Proceedings Paper:
Leite, J.A., Silva, D.F., Bontcheva, K. et al. (1 more author) (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. 10th International Joint Conference on Natural Language Processing - AACL-IJCNLP 2020, 04-07 Dec 2020, Suzhou, China (online). Association for Computational Linguistics (ACL) , pp. 914-924. ISBN 9781952148910

Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis

João A. Leite, Diego F. Silva
Departamento de Computação
Federal University of São Carlos
São Carlos, Brazil
joaoaugustobr@hotmail.com
diegofs@ufscar.br

Kalina Bontcheva, Carolina Scarton
Department of Computer Science
University of Sheffield
Sheffield, UK
k.bontcheva@sheffield.ac.uk
c.scarton@sheffield.ac.uk

Abstract

Hate speech and toxic comments are a common concern of social media platform users. Although these comments are, fortunately, the minority in these platforms, they are still capable of causing harm. Therefore, identifying these comments is an important task for studying and preventing the proliferation of toxicity in social media. Previous work in automatically detecting toxic comments focus mainly in English, with very few work in languages like Brazilian Portuguese. In this paper, we propose a new large-scale dataset for Brazilian Portuguese with tweets annotated as either toxic or non-toxic or in different types of toxicity. We present our dataset collection and annotation process, where we aimed to select candidates covering multiple demographic groups. State-of-the-art BERT models were able to achieve 76% macro-$F1$ score using monolingual data in the binary case. We also show that large-scale monolingual data is still needed to create more accurate models, despite recent advances in multilingual approaches. An error analysis and experiments with multi-label classification show the difficulty of classifying certain types of toxic comments that appear less frequently in our data and highlights the need to develop models that are aware of different categories of toxicity.

1 Introduction

Social media can be a powerful tool that enables virtual human interactions, connecting people and enhancing businesses' presence. On the other hand, since users feel somehow protected under their virtual identities, social media has also become a platform for hate speech and use of toxic language. Although hate speech is a crime in most countries, identifying cases in social media is not an easy task, given the massive amounts of data posted every day. Therefore, automatic approaches for detecting online hate speech have received significant attention in recent years (Waseem and Hovy, 2016; Davidson et al., 2017; Zampieri et al., 2019b). In this paper, we focus on the analysis and automatic detection of **toxic comments**. Our definition of toxic is similar to the one used by the Jigsaw competition,[1] where comments containing insults and obscene language are also considered, besides hate speech.[2] Systems capable of automatically identifying toxic comments are useful for platform's moderators and to select content for specific users (e.g. children). Nevertheless, there are multiple challenges specific to process toxic comments automatically, e.g. (i) toxic language may not be explicit, i.e. may not contain explicit toxic terms; (ii) there is a large spectrum of types of toxicity (e.g. sexism, racism, insult); (iii) toxic comments correspond to a minority of comments, which is fortunate, but means that automatic data-driven approaches need to deal with highly unbalanced data.

Although there is some work on this topic for other languages – e.g. Arabic (Mubarak et al., 2017) and German (Wiegand et al., 2018) –, most of the resources and studies available are for English (Davidson et al., 2017; Wulczyn et al., 2017; Founta et al., 2018; Mandl et al., 2019; Zampieri et al., 2019b).[3] For Portuguese, only two previous works are available (Fortuna et al., 2019; de Pelle and Moreira, 2017) and their datasets are considerably small, mainly when compared to resources available for English.

We present **ToLD-Br** (Toxic Language Dataset for Brazilian Portuguese), a new dataset with Twitter posts in the Brazilian Portuguese language.[4]

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview

[2] This is also similar to the usage of *offensive* comments in OffensEval (Zampieri et al., 2019b, 2020).

[3] A large list of resources is available at http://hatespeechdata.com.

[4] It is important to distinguish the language variant, since

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, pages 914–924
December 4 - 7, 2020.

A total of 21K tweets were manually annotated into seven categories: *non-toxic*, *LGBTQ+phobia*, *obscene*, *insult*, *racism*, *misogyny* and *xenophobia*. Each tweet has three annotations that were made by volunteers from a university in Brazil. Volunteers were selected taking into account demographic information, aiming to create a dataset as balanced as possible in regarding to demographic group biases. This is then the largest dataset available for toxic data analysis in social media for the Portuguese language and the first dataset with demographic information about annotators.[5]

We experiment with Brazilian Portuguese (Souza et al., 2019) and Multilingual (Wolf et al., 2019) BERT models (Devlin et al., 2019) for the binary task of **automatically classifying toxic comments**, since similar models achieve state-of-the-art results for the same task in other languages (Zampieri et al., 2019b). Models fine-tuned on monolingual data achieve up to 76% of macro-$F1$, improving 3 points over a baseline. Besides, BERT-based approaches with multilingual pre-trained models enable transfer learning and zero-shot learning. The OffensEval 2019 OLID dataset (Zampieri et al., 2019a) is then used to experiment with (i) **transfer-learning**: where both OLID and ToLD-Br are used to fine-tune BERT; and, (ii) **zero-shot learning**: where BERT is fine-tuned using only OLID. Results highlight the importance of language-specific datasets, since transfer learning does not improve over monolingual models and zero-shot learning achieves only a macro-$F1$ of 56%.

An error analysis is performed using our best model, where the worst-case scenario, i.e., classifying *toxic* comments as *non-toxic*, is further investigated, taking into account the fine-grained categories. Results show that categories with fewer examples in the dataset (*racism* and *xenophobia*) are more likely to be mislabelled than other classes, with the best performance being achieved by majority classes (*insult* and *obscene*). We also analyse the **amount of data** needed in order to achieve the best performance in binary classification. Models trained with few examples are only accurate in predicting the majority class (*non-toxic*). As the number of instances grow, the performance on the minority class (*toxic*) improves significantly. Finally, we experiment with **multi-label classification**, where each different type of toxicity is automatically classified. This is a considerably harder problem than binary classification, where BERT-based models do not outperform the baseline.

Section 2 presents an overview of relevant previous work. Section 3 shows details about the ToLD-Br dataset. Material and methods are presented in Section 4, whilst results are discussed in Section 5. Finally, Section 6 shows a final discussion and future work.

2 Related Work

Although multiple researchers have addressed the topic of hate speech (e.g. Waseem and Hovy (2016), Chung et al. (2019), Basile et al. (2019)), we focus the literature review on previous work related to toxic comments detection, the topic of our paper. Due to space constraints, we only describe papers that create and use Twitter-based datasets and/or focus on the Brazilian Portuguese language.

English Davidson et al. (2017) present a dataset with around 25K tweets annotated by crowd-workers as containing *hate*, *offensive language*, or *neither*. They build a feature-based classifier with TF-IDF transformation over n-grams, part-of-speech information, sentiment analysis, network information (e.g., number of replies), among other features. Their best model, trained using logistic regression, achieves a macro-$F1$ of 90. Founta et al. (2018) also rely on crowd-workers to annotate 80K tweets into eight categories: *offensive*, *abusive*, *hateful speech*, *aggressive*, *cyberbullying*, *spam*, and *normal*. They perform an exploratory approach to identify the categories that cause most confusion to crowd-workers. Their final, large-scale annotation is done using four categories: *abusive*, *hateful*, *normal*, or *spam*. OffensEval is a series of shared tasks focusing on offensive comments detection (Zampieri et al., 2019b, 2020). The OLID dataset (used in the 2019 edition) has around 14K tweets in English manually annotated as *offensive* or *non-offensive*. The best model for the relevant task A (*offensive* versus *non-offensive*) uses a BERT-based classifier and achieves 82.9 of macro-$F1$.

German A shared task (organized as part of GermEval 2018) aimed to classify tweets in German categorized into *offensive* or *non-offensive* (Wiegand et al., 2018). They make available a manually annotated dataset with approximately 8.5K tweets.

there are multiple differences between Brazilian Portuguese lexicon and other variants of Portuguese.

[5]ToLD-Br is available at: https://github.com/JAugusto97/ToLD-Br

The best system achieved 76.77 of $F1$-score and was a feature-based ensemble approach.

Arabic Mubarak et al. (2017) present a dataset with 1.1K manually annotated tweets into *obscene*, *offensive*, or *clean*. They experiment with lexical-based approaches that achieve a maximum of 60 $F1$-score. Mulki et al. (2019) create a dataset with tweets in the Levantine dialect of Arabic manually annotated into *normal*, *abusive*, or *hate* (with approximately 5K tweets). The authors use feature-based approaches to induce models for ternary and binary scenarios, with best systems achieving 74.4 and 89.6 of $F1$-score, respectively.

Spanish Carmona et al. (2018) present a shared task aiming to detect aggressive tweets in Mexican Spanish. They manually annotate 11K tweets into *aggressive* or *non-aggressive*. The best system is a feature-based approach with macro-$F1$ of 62.

Hindi Mathur et al. (2018) present a dataset of around 3.6K tweets in Hinglish (spoken Hindi written using the Roman script). The dataset was annotated into three classes *not offensive*, *abusive* and *hate-inducing* by ten NLP researchers. A Convolutional Neural Network (CNN) architecture with transfer learning is used, where the model is trained with both Hinglish and English data (from (Davidson et al., 2017)), achieving 71.4% of $F1$-score.

Portuguese de Pelle and Moreira (2017) make available a dataset with 1,250 comments, extracted from comment sessions of `g1.globo.com` website, and annotated them into categories of *offensive* or *non-offensive*. The offensive class was also subdivided into *racism*, *sexism*, *LGBTQ+phobia*, *xenophobia*, *religious in-tolerance*, or *cursing*. They experiment with binary classification, using n-grams as features to SVM and NaiveBayes models. Best results are achieved with SVM reaching a weighted F1 score between 77 and 82, depending on different label interpretations. Fortuna et al. (2019) describe a dataset with 5,668 tweets classified as *hate* vs. *non-hate*, with the *hate* class further classified following a fine-grained hierarchy. Experiments with binary classification show a $F1$ score of 78 using an LSTM-based architecture.

Multilingual HASOC was a shared task aiming to classify hate speech and offensive comments in English, German, and Hindi (Mandl et al., 2019). Their dataset contains around 7K tweets and Facebook posts manually annotated. Sub-task A separates posts into *hate speech* or *offensive* versus *neither*; and, sub-task B separates posts containing *hate speech* or *offence* into three categories: *hate speech*, *offensive* or *profane*. Best performing systems in all languages used deep learning approaches. For OffensEval 2020 (Zampieri et al., 2020), a more extensive training data is available for English (over 9M tweets), although the annotation was made semi-automatically. Arabic, Danish, Greek, and Turkish datasets are also available with manually annotated labels. For all languages, best models are achieved using some variation of BERT.

Our work is different from previous approaches because we (i) release a large-scale dataset for a language other than English, that was created with the aim to reduce demographic biases; (ii) experiment with multilingual approaches, including transfer learning and zero-shot-learning; (iii) perform an analysis of the amount of data needed to train reliable models; and, (iv) experiment with multi-label classification, providing first insights into this challenge task.

3 Dataset

In this section, we describe the procedure adopted to create ToLD-Br and present its main features.

3.1 Data collection

We used the GATE Cloud's Twitter Collector[6] to collect posts on the Twitter platform from July to August 2019. We used two different strategies to select tweets for ToLD-Br, aiming to increase the probability of obtaining posts with toxic content, given that the volume of toxic tweets is significantly smaller than data without offensive language. Our first strategy searches for tweets that mention predefined hashtags or keywords. We chose predefined terms highly likely to belong to a toxic tweet in Brazilian Twitter, such as *gay* (*"Gay tem que apanhar" – "Gay should be beaten up"*), *mulherzinha* (*"Mulherzinha, vai lavar louça" – "Sissy, go wash dishes"*), and *nordestino* (*"Nordestino preguiçoso" – "Lazy Northeastern"*). However, using this strategy alone may hinder learning a model capable of generalising the concept of toxicity beyond the scope of keywords. Consequently, another strategy was adopted: we scraped tweets that mention influential users like Brazil's president Jair Bolsonaro and soccer player Neymar Jr,

[6] `https://cloud.gate.ac.uk/shopfront/displayItem/twitter-collector`

prone to receive abuse (around 50 influential users were monitored). Tweets collected through this method have no restrictions in terms of keywords and should broaden the scope of the data.

We collected more than 10M unique tweets and randomly selected 21K examples to compose the annotated corpus. We note that 12,600 of these posts (60%) comes from the first strategy – predefined keywords – and the remaining are tweets from threads of predefined users. The data was pseudoanonymised before being sent for annotation, with all @ mentions replaced by *@user*.

3.2 Corpus annotation

The annotation process started by choosing volunteers to perform the task of assigning labels for each example. For this, we made a public consultation at the Federal University of São Carlos (Brazil) to find candidate annotators (129 volunteers registered for the task). From these candidates, 42 were selected based on their demographic information, aiming to balance annotation bias as the interpretation of toxicity may vary. Each annotator labelled 1,500 tweets, selecting one of the following categories: *LGBTQ+phobia*, *obscene*, *insult*, *racism*, *misogyny* and/or *xenophobia* (or leaving it blank for *none*). Each tweet was annotated by three different annotators.

To evaluate the diversity among the annotators, we explore their profile. We emphasise that the identity of all annotators has been preserved. At this stage, we only survey general aspects of the volunteers who joined the labelling process. Table 1 presents the distribution of annotators regarding sex, sexual orientation, and ethnicity. To define these categories, we use the same values as the Brazilian Institute of Geography and Statistics,[7] in addition to giving the candidate the option of not declaring a value for each characteristic. Although we tried to keep the demographic aspects as balanced as possible when selecting the annotators, our pool of volunteers was still biased towards people identified as *white* and *heterosexual* (*sex* is a more balanced aspect than the others). The age of the annotators varies between 18 and 37 years, with most of them in the range between 19 and 23. Figure 1 illustrates the age distribution.

We perform different data analysis over the dataset to better understand its properties. Inter-

[7] https://www.ibge.gov.br/en/home-eng.html

	Categories	# annotators
Sex	Male	18
	Female	24
Sexual orientation	Heterosexual	22
	Bisexual	12
	Homosexual	5
	Pansexual	3
Ethnicity	White	25
	Brown	9
	Black	5
	Asian	2
	Non-Declared	1

Table 1: Annotators demographic information.

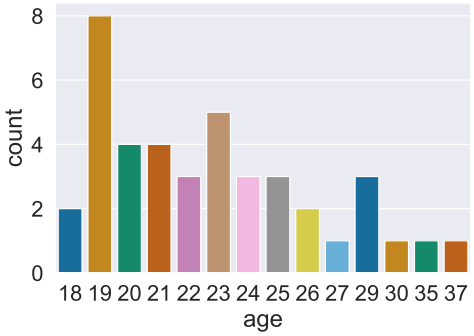


Figure 1: Annotators age distribution.

	α
LGBTQ+phobia	0.68
Insult	0.56
Xenophobia	0.57
Misogyny	0.52
Obscene	0.49
Racism	0.48
Mean	0.55

Table 2: *Krippendorff*'s α for each label.

annotator agreement is calculated in terms of *Krippendorf*'s α (Table 2), since α is robust to multiple annotators, different degrees of disagreement and, missing values (Artstein and Poesio, 2008).

The *LGBTQ+phobia* class shows the highest agreement, which may indicate that comments in this class have a more distinctive lexicon than other classes. The lowest agreement is seem in *obscene* and *racism* classes. Besides, we observed in the annotations many cases in which some examples were labelled as separate classes, although they intend

	Ann 1	Ann 2	Ann 3
*o fdp do filho dela nao parava de tocar auto pra c*****o [...]* *her sob son did not stop to play loud as f**k [...]*	Insult	None	Obscene
*[...] VAI SE F***R IRMÃO VC NÃO É FELIZ PQ NAO QUER* *[...] f**k you brother you are not happy because you do not want to be*	Obscene	Insult	Insult
"Aonde tem um monte que fala mal, mas ninguém vai embora do morro." *acha que alguém mora aqui por que quer, c*****o!? Que idéia. [...]* *"Where there are loads saying bad things, but nobody leaves the slum."* *who thinks that someone lives here because they want, f**k!? What an idea. [...]*	Obscene	Obscene	Insult

Table 3: Example of annotation divergence.

LGBTQ+phobia	Obscene	Insult	Racism	Misogyny	Xenophobia
viado (59)	porra (332)	puta (221)	nego (6)	putinha (38)	sulista (12)
boiola (15)	caralho (317)	caralho (150)	branco (6)	puta (22)	carioca (7)
viadinho (13)	puta (268)	cara (135)	preto (4)	piranha (19)	fala (4)
sapatão (12)	tomar (136)	porra (122)	nada (4)	mulher (11)	paulista (4)
caralho (11)	fuder (98)	lixo (101)	negão (3)	vagabunda (11)	gente (3)
cara (10)	cara (94)	filho (92)	cara (3)	quer (8)	nordestino (3)
quer (9)	merda (90)	burro (87)	falando (3)	vaca (8)	todo (3)
homem (9)	mano (87)	tomar (86)	vida (3)	fica (6)	ainda (3)
todo (9)	toma (85)	merda (78)	segue (2)	onde (5)	sendo (2)
bicha (9)	fazer (77)	idiota (76)	página (2)	tudo (5)	dança (2)

Table 4: The most common words of each class and the number of sentences they occur (within parentheses).

to point the same concept. Classes like *obscene* and *insult* seem to have confused the annotators, which may indicate an intersection in these concepts. Table 3 shows examples of disagreements in the classification of *obscene* and *insult*.

Table 4 presents the ten most frequent words for each class, after removing stopwords. It confirms the intersection between classes *obscene* and *insult*, with six out of ten words in common. For a quantitative analysis, Table 5 presents the *Jaccard* distance between the 100 most frequent words for each class. *Obscene* and *insult* show a considerably lower distance than other pairs (0.57), indicating that they have more words in common.

	a	b	c	d	e	f
a	0.00	0.73	0.78	0.90	0.80	0.94
b	-	0.00	0.57	0.84	0.77	0.90
c	-	-	0.00	0.86	0.75	0.92
d	-	-	-	0.00	0.87	0.95
e	-	-	-	-	0.00	0.94

Table 5: *Jaccard* distance between all pair of classes. (a) LGBTQ+phobia; (b) Obscene; (c) Insult; (d) Racism; (e) Misogyny; (f) Xenophobia.

3.3 Dataset characteristics

For the purpose of training models for automatically classifying toxic comments, we must create aggregated annotations to provide only one binary label for each class. Different rules can be employed to aggregate the annotations, with different semantics. When we set an example as positive for toxicity only when all the annotators consider it to have the same category of offence, we insert bias to the model to not accuse a comment as toxic unless the offence is evident. Since this is very restrictive, we can also use the majority rule, but there must still be a consensus among the annotators. A last option is to consider that only a positive annotation is sufficient to label the example as positive. This procedure acknowledges that annotators may have divergent views about what was said. It is a risky rule if we intend to create rigid systems that classify the tweets and take corrective or prohibitive actions. However, it is beneficial for training a model that "raises a flag" to help moderators to assess the com-

	LGBTQ+phobia	Insult	Xenophobia	Misogyny	Obscene	Racism	Toxic
At least one annotator							
0	20656	16615	20849	20537	14348	20862	11745
1	344	4385	151	463	6652	138	9255
At least two annotators							
0	20824	19131	20958	20867	18597	20967	16566
1	176	1869	42	133	2403	33	4424
Three annotators							
0	20926	20483	20985	20971	20388	20994	19510
1	74	517	15	29	612	6	1490

Table 6: Dataset distribution considering different types of label aggregation.

ments. Table 6 shows the data distribution for each label and each aggregation strategy.

For the sake of reproducibility and further usage, ToLD-Br is split into default training (80%), development (10%) and test (10%) sets using a stratified strategy. Besides, the corpus is released with all the annotations. Thus, future users of ToLD-Br will be able to use it with all the labels and with varying levels of agreement between the annotators. In this paper, we consider the least restrictive case, where if at least one annotator marked any offence category in an example, the example is positive for toxicity. Likewise, if a tweet was not tagged in any of these categories, it is considered non-toxic. We believe that it is essential that if any person feels uncomfortable with a post, it should be flagged as having a certain degree of toxicity. Therefore, a model built with this data must be able to identify offensive posts, even for a specific group of people.

4 Materials and Methods

In this section, we describe the techniques, tools, and other materials used in our experimental evaluation. As mentioned before, we restrict our experiments on the dataset labelled as positive when at least one annotator considers the example as toxic. We then investigate the effects of the number of instances in the training data, different algorithms to train a classification model, various scenarios considering single- and multilingual models, and perform an initial experiment with multi-label classification.

We use Bag-of-Words (BoW) to represent the examples and an AutoML model to build the baseline model (`BoW+AutoML`). For this, we use the `auto-sklearn`[8] library (Feurer et al., 2019). For our BERT-based models, we use the `simpletransformers`[9] library, that allows easy training and evaluation. We use default arguments for parameter tuning and define a seed to allow for reproducibility. Two versions of pre-trained BERT language models are applied: Brazilian Portuguese BERT[10] (Souza et al., 2019), and Multilingual BERT[11] (Wolf et al., 2019).

ToLD-Br is used to fine-tune BERT-based models for our monolingual experiments, with monolingual BERT (`BR-BERT`) and multilingual BERT (`M-BERT-BR`). Although `M-BERT-BR` refers to the multilingual version of BERT, we refer to these two models as "monolingual models," as we trained using the dataset with Brazilian Portuguese sentences alone.

Using the multilingual model, we also carry out experiments in which we add data in English to train the models either through transfer learning or zero-shot learning. For these experiments we use the OLID data, concatenating the training and test splits into a single dataset. For transfer learning, we merged OLID and ToLD-Br to obtain a model with both languages as input, aiming to assess whether extra data in English helps in building better models (`M-BERT(transfer)`). For zero-shot learning, OLID is used alone at training time, building a model that did not have access to any data in Brazilian Portuguese (`M-BERT(zero-shot)`).

[8] https://automl.github.io/auto-sklearn

[9] github.com/ThilinaRajapakse/simpletransformers

[10] huggingface.co/neuralmind/bert-base-portuguese-cased

[11] huggingface.co/bert-base-multilingual-cased

Through these experiments, we can assess the advantages of monolingual models, whether data from another language can directly benefit the classification, and whether a specific monolingual dataset is necessary or not.

We experiment with different sizes of the training set to assess the influence of the volume of data on the classification. For that, we evaluate the results on random subsets of the data. The size of each partition varies in a range between 10% and 100% adding 10% of the data at each iteration. For each step, we repeat the classification three times to minimise the probability of reporting results obtained by chance. Our best model (`M-BERT-BR`) is used for this experiment (c.f. Section 5).

Evaluation for binary classification is done in terms of precision, recall and, $F1$-score per class and macro-$F1$. We also analyse the confusion matrices of our systems in order to better visualise the performance of our models in each class, mainly focusing on an analysis of false negatives.

Although we mainly focus on binary classification, an initial approach for multi-label classification is also presented. We use the adaptation for the multi-label classification scenario available in `simpletransformers`. In this case, the transformer's output consists of six neurons, each representing one of the labels. These neurons are considered independent in the training and prediction process. Thus, when an output neuron is activated, we set the label represented by this neuron to positive. Besides, we evaluate the performance of a baseline based on `BoW+AutoML`, where we train an AutoML model for multilabel classification. Evaluation is done in terms of *Hamming* loss and average precision (Tsoumakas et al., 2009).

5 Results and Discussion

This section shows the results of our experiments in classifying toxic comments using ToLD-Br.

5.1 Binary Classification

For evaluating our models, we are particularly interested in models with high performance in the positive class (classification of *toxic* comments). The worst case scenario are false negatives, i.e. *toxic* comments classified as *non-toxic*. Tables 7 through 11 summarises the results for each model. `BoW+AutoML` is already a competitive model, achieving 74% of macro-$F1$, as shown in Table 7 and Figure 2a.

	Precision	Recall	F1-score
0	0.76	0.75	0.75
1	0.71	0.73	0.72
Macro Avg	0.74	0.74	0.74
Weighted Avg	0.74	0.74	0.74

Table 7: BoW + AutoML

	Precision	Recall	F1-score
0	0.77	0.80	0.79
1	0.76	0.73	0.74
Macro Avg	0.76	0.76	0.76
Weighted Avg	0.76	0.77	0.76

Table 8: BR-BERT

	Precision	Recall	F1-score
0	0.81	0.69	0.75
1	0.69	0.82	0.75
Macro Avg	0.75	0.75	0.75
Weighted Avg	0.76	0.75	0.75

Table 9: M-BERT-BR

	Precision	Recall	F1-score
0	0.80	0.74	0.77
1	0.72	0.79	0.75
Macro Avg	0.76	0.76	0.76
Weighted Avg	0.77	0.76	0.76

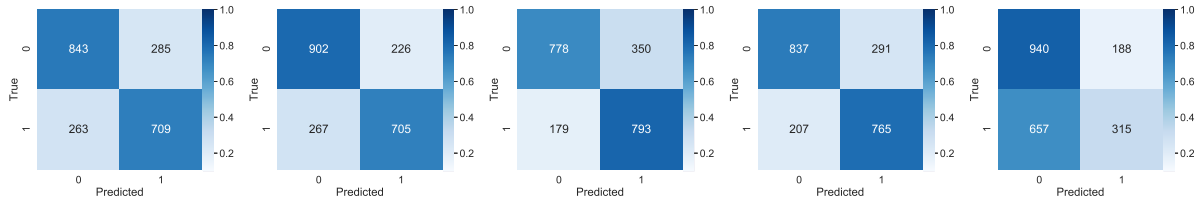
Table 10: M-BERT(transfer)

	Precision	Recall	F1-score
0	0.59	0.83	0.69
1	0.63	0.32	0.43
Macro Avg	0.61	0.58	0.56
Weighted Avg	0.61	0.60	0.57

Table 11: M-BERT(zero-shot)

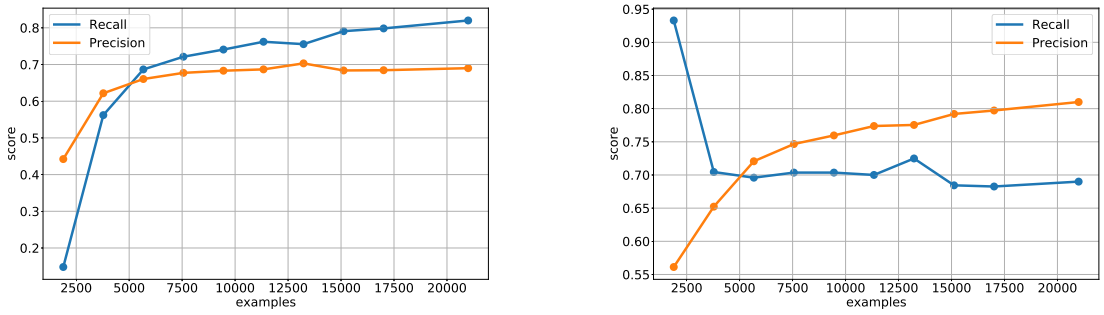
The monolingual models `BR-BERT` and `M-BERT-BR` (Tables 8 and 9, respectively) show very similar performances in all metrics, with `BR-BERT` being slightly better in terms of macro-$F1$. However, `M-BERT-BR` is better in terms of $F1$-score for the positive class and shows fewer false negatives than `BR-BERT` (Figure 2b for `BR-BERT` and Figure 2c for `M-BERT-BR`).

`M-BERT(transfer)` (Table 10) does not out-



(a) (b) (c) (d) (e)

Figure 2: Confusion matrices for each model (a) BoW+AutoML (Baseline); (b) BR-BERT; (c) M-BERT-BR; (d) M-BERT(transfer); (e) M-BERT(zero-shot)



(a) (b)

Figure 3: Precision and recall for different sizes of the training dataset for the (a) *positive* and (b) *negative* classes.

perform the monolingual models and it also shows more false negatives than `M-BERT-BR` (Figure 2e). On the other hand, the number of false negatives in `BR-BERT` (267) is slightly higher than the number of false negatives in `M-BERT(transfer)` (207). Finally, `M-BERT(zero-shot)` (Table 11) is the worst model, as expected. It performs particularly bad when classifying the positive class, achieving only 43% of $F1$-score for this class, mainly caused by its high number of false negatives (Figure 2d).

In summary, transfer learning does not seem to improve over the overall performance of monolingual models. Based on the analysis of false negatives, `M-BERT-BR` appears as our best model. Zero-shot learning shows a very low performance, being particularly bad in the positive class.

Error Analysis We also analyse the performance of our best model (`M-BERT-BR`) in each fine-grained class. The idea is to identify which toxic classes are most difficult to be classified as *toxic* by our binary classifier. As false negatives are a critical type of error in our application, Table 12 shows the false negative rate (false negatives / expected positives) for each toxic class. The ratio of false negatives is inversely proportional to the number of examples for a specific class. *Insult* and *obscene*, the largest classes, show the lowest false negative rate, whilst the highest rates are shown by classes with less examples (*racism* and *xenophobia*). Therefore, in order to improve classification models, these aspects of the imbalanced data need to be taken into account and further studied.

	False negative rate
LGBTQ+phobia	7/35 (0.2)
Insult	67/448 (0.15)
Xenophobia	13/19 (0.68)
Misogyny	7/45 (0.15)
Obscene	117/701 (0.17)
Racism	8/17 (0.47)

Table 12: Error analysis for each label.

5.2 Importance of Large Datasets

In this experiment, we highlight the importance of collecting a considerable amount of examples, as toxicity can be expressed in many different ways. We separated the training data into 10 random splits from 10% to 100% of the data, increasing 10% of data at each step, and trained `M-BERT-BR` with three random samples for each step. Figure 3 shows the mean recall, precision and $F1$-score for the positive and negative classes, respectively, for each

data split. With few training examples, the model only performs well on the majority class, but as the number of instances grows, recall for the negative class starts decreasing while recall for the positive class increases, and precision rises for both classes. At least 6K examples seems to be necessary to achieve reliable results, while previous work for Portuguese reports the largest dataset with only 5,668 examples. This highlights the importance of ToLD-Br, as a large-scale dataset.

5.3 Multi-Label Classification

We experiment with multi-label classification, building a model using the Multilingual BERT (similar to `M-BERT-BR`). Our baseline is a set of `BoW+AutoML` models trained using Binary Relevance (Tsoumakas et al., 2009) for multi-label classification. The BERT-based models adopt a score threshold of 0.5 in the output neuron to deal with multi-label. If the activation for a label in the output layer is higher than the threshold, we consider it positive.

The baseline model obtained 0.08 and 0.20 of *Hamming* loss and average precision, respectively, while `M-BERT-BR` resulted in 0.07 and 0.19 for these measures, respectively. Figure 4 displays the confusion matrices obtained by `M-BERT-BR`.

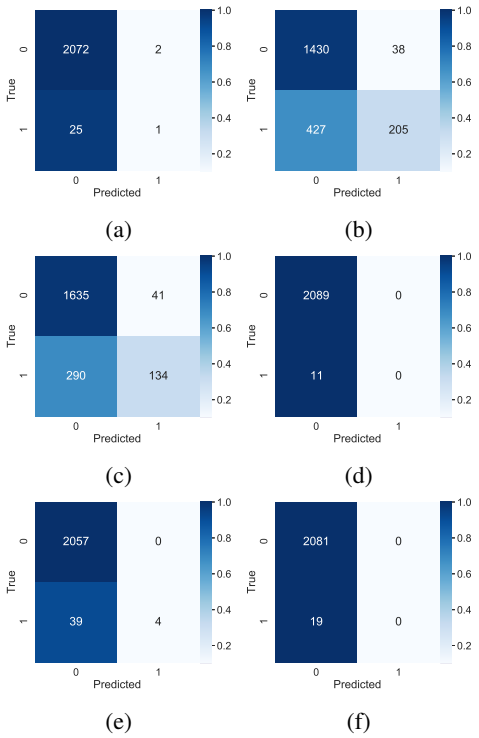


Figure 4: Confusion matrices for each label (a) *LGBTQ+phobia*; (b) Obscene; (c) Insult; (d) Racism; (e) Misogyny; (f) Xenophobia.

This scenario is considerably more challenging than binary classification. The positive class of each label corresponds to a subset of the examples labelled as *toxic*. Thus, it is likely that the number of instances for these classes will be insufficient for the model to learn. Besides, the problem of unbalanced classes becomes evident (c.f. Table 6). As a consequence, it is clear that labels with a small number of positive examples, like *racism*, *misogyny*, *xenophobia*, and *LGBTQ+phobia* were almost entirely classified as negative. In contrast, for *obscene* and *insult*, labels with a considerable amount of positive examples, the model was capable of classifying some examples correctly. In all cases, besides *insult*, the baseline performs slightly better for the positive class (which justify the higher *Hamming* loss). This setback is likely due to the difficulty of the neural model to learn with few examples.

6 Concluding Remarks

In this paper, we present ToLD-Br: a dataset for the classification of toxic comments on Twitter in Brazilian Portuguese. Through a wide and comprehensive analysis, we demonstrated the need for this dataset for studies on automatic classification of toxic comments. We highlight that monolingual approaches for this task still outperform multilingual experiments and that large-scale datasets are needed for building reliable models. Also, we show that there are still challenges to be overcome, such as the naturally significant class imbalance when dealing with multi-label classification.

As future work, in addition to deal with class imbalance, we intend to evaluate if aggregating classes with high divergences between annotators can build more reliable models. Besides, we intend to assess the benefits of adding unlabelled data to ToLD-Br to use semi-supervised techniques.

7 Acknowledgements

We thank the volunteers from UFSCar that made this research possible. The MIDAS group[12] from the Federal University of São Carlos (UFSCar), Brazil, funded the annotation process. The SoBigData TransNational Access program (EU H2020, grant agreement: 654024) funded Diego Silva and João Leite's visits to the University of Sheffield.

[12] `midas.ufscar.br`

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes y Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, volume 2150 of *CEUR Workshop Proceedings*, pages 74–96. CEUR-WS.org.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, pages 512–515. AAAI Press.

Rogers Prates de Pelle and Viviane P. Moreira. 2017. Offensive comments in the Brazilian web: a dataset and baseline results. In *Proceedings of the VI Brazilian Workshop on Social Network Analysis and Mining*, pages 510–519, Porto Alegre, RS, Brazil. SBC.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. 2019. Auto-sklearn: efficient and robust automated machine learning. In *Automated Machine Learning*, pages 113–134. Springer, Cham.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, pages 491–500, Stanford, California. AAAI Press.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, page 14–17, Kolkata, India.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

Fabio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, pages 1–8.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2009. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018

Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 1–10, Vienna, Austria.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, abs/1910.03771:1–11.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, page 1391–1399, Perth, Australia.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 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)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identificationin Social Media (OffensEval 2020). In *To appear in the Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain.