Toxic Content Detection in Online Social Networks: A New Dataset from Brazilian Reddit Communities

Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, Ana Paula Couto da Silva
luiz.quevedo@dcc.ufmg.br, apagano@ufmg.br, ana.coutosilva@dcc.ufmg.br
Universidade Federal de Minas Gerais

Abstract

The proliferation of online social interactions in recent years, with the consequent growth in user-generated content, has brought the escalating issue of toxic language.While automatic machine learning models have been effective in moderating the vast amount of data on online social networks, low-resource languages, such as Brazilian Portuguese, still lack efficient automated moderation tools. We address this gap by creating a high-quality dataset collected from some of the most popular Brazilian Reddit communities. To that end, we manually labeled a sample dataset of 2,500 comments extracted from the most engaging communities. We conducted an in-depth exploratory analysis to gain valuable insights into the language of *toxic* and *non-toxic* content. Our results show a high level of agreement among annotators, attesting to the suitability of this dataset for various downstream machine learning tasks. This research offers a significant contribution to the creation of a safer online environment for users engaging in discussions in Portuguese and paves the way for more effective automatic moderation tools using machine learning.

1 Introduction

With the growth in the number of online social network platforms, increasingly more users are interacting through online media. According to (Statista, 2022), the total number of users of different social networks is 4 billion people. This figure indicates the level of importance and ubiquity of these online platforms in society and their impact, not always beneficial, on people's lives. According to (Vogels, 2021), a study conducted in 2020 with US adults found that around 41% of respondents had experienced some form of online harassment. In addition, abusive comments in discussions propagate toxicity and harmful user engagement, radicalizing discussions (Salehabadi et al., 2022). The consequences of these interactions transcend the virtual world, seriously affecting the lives of real users. According to (Vogels, 2021), 18% of the users who took part in a survey had suffered some kind of abuse considered severe beyond the online environment, including physical threats and stalking.

The manual moderation of user-generated content has long been considered the primary approach to mitigate the negative impact of toxic interactions. However, the scale and speed at which content is generated make manual moderation impractical, prompting the need for automated solutions. Machine learning models have emerged as a promising alternative for automating the moderation of online created content. These models can identify potentially harmful content, enabling platforms to proactively take actions such as banning users and removing harmful content. While machine learning models have proved effective in several languages (Perspective, 2022b), their performance for low resource languages, such as Brazilian Portuguese, is still a concern.

Seeking to address these challenges, this paper introduces a new dataset for toxicity detection in Brazilian Portuguese. The annotated texts were retrieved from one of the most relevant online social networks - Reddit -, which has around 1.5 billion registered users and 430 million active users (Wise, 2023). Reddit is a community that allows users to interact through anonymous posts (submissions) and comments. Users are organized into communities (subreddits) and subscribe to the communities most aligned with their topics of interest. The collection and annotation of these data are motivated by the need to propose new models of toxicity detection and improve existing ones for the unique characteristics of the Portuguese language. Also, the dataset is tailored specifically for online social network data, filling the gap on available models for Portuguese in this domain.

The remainder of this paper is organized as follows. We first review the available literature on

toxicity detection in Portuguese. Next, we introduce the techniques and methodology for our data collection and annotation. We then describe the overall quality of the dataset and report on an experiment comparing our annotation to the one by the Perspective API. Subsequently, we characterize the language used in *toxic* and *non-toxic* comments. Finally, we discuss our findings and their impact, particularly regarding the use of our dataset to fine-tune existing toxicity classification models, seeking to improve automatic content moderation in an ever-growing online environment. By addressing the shortcomings in existinsg resources, we aim to contribute to the efforts to make online social networks safer and more inclusive for all. To allow reproducibility and foster follow-up studies, we have published the annotated dataset for public access.[1]

2 Related work

There are few studies in automatically detecting toxic comments in languages like Brazilian Portuguese, with annotated datasets released for public use and follow-up studies.

Authors in (de Pelle and Moreira, 2017) make available a dataset with 1,250 comments, extracted from comment sessions of g1.globo.com website and annotated for the categories offensive and non-offensive, 32,5% of the total being labeled as offensive. The offensive class was further subdivided into *racism, sexism, homophobia, xenophobia, religious intolerance, and cursing*. Cursing, including vulgar language, was the most frequent category of offensive comments, present in almost 70% of the comments found offensive.

In (Fortuna et al., 2019), the authors describe a dataset with 5,668 tweets, annotated using a hierarchical annotation scheme by annotators with different levels of expertise. Non-experts annotated the tweets with binary labels (*hate vs. no-hate*). Then, expert annotators classified the tweets following a fine-grained hierarchical multiple label scheme with 81 hate speech categories in total.

(Leite et al., 2020) introduce ToLD-Br: a dataset for the classification of toxic comments on Twitter in Brazilian Portuguese. A total of 21K tweets were manually annotated into seven categories: *non-toxic, LGBTQ+phobia, obscene, insult, racism, misogyny and xenophobia*. Each tweet had three annotations made by volunteers from a university in Brazil. Through a wide and comprehensive analysis, they demonstrated the need for building large monolingual datasets for studies of automatic classification of toxic comments.

The performance of the Perspective API for Brazilian Portuguese is assessed in (Kobellarz and Silva, 2022). Comments from two Brazilian news media websites were translated into English and their toxicity was scored by the Perspective API. Human-annotated comments from the news comments dataset were used to assess the scores provided by the Perspective API for the original and the translated versions. Their results show a better performance for texts in their original language.

HateBR corpus was built and shared by the authors in (Vargas et al., 2022). The corpus consists of 7,000 comments from Brazilian politicians' accounts on Instagram, manually annotated by specialists, with a high inter-annotator agreement. The documents were annotated according to three different layers: a binary classification (offensive versus non-offensive comments), offensiveness-level (highly, moderately, and slightly offensive), and nine hate speech groups (*xenophobia, racism, homophobia, sexism, religious intolerance, partyism, apology for dictatorship, antisemitism, and fatphobia*).

(Trajano et al., 2023) introduce OLID-BR, a high-quality NLP dataset for offensive language detection. The dataset contains 6,354 (extendable to 13,538) comments labeled using a fine-grained three-layer annotation schema compatible with datasets in other languages, which allows the training of multilingual models.

Our work contributes to studies on toxic content characterization by exploring Brazilian Portuguese comments posted on Online Social Networks. To the best of our knowledge, this is the first study focused on building and characterizing a Brazilian Portuguese Reddit corpus, manually annotated for toxicity.

3 Methodology

In this section, we first outline our methodology for corpus collection. Then, we describe the annotation process to manually label a sample of comments as *toxic* and *non-toxic*. Last, we present the methods used to analyse the language of the labeled comments.

[1]The dataset is available on https://github.com/luizhenriqueds/reddit-br-toxicity-dataset/.

3.1 Reddit data collection

Reddit is a multilingual Online Social Network founded in 2005 and organized in subcommunities by areas of interest (subreddits). Our dataset consists of user activities (posts and comments) that took place between January and December 2022 in the top-10 Brazilian subreddits with the largest number of subscribers[2] as well as a lifespan of at least five years, which attests to their importance within this online social network.

Table 1 presents the selected subreddtis and some descriptive statistics. We collected a total of 7,348,257 comments and 390,924 posts via Pushshift, a third-party API that aggregates Reddit comments and posts (Baumgartner et al., 2020) Henceforth, we refer to both comments and posts made by the users as *comments*.

Our dataset is restricted to comments in Portuguese only, excluding comments from communities that allow multilingual discussions. Approximately 600k comments, in which the text was replaced with either *deleted* or *removed*, were excluded from the analysis as well as comments containing only emojis or symbols, URLs and laughing text reaction.[3] Finally, we also excluded comments generated by automoderator and bots accounts we detected in our data. These filters reduced our corpus to approximately 6.6M comments. Table 2 presents some statistics for the analyzed subreddits upon applying the filters.

3.2 Annotation process

First, we sampled 2,500 comments from our filtered corpus using a stratified sampling process that preserved the original distribution of the total number of comments by month in each subreddit. This sample of comments was divided into 5 batches of 500 examples each. We then recruited 12 undergraduate and graduate students from Computer Science and Language Studies courses at a Brazilian university as annotators. The students were divided into 4 groups and were instructed to label each Reddit comment with one of four available categories: *Toxic*, *Non-toxic*, *I do not know* or *Insufficient information to label the content*.[4] For annotation purposes, we assumed toxic language involves *a rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion*, as defined by the Perspective API. Each group was assigned a batch and each comment was labelled by three independent annotators. One of the groups was assigned an additional batch of comments, given the high quality of annotation they performed as will be discussed in Section 4.1.

A Reddit comment is classified into one of the categories when there is a majority consensus among the annotators. We applied three metrics to measure inter-rater agreement: Fleiss' Kappa statistic, Krippendorf's alpha and Observed Agreement.

3.3 Language characterization

To investigate whether there are patterns in language choices for toxic content produced by Portuguese language users, we performed the following analysis in our manually annotated dataset.

Automatic Toxic Comments Identification: To measure the correlation between automatic and manual identification of toxic content in the sampled Reddit comments, we chose the Perspective API as our baseline (Perspective, 2022b). The Perspective API is a set of out-of-the-box toxicity classifiers from Google Jigsaw, which has been used extensively in prior research (Almerekhi et al., 2020; Salehabadi et al., 2022; Zannettou et al., 2020; ElSherief et al., 2018). The API takes a comment as input and returns a score from 0 to 1 for several classifiers (e.g., profanity, threats, identity attacks, general toxicity). Regarding the Portuguese language, the authors in (Perspective, 2022a) report an Area Under the ROC-curve (AUC) of 0.89 for the model classification task.

POS Tag Analysis: To characterize the language of *toxic* and *non-toxic* comments, we explored frequency of words used and their POS class. To perform POS tagging (Petrov et al., 2011), we used a pre-trained package model (spaCy, 2022) based on a Universal Dependencies treebank for Portuguese, following the work presented in (Rademaker et al., 2017). The selected model achieves a high accuracy of above 97%.

We computed frequency of POS tags for *toxic* and *non-toxic* comments in order to find out whether this could be a distinctive characteristic of the two types of comments.

Type-Token Ratio (TTR) Analysis: We used TTR as a measure of lexical variety in vocabulary. TTR is calculated as the total number of unique words (types) divided by the total number of words (to-

[2]Following the ranking presented in (Almerekhi et al., 2019)

[3]In Portuguese, laughing text is represented by the character sequence kkkkk.

[4]The last category was included as a category to be further pursued in our future work on toxicity diffusion on Reddit.

Subreddit	Subscribers	Posts	Comments
r/brasil	1,516,433	115,876	2,382,928
r/desabafos	490,049	115,876	1,487,076
r/futebol	369,925	35,826	1,272,009
r/saopaulo	358,681	7,308	88,894
r/eu_nvr	308,064	12,631	221,348
r/botecodoreddit	270,451	7,059	62,999
r/conversas	247,545	21,967	355,761
r/investimentos	232,485	9,756	156,695
r/tiodopave	219,926	2,371	12,106
r/brasilivre	210,582	67,301	1,308,441
	Total	390,924	7,348,257

Table 1: Selected subreddits, number of subscribers, posts and comments for the year of 2022.

Subreddit	Posts	Comments
r/brasil	110,829	2,136,866
r/desabafos	115,876	1,211,643
r/futebol	35,826	1,214,412
r/saopaulo	7,308	81,969
r/eu_nvr	12,631	188,620
r/botecodoreddit	7,059	57,298
r/conversas	21,967	326,061
r/investimentos	9,756	141,823
r/tiodopave	2,371	11,584
r/brasilivre	67,301	1,219265
Total	390,924	6,589,541

Table 2: Subreddits statistics upon the filtering process.

kens) in a given segment of language. We also compared the length of *toxic* and *non- toxic* comments. Differently from other online social networks, Reddit does not restrict text length very much, so this feature allows us to compare the likelihood of users posting a short versus a long text on the platform.

Topic Analysis: To find out the topics of the comments on which annotators agree or disagree the most, we ran BERTopic model (Grootendorst, 2022), which relies on an underlying word embedding representation to cluster similar documents.

Named Entity Recognition: We investigated named entities in the Reddit comments relying on a pre-trained model from Spacy for Named Entity Recognition (NER). The model used was trained for Brazilian Portuguese using the WikiNER dataset (Nothman et al., 2013) and classifies entities into 3 predefined categories: PERSON, LOCATION and ORGANIZATION. Undefined entities are classified as MISCELLANEOUS.

4 Results

In this section, we present the key results obtained from evaluating and characterizing the manually annotated dataset.

Metric	Overall	Binary labels (*Non-toxic* or *Toxic*)
Fleiss kappa	0.31	0.46
Krippendorff's alpha	0.35	0.46
Observed Agreement	0.64	0.80

Table 3: Inter-annotator agreement.

4.1 Annotator Agreement

We first measured the overall degree of inter-annotator agreement across the manually labeled Reddit comments, the results of which are shown in Table 3.

As expected, the *Observed Agreement* metric achieved the highest values, as this measure does not take into account the possibility of agreement occurring by chance. Total agreement and disagreement occurred in 1,594 and 107 comments, respectively. An example of total agreement on a comment as toxic is: *"Como assim? Eu nem sou o OP. Só tô dizendo que ele é retardado de seguir a medicina de gado"*.[5] On the other hand, an example of total disagreement is a controversial comment such as: *"[..] é o lugar do Brasil que mais tem neonazi mesmo ué"*[6]), which points to the high level of subjectivity of the classification task.

Regarding *Fleiss kappa* and *Krippendorff's alpha* metrics, their values indicate fair to moderate agreement in the worst case. Finally, the overall toxicity rated by the annotators was 11.28%, with 88.7% of *non-toxic* comments, which is consistent with the imbalanced nature of this problem.

We then measured inter-annotator agreement of each group of students, named A, B, C and D, for

[5]English translation: What do you mean? I'm not even the OP [original poster]. I'm just saying he's stupid to follow the sheep and take those medications.

[6]English translation: [..] it's the place in Brazil with the biggest number of neo-Nazis

the batches of comments, numbered as 1, 2, 3, 4 and 5. Batches 3 and 5 were annotated by group C, while batches 1, 2 and 4 were annotated by groups A, B and D, respectively. Batch 5 was labeled in a second round of annotation by Group C, selected to do so for being the group with the highest *Fleiss kappa* and *Krippendorff's alpha* inter-agreement values in the first round. Results are displayed in Table 4. Except for Group D, which achieved an agreement none to slight, groups A, B and C achieved fair to moderate agreement.

Next, we examined the labeling done by each annotator, the results of which are shown in Table 5. Group A labeled as *toxic* the lowest percentage of comments. Group B presents the highest variability in labeling toxic content, annotator 2 being the one who labeled more than 21% of comments as *toxic*. Like Group B, Group D achieved a non-negligible level of uncertainty in the classification task, annotator 2 tending to be more tolerant of potential *toxic* content. For the sake of illustration, the comment *"Vamos fingir que não é (você) que posta que quer morrer por ser depressivo. Pick me boy"*[7], was classified as *toxic* by annotators 1 and 3 and as *non-toxic* by annotator 2. Annotators from Group C, who worked on batches 3 and 5, are the ones with the lowest degree of uncertainty.

We further investigated the comments on which annotators held complete disagreement, particularly concerning primary topics extracted using BERTopic model. They have to do with discussions related to specific groups (women, men) and encompass various themes including finance, war, government, and relationships (Table A.1). Words in topic 0 (*feedback, removal*) reveal that some comments were previously moderated by DMCA (Reddit, 2020). Interestingly, the main topics in comments about which annotators held complete agreement also discuss the same themes (Table A.2). However, the topic descriptors include many more offensive (such as curse words) as well as ideologically loaded terms. Due to space limitations, the complete list of topics is shown in Appendix A.

Overall, our results corroborate the high level of subjectivity implicated in the task of classifying content as either *toxic* or *non-toxic*. This is in line with findings in the literature on how perceiving the severity of harmful content is impacted by individual and cultural values (Jiang et al., 2021).

[7]English translation: Let's pretend that you are not the author of those posts saying you want to die because you're depressed. Pick me boy.

4.2 Manual and Perspective API's Labeling

Next, we compared our data annotation performed by the Perspective API. We considered toxic comments which were assigned a score of **severe toxicity** above 0.7 by the Perspective API. This decision prioritizes a good balance between precision and recall, as our intention is to gain a better understanding of the main reasons behind agreement and disagreement in the classification of *toxic* and *non-toxic* content. A threshold value of 0.9 results in only 3% of toxic comments being selected for comparison. In contrast, a value of 0.7 returns approximately 10% of comments as toxic, a similar percentage to the one labeled by our annotators.

Toxicity Percentage: First, we analysed the percentage of comments annotated as *toxic* by our students and the one labeled by the Perspective API. Group A (batch 1) annotated less toxicity than the Perspective API, while one annotator in Group B (batch 2) classified a much higher percentage of comments as *toxic*. Group C (batches 3 and 5) is consistent in overestimating Perspective API's predictions. Group D (batch 4), though showing high disagreement between annotators, also annotated less toxicity that the API. Table 6 shows the performance of the Perspective API on a test sample labeled by the Group C. The goal of this analysis is not to directly compare the agreement between the human annotators and the Perspective, but rather to assess the quality of the Perspective predictions at different thresholds on a curated test set. The results indicate a clear performance trade-off between precision and recall. In practice, by choosing a high precision threshold, we are trading a large portion of recall performance. Therefore, the trained model from Perspective has a large margin of improvement for Brazilian Portuguese texts, considering the selected thresholds. Combining both recall and precision metrics, we get a maximum F1 score of 0.67.

Regarding the topics extracted from comments which all three annotators agreed upon as *toxic* and the Perspective API predicted as *non-toxic* (Table A.3), the main ones have to do with politics, freedom, discrimination and targeted groups. The results indicate that the Perspective API is less context-aware for this specific task for Brazilian Portuguese. For instance, the following comment was labeled as *toxic* by all three annotators, but predicted as *non-toxic* by the Machine Learning

Metric	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
Fleiss kappa	0.46	0.33	0.51	0.17	0.54
Krippendorff's alpha	0.46	0.33	0.51	0.17	0.54
Observed Agreement	0.87	0.81	0.76	0.78	0.77

Table 4: Inter-annotator agreement evaluation metrics per annotation batch.

	Batch 1 (Group A)			Batch 2 (Group B)			Batch 3 (Group C)			Batch 4 (Group D)			Batch 5 (Group C)		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
Non-toxic	84.60%	88.96%	90.60%	83.17%	69.48%	74.95%	75.90%	68.01%	78.51%	72.80%	93.59%	69.14%	84.51%	68.60%	75.20%
Toxic	9.40%	9.84%	7.40%	7.82%	21.29%	4.81%	19.28%	21.73%	17.87%	11.60%	5.21%	9.02%	14.49%	25.00%	19.72%
I do not know	0.60%	1.00%	0.00%	3.81%	3.82%	2.81%	4.02%	7.65%	3.01%	4.20%	0.80%	6.41%	1.01%	4.20%	4.67%
Insufficient Info	5.40%	0.20%	2.00%	5.21%	5.42%	17.43%	0.80%	2.62%	0.60%	11.40%	0.40%	15.43%	0.00%	2.20%	0.41%

Table 5: Annotation labels distribution for each group of annotators.

Threshold	Precision	Recall	F1	# Toxic
0.5	0.65	0.69	0.67	92
0.6	0.69	0.62	0.65	78
0.7	0.8	0.41	0.55	45
0.8	0.81	0.4	0.54	43
0.9	1.00	0.15	0.26	13

Table 6: Perspective API performance on test dataset with different toxicity score thresholds.

Model *"Posso fazer a piada do bebe morto?"*.[8]

Toxic annotation correlation: We computed how the manual labels and the Perspective API's labels correlate with each other. The overall Pearson correlation (Cohen et al., 2009) in the test sample is 0.51 comparing the label of majority vote for each comment. We also computed correlation between groups of annotators and the automated predictions from the Perspective API. Annotators from batches 1, 2 and 3 showed consistent moderate correlation with the API, while annotators from batch 4 presented weak correlation. Finally, annotators from batch 5 showed a consistent and strong correlation with the API.

4.3 Language Characterization of Toxic and Non-toxic Content

We compared language patterns in *toxic* and *non-toxic* content in order to gain a better understanding of how Portuguese speakers employ language to generate toxic content.

TTR Analysis: Regarding comments' length, the average number of tokens and the 95% confidence interval for *non-toxic* comments is 26.34 [24.68, 28.19]. For *toxic* comments, the average is 35.54 [29.41, 42.87]. Therefore, *toxic* comments are on average longer than *non-toxic* ones (p-value < 0.05). Length distribution in *toxic* comments has a larger interval, which might indicate differences within the subreddits themselves.

The mean TTR and the confidence interval for the *non-toxic* comments is 0.78 [0.78, 0.79], while for the *toxic* comments the mean is 0.83 [0.82, 0.84]. The results point to statistical significance, with *toxic* comments considered more diverse. This may vary among subreddits, as some of the communities are more prone to have heavy-interaction type of posts.

POS Tagging Analysis: POS tags diversity for *non-toxic* comments has a mean of 0.51 [0.50, 0.52], while for *toxic* labeled texts the mean is 0.46 [0.43, 0.48]. Even though *toxic* comments are longer in length, they are usually less diverse in terms of POS tags.

To further investigate POS, we compared the distribution of specific tags. First, we compared Adjectives (ADJ) with a mean of 1.68 [1.55, 1.81] for *non-toxic* comments and 2.14 [1.71, 2.66] for *toxic* comments. As the confidence intervals overlap between classes, we conducted a Mann-Whitney statistical test to compare for differences in the distributions. The use of the ADJ tag is statistically different between classes with a p-value < 0.01.

Likewise, we conducted the same test for the NOUN tag. The mean use in *non-toxic* comments is 5.43 [5.07, 5.83], while for *toxic* comments the mean is 7.44 [6.15, 8.94]. This difference is again validated by the Mann-Whitney test with a p-value < 0.01.

An analysis of POS tag distribution in comments is essential to understand the characteristics of the text generated by the Reddit users in Brazilian largest communities. To accomplish that, we used Spacy's pre-trained POS-tagger for Brazilian Por-

[8] English translation: Shall I tell you the joke about the dead baby?

tuguese. Each token in a sentence was classified into one of the existing POS tags. To the list of POS tags, other classes specific to the tag classification problem were added, such as SYM, SPACE and X to denote "symbols", "white space" and "other", respectively, with a cautionary note that, as this is a Machine Learning model trained on corpora pertaining to other domains, the token classification might result in false positives.

The two most common POS tags for *toxic* and *non-toxic* comments are NOUN and VERB. *Non-toxic* comments use more PROPN tags, while a high percentage of *toxic* comments tokens was tagged as PUNCT. Also, *toxic* comments make heavier use of INTJ expressions. We also compared POS tag distributions of both classes through a Chi-square test. The results indicate that the difference observed between the distribution of the POS tags is significant (p-value < 0.05).

To further analyze the differences in word usage by *toxic* and *non-toxic* comments, we calculated the most frequent words by toxicity class for the most frequent POS tags, that is, ADJ, NOUN, and PROPN tags. The results are shown in Table 7. One relevant finding is the term *mulher* (woman) in *toxic* comments. In fact, we conducted a Chi-square test to compare the association of this term with *toxic* and *non-toxic* comments. The results indicate a positive association for some of the Brazilian subreddits (such as r/desabafos) with p-value < 0.05. This result might suggest the presence of misogynous behavior associated with some topics and communities in social networks. Sample comments targeting women in the communities discussions can be found in (Table A.4). A future study will investigate how vulnerable groups are addressed in Brazilian social network communities.

Named Entity Recognition (NER): Table 8 presents the NER analysis performed in our dataset. The most common named entity in both classes is PERSON, representing over 31% of all classified tokens in toxic comments. The second most frequently mentioned entity LOCATION is equally prevalent in both classes. While both *toxic* and *non-toxic* comments mention these entities, their use differs. We conducted a Chi-square test to compare the distribution of POS tags for comments in which at least one named entity is mentioned. The result indicates a significant difference in their POS tags distribution (p-value < 0.01). *Toxic* comments, for instance, use more VERB and NOUN tokens. The following comment is an illustration of named entities being mentioned in users discussion: *"Mais sério que esse tweet só a guerra na Ucrânia"* [9].

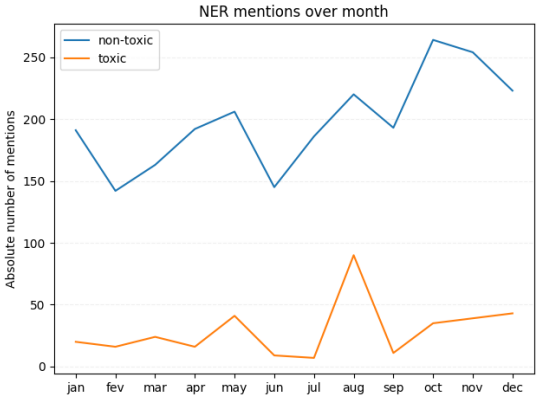

Figure 1: NER mention monthly time series.

It is well-known that online social networks are used as a means for discussing real-life events. We further investigated if our data reveals this behavior by showing the monthly time series of the numbers of NER citations in Figure 1.[10] There are significant spikes in the volume of mentions in August and October, which coincides with the opening month and the two rounds of 2022 Brazilian Elections. Some comments labeled as *toxic* mentioned the presidential candidates: *"Vocês são demasiadamente burros! Esse idiota do Bolsonaro pode até "dar um golpe", eu quero ver sustentar esse ato infame, pois, vejamos na década de 60, por exemplo, o Brasil teve essa porcaria de intervenção graças ao apoio do Tio Sam. [..]"*[11], *"O Lula não vai conseguir ver, pois ele está morto"*.[12]

4.4 Principal Findings

We next summarize our main findings in our study.

Annotation quality. We evaluated the dataset quality by calculating inter-rater agreement, which is in line with similar work (Perspective, 2022b). However, we divided the annotators in groups and our results show that some groups are more sensitive to toxicity comments and also evidence different quality levels. The strong agreement between annotators in group C points to their annotations as

[9] English translation: Only the war in Ukraine is more serious than this tweet.

[10] MISCELLANEOUS was excluded.

[11] English translation: You're too dumb! This idiot Bolsonaro can even "stage a coup", but I doubt whether he will be able to sustain that infamous act, because remember that in the 1960's, for example, Brazil had this crap intervention thanks to the support of Uncle Sam.

[12] English translation: Lula won't be able to witness this, because he's dead.

	ADJ	NOUN	PROPN
Non-toxic	bom (good), melhor (better), mesmo (same), grande (big), mesma (same), pior (worse), fácil (easy), diferente (distinct)	cara (dude), gente (people), pessoas (individuals), coisa (thing), tempo (time), anos (years), vida (life), mundo (world), dinheiro (money)	Brasil, Lula, Bolsonaro, OP (original poster), Deus (God), Flamengo, Landau, Ciro, PT (Workers' Party), STF (Supreme Court)
Toxic	melhor, mesmo, pobre (poor), ruim (bad), primeiro (first), forte (strong), diferente, social, capaz (capable), política (political), rico (rich)	pessoas, cara, mundo (world), mulher (woman), c* (a*s), casa (house), homem (man), m**da (sh*t), pai (father)	Lula, Bolsonaro, Brasil, OP, Ciro, Ucrânia (Ukraine), Flamengo, FDP (s*b), Liberdade, Rússia, Paris

Table 7: Most frequent words by POS tags and toxicity class.

Content	PER	ORG	LOC	MISC
Non-toxic	28.49%	20.26%	26.35%	24.88%
Toxic	31.33%	16.23%	27.92%	24.5%

Table 8: Percentage of NER mentions: PERSON (PER), ORGANIZATION (ORG), LOCATION (LOC) and MIS (MISCELLANEOUS).

a golden sample to evaluate distinct techniques for fine-tuning machine learning models of toxicity detection in Brazilian Portuguese texts.

Agreement with the Perspective API. Our comparison of manual annotation with the Perspective scores shows that some annotators underestimate toxicity, while others are more sensitive to *toxic* generated content. Overall, the average *toxic* comments percentage is close to the one of the API predictions (in the range of 10% to 11%). However, the Perspective API is more sensitive to curse words and lacks the context of the topics being discussed. Moreover, the API fails to detect very specific and nuanced types of targeted attacks in Portuguese (for instance, when specific groups are targeted with offenses in the form of sarcasm or irony).

Language characterization. *Toxic* comments are longer on average. While they have a similar proportion of POS tags to *non-toxic* ones, the most frequent nouns and adjectives evidence differences. A clear upward trend on NER mentions in the subreddits over the months, especially close to the Brazilian election period, shows external events' impact on user interactions. This should be considered when using this dataset for text classification and model creation, as the resulting model might be very sensitive to the available data time window.

Our findings attest to the potential of our dataset for fine-tuning a machine learning model in a downstream task. The high observed agreement among annotators certify the consistency of the labels. With this data, we aim to provide more diverse examples of *toxic* texts from online social network interactions to encourage the development of more robust machine learning models capable of mitigating online offensive behaviors.

Limitations. Regarding limitations of our study, we acknowledge the inherent challenge and subjectivity of the task of labeling toxic content in a contextually limited environment from online social networks. In order to mitigate this issue, we plan to iterate in the labeling experiment specifically providing additional context information to comments with local or limited context. Also, it is worth noting that our sampling procedure may present a bias towards specific external topics that held significant importance both locally and globally during the period of data collection.

5 Conclusion

Even though machine learning models have been successfully deployed as automatic moderation tools for some languages, we still lack support for low resource languages, such as Brazilian Portuguese. Our paper reports a new, manually-annotated dataset of toxic comments in Reddit user interactions from the largest ten subreddits in Brazil. Our results indicate substantial agreement among annotators and strong alignment with external pre-trained models for Portuguese, which supports the utilization of these data for machine learning downstream tasks.

In future works, we aim to integrate this new dataset with pre-trained machine learning models to provide the model with data from real social network interactions. Moreover, we intend to leverage this dataset for more intricate tasks such as detecting toxicity triggers within online conversations in order to be proactive on moderation interventions.

Acknowledgements. The research leading to these results has been partially supported by the Brazilian research agencies CNPq (Grant 313103/2021-6), FAPEMIG and CAPES.

References

Hind Almerekhi, Haewoon Kwak, Bernard J Jansen, and Joni Salminen. 2019. Detecting toxicity triggers in online discussions. In *Proceedings of the 30th ACM*

conference on hypertext and social media, pages 291–292.

Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J Jansen. 2020. Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of the web conference 2020*, pages 3033–3040.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Rogers Prates de Pelle and Viviane P Moreira. 2017. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PLOS ONE*, 16(8):1–22.

Jordan Kobellarz and Thiago Silva. 2022. Should we translate? evaluating toxicity in online comments when translating from portuguese to english. In *Anais do XXVIII Simpósio Brasileiro de Sistemas Multimídia e Web*, pages 95–104, Porto Alegre, RS, Brasil. SBC.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Perspective. 2022a. Perspective api model cards. https://developers.perspectiveapi.com/s/about-the-api-model-cards?language=en_US. Acessado em: 10/12/2023.

Perspective. 2022b. Perspective api training data. https://developers.perspectiveapi.com/s/about-the-api-training-data?language=en_US. Acessado em: 08/04/2023.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy.

Reddit. 2020. What is the dmca? https://support.reddithelp.com/hc/en-us/articles/360043515291-What-is-the-DMCA-. Acessado em: 11/02/2023.

Nazanin Salehabadi, Anne Groggel, Mohit Singhal, Sayak Saha Roy, and Shirin Nilizadeh. 2022. User engagement and the toxicity of tweets. *arXiv preprint arXiv:2211.03856*.

spaCy. 2022. Portuguese models. https://spacy.io/models/pt#pt_core_news_lg. Acessado em: 11/04/2023.

Statista. 2022. Number of social media users worldwide from 2017 to 2027. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. Acessado em: 08/04/2023.

Douglas Trajano, Rafael Bordini, and Renata Vieira. 2023. Olid-br: offensive language identification dataset for brazilian portuguese. *Lang Resources & Evaluation*.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Emily A Vogels. 2021. The state of online harassment. *Pew Research Center*, 13:625.

J. Wise. 2023. Reddit users: How many people use reddit in 2023? https://earthweb.com/how-many-people-use-reddit/. Acessado em: 08/04/2023.

Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and characterizing hate speech on news websites. In *Proceedings of the 12th ACM Conference on Web Science*, pages 125–134.

A Topic modeling

Table A.1 shows topics from comments which all the annotators disagreed upon (total disagreements). The topics include targeted comments to specific groups such as women and men, political and relationships discussions. Also, they present offensive terms such as curse words, used to offend other users in discussions.

Table A.2 shows topics from comments which all the annotators labeled as *toxic* (total toxic agreement). The topics are more fine-grained when sampling only comments on which all three annotators agreed as *toxic*. The discussions on these comments are centered on war, government and ideological issues. Also, they refer to discrimination, targeted groups and use very offensive terms to express users' opinions.

Table A.3 shows topics from comments which all the annotators labeled as *toxic* (total agreement), but the Perspective API labeled as *non-toxic* (false negative). When comparing human annotator labeling with Perspective's labeling, we found some particular cases in which the commercial model predicted wrong outputs for Portuguese. Specifically, the model lacks local context about politics and ideology as well as irony and sarcasm. Finally, the model is very sensitive to curse words. In fact, the mere occurrence of a bad word in a sentence might cause the model to abruptly shift its prediction score.

Table A.4 shows sample comments targeting *mulher* (woman) directly on *toxic* comments. Some of the comments caused total disagreement among annotators. For instance, the comment *"Pelo direito de bater na própria mulher! Uow"* (translation: *For the right to beat your own wife! Wow*), was labeled as "I do not know", "Toxic" and "Non-toxic". One hypothesis is that the text is read as a sarcastic comment or irony. A further experiment with additional context (such as providing the conversation thread to the data annotator) might mitigate disagreement on these cases. Comments requiring contextual clues are hard to label even for human annotators, and even more for machine learning models trained on corpora that do not resemble online social network interactions. In fact, for this specific comment, the Perspective API predicted as *non-toxic* with the pre-defined settings.

Topic	Descriptors
0	video (video), mulher (woman), opinião (opinion), homem (man), dinheiro (money), beleza (beauty), burro (dumb), padrão (standard), feedback, removal
1	guerra (war), liberdade (freedom), post, motivo (reason), país (country), massacres (massacres), massa (mass), atrocidades (atrocities), históricas (historical), democracia (democracy), governo (government), xenofóbico (xenophobic)
2	m**da (sh*t), sexo (sex), maluco (crazy), apoiadores (supporters), preocupado (worried), machão (macho man), malditos (damned), insegurança (insecurity), op (original poster)

Table A.1: Topics and relevant keywords from comments all three annotators disagreed upon (tri-disagreements).

Topic	Descriptors
0	burro (dumb), homem (man), p**ra (fu*k), c* (a**), mulher (woman), mercado (market), gente (people), anos (years), país (country), b**ta (cr*p), criança (child), ódio (hate), sentido (meaning)
1	guerra (war), bolsonaro, ucrânia (Ukraine), realidade (reality), putin, intervenção (intervention), pobre (poor), nuclear (nuclear), bandido (criminal), vergonha (shame), russia (russia)
2	ideologia (ideology), liberdade (freedom), política (politics), mundo (world), cancelamento (cancel culture), expressão (expression), op (original poster), preconceito (discrimination), oprimidos (oppressed), vagabundo (scoundrel), família (family)

Table A.2: Topics and relevant keywords from comments all three annotators labeled as *toxic*.

Topic	Descriptors
0	ideologia (ideology) política (politics) liberdade (freedom), mundo (world), pessoas (people), expressão (expression) mulheres (women) preconceito (discrimination) bolsonaro (bolsonaro) esquerdistas (leftists), apolíticos (apolitical), piada (joke), realidade (reality), oprimidos (oppressed), opiniões (opinions)

Table A.3: Key terms extracted from comments all three annotators labeled as *toxic* and the Perspective API predicted as *non-toxic* (false negatives).

Comment	Text
1	*"A minoria quer realmente ser independente - mas como o universo do /r/brasil é majoritariamente progressista, não irão concordar - as demais estão entre o "mulher tem que ser mulher" e aquelas que usam o discurso de independência, mas acham que quem tem que pagar as coisas é o homem."* ***Translation**:* *The minority really wants to be independent - but since the scenario in /r/brasil is mostly progressive the rest lies somewhere between "women have to be women" and those who adopt the discourse on independence, but think that the ones who have to afford all expenses are men.*
2	*"O mundo é assim, do mesmo jeito que você não quer uma mulher feia, uma mulher não vai querer alguém feio ou sem status, não cai nesse papo de que aparência não importa que em rede social só tem alienado, veja você mesmo pesquisas relacionadas ao assunto ou se tiver coragem crie um perfil com a foto de alguém bonito e veja como as pessoas te tratam diferente."* ***Translation**:* *That's how it works, just as you wouldn't want an ugly woman, a woman wouldn't want [to be with] someone ugly or with no status, don't be misled by the idea that looks don't matter, that there are only alienated people on social networks, get to know some of the surveys on this matter or if you dare do it, create a profile with a photo of someone beautiful and see how people will treat you differently.*
3	*"[..] Mas o homem casa com quem ele quiser. A mulher casa com quem ela consegue."* ***Translation**:* *[..] But a man can marry any woman he wants to. A woman can only marry a man she can manage to.*

Table A.4: Examples of comments mentioning the term "mulher" (woman) in *toxic* comments.