Automated Toxicity in Brazilian Online Political Discourse

Kai Barber-Harris

April 21, 2024

Abstract

An abstract is a short statement about your paper designed to give the reader a complete, yet concise, understanding of your paper's research and findings. It is a mini-version of your paper. A well-prepared abstract allows a reader to quickly and accurately identify the basic content of your paper. Readers should be able to read your abstract to see if the related research is of interest to them. It is a bit like advertising: having read the abstract, do you want to pay to see the whole paper, which might be behind a paywall. Do not include citations in an abstract. 100-200 words.

I certify that all material in this dissertation which is not my own work has been identified.

Signed:

Contents

1 Introduction

1.1 Background and motivation

In the past two decades, the rise of social media has led the world to become increasingly interconnected, with avenues for discussion of daily news, hobbies, music, and politics often no further away than the smartphones in our pockets. One of the most relevant social media platforms for this form of interpersonal discussion is X, formerly Twitter, a microblogging service designed to share short-form statements or questions and quickly spread these posts and ideas via the engagement of other users amplifying the post themselves. While not necessarily the most popular social media site[1], X's post format and engagement options often make it the platform of choice for users seeking discussion along with leaders and figureheads (or at least their PR teams) wishing to more directly engage with their audience.

More recently, many major political events around the world point to a more divided society than ever before, with election and major referendum results often coming down to the wire (*EU Referendum Results - BBC News* [2], *Federal Elections 2016* [3], *Federal Elections 2020* [4], *SIG Eleição - Resultados* [5]). With this polarisation only becoming more pronounced, alongside the paradigm shift of political discourse and news dissemination taking place online especially with younger people[6], it is now common for people to be loosely or even distinctly aware of the possibility of bad actors attempting to influence the zeitgeist of these discussions through the use of automated actors, or "bots", continually posting opinionated messages or simply strengthening another account's postings. Elon Musk, current CEO of X, more than once attempted to use this very concern of bot presence to back out of his deal to buy the company, at the time still Twitter, by including it in the evidence for a countersuit against Twitter's board forcing him to follow through on the signed contract to buy the company [7][8].

The potential payout for such manipulations is extremely high. For example, with Brazil being the 6th-largest audience for X boasting 22 million unique monthly users[9], convincing some of these users to vote a specific way has the potential to sway elections; the most recent Brazilian presidential election was only won by a margin of approximately 6 million votes[5]. Further, the only cost to running these bots is that of some computation time and the creation of the code for them to perform the desired actions - creating accounts on X is free. For any group seeking to influence political discourse or even entire elections, there are very few if any more cost-effective ways to go about it, as evidenced by the interference of Russian groups in the 2016 US presidential elections via X who appear to have used bots to amplify the messaging of manually operated accounts, most of which were "mostly promoting conservative causes and were, specifically, spreading pro-Trump material."[10]

It is for all of these reasons that we present the following work; we hope to further investigate automated posts concerning Brazilian politics on X with a specific focus on the trends of toxicity between different discussions, topics, and communities. While proving a link between any two of political leaning, bot postings, and average toxicity is unlikely, we hope that the exploration of such a subject may provide a more nuanced view on how these automated accounts may attempt to influence discourse in the modern day.

1.2 Specification

This project was undertaken with the overall goal of further investigating a pair of questions put forward in November 2023 as part of the related literature review:

1. Are bots performing in a coordinated manner attempting to influence political discussions to be more incendiary?
2. Do coordinated activities or bot accounts admit a notable difference in average sentiment or stance toward a specific topic when compared to the generic real user base?

The literature review also notes:

> We would consider this project to be successful if either one of the two questions presented

has been definitively answered, or both questions have been explored and further areas of research have been identified to expand upon the findings. A definitive answer to both questions would be considered to be an outstanding success.

These questions and criteria provide us a strong framework with which to continue our report, in which we more so aim to explore these questions rather than answer them outright. However, due to a misunderstanding of the work, the sentiment and stance analysis was abandoned partway through the project (subsection 5.2); as such, we cannot directly answer the second question posed, though we can instead adapt it:

Do coordinated activities or bot accounts admit a notable difference in average toxicity in different communities when compared to the generic real user base?

This adapted question is one we can make an attempt to answer within the scope of the work completed.

2 Data

2.1 Overview

The data used for the project was provided by Diogo Pacheco, who advised on this work and has extensive experience in the field (Pacheco [11], Pacheco, Hui, Torres-Lugo, *et al.* [12], Chen, Pacheco, Yang, *et al.* [13], Pacheco, Flammini, and Menczer [14]). The dataset in total comprises 437 million X posts made by 13 million unique users between August 2018 and March 2023. These tweets are stored in a raw JSON format. A wealth of metadata about each individual tweet is also recorded alongside its original text in these JSON files. The data is recorded in blocks comprising a day each and are labelled appropriately. Pacheco also provided a set of PKL files he had used in his own work before, these files containing a preprocessed PKL version of most (but not all) of the raw JSON files. Not all days with a JSON file have a corresponding PKL file, and vice versa.

2.2 Collection

The data was collected via the Twitter Streaming API before its shutdown in March 2023. The dataset contains tweets collected from a set of keywords themed around Brazilian political discourse, chosen by Pacheco at the beginning of the collection period.

Posts were also collected directly from candidate accounts, from posts containing the full name of any candidate, and from posts containing the official hashtag of each candidate's campaign. In July 2022, the collection was adjusted to include posts made by official party accounts; this is the only time the collection terms changed across the 5 years.

3 Algorithm Design

3.1 Basic data filtering

Some caveats apply to the data that was carried through to the final analysis. As the provided JSON files do not contain a botscore for each post, and the PKL files do not contain the original text of the post, we are limited to only using days which have files of both types due to the nature of our analysis (section 4). Further, the sheer size of the dataset requires us to thin the data in order to perform meaningful analysis within a reasonable time frame (see subsection 3.4). We chose at this point to work with the data beginning July 2022 and ending March 2023, at the single point the collection terms were adjusted. This range of dates contains 122 total days with corresponding PKL and JSON files, representing a total of 86,787,095 posts for analysis; approximately 20% of the entire available dataset.

At this point we removed posts that contained no hashtags at all, along with all "retweets". Posts without hashtags are irrelevant for our analysis, and a "retweet" is an posting action common on X, which effectively shares the original content of the post again. As these retweets are exact copies of

the original, they can be safely removed for the purposes of our analysis. We elected to keep "quote tweets", which are actions similar to retweets but which allow the reposting account to add their own unique comment alongside the original text, which can then itself be retweeted or quoted. These posts were kept as the extra comment may include other hashtags the author has added, thereby making the post relevant for our analysis.

3.2 Collocations

The planned analysis of hashtags and their attributes called for the construction of a collocation network. The algorithm for this work is fairly simple, and naturally involves filtering the dataset to a smaller size. First, we must take each post with 2 or more hashtags in it; this is easily accomplished as the PKL files contain the hashtags within each tweet. We can then sort the hashtags (to prevent reversed instances of the same pairs) and generate the pairwise combinations of hashtags possible from the list of hashtags in each tweet.

Recording the number of occurrences of these pairwise combinations constructs a collocation matrix; at the same time, we can track the number of times each hashtag appears in total to a separate "appearances" matrix. Respectively, these matrices act as edge and node tables, with the appearances of each collocation representing an edge weight and the appearances of each hashtag supporting a deeper analysis.

3.3 Modularity

These edge and node tables can then be loaded into a network analysis program such as Gephi[15]. Using Gephi, we can inspect and analyse the data we have already found. In order to increase the clarity of investigation, at this time we chose to filter out all hashtags (and therefore their collocations) which appeared less than 1,000 times in total. This reduces the size of the network from 134,633 to 1,943 unique hashtags, a 98.5% reduction. This same action also reduces the number of edges from 1,065,768 to 153,444, an 85.6% reduction. With only 1.5% of original nodes visible, nearly 15% of original edges are still visible; these observations together show us that the most important nodes have by far the most connections with one another, and these "smaller" nodes can be filtered out safely. The filter increases the relative "importance" of each node by removing smaller nodes which dilute the analysis; we can now be sure that each and every one of these hashtags is discussed in a significant manner. It should be especially noted that filtering out hashtags which appear less than 100 times across these 122 days results in a 93% reduction in nodes, further proving that the vast majority of hashtags are not especially relevant to everyday discussion in online Brazilian political discourse and can safely be ignored for our analysis.

With this filtering complete, we can run a modularity analysis on the graph with Gephi. Modularity analysis is a methodology for detecting communities in graphs with weighted edges. Gephi implements the Louvain-style modularity analysis, which finds communities by recursively taking the aggregate result of a smaller community-detection function [16]. The result of this process is that each hashtag node will have a modularity class assigned to it, denoting the community it primarily belongs to. For our hashtag network, the Louvain analysis found 18 unique communities.

3.4 Toxicity

We now calculate the probability that any given post made in a hashtag would be considered toxic. The computation of the toxicity score for each hashtag follows an involved process. The steps are as follows:

1. Collect all hashtags which belong to the two largest modularity classes of left-wing and right-wing discussion (see subsection 5.1 for reasoning).
2. For these hashtags, collect all posts in which any of these hashtags appear. In total, there were 968,627 of these posts.

3. Preprocess the original text of each collected post through a standard natural language processing pipeline (tokenisation, lowercasing, removal of punctuation and stopwords, stemming).

4. Run each preprocessed text through a pre-trained machine learning model to determine whether the post is toxic or not.

5. For each hashtag, collect all posts containing that hashtag and calculate the mean toxicity of the posts. This score is then a floating point representing the likelihood of a post containing the hashtag to be toxic.

The pre-trained machine learning model used was ToLD-Br, a "pre-trained Multilingual BERT fine-tuned with ToLD-Br" [17]. ToLD-Br is provided as open-source under the Creative Commons BY-SA 4.0 license, and allows us to use previous work to greatly accelerate our own. The trained BERT model is a CUDA-accelerated[1] binary classification transformer model which, given a text to process, outputs either a "toxic" or "non-toxic" interpretation. The model itself was trained on a dataset of Brazilian Portuguese X posts; this fits our use case perfectly, and we thank Leite, Silva, Bontcheva, *et al.* for the model's contribution to the project.

3.5 Botscore

Finally, we calculate the probability that any post made containing a given hashtag was posted by a suspected bot, or the mean "botscore" of the hashtag. Botscores for most posts in the dataset were already precalculated by the Botometer tool [18] within PKL files. As such, we can simply follow the final step of the toxicity calculation process, collecting the botscores for each post containing a given hashtag and calculating their mean to give an average botscore for this hashtag.

These botscores are extremely valuable when taken in the context of X's recent history. Due to the acquisition and major changes to X since Elon Musk's takeover in late 2022, many useful tools for analysis have been taken offline due to sharp API cost increases and functionality changes. One of these casualties was Botometer, which was used to generate these botscores during Pacheco's earlier work [11]. Botometer today still exists as "Botometer X", but it is severely hampered by these API changes.

4 Results and Analysis

4.1 Hashtag Network

Colour	Topic	% of network
Cyan	Left-wing hashtags	26.7%
Red	Right-wing hashtags	25.8%
Pink	Battleground hashtags	24.5%
Yellow	News hashtags	7.2%
Other	All other hashtags	15.8%

Table 1: Figure 1 and Figure 2's most populous communities

At this time, the data processing is complete and we can begin our analysis. Figure 1 shows the fully filtered network after running the ForceAtlas 2 layout algorithm to convergence, with each node colour-coded to its modularity class[2]. ForceAtlas 2 is an continuous layout algorithm that works by repulsing all nodes from one another while attracting nodes based on edges[19], but this work could easily be recreated with any suitable force-directed layout algorithm. In this graph, we can very clearly

[1]CUDA: NVIDIA-owned and maintained framework for offloading repetitive, simple matrix and arithmetic operations to an attached graphics processing unit which excels at such tasks rather than the central processing unit of the computer, which is comparatively much worse at such operations.

[2]For the sake of presentation, the graph was also run through an algorithm preventing nodes from overlapping.

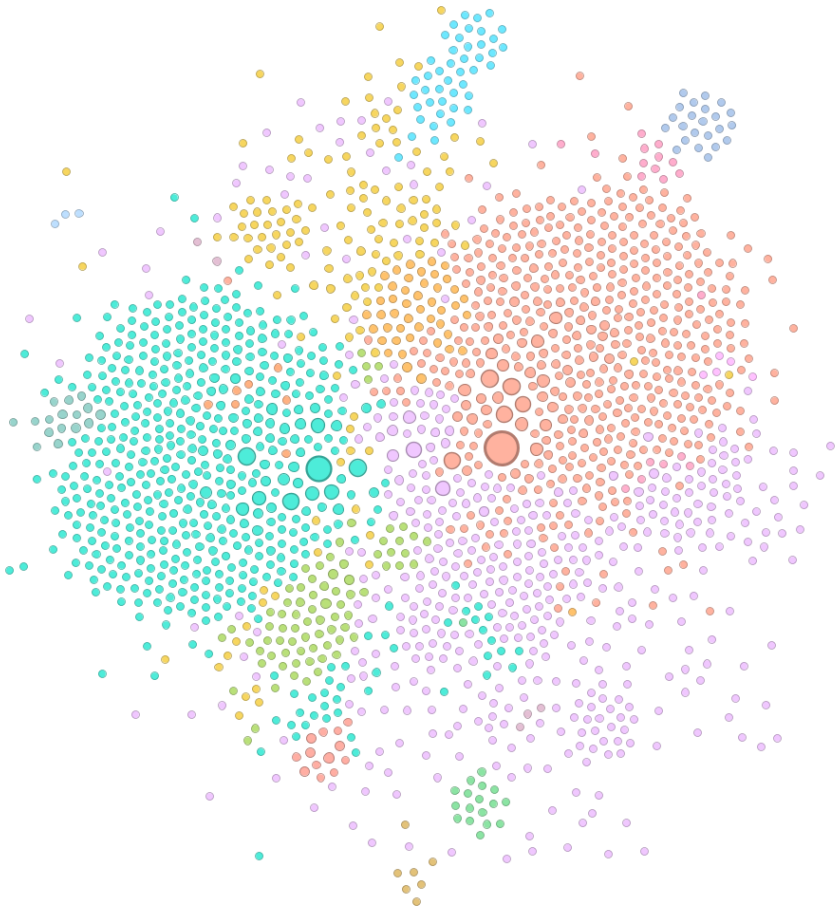


Figure 1: Hashtag network, nodes coloured to modularity class (community) and ForceAtlas 2 layout applied. Notable communities in the network include right-wing terms (red, upper right), left-wing terms (cyan, middle left), "battleground" terms (pink, lower central leaning right), and news-related terms (yellow, upper central leaning left). The size of each node directly correlates to how many times it appears in the data set relative to other hashtags.

see the different communities present in our data, noting the presence of 4 "primary" communities[3], shown in Table 1.

The size of each hashtag is directly taken from the its number of appearances. We can easily identify central hashtags that appear often in their communities; the largest right-wing hashtag is "bolsonaro22", the official campaign hashtag for Jair Bolsonaro. Likewise, the largest left-wing hashtag is "lula13", Lula da Silva's official campaign hashtag. The largest battleground hashtags are those that are relevant to both of the previous communities; terms such as "eleicoes22" (election22), "brasil", "riodejaneiro", and news hashtags include "globonews" (popular Brazilian news network), "nytimes", and "blacklivesmatter".

Figure 3 shows the same network as in Figure 1, but with the node sizes altered to visualise toxicity and bot activity across the network - larger nodes mean a higher likelihood of the hashtag being contained within a toxic post (Figure 3a) or having been made by a suspected bot (Figure 3b).

We can see from a visual inspection that these metrics differ in intriguing ways; toxicity peaks in localised areas of discussion, shown in the many groups of large nodes most visible in the left-wing and right-wing communities separated by gulfs of smaller nodes. Contrastingly, Figure 3b implies that nearly all hashtags in our dataset have a base level of bot activity present, with the vast majority

[3]For more, see Appendix B

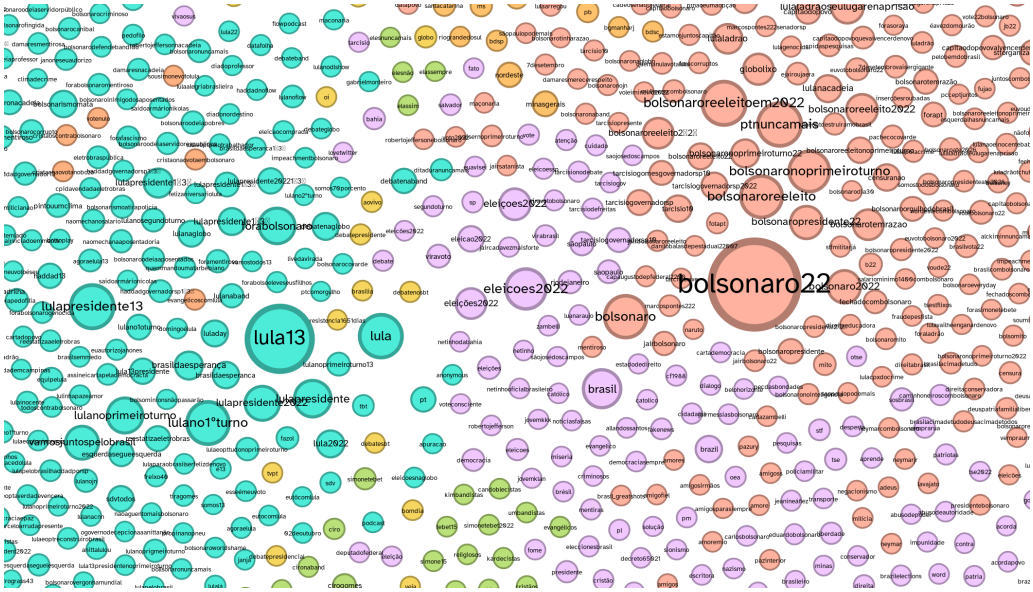
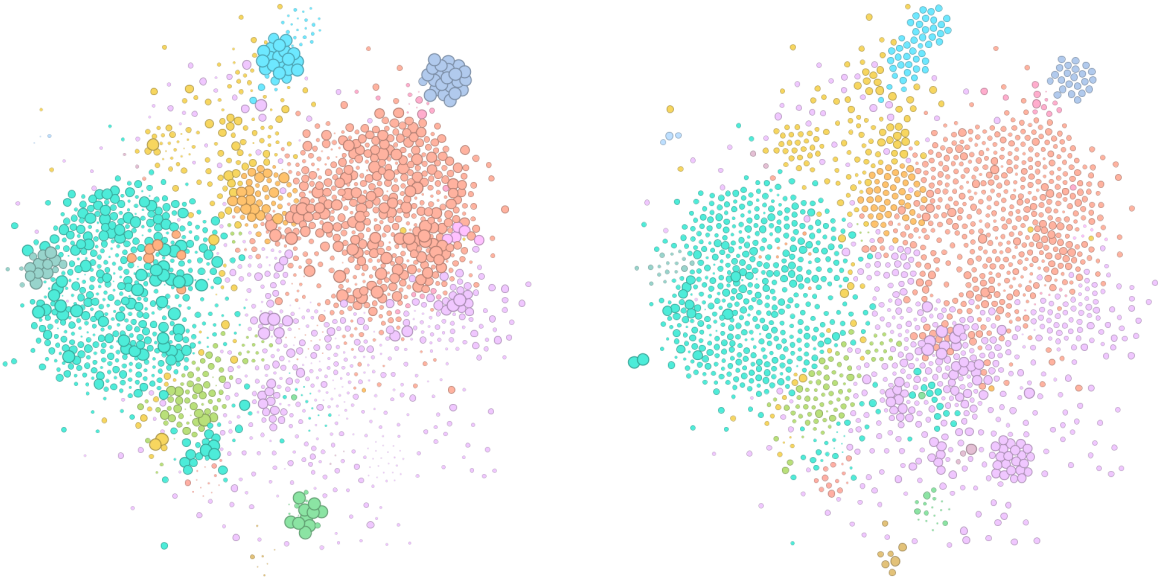


Figure 2: Zoomed view of Figure 1's central region with labels enabled.



(a) Node sizes keyed to average toxicity for posts including hashtag

(b) Node sizes keyed to average botscore for posts including hashtag

Figure 3: Hashtag network displaying different size variables, with layout and colouring inherited from Figure 1.

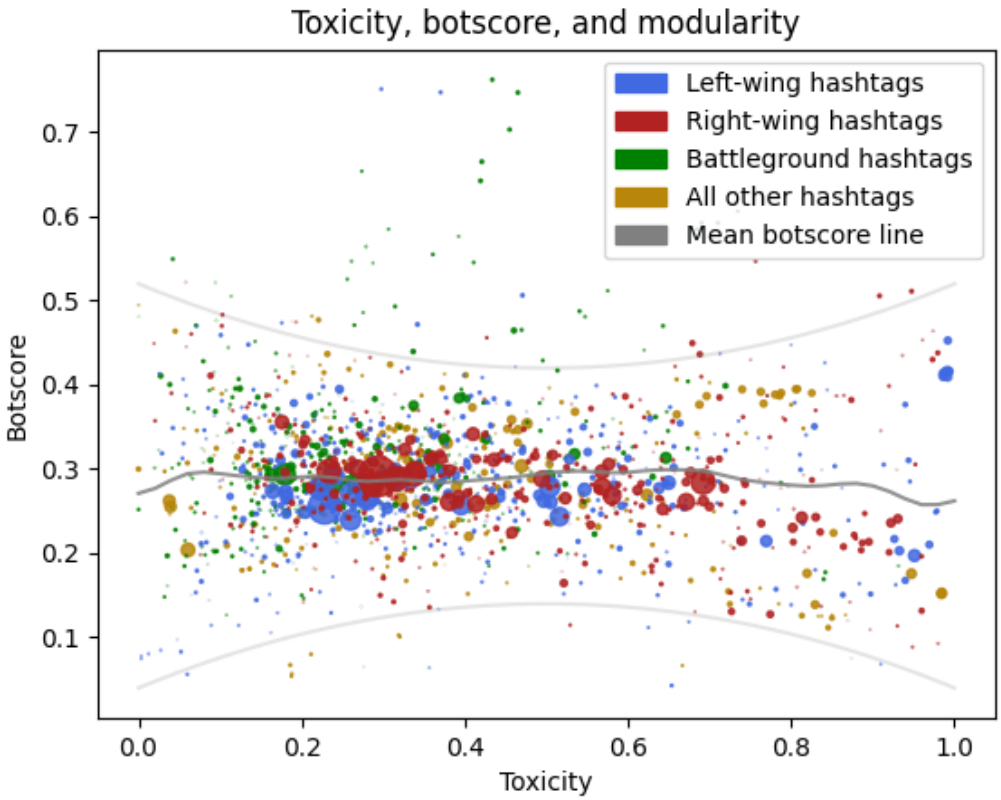


Figure 4: Scatter graph showing the toxicity and botscore of each hashtag. Sizes of nodes correspond to how many posts included the hashtag (the number of "appearances"), and colours show modularity/community.

of nodes notably larger than the minimum size commonly seen in Figure 3a. We can easily confirm this observation, empirically speaking, by comparing the standard deviation of each metric across the whole network. Botscore admits a standard deviation of 0.08, whereas toxicity lies at a much higher 0.22, ratifying our suspicions. We now continue to determine whether this phenomenon may be affected by bots acting as pacifiers and/or agitators.

4.2 Bots and Toxicity

To further explore the data, we simply plot a graph of botscore against toxicity. Sorting all hashtags by their toxicity and calculating an average value, weighted by number of appearances, in a fixed number of regions allows us to directly explore a potential correlation between botscore and toxicity. Figure 4 shows this graph and informs us of a distinct lack of correlation across the dataset, the mean line largely lying flat. We can, however, see that the vast majority of conversation appears within a densely populated zone between 0.20 and 0.40 toxicity, hovering just below a botscore of 0.30. All together, the shape of the graph somewhat resembles a saddle; this saddle is shown in a pair of faintly drawn parabolas, so as not to distract too heavily from the graph. This effect leaves a deadzone at low botscores in hashtags with moderately high toxicities of 0.40 to 0.60; the implication is that moderately toxic hashtags attract a moderate bot presence.

With this saddle shape in mind, we can notice a group of anomalies with notably high botscores at moderate toxicities. These hashtags mostly belong to the "battleground" modularity, with only two from the left-wing modularity. These two are *"13elevoltara"* and *"simboracomlula13"* boasting botscores of 0.75 and 0.74 respectively, meaning *"13 will return"* and *"let's go with lula 13"*. Both of these hashtags are in support of the left-wing presidential candidate Lula da Silva, implying an actor attempted to increase support of the left-wing candidate using bots to amplify their messaging.

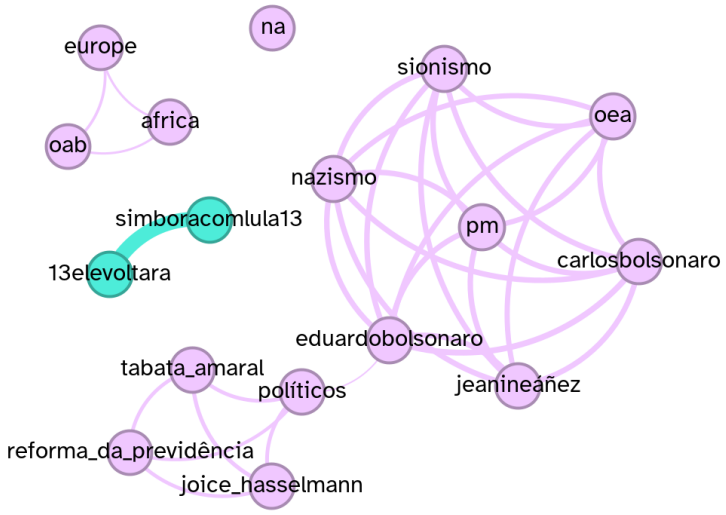
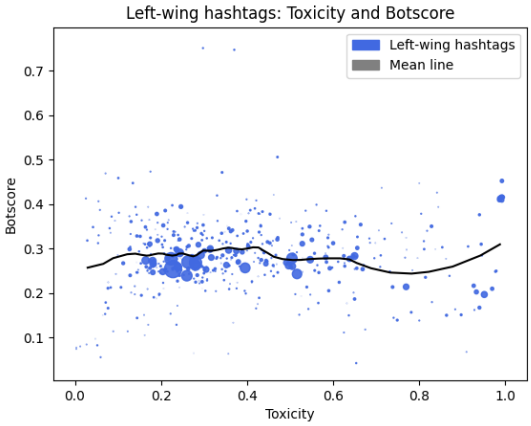


Figure 5: Detail view of the battleground anomalies network.

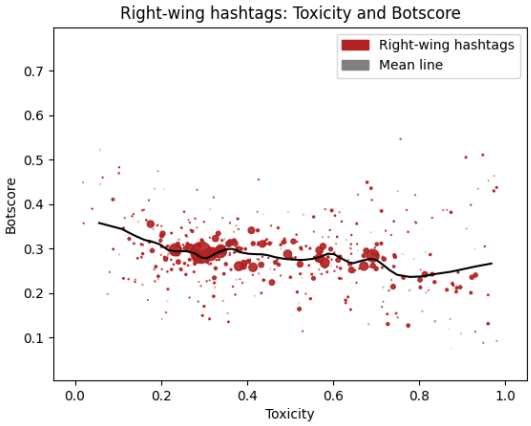
The remaining anomalies are mostly connected to one another in a local network, shown in Figure 5. Some refer to regions, such as *"europe"* or *"africa"*, but most of these anomalously high botscore hashtags concern Jair Bolsonaro or "presidential reform". In the context of when the data was collected, Bolsonaro at the time had been in power for 30-36 months and was approaching the end of his term, with our data including discussion surrounding the election which resulted in his defeat by Lula da Silva. As such, notable connections in this network sit between *"nazismo"* (Nazism), *"zionismo"* (Zionism), and two of Bolsonaro's sons (Eduardo and Carlos), both of whom are also involved in Brazilian politics.

A manual check also reveals that *"nazismo"* and *"zionismo"* both connect to *"bolsonaro"* and *"jair-bolsonaro"*, with many of Figure 5's connections appearing exactly 288 times (or within 10% of 288), shown in the thickness of each edge between nodes. It is fairly likely that these hashtags were levelled as insults against Bolsonaro by a small populace of automated accounts, evidenced by the very high botscore and similar appearance numbers hovering at or just above 288. Many Brazilians dislike Bolsonaro for controversial actions surrounding deforestation of the Amazon rainforest, his handling of the Covid-19 pandemic, and other smaller campaign-related controversies ([20], [21], [22], [23], [24]), though simply existing as a political opponent opens motivation for manipulating opinions as discussed in subsection 1.1.

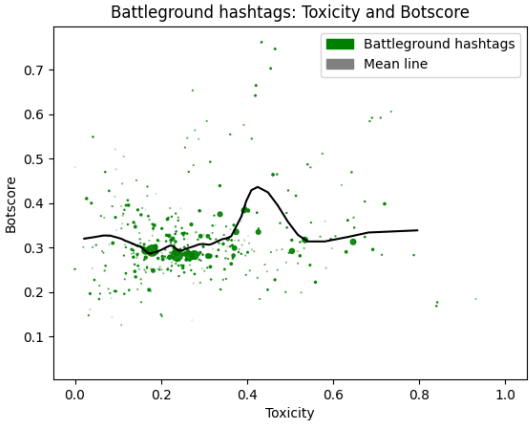
These manipulations may even have been orchestrated by the same actor responsible for amplifying the messaging of *"simboracomlula13"*, using the same techniques to slander Bolsonaro and make him appear a worse fit for the presidential seat. The only reason we have noticed these manipulations is that the botscore of these hashtags appeared notably high, with few "real" people including these hashtags in their discussions; we put forward that similar manipulations are in effect across the dataset, being obscured by the activity of real users. Further, in this case, this anomalous network acts as evidence towards our first hypothesis that bots are being utilised to influence political discussions to become more toxic; hurling insults over the internet is not usually seen as a level-headed gesture.



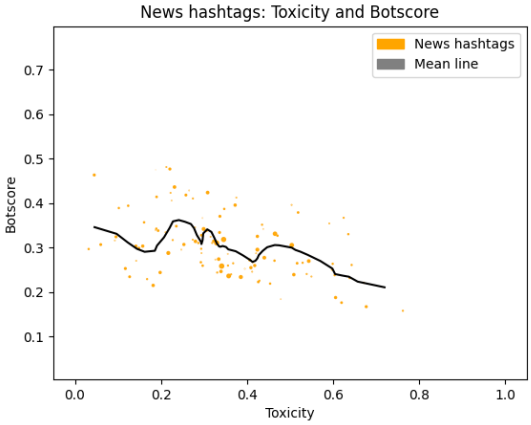
(a) Left-wing community scatter graph



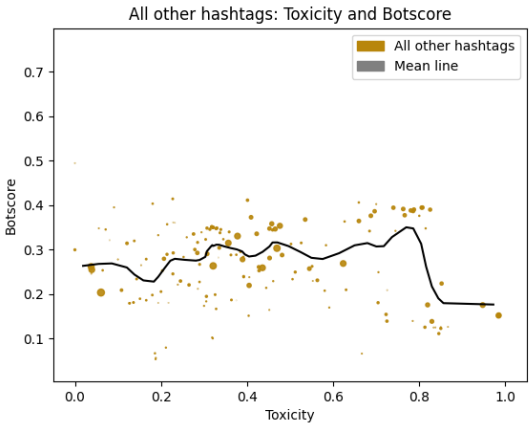
(b) Right-wing community scatter graph



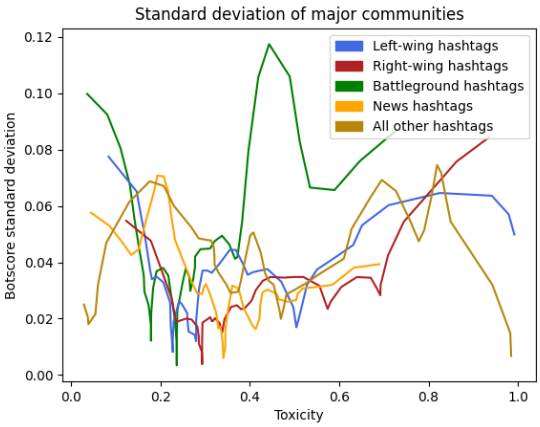
(c) Battleground community scatter graph



(d) News community scatter graph



(e) All other hashtags scatter graph



(f) Standard deviation of all major communities, weighted to number of appearances

Figure 6: Plots of each major community, demonstrating the spread of each community's hashtags along with the local mean weighted for the influence of each hashtag. Figure 6f shows the botscore's standard deviation in each community as toxicity varies, accounting again for the influence of each hashtag.

Figure 7: Detail view of the anomalous hashtags in Figure 6a.

4.3 Partitioning by Modularity

In this section, we partition the dataset by modularity and perform a similar analysis as in subsection 4.2, but within each community rather than across all processed data. Figure 6 shows a scatter graph of each major community's points as in Figure 4 but split out to their own graphs with their own mean lines applied. Each mean is calculated as a weighted average, using each hashtag's appearances relative to others as the weighting. Each community has its own characteristics shown in the mean.

4.3.1 Left-wing community analysis

Left-wing hashtags stay at roughly the same botscore across toxicities, but with a curve upwards at the end likely influenced by the single large hashtag at near-max toxicity. A closer inspection reveals that this is not a single hashtag, but a group of abnormally high botscore hashtags at the highest toxicity values present in the whole dataset.

This local network is shown in Figure 7, and shows how deeply connected these hashtags are. With the size of each node keyed to its appearances, we even notice that all of these hashtags except one boast between 23,000 and 24,000 appearances, the exception having 16,175. Further, many of these connections have weights of exactly 1,753 or 1,754 appearances together, implying that when two of these hashtags were used the others were extremely likely to follow. These close numbers, deeply linked connections, and the high botscore when compared against the rest of the left-wing data together suggest a discussion being deeply influenced by bots. Not only this, but the very high toxicity of these hashtags further suggests these bots may be acting to reinforce a toxic discussion already occurring; at the very least, they aren't helping, otherwise the toxicity of the local network would be lower - the average toxicity of this local network is 0.991, with each hashtag in the top 1% most toxic in the dataset.

The overall topic of these hashtags lends itself to a toxic discussion. The overall tone is assertive and accusatory, many of these posts likely made during the late 2022 presidential election campaigns: *"Bolsonaro hates the poor"*, *"pedophilia out"*, *"fascism out"*, *"don't mess with retirement"*, and *"he was a liar"* are some translated examples from this network. Our earlier observations indicated a discussion

Community	Toxicity		Botscore	
	Mean	Median	Mean	Median
Left-wing	0.38	0.35	0.28	0.28
Right-wing	0.43	0.42	0.29	0.29
Battleground	0.28	0.25	0.31	0.30
News	0.35	0.33	0.30	0.30
Other	0.46	0.36	0.28	0.28

Table 2: Statistics comparison table for all major communities. Means are weighted to appearances.

highly influenced by bot activity, though these discussion topics and accusations can be expected to begin organically in a modern day election cycle online. Furthermore, the disparity between edge weights and appearances implies these hashtags are often used outside of this deeply linked network in other discussions elsewhere in the network. As such, we can conclude that this is most likely a case of bots being used to amplify a toxic discussion and spread these opinions further rather than bots being utilised to begin one from scratch, potentially leveraging these hashtags to gain a greater reach and influence. These findings further support our hypothesis in subsection 4.2 that bot manipulations are easily obscured by the postings of real users, along with reinforcing our earlier suspicion that bots are indeed influencing political discussions to become more toxic.

4.3.2 Right-wing community analysis

In contrast to the flat mean of the left-wing network previously discussed, the right-wing network demonstrates a moderate negative correlation between toxicity and botscore with a slight upturn in botscore at the highest toxicities. On a visual inspection, the right-wing network appears to contain more hashtags at the higher end of the toxicity scale. We can investigate this difference by calculating the mean and median of both botscore and toxicity for each community, shown in Table 2.

As suspected, when compared to other focused communities, the right-wing community has a notably higher mean and median toxicity, with mean and median botscores in the expected range. Note in this case that "other" hashtags also have a very high toxicity; this is due to the varied and unfocused discussion occurring across the remaining 14 communities which all in all comprise only 15.8% of the data (Table 1); furthermore, the right-wing community has a direct counterpart in the left-wing, and as such we continue our inspection of this comparatively high toxicity. While a similar upturn occurs at the high end of toxicity as in the left-wing community, the intensity of the right-wing's equivalent is much lower with a manual inspection revealing many much less popular hashtags appearing to bring this line upwards. With this finding in mind, along with the mean botscore steadily lowering as toxicity increases, we show that these right-wing discussions appear to simply trend towards toxicity without any notable bot influence attempting to coerce the discussion towards it; in fact, these toxic discussions consistently contain less bots the more toxic they become.

We now check this finding by inspecting the weighted standard deviations of each community across toxicities. Figure 6f shows this plot, demonstrating that the right-wing community is overall one of the closest-fitting to the baseline data. However, as toxicity rises we also see the standard deviation rise steadily. This curve shows us that, at the highest toxicities, there is a wide range of hashtags with both high and low botscores active in the community. For this reason, we can say in this case that fringe communities of bots are attempting to influence political discussions on the right to become more toxic.

The affected hashtags at this highest end of toxicity include *"capitaobolsonaro"*, *"stfmilitarjá"*, *"eleicoessemsacanagem"*, and *"capitaoparaobemdanacao"*, all of which have appearances totalling between 4,000 and 5,000. While not obvious or egregious as the similar anomaly network in the left-wing graph, we can consider this a very similar case of bots amplifying a toxic message. The translated hashtags are *"captain Bolsonaro"*, *"STF military now"*, *"elections without bullshit"*, and *"captain for the good of the country"*, clearly referring to Bolsonaro again.

"Elections without bullshit" is reminiscent of slogans adopted by western alt-right political movements concerning election fraud such as *"stop the steal"*, popularised by the high-profile defeat of Donald Trump in the American 2020 presidential election[4] and his following campaign to overturn the result on claims of voter fraud[25], [26], eventually culminating in the attack on the American political centre of the Capitol on January 6th, 2021 [27], [28]. Similarly, *"STF"* refers to the Brazilian Supreme Federal Court[29]; due to the extremely high toxicity of the hashtag, it's possible these posts were attempting to incite violence against members of the court or prompt the STF to "use" the military in some degree, a line of thinking potentially borrowed once again from Trump and his supporters' militarisation [30],[31].

With the topic of these most common toxic right-wing hashtags in mind, we can conclude once again that communities of bots are attempting to amplify incendiary ideas and topics across the wider spectrum of discussion.

5 Limitations and Further Work

5.1 Computation time

The primary problem encountered during this project was the scale of computation required for a satisfactory analysis. The largest computation was by far the requirement of running the toxicity prediction model, ToLD-Br[17], on just under 1 million unique texts.

The strain of running this toxicity prediction model is the primary reason for the aggressive nature of our data filtering. Each text prediction took on average 0.2 seconds to perform on the available computing power (see Appendix A) - the dataset as a whole contains 437 million posts. This quantity of data is beyond enormous, and processing all of it is clearly unfeasible for the time allocated to the project[4].

The filters applied, in order, are as follows:

1. Full dataset, August 2018 to March 2023 (approx. 437 million posts)
2. Recent data after collection term adjustment, July 2022 to March 2023 (80,688,513 posts)
3. Only posts containing hashtags (7,806,357 posts)
4. Removed all "retweets" (3,523,010 posts)
5. Only posts which contain a hashtag either:
    - directly appearing in the modularity class related to left-wing or right-wing posts (subsection 3.3),
    - appearing alongside one of the previous hashtags (968,627 posts)

Implicit in step 5 of this filtering is the fact that the hashtags are extracted "backwards" from posts which appear more than 999 times in the data collected in step 2; a small number of posts only containing these "rare" hashtags will have also been removed from the search in this step.

These 968,627 posts were then taken as the final set of posts to be processed through ToLD-Br. Multiplying the expected time of 0.2s each with the 968,627 posts to process, the processing of the data was predicted to take approximately 54 machine hours. Due to bugfixes and memory-full errors causing unexpected crashes, this process took just under 6 days to complete.

If we had not performed the extra pruning in step 5, we would have run the toxicity model for step 4's 3,523,010 total texts. Due to the nature of the data, removing texts not directly linked to the primary modularities for processing would be unlikely to greatly affect conclusions; pruned texts would by

[4]For completeness, this calculation would take an estimated 3 years (`0.2 * 437,000,000` seconds) on the hardware available.

definition be unrelated to the primary analysis target of partisan discourse, and so a 73% reduction in expected computation time was accepted as a worthwhile trade.

5.2 Sentiment and stance analysis

At the same time as the toxicity analysis was performed (subsection 3.4), a sentiment analysis was also attempted on the final set of filtered posts. Our sentiment analysis corpus was SentiLex-PT02, a well-respected corpus for Portuguese specialised towards "opinions targeting human entities"[32]; a corpus in this sense is effectively a lookup table allowing us to correlate words to opinions. SentiLex-PT02 is provided under the Creative Commons Attribution (CC-BY) license, and required some modifications before it was appropriate for our usage.

The final sentiment analysis algorithm iterated through the following steps:

1. Preprocess the original text of each collected post through a standard natural language processing pipeline (tokenisation, lowercasing, removal of punctuation and stopwords, stemming)
2. Iterate over each word in the text and check for its existence in the sentiment analysis corpus
3. If the item does not appear, move to the next word. If it does, determine the word's polarity ($+1$ or -1)
4. Add the polarity to a tracker variable, beginning at 0
5. For each hashtag, collect all posts containing that hashtag and calculate the mean sentiment of the posts.
6. Normalise each result between 0 and 1, 1 being "positive" and 0 being "negative"

This approach yielded results that were considered near-useless, despite our best understanding of how to tackle the problem. Extreme outliers and many "neutral"[5] texts contained in the data meant the vast majority of hashtags held a sentiment value between 0.48 and 0.52, very close to the neutral value of 0.5.

The purpose for the development of this sentiment analysis algorithm was to support the further development of a stance analysis algorithm. A stance analysis is a form of opinion classification with the goal of using a text to interpret the author's true feelings towards the target (subject). ALDayel and Magdy [33] concisely demonstrates the delineation between sentiment and stance in **??**, alongside the quote found under the table.

In the interest of overall project success and with the knowledge that the toxicity and botscore analyses were producing appropriate results (i.e. behaving correctly), the work to correct the sentiment analysis engine was abandoned, as it would only support the following creation of a stance analysis engine; all in all, a much larger undertaking. While the success of this algorithm and the following results surely would have allowed a more nuanced inspection of the data (and the exploration of the original question as shown in subsection 1.2), the results found from the toxicity and botscore analyses (section 4) stand on their own and should be considered complete in themselves.

The further development of a functioning sentiment and stance analysis can be considered as a potential extension to our work, and a ripe area for further research in this dataset.

5.3 "Self-selection" bias

There is a notable potential error in that by filtering to only posts containing hashtags, we are not fully considering the landscape of discussion. As seen in subsection 5.1, approximately 10% of the dataset contains hashtags; what of the other 90%? Could it be possible that bots are more or less

[5] It is possible that some texts contained equal amounts of positive and negative polarity words, resulting in a neutral assessment.

Table 3: Table 1 verbatim from ALDayel and Magdy, showing the sentiment polarity of the expressed stance. Note that positive or negative stances can have positive, negative, or neutral sentiments.

#	Tweet	Target	Sent.	Stance
1	It is so much fun having younger friends who are expecting babies. #beentheredonethat #chooselife	Legalisation of abortion	+	-
2	Life is sacred on all levels. Abortion does not compute with my philosophy. (Red on #OITNB)	Legalisation of abortion	0	-
3	The biggest terror threat in the World is climate change #drought #floods	Climate change is a real concern	-	+
4	I am sad that Hillary lost this presidential race	Hillary Clinton	-	+

> "...several studies have demonstrated that it is insufficient to use sentiment as the only dependent factor to interpret a user's stance (Aldayel and Magdy, 2019b; Elfardy and Diab, 2016; Mohammad et al., 2017; Sobhani et al., 2016; Somasundaran and Wiebe, 2009). This is due to the complexity of interpreting stances from a given text, as they are not always directly aligned with the polarity of a given post."
>
> (ALDayel and Magdy, 2021)

likely to include hashtags in their postings at a baseline, and does that mean the overall discourse is differently prone to bot manipulation than we have found?

Finding empirical answers to these questions is out of the scope of this work, leaving our conclusions vulnerable to this issue, but does provide an exciting opportunity for further research in the field. With this in mind, the findings presented here should be understood with the caveat that the final analysis is purposely limited to posts including hashtags in order to accommodate the earlier modularity analysis (subsection 3.3); the work does not prove or disprove any conjecture in particular because of this limitation, but does provide obvious pathways for investigation in order to more concretely verify these findings by developing methods to expand the community detection beyond hashtag collocations and perform a similar analysis, ideally on an even larger dataset by the use of a more powerful computer.

6 Conclusion

The original aim of this project was to investigate possible correlations between toxicity, botscore, and community in online political discussion specifically taking place on the microblogging site X, formerly Twitter. In the course of this investigation, we

Future analyses are advised to approach the task in a more granular way, investigating suspected coordination of bot accounts rather than overarching statistics across a network.

Appendix A Workstation

All computation for this dissertation was performed using the author's personal machine on a Linux/GNU installation dedicated specifically for the task. The machine's relevant specifications are as follows:

- CPU: AMD Ryzen 5 5600G (3.7GHz, 4.6GHz boost)
- GPU: NVIDIA RTX 2070 SUPER (2560 CUDA cores, 8GB VRAM)
- RAM: 16GB Corsair "Vengeance" (3200MHz)
- OS: Linux Mint 21.3 "Victoria", kernel version 5.15.0

These specifications were sufficient for the computations being performed, but given a more powerful computer it is possible that the analysis could have been deepened. All data was stored on a high quality NVME M.2 solid-state drive connected directly to the motherboard.

It is possible that with more foresight, access could have been gained to a more powerful device such as Exeter University's ISCA supercomputer. However, it is the author's belief that by the time they realised it would be useful, it would take longer to gain access and learn to use ISCA's systems than it would to simply run the calculations locally.

Appendix B Modularity classes

This appendix contains the full list of modularity classes discovered by the Louvain analysis and the author's best interpretation of what each community represents. Note that the actual modularity class value is arbitrary and assigned during the Louvain analysis.

Class	Colour	Topic	% of network
7	Cyan	Left-wing hashtags	26.7%
9	Red	Right-wing hashtags	25.8%
1	Lilac	Battleground hashtags	24.5%
14	Yellow	News hashtags	7.2%
3	Green		4.2%
0	Orange		2.4%
16	Light blue		2.3%
12	Pink		1.2%
13	Light green		1.2%
17	Slightly darker than light blue		1.2%
2	Aqua		0.8%
5	Salmon		0.8%
6	Tan		0.6%
15	Dark orange		0.4%
4	Dusty purple		0.3%
8	Light pink		0.3%
11	Grey		0.3%
10	Lighter blue		0.2%

References

[1] *Biggest social media platforms 2024*, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. (visited on 04/12/2024).

[2] *EU Referendum Results - BBC News*, https://www.bbc.co.uk/news/politics/eu_referendum/results. (visited on 04/12/2024).

[3] *Federal Elections 2016*, https://www.fec.gov/introduction-campaign-finance/election-results-and-voting-information/federal-elections-2016/. (visited on 04/12/2024).

[4] *Federal Elections 2020*, https://www.fec.gov/introduction-campaign-finance/election-results-and-voting-information/federal-elections-2020/. (visited on 04/12/2024).

[5] *SIG Eleição - Resultados*, https://sig.tse.jus.br/ords/dwapr/r/seai/sig-eleicao-resultados/home. (visited on 04/12/2024).

[6] E. Shearer, *Social media outpaces print newspapers in the U.S. as a news source.* (visited on 11/07/2023).

[7] *Musk - Public Version of Counterclaims & Answer (w COS)*, https://www.documentcloud.org/documents/22 musk-public-version-of-counterclaims-answer-w-cos. (visited on 04/12/2024).

[8] Elon Musk [@elonmusk], *@Teslarati 20% fake/spam accounts, while 4 times what Twitter claims, could be *much* higher. My offer was based on Twitter's SEC filings being accurate. Yesterday, Twitter's CEO publicly refused to show proof of <5%. This deal cannot move forward until he does.* Tweet, May 2022. (visited on 04/12/2024).

[9] *X/Twitter: Global audience 2024*, https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. (visited on 04/12/2024).

[10] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2018, pp. 258–265. DOI: 10.1109/ASONAM.2018.8508646. (visited on 04/12/2024).

[11] D. Pacheco, *Bots, Elections, and Controversies: Twitter Insights from Brazil's Polarised Elections*, Oct. 2023. DOI: 10.48550/arXiv.2310.09051. arXiv: 2310.09051 [cs]. (visited on 11/06/2023).

[12] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, *Uncovering Coordinated Networks on Social Media: Methods and Case Studies*, Apr. 2021. DOI: 10.48550/arXiv.2001.05658. arXiv: 2001.05658 [physics]. (visited on 11/07/2023).

[13] W. Chen, D. Pacheco, K.-C. Yang, and F. Menczer, "Neutral bots probe political bias on social media," *Nature Communications*, vol. 12, no. 1, p. 5580, Sep. 2021, ISSN: 2041-1723. DOI: 10.1038/s41467-021-25738-6. (visited on 11/06/2023).

[14] D. Pacheco, A. Flammini, and F. Menczer, "Unveiling Coordinated Groups Behind White Helmets Disinformation," in *Companion Proceedings of the Web Conference 2020*, Taipei Taiwan: ACM, Apr. 2020, pp. 611–616, ISBN: 978-1-4503-7024-0. DOI: 10.1145/3366424.3385775. (visited on 11/06/2023).

[15] *Gephi - The Open Graph Viz Platform*, https://gephi.org/. (visited on 04/20/2024).

[16] *Findcommunities*, https://sites.google.com/site/findcommunities/. (visited on 04/07/2024).

[17] J. A. Leite, D. Silva, K. Bontcheva, and C. Scarton, "Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis," in *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*, K.-F. Wong, K. Knight, and H. Wu, Eds., Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 914–924.

[18] M. Sayyadiharikandeh, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "Detection of Novel Social Bots by Ensembles of Specialized Classifiers," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, ser. CIKM '20, New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 2725–2732, ISBN: 978-1-4503-6859-9. DOI: 10.1145/3340531.3412698. (visited on 04/07/2024).

[19] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software," *PLOS*

ONE, vol. 9, no. 6, e98679, Jun. 2014, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0098679. (visited on 04/08/2024).

[20] *Brazilian Amazon deforestation up 150% in Bolsonaro's last month*, https://www.aljazeera.com/news/2023/1 amazon-deforestation-up-150-in-bolsonaros-last-month. (visited on 04/17/2024).

[21] P. Trevisani and J. T. Lewis, "Brazil's Bolsonaro Fires Key Adviser Amid Accusations of Campaign-Finance Fraud," *Wall Street Journal*, Feb. 2019, ISSN: 0099-9660. (visited on 04/17/2024).

[22] B. Dupeyron and C. Segatto, *Just like Trump, Brazil's Bolsonaro puts the economy ahead of his people during coronavirus*, http://theconversation.com/just-like-trump-brazils-bolsonaro-puts-the-economy-ahead-of-his-people-during-coronavirus-136351, Apr. 2020. (visited on 04/17/2024).

[23] *Brazil military chiefs quit as Bolsonaro seeks their support*, https://apnews.com/article/brazil-rio-de-janeiro-cabinets-health-coronavirus-pandemic-fdeb61a84563bca5ed9ddf9dd437a8c9, Mar. 2021. (visited on 04/17/2024).

[24] T. Phillips and T. P. L. A. correspondent, "Brazil's Jair Bolsonaro threatens purge of leftwing 'outlaws'," *The Guardian*, Oct. 2018, ISSN: 0261-3077. (visited on 04/17/2024).

[25] J. Rutenberg, J. Becker, E. Lipton, *et al.*, "77 Days: Trump's Campaign to Subvert the Election," *The New York Times*, Jan. 2021, ISSN: 0362-4331. (visited on 04/20/2024).

[26] A. Rupar, *Trump's desperate "STOP THE COUNT!" tweet, briefly explained*, https://www.vox.com/2020/11 tweet-stop-the-count-votes-presidential-election, Nov. 2020. (visited on 04/20/2024).

[27] *The January 6 Attack on the U.S. Capitol - American Oversight*, https://www.americanoversight.org/investig january-6-attack-on-the-u-s-capitol. (visited on 04/20/2024).

[28] *January 6 U.S. Capitol Attack— Background, Events, Criminal Charges, & Facts — Britannica*, https://www.britannica.com/event/January-6-U-S-Capitol-attack, Apr. 2024. (visited on 04/20/2024).

[29] *STJ International - Superior Tribunal of Justice in Brazil*, https://international.stj.jus.br/en/Brazilian-Judicial-Branch/Supreme-Federal-Court. (visited on 04/20/2024).

[30] M. Loh, *'Stop the Steal' leader Ali Alexander calls for a military coup in Brazil to intervene in its presidential election after Jair Bolsonaro's defeat*, https://www.businessinsider.com/ali-alexander-calls-for-brazilian-military-coup-bolsonaro-defeat-2022-10. (visited on 04/20/2024).

[31] F. Sonmez, J. Dawsey, D. Lamothe, and M. Zapotosky, "A frustrated Trump redoubles efforts to overturn election result," *Washington Post*, Dec. 2020, ISSN: 0190-8286. (visited on 04/20/2024).

[32] P. Carvalho and M. J. Silva, *SentiLex-PT 02*, 1970. DOI: 10.23728/B2SHARE.93AB120EFDAA4662BAEC6ADEE8E7585F.

[33] A. ALDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Information Processing & Management*, vol. 58, no. 4, p. 102 597, Jul. 2021, ISSN: 0306-4573. DOI: 10.1016/j.ipm.2021.102597. (visited on 11/13/2023).