Language Resources and Evaluation (2024) 58:1263–1289
https://doi.org/10.1007/s10579-023-09657-0

ORIGINAL PAPER

OLID-BR: offensive language identification dataset for Brazilian Portuguese

Douglas Trajano[1] · **Rafael H. Bordini**[1] · **Renata Vieira**[2]

Accepted: 27 March 2023 / Published online: 3 May 2023

Abstract

Social media has revolutionized the manner in which our society is interconnected. While this extensive connectivity offers numerous benefits, it is also accompanied by significant drawbacks, particularly in terms of the proliferation of fake news and the vast dissemination of hate speech. Identifying offensive comments is a critical task for ensuring the safety of users, which is why industry and academia have been working on developing solutions to this problem. Prior research on hate speech detection has predominantly focused on the English language, with few studies devoted to other languages such as Portuguese. This paper introduces the Offensive Language Identification Dataset for Brazilian Portuguese (OLID-BR), a high-quality NLP dataset for offensive language detection, which we make publicly available. 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/cross-lingual models. The five NLP tasks available in OLID-BR allow the detection of offensive comments, the classification of the types of offenses such as racism, LGBTQphobia, sexism, xenophobia, and so on, the identification of the type and the target of offensive comments, and the extraction of toxic spans of offensive comments. All those tasks can enhance the capabilities of content moderation systems by providing deep contextual analysis or highlighting the spans that make a text toxic. We further experiment with and evaluate the dataset using state-of-the-art BERT-based and NER models, which demonstrates the usefulness of OLID-BR for the development of toxicity detection systems for Portuguese texts.

Keywords Hate speech · Offensive comments · OLID-BR · Dataset · NLP · Natural language processing · Offensive language detection · Content moderation systems · Toxicity detection systems · Toxic spans detection · NER · BERT

✉ Douglas Trajano
douglas.trajano@edu.pucrs.br

Extended author information available on the last page of the article

1 Introduction

Social media has transformed our society in many ways. Such platforms enable virtual human interaction, connecting people from different social groups, regions, and cultures. It can also be the most democratic and important instrument of freedom of speech these days as almost all people can share their ideas with almost no restrictions (Siddiqui et al., 2016).

The most common way that users interact on these platforms is through text messages such as comments or direct (private) messages. Thousands of comments are published every day. These comments can express ideas, opinions, and support messages, but they can also contain hate speech. Hate speech or toxic language can be defined as any communication that disparages a person or a group based on the characteristics such as race, color, ethnicity, gender, sexual orientation, ideology, religion, and so on (Levy et al., 2000). Due to the large amount of data generated every day on social media platforms, manual moderation is not viable, leading to an increase in demand for automated systems that moderate inappropriate content (Alonso et al., 2020).

Almost all toxicity detection systems use supervised learning models that require a large amount of labeled data. In our analysis, we identified more studies for English than Portuguese, we also noted that the most advanced techniques focused on English due to the easy accessibility to high-quality datasets for this language.

In this context, this work proposes a high-quality dataset named **Offensive Language Identification Dataset for Brazilian Portuguese** (OLID-BR), which supports several Natural Language Processing (NLP) tasks related to toxic language. **Task A** is a binary classification problem that predicts if a given comment is offensive or not. **Task B** is a multi-label problem that predicts the possible types of toxicity present in a given comment. **Task C** is a binary classification problem that predicts if a given toxic comment is targeted or not. **Task D** is a multi-class problem that detects the type of the target in a given targeted toxic comment. **Task E** detects the possible toxic spans in a given toxic comment. As far as we know, OLID-BR is the first Brazilian-Portuguese dataset labeled at the span level. It also uses an annotation schema followed by OLID-based datasets for other languages, which means that can be used to train cross multilingual/cross-lingual models.

We also developed machine learning experiments to assess the proposed dataset. The experiments reported here are for **Task B** (Toxicity Type Detection), **Task C** (Targeted Toxicity Detection), and **Task D** (Targeted Toxicity Type Detection). The experiments demonstrate the efficacy of utilizing the OLID-BR dataset for training supervised learning models to address various NLP tasks associated with toxic language. These models can be applied in multiple ways, such as highlighting the toxic spans in offensive comments in semi-supervised content moderation or automatically removing them while providing comprehensive contextual information like the target of the offense and the type of toxic language.

The remainder of this paper is structured as follows. Section 2 discusses related work. Section 3 provides a detailed explanation of the dataset creation process.

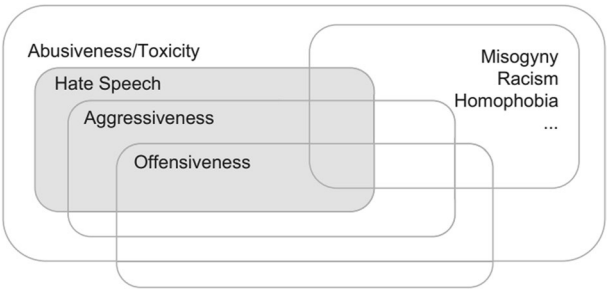

Fig. 1 Relations between HS and related concepts. Source: (Poletto et al., 2021)

Section 4 explains the experiments and their results. Finally, Sect. 5 discusses our main findings and possibilities for future work.

2 Related work

As the prevalence of hate speech on social media platforms continues to rise, researchers are actively developing studies to better understand and automatically detect such offensive content. To gain a deeper understanding of the terminology used in this field, a study by Poletto et al. (2021) examines the terms and concepts related to hate speech. The research suggests that hate speech is a subset of the broader category of abusiveness/toxicity language (Fig. 1).

Fortuna and Nunes (2018) analyzed several hate speech definitions from different sources and propose that hate speech is the language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used. However, deciding whether a comment contains hate speech is a complex task, even for humans. Due to the different experiences and personal relationships of each person, the very nuances of the language also affect the perception of what constitutes hate speech. Based on those factors, the authors warn of a low agreement between annotators when creating datasets for hate speech detection.

In Zampieri et al. (2019a), the authors create a dataset named OLID with 14,100 tweets in English following a three-level hierarchical annotation schema proposed by the same authors. Each comment was labeled by two different annotators, in case of disagreement, a third annotator was used and then the majority vote technique was employed to determine the final label. The annotation schema encompasses the following categories: **Offensive Language Detection**, which distinguishes whether a given comment contains toxic language or not. **Categorization of Offensive Language**, which detects if a given comment is targeted or not. **Offensive Language Target Identification**, which identifies the type of a targeted toxic comment. Figure 2 illustrates the annotation schema described above.

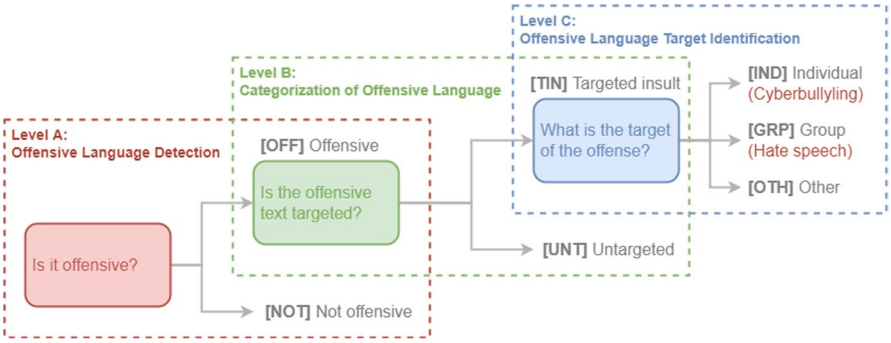

Fig. 2 OLID annotation schema. Source: Prepared by the author, adapted from (Zampieri et al., 2019a)

OLID was first used in the OffensEval-2019 (Zampieri et al., 2019b), a scientific competition with a task named "Identifying and Categorizing Offensive Language in Social Media". A huge amount of relevant contributions were produced based on the submissions for this competition. Some of them produce new OLID-based datasets that were used in the OffensEval-2020 (Zampieri et al., 2020) in the "Multilingual Offensive Language Identification in Social Media" task. The Pitenis et al. (2020) presented a Greek dataset with 4,779 posts from Twitter. A Turkish dataset was proposed in Çöltekin (2020) with 36,232 posts from Twitter. Sigurbergsson and Derczynski (2020) presented a Danish dataset with 3,600 from multiple social media platforms (Twitter, Facebook, and Reddit). Finally, a large-scale dataset named SOLID was proposed in Rosenthal et al. (2021), which contains over nine million English tweets labeled using a semi-supervised learning approach from previous works on the OLID dataset. The models trained using the SOLID and OLID demonstrate better performance when compared with only OLID when evaluated on the OLID test set.

Unfortunately, the majority of research making significant advances in this field concentrates on the English language, leaving other languages, such as Portuguese, with a lack of resources such as datasets. A systematic review of hate speech detection was carried out by Poletto et al. (2021); the authors pointed out that there exists more datasets for English than other languages, 37 out of 64 are English datasets, for Portuguese, there are only 2 datasets found by those authors. A systematic review of hate speech detection was conducted by Poletto et al. (2021) where the authors noted a disparity in the availability of datasets for non-English languages, 37 out of 64 being in English while only 2 datasets were found for Portuguese. The first dataset, called NCCVG,[1] was published in Nascimento et al. (2019); it contains 7,671 entries from two data sources: Twitter and 55chan. The second dataset, called OFFCOMBR,[2] published in de Pelle and Moreira (2017), contains 1,250 comments posted on a news site in Brazil https://g1.globo.com/. Both datasets were

[1] https://github.com/LaCAfe/Dataset-Hatespeech.
[2] https://github.com/rogersdepelle/OffComBR.

labeled with a binary task ("hate" or "no-hate"). Additionally, the paper highlights the importance of shared tasks in the advancement of research in non-English languages, as they serve as a good place for the creation of resources. For example, the HatEval dataset proposed by Basile et al. (2019), which is a context-specific dataset, was used in the SemEval 2019 Task 5 named "Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter". The HatEval dataset is composed of 13,000 English tweets and 6,600 Spanish tweets, totaling 19,600 tweets in both languages. The Pavlopoulos et al. (2021) proposed a "Toxic Spans Detection" task in SemEval-2021 with the goal to identify the toxic spans within each text that were responsible for the toxic label of the posts. The organizers provided a dataset labeled at the span level with 10,629 English toxic texts.

The Brazilian-Portuguese datasets analyzed in our literature review focus on a binary task ("hate" or "no-hate") or a multi-label task with few categories of toxic language. The ToLD-Br dataset, described in Leite et al. (2020), contains 21,000 tweets in Portuguese manually annotated into seven categories: *non-toxic*, *LGBTQ+phobia*, *obscene*, *insult*, *racism*, *misogyny* and *xenophobia*. The small number of Portuguese datasets available for the research community was also cited by authors as a challenge and a motivation to create the dataset. In Fortuna et al. (2019) the authors developed a dataset composed of 5,668 tweets in Portuguese using a hierarchical annotation schema. The dataset provides a binary classification ("hate" or "no-hate") and a multi-label classification with the following labels: sexism, body, origin, homophobia, racism, ideology, religion, health, and other-lifestyle.

Almost all of those papers did not provide detailed information on inter-rater reliability. Zampieri et al. (2019a) used Fleiss' Kappa on 21 tweets labeled by five annotators, only for the binary task (offensive vs. non-offensive), the value was 0.83 (high agreement). Fortuna et al. (2019) also used Fleiss' Kappa in the binary task, the value was 0.17 (poor agreement) considering all 5,668 messages labeled by three annotators; the authors suggest that the usage of non-expert annotators was the reason for the low agreement on this task. For the multi-label task (toxicity labels), expert annotators labeled 500 messages, and those annotations were evaluated using Cohen's Kappa, which resulted in 0.72 (substantial agreement). de Pelle and Moreira (2017) calculated Fleiss' Kappa for all the 1,250 messages labeled by the three annotators, the value was 0.71 (substantial agreement). In our study, Leite et al. (2020) demonstrated the most mature inter-rater reliability study, as it used Krippendorff's alpha for all the 21,000 tweets, every message was labeled by three annotators, the mean for all labels was 0.55 (moderate agreement), and the authors also provided the annotations for each instance, which we believe can be very important to support more use-cases. In summary, all related work demonstrated that achieving good agreement between annotators is a challenge in this area.

In this work, we present OLID-BR, a Brazilian-Portuguese dataset using an annotation schema compatible with the OLID-based datasets and inspired by the recently shared tasks. The OLID-BR can be used for modeling 5 NLP tasks and in multilingual experiments together with other OLID-based datasets. To ensure the high quality of the dataset and mitigate origin and annotation bias, we collected data from multiple data sources and followed a detailed annotation process described in Sect. 3.

3 OLID-BR dataset

In this section, we detail how the data was collected, selected, and anonymized. We also describe the annotation process, the inter-rater reliability experiment, the data format, and provide an overview of the dataset statistics.

3.1 Data collection

To take into account different user behaviors and avoid bias from the data origin, we collected data from multiple data sources.

3.1.1 Twitter

Twitter is a microblogging and social networking platform where users post messages known as "tweets". We collected tweets from Twitter using two different approaches. The first approach used a set of profiles of public persons selected by the first author of this paper; the profiles were organized into the following categories: *politics*, *news*, *enterprise*, *artists*, *influencers*, *sports*, and *entertainment*. The second approach used the toxic words labeled in the first annotation process iteration as the keywords for a search using the Twitter API.[3]

3.1.2 YouTube

YouTube is an online video-sharing platform that enables users to upload and share content, as well as engage in discourse through commenting. Based on our domain expertise, we manually selected a set of videos that were likely to have offensive comments using the following criteria: (1) the video had a high volume of comments, (2) a high volume of views, and (3) the presence of controversial or polemic subject matter in the video title or description. We also followed the same categories defined in Sect. 3.1.1, then we extracted the comments and comment replies from each selected video.

3.1.3 Related datasets

We consider related datasets the datasets with Portuguese texts, but with a different annotation schema which is incompatible with this project. Four datasets were used: *OffComBR* from de Pelle and Moreira (2017), *HLPHSD* by Fortuna et al. (2019), *NCCVG* by Nascimento et al. (2019), and *ToLD-Br* from Leite et al. (2020). We can use the texts from these datasets as a data source, but we need to adapt the annotation schema to our needs, so only the raw text was used, and our annotators would then label the text using our annotation schema without seeing the old labels.

[3] https://developer.twitter.com/.

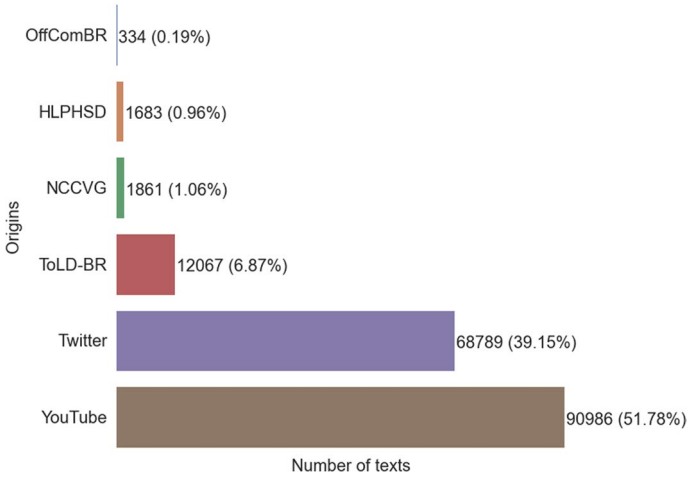

Fig. 3 OLID-BR collected texts by origin

3.2 Data selection

For all comments collected from our data sources (Twitter, YouTube, and related datasets), we evaluated if the comment was offensive or not. We then filtered only the offensive comments and used them as input data for the labeling task. We have some reasons for that. First, almost the all tasks proposed in this work are concentrated on offensive comments as we identified more significant studies and industry solutions for **Task A** than the other tasks. Second, as described in Sect. 3.4.3, the annotators were remunerated on a per-annotation basis, so we prioritized offensive comments since we needed to maximize the number of annotations for tasks dependent on offensive comments. Third, non-offensive comments are a lot more common than offensive comments, and they can be easily found on the Internet, so we believe that providing more offensive comments is a better contribution to the community than providing more non-offensive comments.

We used the Perspective API[4] to get the toxicity score for each comment. The toxicity score is a value between 0 and 1, where 0 is non-toxic and 1 is completely toxic. Initially, we set the threshold to 0.5, but after the first annotation iteration, we found that the threshold was too permissive as our annotators relabeled 47% of the comments as non-offensive. We then decided to increase the threshold to 0.7.

We collected a total of **249,162** comments. After filtering the comments with the toxicity score explained above and applying some cleaning techniques (removing misunderstood texts, non-Portuguese texts, duplicates, etc.), we had a total of **153,559** offensive comments available for annotation, which was enough for our purposes. Figure 3 shows the distribution of the unlabeled data by the origin of the text.

[4] https://www.perspectiveapi.com/.

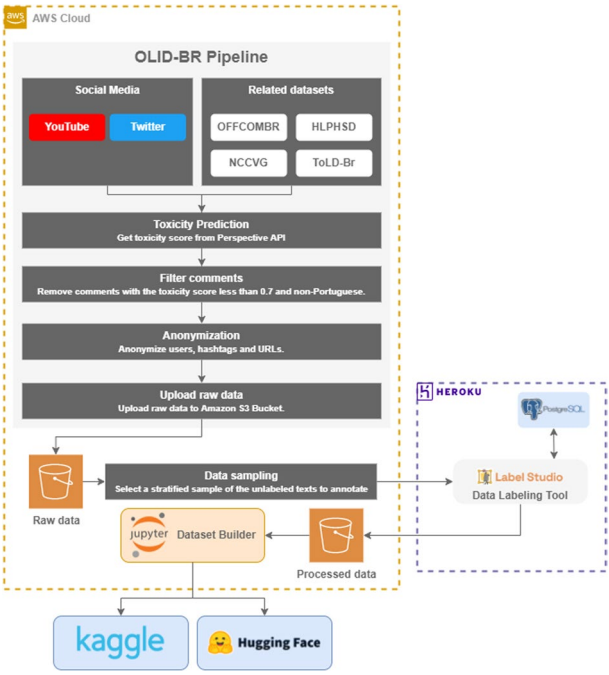

Fig. 4 OLID-BR data pipeline

For each annotation iteration, we selected a subset of the unlabeled texts to annotate stratified by the text origin, which allows us to have a balanced dataset with examples from different user behaviors. Figure 4 shows the entire data pipeline for the OLID-BR dataset.

3.3 Data anonymization

Data anonymization is the process of de-identifying data while preserving the format and type of the data. The anonymized data is altered to have a special sequence of characters or random values in place of the information to be anonymized. For example, replacing the name "Don Quixote" with "Ron Edwards" in a sentence is an example of data anonymization, and replacing "Don Quixote" with "XXXXX" is an example of data masking. However, both the terms data masking and data anonymization are interchangeable (Raghunathan, 2013). Identifying aggressors or victims is not our goal when developing this dataset. Our purpose was to create a dataset to be used for scientific research or open competitions. So, we apply some anonymization techniques in order to mask data that can be used to identify a person. We used regular expressions to detect usernames, hashtags, and URLs and replace them with "USER", "HASHTAG", and "URL" respectively. To remove given names, we used the SpaCy library to detect the given names using morphological analysis.

3.4 Annotation

3.4.1 Annotation process

An annotation process is a systematic series of actions based on rules or guidelines directed to associate classes or labels to the data in order to use the dataset in the training of supervised and semi-supervised learning algorithms (Han et al., 2011). There are several approaches to label data; Poletto et al. (2021) investigated the most used approaches specifically for hate speech. The first approach uses experts in the domain knowledge (annotators or researchers who have knowledge in the area) to take the annotations. Another commonly used approach is the annotation by non-specialists in the field, usually students. Some other approaches used crowd-sourcing companies such as Figure Eight[5] and Amazon Mechanical Turk.[6] Finally, automatic classification was also one of the techniques used by the researchers. In this work, we combine several approaches based on the task or the phase of the process to meet our annotation schema, for example, we used automatic classification for **Task A** and experts in the domain knowledge trained by the first author of this paper as described in 3.4.3. To validate and continuously improve our process, we worked following an iterative and incremental methodology, which allows us to use outputs from previous iterations to adjust or guide the work of the next iteration.

3.4.2 Annotation schema

Our annotation schema is driven by the needs of the tasks described in Sect. 1. We also optimized the questions asked to annotators by combining information that can be used to determine more than one question at the same time. To categorize if a given comment is toxic or not (Task A), we used the toxicity score provided by the Perspective API[7]; this score is a number between 0.0 and 1.0, the closer to 1.0 the more probable the comment is toxic. We also used the toxicity score to filter non-toxic comments, maximizing the annotators' productivity by showing only comments that are likely to be toxic in the labeling task. The annotators were able to reclassify the label to "No" if they considered that a given comment was not toxic. The OLID-BR dataset contains annotations for 5 tasks as follows:

- **Toxic Comment Classification** Binary classification task that is used to identify whether a comment is toxic or not.
- **Toxicity Type Detection** Multi-label classification task that identifies the toxicity labels present in a toxic comment.
- **Toxicity Target Classification** Binary classification task that predicts whether a toxic comment is targeted or not.

[5] https://appen.com/.
[6] https://www.mturk.com/
[7] https://www.perspectiveapi.com/.

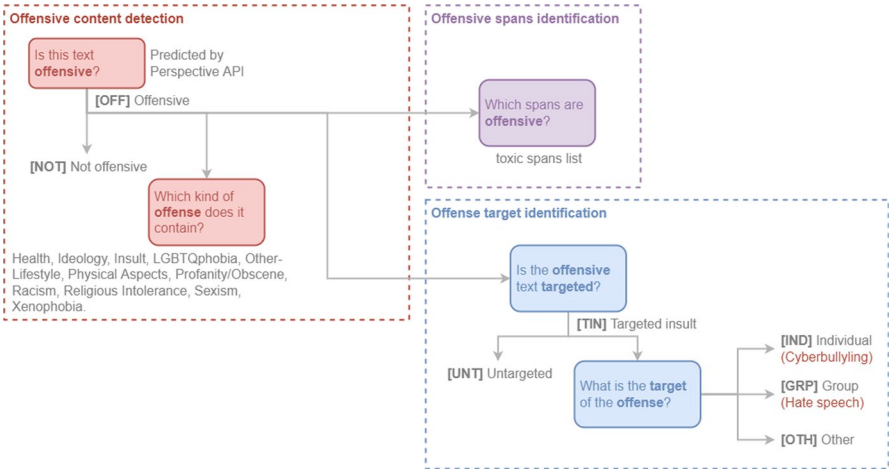

Fig. 5 OLID-BR annotation schema

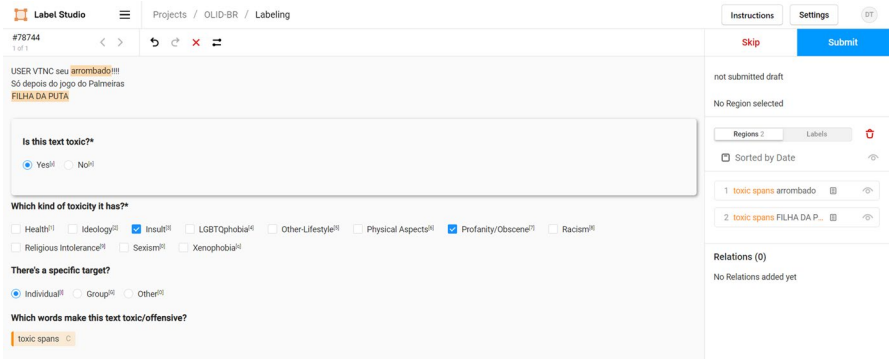

Fig. 6 Labeling interface

- **Toxicity Target Type Identification** Multi-class classification task that identifies the target type of a targeted comment.
- **Toxic Spans Detection** Span Categorization task that extracts the toxic spans from a toxic comment.

Figure 5 illustrates the annotation schema that was used to build the labeling interface shown in Fig. 6.

3.4.3 Qualified annotators

Using qualified annotators is an important aim in this project. Annotation agreement is a major challenge in this topic due to the complexity and subjectivity of this task (Fortuna et al., 2019). To mitigate the disagreement between annotators, we defined

some requirements to consider an annotator to be qualified for this task. A qualified annotator must be a native Portuguese speaker to understand the nuances of the language, it also required basic communication in English because the annotation tool was written in English. We also required them to take the course Non-Violent Communication[8] that takes an average of 4 h to be completed. The course contains the following content program: (I) differences between negativity and toxicity in communication and behavior; (II) toxic people and behavior; (III) assertive behavior and communication; (IV) Non-violent communication, awareness, and non-judgment. Additionally, an online meeting with the first author of this paper presenting the annotation guidelines and explaining the main concepts and terms to be used was done with all annotators in this project. To select the annotators for the labeling task, we created an online form with personal information questions such as gender and age, and the educational background area such as Computer Science and Social Science, we then shared the form with the local academic community. The annotators were chosen based on the requirements and their diversity to mitigate the annotation bias in the dataset. The contracted annotators were paid based on the number of annotations they produced.

3.4.4 Annotation guidelines

We created a document with annotation guidelines, including an explanation of the annotation process, concepts, task description, and provided some labeled examples, to guide annotators during their task. The document was written by the first author of this paper and is available at https://dougtrajano.github.io/olid-br/annotation/guidelines.html. The annotation guidelines were also presented in a training session with the annotators. Based on the related work and our experience, the topic can have high subjectivity, and having a clear and consistent understanding of the task is important to have high-quality data. We believe that the guidelines are a good way to increase agreement between annotators.

3.4.5 Labeling interface

The labeling interface is the UI (user interface) we developed to be used by the annotators to label each sentence in the dataset. It helps annotators in answering the following questions:

- **Question 1** "Is this text toxic?" aims to identify if a given comment is toxic or not; by default, the question comes checked with "Yes" and the annotators have the option to switch it to "No". If "No" is selected, the other questions are automatically disabled.
- **Question 2** "Which kind of toxicity it has?" allow annotators to select the toxicity labels ("Health", "Ideology", "Insult", "LGBTQphobia", "Other-Life-

[8] https://www.fecap.br/curta-duracao/comunicacao-nao-violenta/.

style", "Physical Aspects", "Profanity/Obscene", "Racism", "Religious Intolerance", "Sexism", and "Xenophobia") that apply to a given text.
- **Question 3** enables annotators to select the type of target present in a given targeted toxic comment; the annotators were instructed to keep this question blank if no target is present in the text.
- **Question 4** can be used to select the spans within the text that the annotators detected as toxic.

3.4.6 Inter-rater agreement

Inter-rater agreement is a measure of consistency used to evaluate the agreement between raters (i.e., annotators) that independently perform the assignment of predefined classes or categories to the same set of objects (Gwet, 2014). Related research applied different techniques and statistics to measure the agreement between raters, making it hard to compare those studies. Gwet (2014) proposed a diagram that helps to choose the correct agreement coefficient for each study. All our ratings are nominal data, so we use unweighted coefficients as suggested in the diagram. Using multiple reliability metrics with different methods for computing chance-corrected agreement can be more revealing than a single coefficient (Eugenio & Glass, 2004). With these thoughts, we selected the following coefficients in our inter-rater reliability analysis: **Percent Agreement** is a simple measure of the agreement between the raters. It is calculated by diving the number of agreeing ratings by the total number of ratings. As **Gwet's AC_1** does not depend upon the assumption of independence between raters, it can be used in more contexts than kappa. **Krippendorff's alpha** α is a chance-corrected agreement that can be applicable to several data types, missing values, and different levels of measurement.

3.4.7 Label assignment

Label assignment is the process of aggregating all the annotations in the form of one label for each feature of the dataset. We can apply different strategies to aggregate the annotations, with different semantics (Leite et al., 2020). If we set an instance as positive for a given feature only when all the annotators agreed on that, we can insert a bias into the model in the sense that it can only predict the feature if it is very evident. This option can be useful for models that will be used to take prohibitive actions. We can also use a majority voting strategy, as it represents the majority view among the annotators, it is less restrictive. We can also consider an instance as positive if at least one annotator tagged the text as positive. This strategy can be useful for training a model that supports content moderators when performing the moderation manually, for example. There is a trade-off that we need to consider when aggregating the annotations. In this work, we applied different strategies in the label assignment based on each feature; the label assignment strategies are described in Table 1.

Table 1 Label assignment strategy

Label	Label assignment strategy
is_offensive	majority vote
is_targeted	majority vote
targeted_type	majority vote
toxic_spans	all labeled spans
health	at least one
ideology	at least one
insult	at least one
lgbtqphobia	at least one
other_lifestyle	at least one
physical_aspects	at least one
profanity_obscene	at least one
racism	at least one
religious_intolerance	at least one
sexism	at least one
xenophobia	at least one

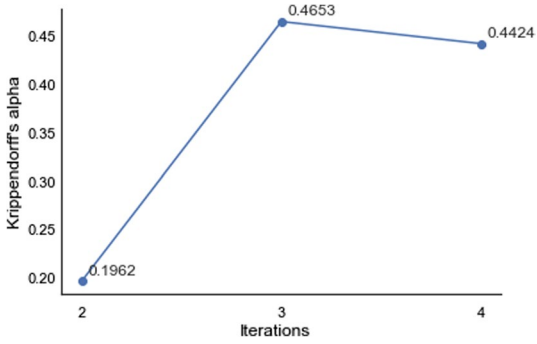

Fig. 7 Change on α over iterations of the annotation process

3.5 Inter-rater reliability

We carried out an inter-rater reliability analysis at the end of each iteration and in the dataset build stage; the coefficients used to do that analysis are detailed in Sect. 3.4.6. It was fundamental to refine our annotation process and define the strategy for the next iteration. We saw a significant increase in the agreement between annotators over iterations. For example, Krippendorff's Alpha of toxicity labels in the first iteration was 0.1962, and in the last iteration was 0.4424. Figure 7 shows the coefficient trends over iterations.

Only a few toxicity dataset papers included inter-rater reliability results as mentioned in Sect. 2, and for those that did include, some used different coefficients, hence cannot be compared to our results. Leite et al. (2020) labeled some toxicity labels (Task B) such as *LGBTQ+phobia*, *Insult*, *Xenophobia*, *Misogyny*, *Obscene*, and *Racism*. Those toxicity labels were evaluated using Krippendorff's Alpha, the

Table 2 Inter-rater reliability for OLID-BR v1.0

Feature	Percent agreement	α	Gwet's AC_1
Is_offensive (Question 1)	0.6641	0.1733	0.6929
Is_targeted (Question 3)	0.3000	0.0355	0.0960
Targeted_type (Question 3)	0.1505	0.4149	0.5689
Toxic_spans (Question 4)	0.1679	0.3918	N/A
Toxicity labels (Question 2)	0.2435	0.3648	N/A

results showed a range from 0.48 to 0.68 with an average of 0.55. Table 2 shows the coefficient results for version 1.0 of the OLID-BR dataset.

Our inter-rater reliability analysis demonstrates moderate agreement between annotators, which can be explained by the subjectivity of the task. The low number of observations that all raters agreed on the same label can lead to a high disagreement. Additionally, Krippendorff's Alpha is calculated on observed and expected disagreements, the expected disagreement is strongly influenced by the class distribution and an expected disagreement low can lead to a low Krippendorff's Alpha. We also selected our annotators from different genders, age groups, and educational backgrounds precisely to reduce possible bias, which can be another source of disagreement. As discussed previously, the related work on toxicity detection uses different coefficients for the Inter-Rater Reliability such as Cohen's and Fleiss' Kappa that have limitations and suffer from paradoxes that can lead to misleading results as explained in Feinstein and Cicchetti (1990). After an extensive study, we selected the coefficients used in the work that is most relevant to our scenario as explained in Sect. 3.4.6.

3.6 Sampling and data format

The dataset is split into training, public test, and private test sets, with the following sizes: 4,765, 1,589, and 1,589 samples, respectively. We stratify the dataset considering all labels, except *toxic_spans*, and we ensure that the distribution of labels in each set is similar to the distribution in the whole dataset. The training and public test sets are available for download, while the private test set is reserved for use in possible future competitions.

The dataset is available in two formats: **CSV files** containing the consolidated labels (following the label assignment strategy described in Sect. 3.4) and a **JSON file** containing the raw data (text, metadata, and all the three annotations).

In the CSV option, the dataset contains three files: *train.csv* (training set), *test.csv* (test set), and *metadata.csv* (metadata for all texts in the training and test sets). The *train.csv* and *test.csv* files contain the following columns: *id*, *text*, *is_offensive*, *is_targeted*, *targeted_type*, *toxic_spans*, *health*, *ideology*, *insult*, *lgbtqphobia*, *other_lifestyle*, *physical_aspects*, *profanity_obscene*, *racism*, *religious_intolerance*, *sexism*, and *xenophobia*. The *metadata.csv* file contains the following columns: *id*,

source, *created_at*, *collected_at*, *toxicity_score*, *annotator_id*, *gender*, *age*, *education_level*, and *annotator_type*.

In the JSON option, the dataset contains two files: *train.json* (training set) and *test.json* (test set). Each file contains a list of dictionaries, where each dictionary is a data point using the following schema:

```
{
  "id": "string",
  "text": "string",
  "metadata": {
    "source": "string",
    "created_at": "string",
    "collected_at": "string",
    "toxicity_score": "float"
  },
  "annotations": [
    {
      "annotator_id": "int",
      "is_offensive": "string",
      "is_targeted": "string",
      "targeted_type": "string",
      "toxic_spans": ["int"],
      "health": "bool",
      "ideology": "bool",
      "insult": "bool",
      "lgbtqphobia": "bool",
      "other_lifestyle": "bool",
      "physical_aspects": "bool",
      "racism": "bool",
      "religious_intolerance": "bool",
      "sexism": "bool",
      "xenophobia": "bool"
    }
  ]
}
```

Additionally, we provide an *additional_data.json* file containing 7,184 comments with incomplete annotations (i.e., less than 3 annotators) that were not used in the dataset. The file can be used to increase the number of training samples, but it requires careful analysis of the annotations to ensure that the data is reliable.

3.7 Data analysis

As explained in Sect. 3.6, the dataset is split into training, public test, and private test sets. In this section, we analyze the training set to understand the dataset characteristics. As we stratified all the sets, we expect the distribution of labels in each set to be similar to the label distribution in the whole dataset. The training set contains 4,765 sentences and 15 available labels. 11 labels are binary (toxicity labels) and

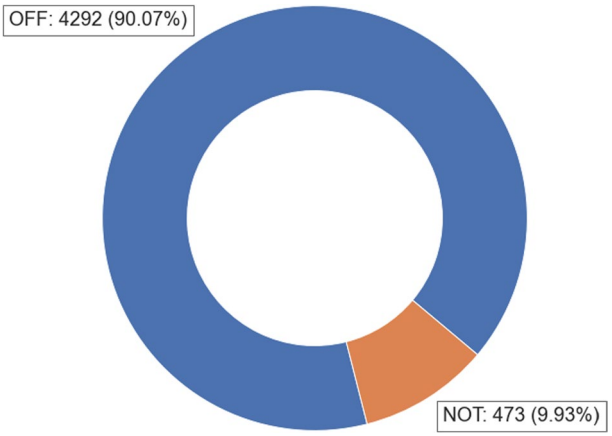

Fig. 8 Offensive comments distribution

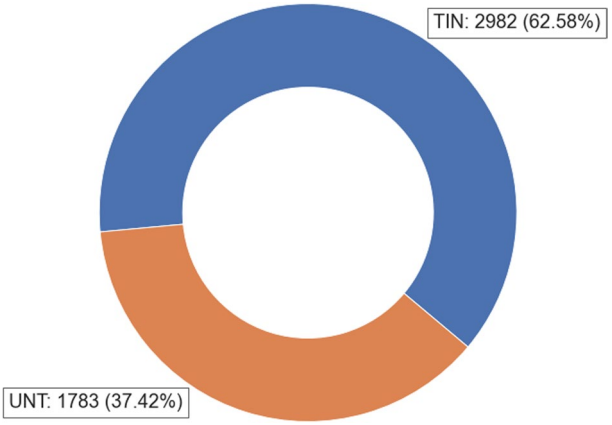

Fig. 9 Targeted offensive comments distribution

4 labels are categorical (is_offensive, is_targeted, targeted_type, and toxic_spans). The dataset is highly imbalanced in almost all labels as we can see below.

The *is_offensive* field contains 4,292 instances for "OFF" (offensive) and 473 instances for "NOT" (not offensive). As all selected texts for the labeling task had been previously determined as offensive by the Perspective API, we can say that 473 texts were reclassified to non-offensive, which suggests an overall accuracy of 90.07% for the Perspective API. We also analyzed the original labels for **Task A** in texts from related datasets with the labels associated in OLID-BR, 253 from 352 texts have the same label, which is 72% of them, the label for **Task A** for texts from related datasets were reclassified using the Perspective API and reviewed by our annotators. Figure 8 shows the distribution of the offensive and non-offensive comments.

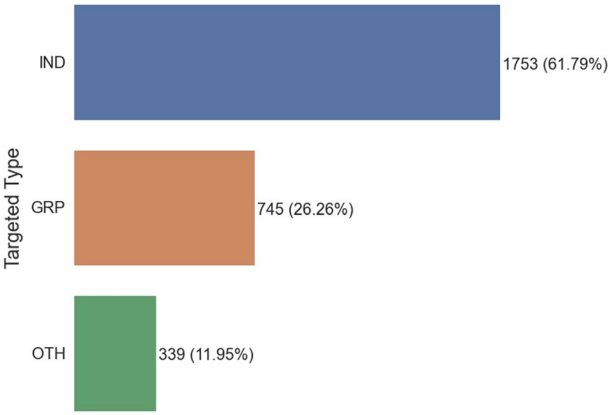

Fig. 10 Distribution of targeted offensive comment types

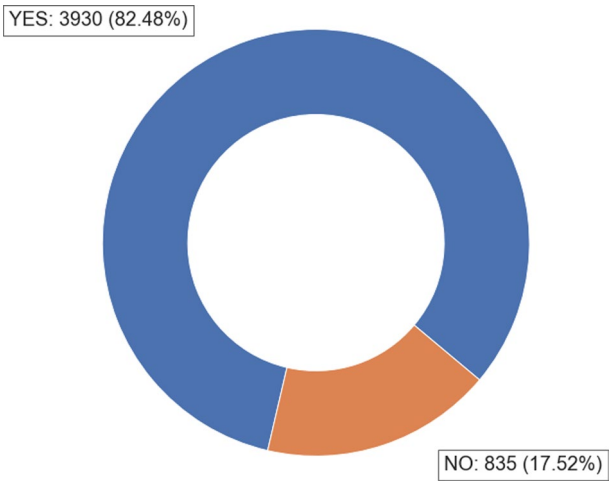

Fig. 11 Toxic spans distribution

The *is_targeted* field contains 2,982 instances for "TIN" (targeted insult) and 1,783 instances for "UNT" (untargeted). Figure 9 shows the distribution of the targeted and untargeted offensive comments.

The *targeted_type* field contains 1,753 instances for "IND" (individual), 745 instances for "GRP" (group), and 339 instances for "OTH" (other). Figure 10 shows the distribution of the targeted toxic comments.

The *toxic_spans* contains 835 missing values, which means we have the toxic spans for more than 80% of the entries in the dataset. Figure 11 shows the distribution of comments with toxic spans assigned.

Toxicity labels are the most imbalanced labels in the dataset. The *health*, *lgbtqphobia*, *other_lifestyle*, *physical_aspects*, *racism*, *religious_intolerance*, *sexism*, and *xenophobia* labels are highly imbalanced. The *ideology*, *insult*, and

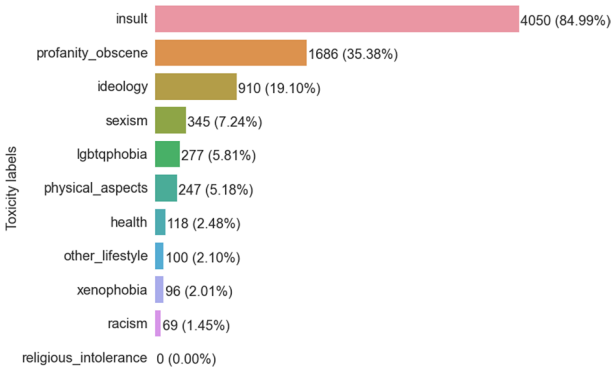

Fig. 12 Toxicity labels distribution

profanity_obscene labels are moderately imbalanced. Figure 12 shows the distribution of the toxicity labels.

4 Experiments

This section details the experiments developed to validate the usefulness of the OLID-BR dataset, despite the primary focus of this paper is the presentation of the dataset itself. We trained models for all NLP tasks available in the OLID-BR. The Toxic Comment Classification 4.1, Toxicity Type Detection 4.2, Toxicity Target Classification 4.3, and Toxicity Target Type Identification 4.4 are based on the pre-trained BERT model provided by Souza et al. (2020), a transformer-based model proposed by Vaswani et al. (2017), which is considered the state-of-the-art for various NLP tasks. The model contains 12 layers with 110 million parameters and slight modifications were made by us to the model to fit the specific problem type of each task, as outlined in subsequent sections. The Toxic Spans Detection task uses the SpaCy[9] Named Entity Recognition model adapted to work at the span level. All experiments had a hyperparameter tuning job that uses a Bayesian search strategy to find the best values for hyperparameters. The models were evaluated using a set of metrics: precision, recall, and F1-score.

4.1 Toxic comment classification

The Toxic Comment Classification is a binary classification task that predicts if a given comment is toxic or not. We added the Softmax as the activation function to the model output and used the BCEWithLogitsLoss loss function to get the probability of each class and calculate the loss using weights to handle the imbalanced classes. Different from other tasks, OLID-BR contains almost all comments

[9] https://spacy.io/.

Table 3 Toxic comment classification: overall metrics

Metric	Score
Precision (weighted)	0.8567
Recall (weighted)	0.8568
F1-score (weighted)	0.8568

Table 4 Toxic comment classification: class metrics

Class	Precision	Recall	F1-score	Support
NOT	0.8679	0.8738	0.8709	1,775
OFF	0.8429	0.8359	0.8394	1,438

classified as offensive ("OFF"), we explained the reasons in Sect. 3.2, but to train this task, we need to add some non-offensive comments ("NOT"). To do that, we randomly selected 8,117 non-offensive comments from related datasets described in Sect. 3.1.3. We then split the training set into 80% for training and 20% for validation, and the test set was used to evaluate the model. We had 9,006, 2,252, and 3,213 examples in the training, validation, and test sets, respectively. To identify the best values for the hyperparameters, we used the Bayesian search strategy with the following hyperparameters and ranges: *learning_rate* (1e-5 to 1e-3), *weight_decay* (0.0 to 0.1), *adam_beta1* (0.8 to 0.999), *adam_beta2* (0.8 to 0.999), *adam_epsilon* (1e-8 to 1e-6), *label_smoothing_factor* (0.0 to 0.1), and *optimizer* (adamw_hf, adamw_torch, adamw_apex_fused or adafactor). We defined some static hyperparameters such as *num_train_epochs* (30), *early_stopping_patience* (2), *batch_size* (8), and *seed* (1993). The objective metric was the weighted F1-score and the number of iterations was 18 with 3 jobs running in parallel. The best hyperparameters found were: *learning_rate* = 3.255788747459486e−05, *weight_decay* = 0.031031065174245122, *adam_beta1* = 0.8445637934160373, *adam_beta2* = 0.8338816842140165, *adam_epsilon* = 2.527092625455385e−08, *label_smoothing_factor* = 0.07158711257743958, and *optimizer* = adamw_hf. We then trained the model with the best hyperparameters in the training and validation sets and evaluated the model in the test set. After 6 epochs, the model stopped training because the weighted F1-score did not improve for 2 epochs. Table 3 shows the results for the Toxic Comment Classification task.

We also generated the precision, recall, and F1-score for each class as presented in Table 4.

4.2 Toxicity type detection

Toxicity Type Detection is a multi-label classification task that is used to detect the toxicity labels of a given toxic comment. We added the Sigmoid as the activation function to the model output and used the BCEWithLogitsLoss loss function to get the probability of each label and calculate the loss using weights to handle the imbalanced labels. We filtered the dataset to remove the instances that have no

Table 5 Toxicity type detection: overall metrics

Metric	Score
Precision (weighted)	0.8127
Recall (weighted)	0.6710
F1-score (weighted)	0.7250

Table 6 Toxicity type detection: label metrics

Toxicity label	Precision	Recall	F1-score	Support
Health	0.2419	0.3846	0.2970	39
Ideology	0.7390	0.6612	0.6979	304
Insult	0.9860	0.6765	0.8025	1351
Lgbtqphobia	0.6634	0.7283	0.6943	92
Other_lifestyle	0.4000	0.3529	0.3570	34
Physical_aspects	0.4286	0.4699	0.4483	83
Profanity_obscene	0.7124	0.7669	0.7386	562
Racism	0.4545	0.4348	0.4444	23
Sexism	0.3438	0.5739	0.4300	115
Xenophobia	0.4643	0.4062	0.4333	32

toxicity labels. We also removed the *religious_intolerance* label as we have only one instance with this label in the whole dataset, we then got 10 labels for the task. The training set was split into 80% for training and 20% for validation, and the test set was used to evaluate the final model. We had 3,417, 855, and 1,438 examples in the training, validation, and test sets, respectively. To identify the best values for the hyperparameters, we used the Bayesian search strategy with the following hyperparameters and ranges: *learning_rate* (1e-5 to 1e-3), *weight_decay* (0.0 to 0.1), *adam_beta1* (0.8 to 0.999), *adam_beta2* (0.8 to 0.999), *adam_epsilon* (1e-8 to 1e-6), and *optimizer* (adamw_hf, adamw_torch, adamw_apex_fused or adafactor). We defined some static hyperparameters such as *num_train_epochs* (30), *early_stopping_patience* (2), *batch_size* (8), and *seed* (1993). The objective metric was the weighted F1-score and the number of iterations was 18 with 3 jobs running in parallel. The best hyperparameters found were: *learning_rate* = 7.044186985160909e−05, *weight_decay* = 0.02426675806866223, *adam_beta1* = 0.9339215524915885, *adam_beta2* = 0.9916979096990963, *adam_epsilon* = 3.4435900142455904e−07, and *optimizer* = adamw_apex_fused. We then trained the model with the best hyperparameters in the training and validation sets and evaluated the model in the test set. After 7 epochs, the model stopped training because the weighted F1-score did not improve for 2 epochs. The model achieved a weighted F1-score of 0.7250, which is a good result for a multi-label classification task. Table 5 shows the overall metrics for the toxicity type detection task.

We also generated the precision, recall, and F1-score for each toxicity label as presented in Table 6. In summary, we can see that the model performs well

Table 7 Toxicity target classification: overall metrics

Metric	Score
Precision (weighted)	0.6744
Recall (weighted)	0.7003
F1-score (weighted)	0.6767

Table 8 Toxicity target classification: class metrics

Class	Precision	Recall	F1-score	Support
UNT	0.5219	0.3228	0.3989	443
TIN	0.7423	0.8683	0.8004	995

on labels that we have more examples of, which is why in future work we aim to obtain more annotated samples for the dataset.

4.3 Toxicity target classification

The Toxicity Target Classification is a binary classification task that is used to detect if a given toxic comment is targeted at someone or some group. We added the Softmax as the activation function to the model output and used the BCEWithLogitsLoss loss function to get the probability of each label and calculate the loss using weights to handle the imbalanced classes. We filtered the datasets to consider only offensive comments. The training set was split into 80% for training and 20% for validation, and the test set was used to evaluate the model. We had 3,433, 859, and 1,438 examples in the training, validation, and test sets, respectively. To identify the best values for the hyperparameters, we used the Bayesian search strategy with the following hyperparameters and ranges: *learning_rate* (1e-5 to 1e-3), *weight_decay* (0.0 to 0.1), *adam_beta1* (0.8 to 0.999), *adam_beta2* (0.8 to 0.999), *adam_epsilon* (1e-8 to 1e-6), *label_smoothing_factor* (0.0 to 0.1), and *optimizer* (adamw_hf, adamw_torch, adamw_apex_fused or adafactor). We defined some static hyperparameters such as *num_train_epochs* (30), *early_stopping_patience* (2), *batch_size* (8), and *seed* (1993). The objective metric was the weighted F1-score and the number of iterations was 18 with 3 jobs running in parallel. The best hyperparameters found were: *learning_rate* = 4.174021560583183e−05, *weight_decay* = 0.05595810634526813, *adam_beta1* = 0.9360294728287728, *adam_beta2* = 0.9974781444436187, *adam_epsilon* = 8.016624612627008e−07, *label_smoothing_factor* = 0.09936835309930625, and *optimizer* = adamw_hf. We then trained the model with the best hyperparameters in the training and validation sets and evaluated the model in the test set. After 6 epochs, the model stopped training because the weighted F1-score did not improve for 2 epochs. Table 7 shows the results for the Toxicity Target Classification task.

We also generated the precision, recall, and F1-score for each class as presented in Table 8.

Table 9 Toxicity target type identification: overall metrics

Metric	Score
Precision (weighted)	0.7831
Recall (weighted)	0.7748
F1-score (weighted)	0.7783

Table 10 Toxicity target type identification: class metrics

Class	Precision	Recall	F1-score	Support
IND	0.8786	0.8440	0.8610	609
GRP	0.6544	0.6667	0.6605	213
OTH	0.5347	0.6210	0.5746	124

4.4 Toxicity target type identification

The Toxicity Target Type Identification is a multi-class classification task that is used to identify the type of targeted toxic comments. We added the Softmax as the activation function to the model output and used the BCEWithLogitsLoss loss function to get the probability of each label and calculate the loss using weights to handle the imbalanced classes. We filtered the dataset to consider only the targeted toxic comments (targeted_type not null), which are distributed in individual (IND), group (GRP), and other (OTH) classes. The training set was split into 80% for training and 20% for validation, and the test set was used to evaluate the model. We had 2,269, 568, and 946 examples in the training, validation, and test sets, respectively. To identify the best values for the hyperparameters, we used the Bayesian search strategy with the following hyperparameters and ranges: *learning_rate* (1e-5 to 1e-3), *weight_decay* (0.0 to 0.1), *adam_beta1* (0.8 to 0.999), *adam_beta2* (0.8 to 0.999), *adam_epsilon* (1e-8 to 1e-6), *label_smoothing_factor* (0.0 to 0.1), and *optimizer* (adamw_hf, adamw_torch, adamw_apex_fused or adafactor). We also defined some static hyperparameters such as *num_train_epochs* (30), *early_stopping_patience* (2), *batch_size* (8), and *seed* (1993). The objective metric was the weighted F1-score and the number of iterations was 18 with 3 jobs running in parallel. The best hyperparameters found were: *learning_rate* = 3.952388499692274e−05, *weight_decay* = 0.1, *adam_beta1* = 0.9944095815441554, *adam_beta2* = 0.8750000522553327, *adam_epsilon* = 1.8526084265228802e−07, *label_smoothing_factor* = 0.047566123672759336, and *optimizer* = adafactor. We then trained the model with the best hyperparameters in the training and validation sets and evaluated the model in the test set. After 5 epochs, the model stopped training because the weighted F1-score did not improve for 2 epochs. Table 9 shows the results for the Toxicity Target Type Identification task in the test set.

We also generated the precision, recall, and F1-score for each class as presented in Table 10.

Table 11 Toxic spans detection: overall metrics

Metric	Score
Precision	0.6876
Recall	0.4918
F1-score	0.5734

4.5 Toxic spans detection

Toxic Span Detection is a span classification task that is used to identify the part of the text that is toxic. Shelar et al. (2020) conducted a review of NLP tools, including SpaCy, Apache OpenNLP, and TensorFlow, for the Named Entity Recognition (NER) task and compared their results. The review found that SpaCy was the most accurate tool for the NER task and offered more flexibility in customizing the model. Then, we used the Portuguese pre-trained model from the SpaCy library with the NER pipeline to extract the toxic spans. We filtered the dataset to consider only offensive comments that had at least one toxic span (toxic_spans not null). The training set was split into 80% for training and 20% for validation, and the test set was used to evaluate the model. We had 3433, 859, and 1438 examples in the training, validation, and test sets, respectively. The target label is represented by a list of integers, and each integer represents the index of the span of a given text. To identify the best values for the hyperparameters, we used the Bayesian search strategy with the following hyperparameters and ranges: *learning_rate* (1e-4 to 1e-2), *droupout* (0.0, 0.1, 0.2, 0.3, 0.4, or 0.5), *weight_decay* (0.0 to 0.1), *adam_beta1* (0.8 to 0.999), *adam_beta2* (0.8 to 0.999), *adam_epsilon* (1e-8 to 1e-6), and *optimizer* (adam or radam). We defined some static hyperparameters such as *num_train_epochs* (30), *early_stopping_patience* (5), *batch_size* (start = 8, stop = 64, step = 1.01), and *seed* (1993). The objective metric was the F1-score and the number of iterations was 18 with 3 jobs running in parallel. The best hyperparameters found were: *learning_rate* = 0.00038798590315954165, *droupout* = 0.3, *weight_decay* = 0.1, *adam_beta1* = 0.9978242993498763, *adam_beta2* = 0.9988901284249041, *adam_epsilon* = 3.12576102525027e−08, and *optimizer* = adam. We then trained the model with the best hyperparameters in the training and validation sets and evaluated the model in the test set. After 10 epochs, the model stopped training because the F1-score did not improve for 5 epochs. Table 11 shows the results for the Toxic Spans Detection task in the test set.

With an F-Score of 57.34%, the model presented moderate results, as seen in Table 11, The model had a higher precision than recall, suggesting a difficulty in identifying all toxic spans in the test set.

5 Conclusion

In this paper, we presented the Offensive Language Identification Dataset for Brazilian Portuguese (OLID-BR), a carefully curated dataset for fine-grained toxicity detection. We gathered the data from different data sources such as social media

(Twitter and YouTube) and other Brazilian datasets. All labels were manually annotated by contracted annotators. The dataset is publicly available in both CSV and JSON formats. The dataset is publicly accessible in both CSV and JSON formats, with the CSV file containing labels assigned using our label assignment strategy, and the JSON file including all annotations for each text. We conducted a detailed inter-rater reliability analysis, demonstrating the reliability of the annotations for the proposed hierarchical taxonomy. OLID-BR is compatible with the annotation schema used by other OLID-based datasets available in other languages such as English, Greek, Turkish, and Danish, which means that it can be used to train multilingual/cross-lingual models such as XLM-R and mBERT. To our knowledge, this is the first time a Brazilian-Portuguese dataset has been annotated at the span level. We also built baseline models using the dataset, which yielded promising results with ample room for improvement. We make our dataset freely available to promote further research in this area.

The subjectivity of the labeling task can be defined as the main challenge in this work. We noticed that our annotators had different interpretations of the same text, which can be motivated by the different educational backgrounds, cultural and social contexts, and the different experiences of each annotator. We made an effort to reduce the subjectivity of the task by providing a detailed description of the task and the taxonomy of terms to be used in the annotation process, and by providing examples of toxic language. Our iterative process of annotation and validation also helped to reduce the subjectivity of the task and decrease the disagreement between the annotators. The experiments also showed that the toxicity labels that still have a comparatively small number of instances are more difficult to classify, as expected. This can be explained by the imbalanced data problem, which might be alleviated by using more sophisticated techniques such as data augmentation.

We believe that the OLID-BR dataset will be a valuable resource for the academic community, and we are looking forward to seeing the results of future research. We aim to develop more sophisticated models or use advanced techniques such as data augmentation, hyperparameter tuning, and imbalanced data handling. In addition, future work can also extend the proposed taxonomy to other languages. Beyond the aforementioned research topics, our dataset enables researchers to study annotators' behavior based on their profile data also published with the dataset metadata. Finally, extending the dataset to include more entries, in particular of the types that currently have the fewest examples, would certainly improve the use of the dataset in future research.

Data Disclaimer: We are aware that the dataset contains biases and is not representative of global diversity. We are aware that the language used in the dataset may not represent the language used in different contexts. Potential biases in the data include: Inherent biases in the social media and user base biases, the offensive/vulgar word lists (labeled in iteration 1) used for data filtering, and inherent or unconscious bias in the assessment of toxicity content. All these likely affect labeling, precision, and recall for a trained model. The baseline models were trained on examples

reviewed by the first author of this paper. Anyone using this dataset should be aware of these limitations.

Acknowledgements We gratefully acknowledge the financial support of Uol EdTech, Brazilian National Council for Scientific and Technological Development (CNPq), and Portuguese Foundation for Science and Technology (FCT) under the projects CEECIND/01997/2017, UIDB/00057/2020.

Author contributions All authors contributed to the conception and design of the work. DT contributed to the data collection, data processing, annotation process, data analysis, and performed the experiments. All authors reviewed the work, contributed to the writing of the manuscript, and approved the final manuscript.

Availability of data and materials The dataset created in this work is available on Kaggle (https://www.kaggle.com/dougtrajano/olidbr) and HuggingFace (https://huggingface.co/datasets/dougtrajano/olid-br).

Code availability The source code and experiments for this paper are available on the GitHub platform at https://dougtrajano.github.io/olid-br/ and https://dougtrajano.github.io/ToChiquinho/.

Declarations

Conflict of interest The authors declared no potential conflicts of interest concerning this article's research, authorship, and publication.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the Research Ethics Committee of the Pontifical Catholic University of Rio Grande do Sul (PUCRS) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Alonso, P., Saini, R., & Kovács, G. (2020). Hate speech detection using transformer ensembles on the hasoc dataset. In *International Conference on Speech and Computer* (pp. 13–21). Springer

Basile, V., Bosco, C., & Fersini, E., et al. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 54–63).

Çöltekin, Ç. (2020). A corpus of turkish offensive language on social media. In *Proceedings of the 12th language resources and evaluation conference* (pp. 6174–6184).

de Pelle, R. P., & Moreira, V. P. (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.

Eugenio, B. D., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics, 30*(1), 95–101.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR), 51*(4), 1–30.

Fortuna, P., da Silva, J. R., & Wanner, L., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 94–104).

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD, USA: Advanced Analytics, LLC.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Leite, J. A., Silva, D., & Bontcheva, K., et al. (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 Chapter of the Association for Computational Linguistics and the 10th International*

Joint Conference on Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp. 914–924, https://aclanthology.org/2020.aacl-main.91

Levy, L., Karst, K., & Winkler, A. (2000). Encyclopedia of the American Constitution. No. v. 6 in Encyclopedia of the American Constitution, Macmillan Reference USA, USA.

Nascimento, G., Carvalho, F., & Cunha, A. M. d., et al. (2019). Hate speech detection using brazilian imageboards. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web* (pp. 325–328).

Pavlopoulos, J., Sorensen, J., & Laugier, L., et al. (2021). SemEval-2021 task 5: Toxic spans detection. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). Association for Computational Linguistics, Online, pp. 59–69, https://doi.org/10.18653/v1/2021.semeval-1.6, https://aclanthology.org/2021.semeval-1.6

Pitenis, Z., Zampieri, M., & Ranasinghe, T. (2020). Offensive language identification in Greek. In *Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association* (pp. 5113–5119). Marseille, France. https://aclanthology.org/2020.lrec-1.629

Poletto, F., Basile, V., Sanguinetti, M., et al. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation, 55*(2), 477–523.

Raghunathan, B. (2013). *The complete book of data anonymization: From planning to implementation*. Auerbach Publications.

Rosenthal, S., Atanasova, P., Karadzhov, G., et al. (2021). Solid: A large-scale semi-supervised dataset for offensive language identification. *Findings of the Association for Computational Linguistics: ACL-IJCNLP, 2021*, 915–928.

Shelar, H., Kaur, G., Heda, N., & Mai. (2020). Named entity recognition approaches and their comparison for custom ner model. *Science & Technology Libraries, 39*, 324–337. https://doi.org/10.1080/0194262X.2020.1759479

Siddiqui, S., Singh, T., et al. (2016). Social media its impact with positive and negative aspects. *International Journal of Computer Applications Technology and Research, 5*(2), 71–75.

Sigurbergsson, G. I., & Derczynski, L. (2020). Offensive language and hate speech detection for danish. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 3498–3508).

Souza, F., Nogueira, R., & Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian Conference on Intelligent Systems* (pp. 403–417). Springer.

Vaswani, A., Shazeer, N., & Parmar, N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008). Curran Associates, Inc.

Zampieri, M., Malmasi, S., & Nakov, P., et al. (2019a). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 1415–1420, https://doi.org/10.18653/v1/N19-1144, https://aclanthology.org/N19-1144

Zampieri, M., Malmasi, S., & Nakov, P., et al. (2019b). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 75–86).

Zampieri, M., Nakov, P., & Rosenthal, S., et al. (2020). Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1425–1447).

Authors and Affiliations

Douglas Trajano[1] · Rafael H. Bordini[1] · Renata Vieira[2]

Rafael H. Bordini
rafael.bordini@pucrs.br

Renata Vieira
renata@uevora.pt

[1] School of Technology, Pontifical Catholic University of Rio Grande do Sul - PUCRS, Porto Alegre, Brazil

[2] CIDEHUS, University of Evora, Évora, Portugal