Characterizing Toxicity in Facebook Comments

Samuel S. Guimarães  
UFMG, Brazil  
samuelegu@ufmg.br

Jisun An  
QCRI, Qatar  
jisun.an@acm.org

Julio C. S. Reis  
UFMG, Brazil  
 julio.reis@dcc.ufmg.br

Haewoon Kwak  
QCRI, Qatar  
 haewoon@acm.org

Filipe N. Ribeiro  
UFOP, Brazil  
 filipe.ribeiro@ufop.edu.br

Filipe N. Ribeiro  
UFOP, Brazil  
 filipe.ribeiro@ufop.edu.br

1 INTRODUCTION

Today, social media sites like Facebook and Twitter are popular destinations for users to find, share, and discuss real-time news about the world around them. A recent survey by Pew Research Center estimates that 68% of the U.S. adults consume news primarily from social media sites [20]. Since 2018, social media sites have surpassed print newspapers as a news source for North Americans [19].

Social media sites brought new features that helped to promote this kind of online environment to become a place for people to share news, affecting the way news are produced, disseminated, and discussed. First, anyone can register as a news producer on a social media site. For example, anyone can create a Facebook page claiming to be a news media organization. Indeed, a recent study has identified more than 20,000 Facebook pages self-declared as news media only in the U.S. [17]. Second, social media sites offer a novel dissemination strategy, in which users help to share news pieces as an attempt to influence their friends. Finally, social media sites provide ways for users to interact and discuss with each other, through comments sections associated with each post. Among people that comment on news, a survey found that 77.9% of them comment on social media [22].

On one hand, these comments allow users to engage in discussions with the author and open space for different perspectives and to constructive criticism. Comments also help to transform users from passive readers into active ones, allowing news outlets to get their audience insights on how to be better [8]. However, on the other hand, the most common motive for people that comment online to interact is the desire to express an emotion or an opinion [22], and despite the advantages that commenting have, the sentiment expressed can be very negative [16]. Independently of the headline of the article, online news very frequently attracts comments with a negative sentiment [16]. This can transform comment sections into a home for hostility and toxicity, similar to what happened to comment sections in news websites, which prompted a lot of journals to disable the comments section [5].

Although such high toxicity we see in the comment sections in newspapers and in comments associated with news, this situation observed is not in many other contexts and inside different online social media platforms like Twitter [14, 21], YouTube [18], Gab [13] and Facebook [7, 12], However, when we focus on toxic comments, especially towards news and similar content, Facebook stands out. It is the largest social network worldwide, with 2.5 billion monthly active users as of December 2019 [3], and with 67% of U.S. Facebook’s users getting their news there [20]. Daily, those news are flooded with toxic comments [9, 23]. Nonetheless, despite the huge importance of this part of our current news ecosystem, little is still known about it.

In this paper, we aim at filling this gap. Our contribution is to provide a large scale diagnostic about the toxicity in the comments associated with posts in news shared on Facebook. More specifically we expect to answer these following Research Questions (RQs):

- RQ1: What type of pages receive more toxic comments?
- RQ2: Which factors boost the proportion of toxic comments?
- RQ3: What is the common characteristics of the toxic comments?
To answer these questions we took the opportunity of a major event in politics in Brazil and collected all posts and comments from relevant pages during this event. After 580 days in prison for corruption charges, Former President Lula was released on November 8, 2019. Being a controversial figure that the current President of Brazil, Jair Bolsonaro, sees as his major political enemy, this was a major news story, shared by a polarized nation with different narratives. As the event unfolded we collected posts and comments from one week before and one week after his release, focusing on pages that were related to news and politics. We then used the Perspective API from Google to measure the toxicity of the comments and posts. Our analysis of the toxicity in different types of pages, posts, and comments posted in different moments constitute the main results of this paper.

In our results, we discover that pages from news media receive more toxicity than pages from public figures. However, public figures increase bad comments when they become the topic in a post. While the political affiliation is irrelevant for them in terms of the level of toxic content, when we look at media, mainstream gets more incivility than Alternative media. In general, replies are more toxic than just comments to a post, and pages with more toxic posts receive more toxic comments. This last correlation is higher in public figures than in media, higher for Right-wing than in the Left-wing and higher for alternative media than in the mainstream, with no correlation for the mainstream. Another finding is that the toxic comments and toxic posts found are concentrated, with circa 20% of pages responsible for 60% of toxic comments and 56% of all toxic posts. Finally, we find that Lula’s Release shows that a certain political event can also increase toxicity but the effect may be marginal. We hope our findings may shed a light on the key factors that influence toxicity in comments associate with news and may help social media platforms to design better content policies able to deal with this important issue.

The rest of the paper is organized as follows. The second section presents background and related work. We then describe our experimental methodology for finding posts, pages, and comments, and the characteristics of our final data set. Subsequently, we describe our results to answer the three research questions. And finally, we discuss future directions and our conclusions.

2 BACKGROUND AND RELATED WORK

Hate speech, political polarization, and analysis of user interactions in social networks have been hot topics in recent years, especially after some countries passed laws intended to reduce the amount of hate speech in online media, and political polarization reached an apparent peak [6]. In the topic of hateful and toxic speech, some of the recent researches in this trend used a grading of how much hate there is on online messages, some other studies focused on authors or targets of the hate content, and few others analyzed the general picture. Between the efforts in measuring comments, [23] found that impoliteness and incivility were more prevalent on comments on Facebook pages from conservative and local news sites, with around 20% to 40% of user comments on those news stories consisting of uncivil comments. In a similar fashion, [16] showed that comments’ sections from major newspapers are becoming home for hostility, trolls, and negativism, by collecting comments from news websites and using sentiment analysis to confirm this trend. We also focus on comments of news pages but complement the analysis with a comparison to other relevant page categories, like politicians and political activists.

In a different approach, [2] is an example of research interested in the authors of hate. In this work, the authors dissect the #Gamergate controversy, where some blog post turns out to generate a polarizing issue involving social justice, sexism, and feminism in the gaming industry. The analysis they made showed that new and popular accounts were generally more engaged, posted greater negative sentiment and more hate than random users, showing how the authors of toxic content operated. [12] searched data from Youtube and Facebook related to terrorist groups, looking for whether hate speech was used to justify or promote the ideologies of the organization, or their tactics, besides denigrating their targets.

Besides the analysis of the authors of denigrating messages, [14, 21] are examples of authors providing a deeper of who are the most common targets of hate speech in these systems, studying easily identifiable hateful messages, found by using the grammatical structure of the phrase to evaluate if a message was conveying hatred. As we focus on politics, the targeted are more clear in our work, and that facilitates an examination of the extent they are attacked.

Looking in the general picture, few articles try to create or use databases of Facebook news comments, like we intend to use. In [10] the authors created a dataset from Facebook data. In their work, the Amazon Facebook page had its posts collected during 5 years using the Graph API, and after feature extraction on the information, they used Deep neural network (DNN), Extreme Learning Machine (ELM) and Long Short-Term Memory (LSTM) to try to predict the number of interactions of different types the post would receive, based on the post content, form of content (video, link, picture, and text) and some time-related information. [11] built a dataset focused on the notion of the constructiveness of news comments, evaluating the result with a deep learning approach. Their notion of constructiveness was that: “Constructive comments intend to create a civil dialogue through remarks that are relevant to the article and not intended to merely provoke an emotional response. They are typically targeted to specific points and supported by appropriate evidence.” With the database having a reasonable result, an analysis of the relationship of the toxicity with constructiveness was done with the Perspective API. This relates to our approach, as we use machine learning to both calculate ideological bias, as to measure toxicity using the same API.

In summary, calculating toxicity based online data is becoming accurate and useful. On the line of these studies, we conduct a large-scale analysis in Brazilian context and we focus on Facebook, the largest social media platform used for news sharing.

3 METHODOLOGY

In this section, we briefly describe the methodology adopted for the analysis, including our strategy to select and group news outlets and political pages on Facebook and, to infer toxicity of comments and posts associated with them.
3.1 Finding News and Political Facebook Pages in Brazil

The first step in our methodology is selecting the News and Politicians Facebook pages to monitor. We start from a list of 22 Facebook pages introduced in [15], which includes Brazilian news outlets from mainstream and alternative media outlets as well as other meta-information such as their political affiliation and reach. As, there were only 14 pages still active, we used the Facebook Audience Insights Tool\(^1\) to expand the initial set.

This tool helps advertisers to refine the audience they want to show an ad by defining a set of attributes such as age, location, gender, and Interests. The interests are topics the users are interested in, including public figures, politicians, political parties, types of food, restaurants, activities, etc. One of Audience Insights functions helps to specialize the targeted public by suggesting related topics. So given an input Interest, the tool suggests related Facebook Pages with similar audience (‘Page Likes’ menu).

Thus, we choose four interests related to the Brazilian politics scenario as seed and manually searched for related pages. The following interests were used as seed: 1) Jair Messias Bolsonaro, the current Right-wing President; 2) Lula, the left-wing Former President; 3) the Social Liberal Party (PSL), a Brazilian right-wing political party\(^2\); and 4) the Worker’s Party (PT), a Brazilian left-wing political. From the suggested pages matching to each of the four interests, we included those from the following categories\(^3\):

- Public Figures, Politicians, Government Officials, Authors, Political Organizations, Political Parties, Media, News & Media Websites, Media/News Companies, Broadcasting & Media Production Companies, Magazines, Journalists, TV Programs (News related) and Newspapers.

As a result, we ended up with 63 Brazilian News and Political Facebook pages from various categories\(^4\) but they are mainly divided into two categories: Public figures and Media.

For our analyses, we further regroup them into the following five sub-groups:

1. Public Figures:
   - Right-Wing Figures
   - Centrist Figures
   - Left-Wing Figures
2. Media:
   - Mainstream Media
   - Alternative Media

Since our list includes not only politicians, but also political activists, we use the generic term ‘public figure’ instead of ‘politicians’ in our study. For public figures, we use their claimed political leanings for assigning the subgroup, and we ignore pages if they do not not clearly state their political leanings. For Media, we consider pages as Alternative Media only when they identify themselves as an alternative to other media, and we consider pages as Mainstream Media only if they are included in the initial list as mainstream media [15]. Alternative Media tend to present online-only or sometimes they only have Facebook Pages and publish only on Facebook. Next, we present the details of our data collection of posts and comments.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOXICITY</td>
<td>Rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.</td>
</tr>
<tr>
<td>SEVERE_TOXICITY</td>
<td>A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. Less sensitive to comments that include positive uses of curse words.</td>
</tr>
<tr>
<td>IDENTITY_ATTACK</td>
<td>Negative or hateful comments targeting someone because of their identity.</td>
</tr>
<tr>
<td>INSULT</td>
<td>Insulting, inflammatory, or negative comment towards a person or a group of people.</td>
</tr>
<tr>
<td>PROFANITY</td>
<td>Swear words, curse words, or other obscene or profane language.</td>
</tr>
<tr>
<td>THREAT</td>
<td>Describes an intention to inflict pain, injury, or violence against an individual or group.</td>
</tr>
</tbody>
</table>

Table 1: Overview of metrics from perspective API.

3.2 Gathering Dataset

Once we obtain the list of news and politics Facebook pages, we use the Facebook Graph API to collect posts and comments of the pages in our list. The posts and comments data includes textual content, the number of likes, published date, and mentions of other Facebook users in the text. We also collect all the replies of the comments. Using the API, we found posts ranging from mid-2018 to end-2019 in the pages we selected, but we focused on the posts from the period of 31 October to 15 November 2019 to compose our data set, a week prior and after Lula was released, capturing all posts.

3.3 Inferring Toxicity

We use Google’s Perspective API\(^6\) to infer toxicity of the posts and comments in the Facebook pages. There are various models provided by the Perspective API which are described in Table 1. All this metrics exist for English, but an experimental version for Portuguese was release last year. In all cases, given a text, the models return a probability of the text being toxic or an attack, which we call as their scores. When the text is confusing or misspelled, the model returns no score. For 8.17% of the posts and 9.13% of the comments in our dataset, the API were not be able to measure the toxicity. We used all models to measure the toxicity of posts and comments, but we found that results are highly correlated to each other (the Pearson’s correlation coefficient is larger then 0.8

\(^1\)https://www.facebook.com/ads/audience-insights/

\(^2\)This was Bolsonaro’s Party during the search, but he latter left the party.

\(^3\)The administrator of Facebook page assigns the categories to the page among the predefined list of Facebook categories.

\(^4\)Those 63 pages are from various categories–1 Author, 4 Broadcasting & Media Production Company, 5 Magazines, 1 Media, 16 Media/News Companies, 9 News & Media Websites, 1 Political Organization, 15 Politicians and 11 Public Figures.

\(^6\)https://www.perspectiveapi.com
### Table 2: Examples of comments and their toxicity score.

<table>
<thead>
<tr>
<th>Comment (Translated to English)</th>
<th>Comment (Original in Portuguese)</th>
<th>Toxicity Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PUTINh* (A play with the words Putin and Wh’re)</td>
<td>PUTINh*</td>
<td>0.926</td>
</tr>
<tr>
<td>This M’T-o gets nothing more, he is a F’got</td>
<td>F’P não ganha mais nada, é vi’d0</td>
<td>0.905</td>
</tr>
<tr>
<td>Fascist scumbag.</td>
<td>Fascista canalha.</td>
<td>0.905</td>
</tr>
<tr>
<td>How about ... killing Lula and all the supreme court bad guys</td>
<td>tal ... matar-mos Lula e tds os bandidos do STF</td>
<td>0.884</td>
</tr>
<tr>
<td>Do not share anything about the hobo the convicted thief, neither give him a like</td>
<td>Não compartilhem nada do vagabundo do condenado ladrão Nem curtir</td>
<td>0.839</td>
</tr>
<tr>
<td>Cattle can’t read</td>
<td>Os gados não sabem ler</td>
<td>0.736</td>
</tr>
<tr>
<td>They will all go to jail</td>
<td>Vão tudo pra cadeia</td>
<td>0.506</td>
</tr>
<tr>
<td>U know he doesn’t like the poor</td>
<td>sabe ele não gosta de pobre</td>
<td>0.673</td>
</tr>
<tr>
<td>Guuuuuuys ... nobody does anything to stop this scum?</td>
<td>Genteeeee... ninguém faz nada pra cessa essa corja?</td>
<td>0.369</td>
</tr>
<tr>
<td>Rio needs this a lot! It also needs tubal ligation and vasectomy handouts in the communities.</td>
<td>Rio precisa muito disso! Tbm precisa de bolsa laque-adura e bolsa vasectomia nas comunidades.</td>
<td>0.181</td>
</tr>
<tr>
<td>Don’t let them set Barabas out again. Can’t make this mistake again ...</td>
<td>Não deixem soltar Barabás novamente Não podem cometerem este erro de novo ...</td>
<td>0.199</td>
</tr>
<tr>
<td>of course it is ... he wants it to remain state owned to get his hands on the money that comes in . very simple</td>
<td>claro que ta ...ele quer que continue sendo estatal pra meter a mao na grana que entra ..muito simples</td>
<td>0.128</td>
</tr>
<tr>
<td>AI-5 SUPPORTED</td>
<td>AI-5 APOLADO</td>
<td>0.042</td>
</tr>
<tr>
<td>AI 5, I support</td>
<td>AI 5 apoio</td>
<td>0.037</td>
</tr>
</tbody>
</table>

As collecting all the data took a few days, we expect our data to not to represent the situation of the comments as they were posted, considering that Facebook has deleted part of the most toxic ones. After measuring how many comments were deleted two months after our initial search on a random sample, we found that near 1% of them got deleted.

**Toxic comments detection by Facebook itself.** Another limitation of our data set is related that Facebook is actively trying to diminish the toxic environment that can occur on comment threads. As collecting all the data took a few days, we expect our data to not to represent the situation of the comments as they were posted, considering that Facebook has deleted part of the most toxic ones. After measuring how many comments were deleted two months after our initial search on a random sample, we found that near 1% of them got deleted.

Even with those limitations, we believe that our data set can provide interesting insights on toxicity in Facebook comments. Next section we explore. In the following section, we present and discuss the main results from characterizing toxicity in Facebook comments.

### 3.4 Final Dataset

Our list includes 63 news and politics Facebook pages. Using the Graph API, we collected 4,375,019 comments from 15,075 posts. Table 3 shows some aggregate statistics on our data set by different categories. We note that considering popularity in our group division, Media have a bigger public than Public Figures in all popularity metrics we gathered. In the proportion of toxic comments, Media also has a bigger presence of toxic comments in their comment section. In the next major section we will give more details about that difference. The average toxicity of a comment also follows that trend, but the average toxicity of the comments grouped by post varies more in contrast to the Average of all posts for Media than for the posts from Public Figures, considering the metrics of stability of the post toxicity compared to all the group. Differently, the proportion of toxic posts is 1.52% for all Public Figures and 1.07% for Media, with Mainstream Media and Centrist Public Figures with the smallest proportions.

### 3.5 Potential Limitations

There are a few limitations of our data, discussed next.

**Accuracy of the inference of toxicity by Perspective API models for Portuguese.** Measuring toxicity of a text is still a research topic that is in development. The Perspective API represents one of the first ‘off-the-shelf tools available and there are not current studies available about its accuracy in the Portuguese language. Therefore, to estimate its accuracy, we manually labeled a sample of the data and measured the Cohen’s Kappa coefficient of agreement of our labeling, and the agreement of our labeling with the results provided by Perspective API. Two volunteers labeled randomly selected 2,000 comments as toxic or non-toxic, reaching a kappa coefficient of 0.45, with a confidence interval of [0.36-0.55]. The volunteers then discussed the social media content and reached a final verdict on the labelling, that was then compared with the classification given by the Perspective API model using the value of 0.8 as cutoff. The kappa coefficient between the human label and the API was 0.45, with a confidence interval of [0.36-0.53]. Similarly using translation and then using the Perspective API in English, using the metrics that are not experimental, gives a kappa of 0.43, with a confidence interval of [0.34-0.52], which makes the use of the experimental version better than using what is available with translation. These results also show that measuring toxicity of texts is a difficult test and Perspective API can be as good as a Human.
4 TOXICITY ANALYSIS

In this section, we aim at analyzing the toxicity level of comments and posts and checking the extent to which the toxicity is correlated with different categories and ideological leaning.

Figure 1: Distribution of toxicity for all comments and posts.

Figure 1 shows the distribution of the toxicity scores for all comments and posts. In this work, we consider a comment or post to be toxic if the toxicity score is above 0.8, following our validation of the Toxicity measure, and the previous work [4]. By using the threshold of 0.8, we find that 12.1% of the comments and 1.1% of the posts are considered to be toxic. Although the overall fractions of toxic comments and posts may seem to be low, the level of toxicity varies across pages by the factor of two to ten compared with the average values. Figure 2 shows such large variation. We can see that 20% of all pages responsible for 60% of toxic comments and 56% of all toxic posts. A large number of toxic comments on a page may indicate that the page is being attacked by a group of angry users, or there is a fight occurring between users created by discussion in the comment section. Thus, in the following experiments, we examine which type of pages are more likely to have toxic comments and posts.

4.1 Toxicity in Brazilian Pages

What types of pages are more likely to receive more toxic comments? In this section, we look into how various characteristics of the pages (media type, political leaning, the level of toxicity in posts) are relating to the level of the toxicity in the comment sections. To analyze the toxicity of various pages, we start by calculating the proportions of toxic comments for each of all posts.

We first investigate how the political leaning of public figures (left, center, vs right leaning) and the type of media (mainstream vs alternative) of pages are relating to the level of toxicity in their comment sections. For each sub-group, we aggregate all the comments of the corresponding pages. Figure 3 shows the distribution of the comments by their toxicity scores for each sub-group. We first observe that the average toxicity score for the public figures (0.342) is slightly lower than those for media (0.429). In terms of the proportion of toxic comments (i.e., comments whose toxicity score is above 0.8), the medians of the proportions of toxic comments on public figure posts and media posts were 9.98% and 16.50%, respectively. We ran a Mann-Whitney's U test to evaluate the difference in the proportions of toxic comments of posts in the two groups. We found a significant effect of the type of pages (U = 12380000, p < 0.005), showing that media posts receive statistically more toxicity than public figures posts. One possible reason is that people may consider the pages of public figures as homogeneous political discussion spaces where they support for the corresponding figure while consider the media pages as cross-cutting political discussion spaces where people with various political leanings come together and argue or discuss [1]. Then, in homogeneous spaces, since they talk with like-minded people, there may be less toxicity. On the other hand, in cross-cutting spaces, there may be higher chance to meet people whose opinions are opposing from their own, which then may trigger to use toxic words in their comments. In fact, all media pages, except one, have more than 10% toxic comments, while half the public figure pages such level of toxicity in their comment sections. The pages by Lula, Michel Temer, and Jair Bolsonaro, who are the former and current presidents of Brazil, are the second, third, and sixth pages by the lowest proportions of the toxic comments in their pages.

With Lula and Bolsonaro being the most polarizing figures of the recent years, it is unexpected that they have such low values. Beyond that being homogeneous political discussion spaces, another possible explanation is that, as politicians, their accounts may be maintained by the professionals. As the public figure’s online presence is extremely important, their assistants could flag the toxic comments and try to remove them. In contrast, news media might not have the same amount of man power, specially for alternative media, and in some cases they might even benefit if people come to their page to fight about current events, so their effort to flag comments and moderate them might be lesser extent.
Next, we examine whether the toxicity in the posts affect the toxicity in the comments. When comparing the proportions of toxic comments and toxic posts across pages, we find a moderate positive correlation with the Spearman’s ranking correlation coefficient ($\rho$) of 0.463 and $p < 0.05$. For the pages of the public figures, the correlation between the proportion of toxic posts and the proportion of toxic comments is even higher ($\rho=0.660$). However, it varies by the political leanings of public figures. For the pages of the centrist public figures, the correlation is not significant, and for the pages of the left-wing public figures, the correlation is higher ($\rho=0.699$) and significant. Interestingly, the pages of the right-wing public figures show a very high positive correlation ($\rho=0.808$). For media pages, we observe a moderate positive correlation of 0.531. However, when looking at by different media type (mainstream vs. alternative), we find the same tendency only for the pages of alternative media ($\rho=0.593$), but not for the pages of mainstream media. We observe a clear pattern of using more or less toxic posts attracts more or less toxic comments for right-wing and left-wing public figure pages and alternative media pages, but not for centrist public figure pages and mainstream media pages. The results indicate that the users active in certain pages (e.g., right/left-wing public figure pages and alternative media pages) are more reactive to the level of toxicity of the posts.

### 4.2 Toxicity in Posts

Next, we examine the proportion of toxic comments to select the worse posts and also analyze to see if it is related to politics. We create a list of posts with more than 20 comments and another with the posts with more than 500 comments to examine the proportion of toxic comments in them. The results are in Tables 4 and 5.

![Figure 3: Distribution of toxic comments by political leaning and media type.](image)

#### Table 4: Top 10 Posts with 20 comments or more.

<table>
<thead>
<tr>
<th>Page</th>
<th>Post Content</th>
<th>Post Toxicity</th>
<th>Toxic Comments</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brasil de Fato</td>
<td>Politics and Racism</td>
<td>0.6136</td>
<td>59.09%</td>
<td>22</td>
</tr>
<tr>
<td>O Globo</td>
<td>Music Award</td>
<td>0.2959</td>
<td>56.82%</td>
<td>22</td>
</tr>
<tr>
<td>Gente de O Estado</td>
<td>Return of PT</td>
<td>0.2831</td>
<td>55.93%</td>
<td>115</td>
</tr>
<tr>
<td>Diário do Brasil</td>
<td>Politics and Police</td>
<td>0.2818</td>
<td>53.33%</td>
<td>120</td>
</tr>
<tr>
<td>Jornal da Cidade Online</td>
<td>Politics and Racism</td>
<td>0.3604</td>
<td>55.25%</td>
<td>92</td>
</tr>
<tr>
<td>Diário do Brasil</td>
<td>Politics and Fake News</td>
<td>-</td>
<td>51.00%</td>
<td>47</td>
</tr>
<tr>
<td>Jornal da Cidade Online</td>
<td>Politics and Internet</td>
<td>0.3204</td>
<td>51.02%</td>
<td>49</td>
</tr>
<tr>
<td>Jornal do Brasil</td>
<td>Politics and Racism</td>
<td>0.4781</td>
<td>58.21%</td>
<td>283</td>
</tr>
<tr>
<td>UOL</td>
<td>Politics and Internet</td>
<td></td>
<td>46.88%</td>
<td>79</td>
</tr>
<tr>
<td>Falando Verdades</td>
<td>Glenn Greenwald is punched</td>
<td>0.7269</td>
<td>46.15%</td>
<td>104</td>
</tr>
</tbody>
</table>

#### Table 5: Top 10 Posts with 500 comments or more.

<table>
<thead>
<tr>
<th>Page</th>
<th>Post Content</th>
<th>Post Toxicity</th>
<th>Toxic Comments</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eduardo Bolsonaro</td>
<td>Return of PT</td>
<td>0.4260</td>
<td>42.54%</td>
<td>6279</td>
</tr>
<tr>
<td>Diário do Brasil</td>
<td>Politics and Money</td>
<td>0.1630</td>
<td>16.30%</td>
<td>166</td>
</tr>
<tr>
<td>Marcelo Frizzo</td>
<td>Marielle Franco and Bolsonaro</td>
<td>0.2492</td>
<td>24.92%</td>
<td>3646</td>
</tr>
<tr>
<td>Jornal da Cidade Online</td>
<td>Politics</td>
<td>0.7940</td>
<td>79.40%</td>
<td>1730</td>
</tr>
<tr>
<td>Mário da Rosário</td>
<td>Bolsonaro Attack</td>
<td>0.6878</td>
<td>68.78%</td>
<td>3578</td>
</tr>
<tr>
<td>Jornal da Cidade Online</td>
<td>Bolsonaro</td>
<td>0.6278</td>
<td>62.78%</td>
<td>1505</td>
</tr>
<tr>
<td>Eduardo Bolsonaro</td>
<td>Politics and Fake News</td>
<td>0.0500</td>
<td>0.50%</td>
<td>3488</td>
</tr>
<tr>
<td>Eduardo Bolsonaro</td>
<td>Politics and Aggression</td>
<td>0.0824</td>
<td>8.24%</td>
<td>3549</td>
</tr>
<tr>
<td>Eduardo Bolsonaro</td>
<td>Politics and Money</td>
<td>0.6329</td>
<td>63.29%</td>
<td>2892</td>
</tr>
<tr>
<td>Jornal da Cidade Online</td>
<td>Politics and Money</td>
<td>0.5253</td>
<td>52.53%</td>
<td>3283</td>
</tr>
</tbody>
</table>

Interestingly, we can see a little overlap between two tables. When considering posts with more than 20 comments, more left-leaning pages appear in top 10. By contrast, in the posts with more than 500 comments, right-wing pages become more prevalent. Such trend indicates that partisan pages from the left can ignite or suffer with more toxicity relative to their sizes. By contrast, right-leaning pages attract a larger audience and high amounts of toxic messages. From the perspectives of content, we also see that Racism is a common topic attracting toxic comments along with politics. As
more than half the population of Brazil is non-white, and slavery was abolished very late, racial tensions are one of the social issues in Brazil. News about celebrities are likely to attract toxic comments, being more present when only considering posts with more than 20 comments. The only posts created by a politician or public figure in the list are written by Congressman Eduardo Bolsonaro, an deputy son of the current President, with a post about Lula and his father, and Deputies Marcelo Freixo and Maria do Rosário, both harsh critics of Bolsonaro, with posts about him. These three politicians are all present in the top 5 of pages with biggest proportion of toxic comments, with Rosário and Freixo in the first and second position, respectively. That post shows that Lula and Bolsonaro are part of the polemic topics that creates hate in the comment section when is posted in other places besides their page.
4.3 Toxicity in Comments

To delve deeper into the relationship between what causes a user to publish toxic comments, we analyze all the comments whose toxicity score is 1 (the maximum score). We first look into where those comments are belonging to among the five sub-groups (Figure 8). We see that right-wing pages have the largest proportion of the most toxic comments, and alternative media have slightly higher number of the most toxic comments than mainstream media.

We find two interesting patterns by manual examination. First, as we said previously, Bolsonaro has a low proportion of toxic comments, but he still receives the extreme toxic messages (i.e., toxicity score is 1). It might imply that he does not actively remove the toxic comments, but due to his large audience, the proportion of toxic comments stays at low-level. Or, those comments might not aim at Bolsonaro and thus are not flagged. Second, we also see that Jornal da Cidade Online, an alternative media page, has the most toxic comments, more than the President.

We also examine content of the top 10 toxic comments with toxicity less than 1, which is presented in Table 6. As we can clearly see, the topic of politics is predominant, with Bolsonaro and Lula being in 8 of the 10 comments, with only one being about soccer instead of politics, but still with fascist cited as insult. Given that Lula’s page is with very few toxic comments, it is surprising to see him appeared in the comments with the highest toxicity. Carlos Bolsonaro and Caneta Desesquerdizadora, the alderman son of the president and an alternative media website, are interesting presences in both the top 10 toxic comments equal to 1, and top worse comments below one. They are distinguished from other pages in terms of toxicity of contents. About 10.00% of Bolsonaro’s posts, and 8.00% of the alternative media, have the toxicity score of higher than 0.8. This might indicate that they talk about polemic topics and create more polemic content. As our data collection covers the period of that the Ex-President Lula was release from jail, we search if he is a topic in the toxic posts by looking for his complete name and common nickname in the posts text. We find that 7.85% of the posts of all pages are about Lula. Figure 9 shows that the distribution of the proportion of toxic comments for posts about Lula is slightly higher, with a mean of 11.69% compared to a 10.38% mean for posts not related to him. This indicates that people leave toxic comments about him much more frequently in posts about him, and he becomes a controversial topic that attracts a lot of hate.

Figure 10 shows that Bolsonaro has a similar effect; the average proportion of toxic comments in the posts mentioning him is 14.20%, which is higher than 10.08% from the posts without mentions about him. Moreover, we can see the names of Lula and Bolsonaro in the most common words in the comments with toxicity equal to 1, the only names present, shown in Figure 11.

4.3.1 How Lula’s Release affected the comments. Finally, we answer the question about how a single political incident can change the toxicity in the comment section. As we mentioned earlier, our data collection is a valuable resource to see how online discussions were made around the release of Lula, following 580 days after his controversial arrest. From our data collection, we extract posts from one week before and one week after his release. We compare the toxicity of posts and comments before and after the release of Lula. The results are shown in Figures 5 and 6. We can see the release of Lula increases the toxicity levels and proportion, and the increase is statistically significant. Doing the Mann-Whitney U Test, we find that before the release the posts have in average 14.49% of toxic comments, which increase to 15.14%. Similarly, the average toxicity level in a comment was 0.3738 and it increase to 0.3980, in both cases with a confidence level of 99.9%, and U value of $2.0949 \times 10^{12}$. This shows that a single political incident can cause a significant change in the toxicity of the comments. However, comparing this difference with the previous analyses, the difference in toxicity between the page types is bigger than the increase caused by the political incident. In the same way, as we see more toxic comments from the posts about Bolsonaro or Lula, the content is impactful.
<table>
<thead>
<tr>
<th>Facebook Page</th>
<th>Comment (Translated to English)</th>
<th>Toxicity</th>
<th>Post Content (Translated to English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBC News Brasil</td>
<td>F<em>ck the f</em>cked!!</td>
<td>0.99999</td>
<td>News about Lula lawyers gaining an appeal</td>
</tr>
<tr>
<td>BBC News Brasil</td>
<td>The big fool took a foot up his *ss from Trump</td>
<td>0.99998</td>
<td>News about Jair Bolsonaro talking about China</td>
</tr>
<tr>
<td>Magno Malta</td>
<td>Thug Son Of A B<em>tch! Fascist pastor! F</em>ck you!</td>
<td>0.99998</td>
<td>Magno Malta defends Jair Bolsonaro</td>
</tr>
<tr>
<td>Portal R7</td>
<td>Already sh<em>t his pants!! Neither the people nor the military!! A C</em>ck up the <em>ss of this sh</em>tty comunist!!</td>
<td>0.99995</td>
<td>News about the Vice president of Bolivia renouncing</td>
</tr>
<tr>
<td>Eduardo Bolsonaro</td>
<td>This <em>sshole just talks sh</em>t.</td>
<td>0.99993</td>
<td>Eduardo defends himself after talking about the 1964 Brazilian coup d’état</td>
</tr>
<tr>
<td>Eduardo Bolsonaro</td>
<td>You <em>sshole son of a b</em>tch you’ll go to jail with your dad and this whole militia gang</td>
<td>0.99990</td>
<td>Video of Jair Bolsonaro on accusations about a killed opposition politician</td>
</tr>
<tr>
<td>Carlos Bolsonaro</td>
<td>Put the f<em>cking monarchy in charge, and f</em>ck it</td>
<td>0.99989</td>
<td>Jair Bolsonaro talking about Lula being released</td>
</tr>
<tr>
<td>Brasil 247</td>
<td>Dumb *ss. Shut up!!!</td>
<td>0.99986</td>
<td>News article about Jair Bolsonaro condoning the 1964 Brazilian coup d’état</td>
</tr>
<tr>
<td>Eduardo Bolsonaro</td>
<td>An idiot sh*tty journalist</td>
<td>0.99986</td>
<td>Eduardo talking about a news piece about his wife</td>
</tr>
<tr>
<td>Caneta Desesquerdizadora</td>
<td>Sh*tty Supreme Court , gang of sons of bitches</td>
<td>0.99985</td>
<td>Satirical cartoon about Lula and the Brazilian Supreme Court</td>
</tr>
</tbody>
</table>

Table 6: Top 10 Worst Comments on the data set with toxicity below 1.

Figure 9: Cumulative distributions of toxic comments by Lula related content.

Figure 10: Cumulative distributions of toxic comments by Bolsonaro related content.

5 CONCLUSION
In this paper, we presented an in-depth characterization of the toxicity score from 4,375,019 online comments in 15,075 posts of 63 Facebook pages, collected during two weeks, from 31 October to 15 November 2019. After dividing the pages in Right-wing, Centrist or Left-wing Public Figures, and Mainstream Media or Alternative Media, depending on if the page is from Media or Public Figures, we analyzed their toxicity score in various levels of granularity.

Our results disclosed a series of interesting trends. We find that the toxic comments and toxic posts are the minority, 12.11%, and 1.13% respectively, but are concentrated, with about 20% of pages responsible for near 60% of toxic comments and 56% of all toxic posts. In general, replies are more toxic than just comments to a post, and there is a positive correlation between the page proportion of toxic posts and toxic comments, even that directly toxicity of the post does not correlate to the comments’ score. In our division,
news media pages receive more toxicity than pages from public figures, but when public figures are cited in a post, the proportion of toxic comments increases. The political affiliation of the public figures did not affect the proportion of toxic comments or posts, but Mainstream media receives more toxic comments than Alternative Media. Finally, we found that Lula’s Release shows that a certain political event can also increase toxicity but the effect may be marginal. We expect that these insights can be used to improve and guide Content Policies for Facebook, as well as comment sections in general because it shows where the toxicity is concentrated, how it is boosted, and common characteristics of its content.

Our findings provide a better understanding of this complex ecosystem and identify key factors that may influence toxicity in comments associate with the news. We hope this can be valuable for the design of new features to social media systems and these platforms to design better content policies able to deal with a toxic online environment.

REFERENCES


Figure 11: Words find in the worse comments tied for first place in the top 10.