda e 'santo' da Odebrecht. O chuchu não tem moral.	@geraldoalckmin attacks Bolsonaro because he told an idiot that she was an idiot. He replied to Maria do Rosário's attack as he should have. Geraldo attacks, but if he looks in the mirror he'll see that only one is corrupt, a school-meal thief and Odebrecht's 'saint.' The "chuchu" has no morals.

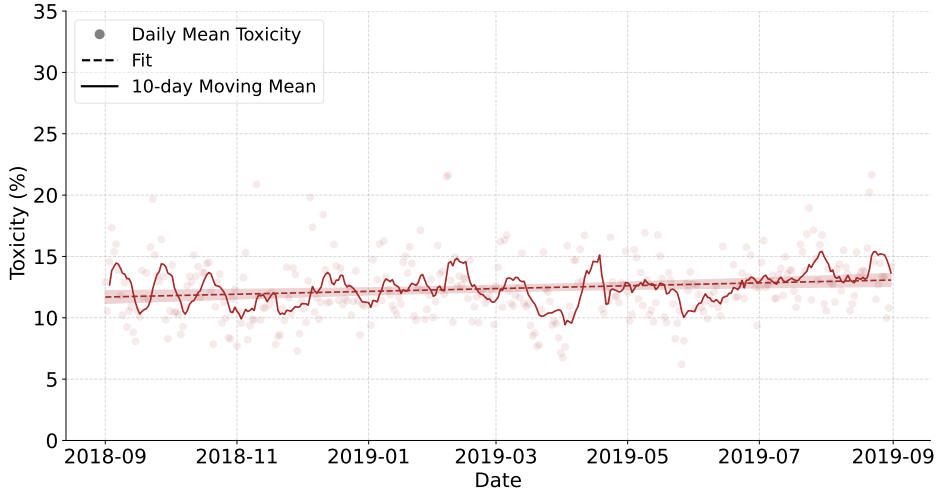


Figure 2: There is a small but statistically significant upward trend in the percentage of toxic posts (coef. = 0.0038, p-value = 0.003). Daily percentage of toxic posts, 10-day rolling average, and linear regression trend line. Each dot represents the percentage of toxic posts for a single day.

come from engagement with existing posts. Retweets are the most common engagement type, in which a user shares someone else's post without adding new content; they account for approximately 60% of all records (Figure 4). Because retweets repeat the content of the original tweets, their toxicity is identical to that of the original. Replies are the second most frequent type; they are a direct response to another user's post within a conversation thread. Finally, quotes form the third category, in which a user shares another post while adding their own comment or context.

When we focus on toxic posts, we find that the overall ranking of engagement types remains the same. However, their toxicity composition differs: retweets contribute a smaller share of toxic content than expected from their volume, while replies contribute a larger share (see Figure 4).

To quantify the significance of these differences, we use relative risk (RR), which compares the probability of a post being toxic within a given category (the exposed group) to the probability of a post being toxic in all other categories (the unexposed group) (see Methods). We find clear differences across post types. Replies have the highest relative risk (around 2 at a 0.5 toxicity threshold), meaning they are twice as likely to be toxic compared to other types of posts. In contrast, retweets have the lowest relative risk (around 0.5), indicating they are only half as likely to be toxic (Figure 5). The remaining categories have relative risk values close to 1, suggesting toxicity levels similar to the overall average. We repeated the analysis using a higher toxicity threshold, and although the absolute values changed, the ranking of post types remains robust.

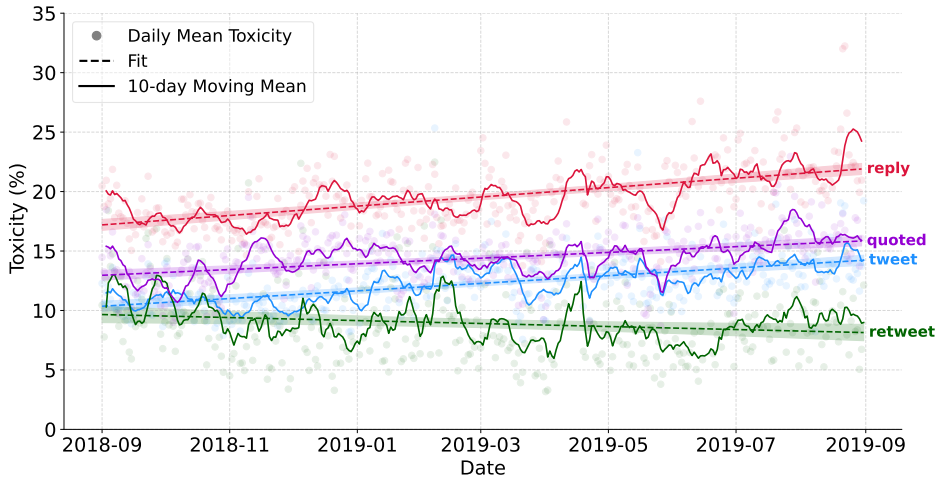


Figure 3: The overall trend for retweets is negative, which is opposite the rest. The 10-day moving mean toxicity of retweets is consistently lower than that of tweets, while replies are more toxic.

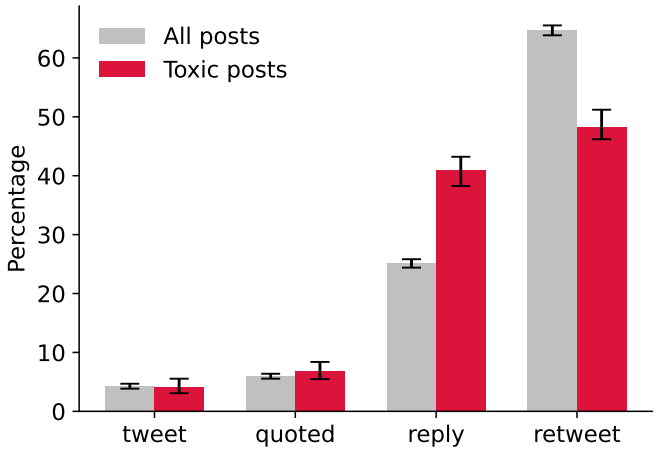


Figure 4: Replies, the second most common post type, contain a disproportionately high share of toxic content, whereas retweets contribute less toxicity than expected. Tweets and quotes show toxic shares matching their overall proportions.

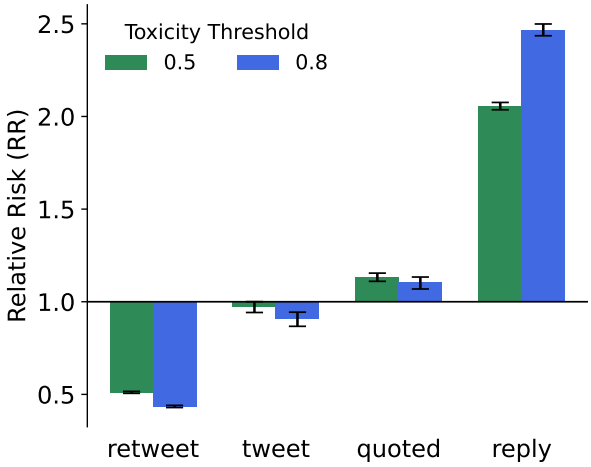


Figure 5: Replies are relatively more toxic, with a relative risk of around 2 at a toxicity threshold of 0.5, whereas retweets are less toxic, with a relative risk of approximately 0.5. The patterns are similar for different toxicity thresholds.

Given the prominence of replies, we examine the toxicity of the posts they reference. We find that replies are twice as likely to be toxic as the parent posts (20% for replies and 10% for the parent posts). Both toxic and non-toxic parents can attract toxic replies, but not to the same extent: 29% of replies to toxic posts are toxic, compared with 18% of replies to non-toxic posts are toxic (Figure 6.A) . Relative risk confirms this asymmetry: replies to toxic posts are substantially more likely to be toxic (RR $\approx$ 1.6), whereas replies to non-toxic posts are less likely (RR $\approx$ 0.6) (Figure 6.B). Taken together, these findings suggest that toxicity in parent posts increases the likelihood that toxicity will propagate into the reply thread.

More automated accounts are disproportionately toxic in replies

The results for RQ1 showed that toxicity varies systematically by engagement type: retweets, which constitute the majority of posts, are consistently less toxic, while replies are substantially more likely to contain toxic content. This pattern suggests that toxicity is not uniformly distributed across interactions, but instead concentrated in conversational exchanges.

Building on this finding, we examine whether account automation is associated with different engagement styles and, consequently, different toxicity patterns. Accounts were grouped into three categories according to their mean bot-score (see Methods).

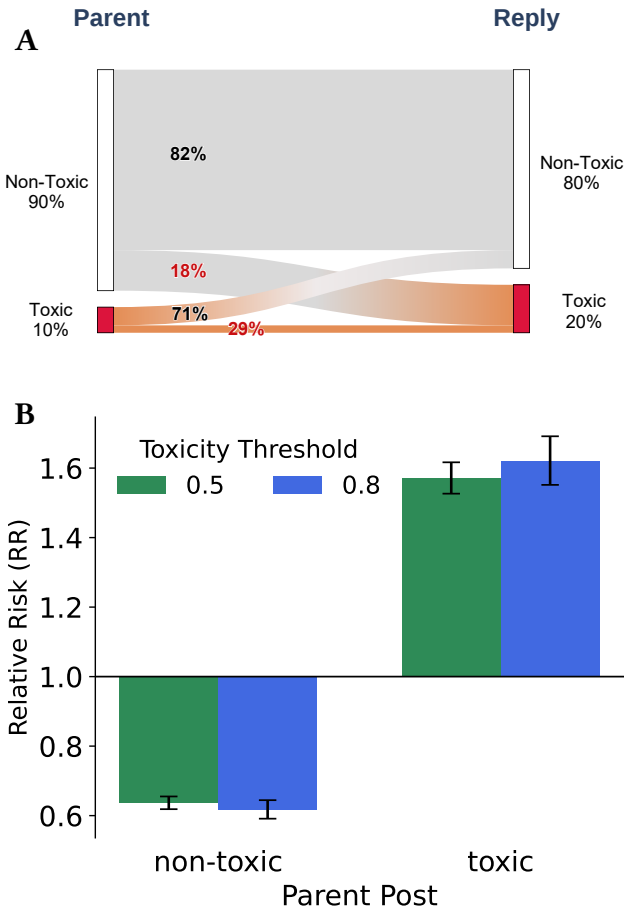


Figure 6: (A) Replies that are responses to toxic parent posts are more likely to be toxic compared to replies to non-toxic posts. Both toxic and non-toxic parent posts can receive toxic replies, although not in the same proportions. (B) The relative risk of toxicity for replies to toxic and non-toxic parent posts (sample size = 100,000, 100 iterations) shows that toxic content evokes more toxicity, with relative risk values of approximately 1.6 for toxic parents and 0.6 for non-toxic parents (for both toxicity thresholds).

The majority of accounts (approximately 71%) have an average bot-score below 0.33 and are classified as human-like, whereas only about 2% fall above 0.66 and are classified as bot-like.

We first observe that bot-score levels are associated with distinct engagement patterns. As the average bot-score increases, accounts tend to produce proportionally fewer retweets and relatively more original tweets and replies (see Figure 8.A). In other words, more automated accounts are less likely to amplify existing content and more likely to participate directly in conversational exchanges. This finding aligns with prior work showing that bots disproportionately participate in reply-based interactions [30].

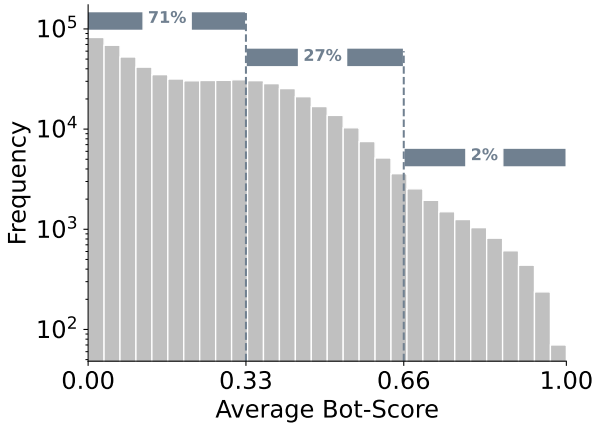


Figure 7: Distribution of average bot-score. Most accounts have an average bot-score below 0.33, suggesting the majority of them are human-like accounts.

Given that replies are the most toxic engagement type (RQ1), this shift in engagement style has important implications for toxicity exposure. To assess whether automation is associated with a higher likelihood of toxic posting, we compute the relative risk (RR) of toxicity within each engagement type (see Figure 8.C). We find that the association between automation and toxicity is not uniform across post types. For retweets, toxicity remains comparatively low across all bot-score categories. However, within replies, the likelihood of posting toxic content increases with bot score. Accounts with higher average bot scores exhibit substantially elevated relative risk of toxicity in replies.

These results indicate that automation is linked to toxicity primarily through engagement behaviour. More automated accounts are both more likely to engage in replies, the interaction type most prone to toxicity, and more likely to post toxic content within that context. Rather than suggesting that bots are uniformly more toxic, the findings reveal an interaction between automation level and engagement type: toxicity differences are concentrated in conversational exchanges rather than in content amplification.

High activity shifts behaviour toward retweets, relatively less toxic in replies

More active accounts engage more frequently with content, which raises the possibility that they may also share more toxic content. Conversely, accounts that post less often may be more selective and avoid toxic content. To assess this, we examine the relationship between user activity and the toxicity of their posts. We define activity as the total number of posts shared by a user (see Methods section "User Activity Measurement"). Following our analysis of automation, we examine the relationship between account activity level and engagement styles, and how it is linked to toxicity.

Most of the accounts (74%) are those with a low level of activity (20 to 100 posts), while a small fraction of accounts (approximately 2%) exhibit very high activity with over 1,000 posts, sometimes reaching nearly 10,000 total shares (see Figure 9).

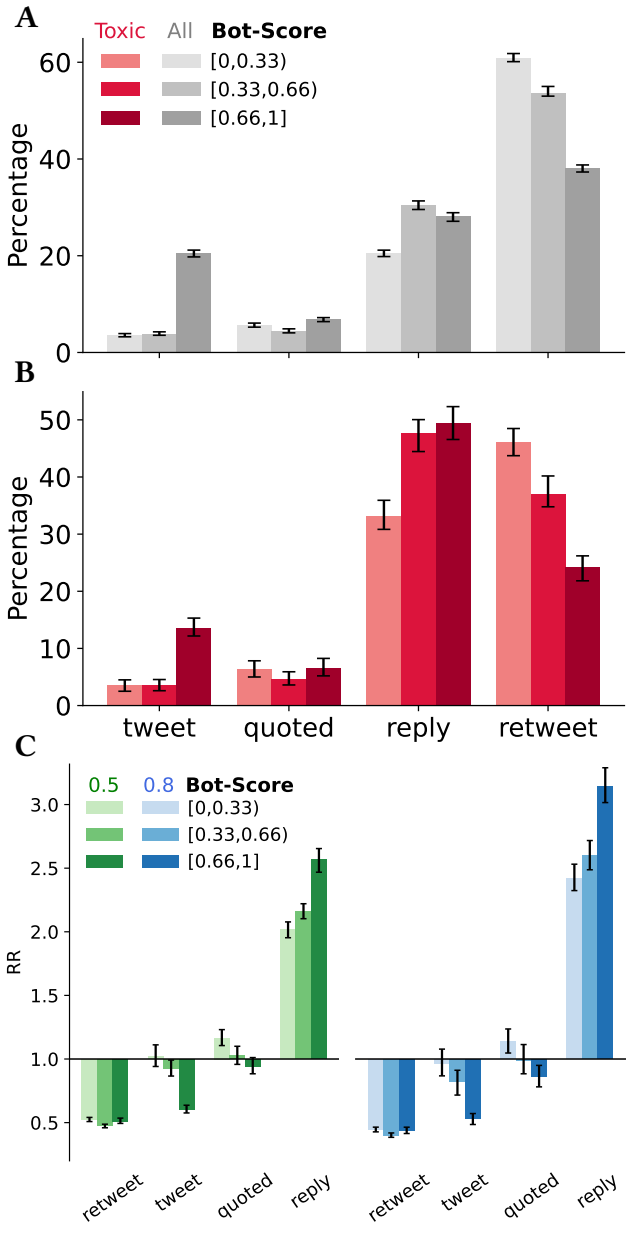


Figure 8: (A) Bot-like accounts are relatively more active in original tweets and replies, and less in retweets. (B) The share of toxic posts follows a pattern similar to that of all posts. (C) Higher bot scores are linked to an increased relative risk of toxicity in replies and a decrease in tweets. Toxicity in retweets remains low across all bot-score categories (sample size = 100,000, 100 iterations).

By comparing the engagement patterns of accounts, we notice that high-activity accounts tend to produce proportionally more retweets, and fewer replies and tweets. This suggests that highly active accounts mainly amplify existing content rather than engage in direct conversational exchanges (see Figure 10.A). For the share of toxic content across different post categories, the pattern is similar to all posts (see Figure 10.B).

To compare the behaviour of different activity levels in each post category in terms of sharing toxic content, we calculate the relative risk (RR) of toxicity for each type of engagement (see Figure 10.C). Although replies have the highest RR overall, this value is lower in the high-activity group. For the rest of the post categories, the pattern is the opposite: as activity increases, their relative risk increases.

These results suggest that while high-activity accounts generate a large volume of content, their contribution to toxicity is primarily moderated by their engagement behaviour. Because these accounts prefer retweeting, the engagement type least likely to be toxic, their overall impact on toxicity may be less than their posting volume suggests. Toxicity continues to be most concentrated in replies, which constitute a larger relative portion of the activity for accounts that post less frequently.

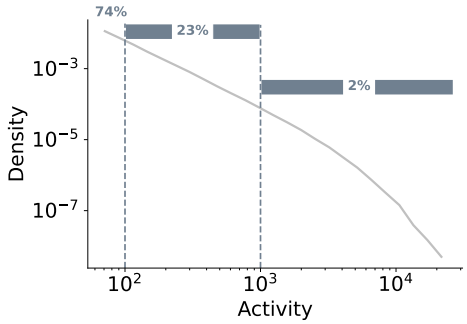


Figure 9: Distribution of account activity. Most of the accounts (74%) have 20-100 posts.

Discussion

This study explores the dynamics of toxicity on Twitter over the course of a year, focusing on patterns of user engagement and account activity. Our analysis addressed three key research questions: (1) How does online toxicity evolve in original tweets compare with different engagement types (i.e., retweets, replies, and quotes)? (2) How does the rate of toxic content differ between posts shared by bot-like and human-like accounts? (3) Does the frequency of user activity correlate with the toxicity of the posts they share?

We observed a slight upward trend in the percentage of toxic posts, which is roughly three times as strong for original tweets as for all posts. This difference is explained by trends in retweets: since retweets constitute the majority of shared posts and show a small decreasing trend, they tend to mask the increase observed in original tweets. As retweets replicate the content of the originals, their lower toxicity suggests a bias toward reposting, with lower-toxicity tweets more likely to be shared.

Replies, the second most common type of engagement, showed the highest percentage of toxic posts, with an upward trend over time. This pattern may be because users are more likely to reply when they disagree with a post, leading to more aggressive responses. More careful analysis of the parent posts on the replies

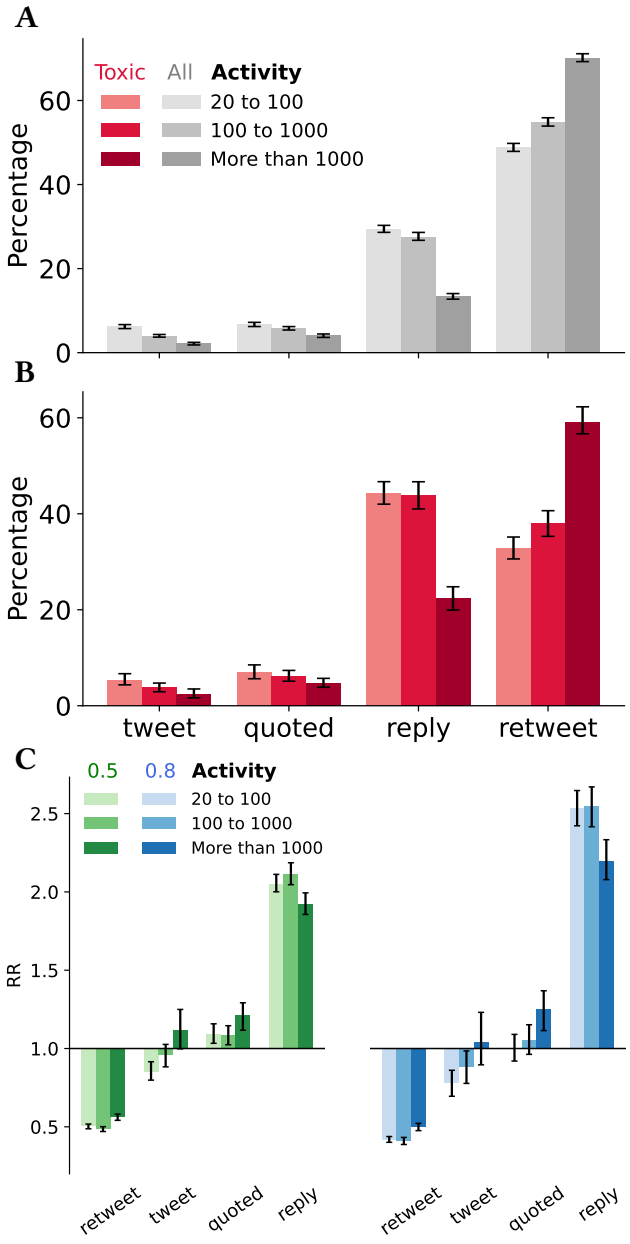


Figure 10: (A) Highly active accounts prefer retweeting (content amplification) over original posting or replying. We included the accounts with at least 20 posts in our analysis. (B) Toxic content follows a distribution similar to that of total posts. (C) High activity correlates with a slightly higher risk of toxicity in original tweets and quotes. Toxicity risk in replies remains high for all activity levels, though most active users show a slightly lower risk at the 0.8 threshold (sample size = 100,000, 100 iterations).

revealed that replies to toxic parent posts are more likely to be toxic than replies to non-toxic parents. Future analyses could examine whether replies are concentrated toward specific account clusters, as network analysis may reveal structural patterns associated with toxic replies.

Our results also indicate that bot-like accounts are more likely to post tweets in comparison to human-like accounts, while relatively more toxic in replies. Moreover, our analysis on different activity levels shows that the most active accounts retweet more, which is the least toxic category of the posts, and the relative risk of replies is less in comparison to the rest. These findings show that social media toxicity is complex, highlighting the importance of considering both user engagement patterns and account types, human or bot, when assessing harmful content.

Despite using a model fine-tuned on the ToLD-Br dataset, which is specific to Brazilian Portuguese on Twitter and suitable for our study, limitations remain, as context-dependent irony, evolving slang, and subtle discourse nuances may still lead to misclassification of toxicity.

Finally, for future work, we intend to analyse the network features influencing the spread of toxicity, as simulation results suggest that misinformation tends to diffuse in dense communities [31]. Further steps in our study include analysing the trigger events that lead to sharp increases in online toxicity, and the sources of more toxic replies. Understanding the interplay between original content and reply behaviour will deepen our insight into the mechanisms driving online toxicity.

Conclusion

In conclusion, our results show that engagement structure plays a central role in shaping observed toxicity on Twitter. Although aggregate toxicity appears relatively stable, this masks diverging dynamics across interaction types: toxicity increases in original tweets, quotes, and especially replies, while retweets, the dominant form of engagement, exhibit a slight decline in toxicity. As a result, large volumes of low-toxicity retweets obscure rising hostility within conversational exchanges.

We further show that toxicity propagates within discussions: replies to toxic parent posts are substantially more likely to be toxic than replies to non-toxic parents. Automation contributes to this pattern not through amplification, but through engagement style. More automated accounts are disproportionately active in replies and exhibit higher relative toxicity within that interaction type compared to human-like accounts. In contrast, highly active accounts predominantly engage in retweeting behaviour and are therefore not the primary drivers of toxicity.

By analysing toxicity across engagement categories rather than relying solely on account-level classifications, we provide a more robust framework for understanding how hostility emerges and spreads. These findings highlight the importance of focusing on conversational dynamics and interaction patterns when assessing and addressing online toxicity.

References

[1] Suler, J. The online disinhibition effect. *Cyberpsychology & behavior* **7**, 321–326 (2004).

[2] Lupu, Y. *et al.* Offline events and online hate. *PLoS one* **18**, e0278511 (2023).

[3] Tufa, W. T., Markov, I. & Vossen, P. Grounding toxicity in real-world events across languages. In *International Conference on Applications of Natural Language to Information Systems*, 197–210 (Springer, 2024).

[4] Müller, K. & Schwarz, C. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association* **19**, 2131–2167 (2021).

[5] Dreißigacker, A., Müller, P., Isenhardt, A. & Schemmel, J. Online hate speech victimization: consequences for victims' feelings of insecurity. *Crime Science* **13**, 4 (2024).

[6] Aleksandric, A., Roy, S. S. & Nilizadeh, S. Twitter users' behavioral response to toxic replies. *arXiv preprint arXiv:2210.13420* (2022).

[7] Avalle, M. *et al.* Persistent interaction patterns across social media platforms and over time. *Nature* **628**, 582–589 (2024).

[8] Giacomozzi, A. I., Gomes Fiorott, J., Bertoldo, R. & Contarello, A. Social representations of political polarization through traditional media: a study of the brazilian case between 2015 and 2019. *Human Affairs* **33**, 67–81 (2023).

[9] Salles, D., de Medeiros, P. M., Martins, B., Regattieri, L. & Santini, R. M. The role of social bots in the brazilian environmental debate:: an analysis of the 2020 amazon forest fires in twitter. *The International Review of Information Ethics* **33** (2024).

[10] Caldarelli, G., De Nicola, R., Del Vigna, F., Petrocchi, M. & Saracco, F. The role of bot squads in the political propaganda on twitter. *Communications Physics* **3**, 81 (2020).

[11] Stella, M., Ferrara, E. & De Domenico, M. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* **115**, 12435–12440 (2018).

[12] Rossetti, M. & Zaman, T. Bots, disinformation, and the first impeachment of us president donald trump. *Plos one* **18**, e0283971 (2023).

[13] Santini, R. M., Salles, D. & Tucci, G. When machine behavior targets future voters: the use of social bots to test narratives for political campaigns in brazil. *International Journal of Communication* **15**, 1220–1223 (2021).

[14] Ribeiro, M., Calais, P., Santos, Y., Almeida, V. & Meira Jr, W. Characterizing and detecting hateful users on twitter. In *Proceedings of the international AAAI conference on web and social media*, vol. 12 (2018).

[15] de Albuquerque, A. & Alves, M. Bolsonaro's hate network: From the fringes to the presidency. *86272* **12**, 27–42 (2023).

[16] Saveski, M., Roy, B. & Roy, D. The structure of toxic conversations on twitter. In *Proceedings of the web conference 2021*, 1086–1097 (2021).

[17] Mamakos, M. & Finkel, E. J. The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS nexus* **2**, pgad325 (2023).

[18] Törnberg, P. & Chueri, J. Elite political discourse has become more toxic in western countries. *arXiv preprint arXiv:2503.22411* (2025).

[19] Szabó, G., Kmetty, Z. & Molnár, E. K. Politics and incivility in the online comments: what is beyond the norm-violation approach? *International Journal of Communication* **15**, 26 (2021).

[20] Goovaerts, I. & Turkenburg, E. How contextual features shape incivility over time: An analysis of the evolution and determinants of political incivility in televised election debates (1985–2019). *Communication research* **50**, 480–507 (2023).

[21] Sun, Q., Wojcieszak, M. & Davidson, S. Over-time trends in incivility on social media: Evidence from political, non-political, and mixed sub-reddits over eleven years. *Frontiers in Political Science* 130 (2021).

[22] Xu, W. *et al.* Social media warfare: investigating human-bot engagement in english, japanese and german during the russo-ukrainian war on twitter and reddit. *EPJ Data Science* **14**, 10 (2025).

[23] Uyheng, J. & Carley, K. M. Bots and online hate during the covid-19 pandemic: case studies in the united states and the philippines. *Journal of computational social science* **3**, 445–468 (2020).

[24] Blumer, K. & Kleinberg, J. Tracking patterns in toxicity and antisocial behavior over user lifetimes on large social media platforms. *arXiv preprint arXiv:2407.09365* (2024).

[25] Jiang, J. & Ferrara, E. Social-llm: Modeling user behavior at scale using language models and social network data. *arXiv preprint arXiv:2401.00893* (2023).

[26] Pacheco, D. Bots, elections, and controversies: Twitter insights from brazil's polarised elections. In *Proceedings of the ACM Web Conference 2024*, 2651–2659 (2024).

[27] Leite, J. A., Silva, D., Bontcheva, K. & Scarton, C. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In Wong, K.-F., Knight, K. & Wu, H. (eds.) *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 914–924 (Association for Computational Linguistics, Suzhou, China, 2020). URL https://aclanthology.org/2020.aacl-main.91.

[28] Sistrom, C. L. & Garvan, C. W. Proportions, odds, and risk. *Radiology* **230**, 12–19 (2004).

[29] Yang, K.-C., Varol, O., Hui, P.-M. & Menczer, F. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 1096–1103 (2020).

[30] Bal, M. I. & Pacheco, D. Echoes of automation: How bots shaped political discourse in brazil. In Cherifi, H., M. Rocha, L., Cherifi, C. & Ertem, M. Z. (eds.) *Complex Networks & Their Applications XIV*, 171–181 (Springer Nature Switzerland, 2026).

[31] Karimi, S., Oliveira, M. & Pacheco, D. Misinformation dissemination: Effects of network density in segregated communities. *arXiv preprint arXiv:2411.19866* (2024).