Classifying toxic posts on the ToLD-Br dataset

Pedro Fracassi *Computer Engineering Student, Insper*

***Abstract*—The abstract goes here.**

***Index Terms*—IEEE, IEEEtran, journal, LaTeX, paper, template.**

Introduction

CONTENT MODERATION is a topic ever increasing in importance in the modern days, given the amount of hateful posts made on Social Media. Although necessary, human-based moderation comes at an emotional cost for moderators [1], making automated algorithmic moderation a necessity. One of the first steps of automated content moderation is detection and classification of posts in different categories, so that actions can be taken.

I. Dataset

The Toxic Language Detection for Brazilian Portuguese (ToLD-Br) is a dataset with tweets in Brazilian Portuguese annotated according to different toxic aspects [2], and has been used in recent years to train models that detect toxic comments and hate speech in Brazilian Portuguese [3].

The dataset is provided both in multi-label and binary forms. The binary form was chosen for classification, with the possible classes being **toxic** and **non toxic**.

TABLE I
Dataset Distribution

Label	Train.	Valid.	Test	Prop.
Toxic	7,375	921	972	44%
Non-toxic	9,425	1,192	1,128	56%

II. Classification Pipeline

A. Pre-processing and Feature Engineering

Before fitting the classifier to the ToLD-Br corpus, we pre-processed its contents. All text was converted to lowercase, punctuation was removed, multiple whitespaces were stripped, and URLs and `@user` mentions were removed.

After cleaning up the dataset, some feature engineering was also done. Stop words like ”o”, ”a”, etc. were removed. Lemmatization were applied with , reducing words to their root forms.

B. Classifier Pipeline

A simple scikit-learn Pipeline with a TFIDF vectorizer and a Random Forest classifier was used. The vectorizer was set to include ngrams of from 1 up to 3 words, to capture multi-word expressions that might represent toxicity.

As this is a bag-of-words model, it might be easily exploited and confused by irony, sarcasm or citations, where the ”toxic” words do appear, but aren’t necessarily toxic in that context.

III. Evaluation

After fitting, and evaluating the model with a re-shuffled train/test split of the dataset, a balanced accuracy of **72.27%** was achieved. This is better than random-guessing, and slightly below state-of-the-art back in 2020 when the original dataset paper was published.

The 20 most important ngrams for classification were extracted, as seen on Table II. It is very easy to see that everything listed there are common portuguese curse words, which leads us to believe that the classifier is working correctly, instead classifying coincidences.

IV. Downsampling

The training and test errors were assessed at various levels of dataset downsampling, specifically at 10% to 90% of the dataset’s size, as show, by Figure 1. The results of the assessment suggest that there is still room for improvement by increasing dataset size, as the error steadily decreases when a larger percentage of the dataset is used.

Expanding the dataset is feasible, but resource-consuming: collecting new posts would be trivial through APIs, but classification would require more time, money and volunteers.

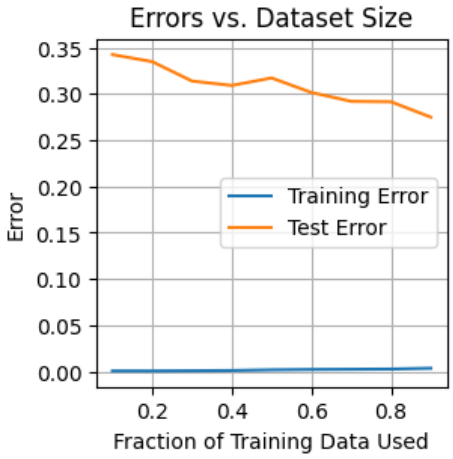

Fig. 1. Analisis of classification errors at different levels of downsampling

V. Topic Analysis

A topic analysis was done to separate the corpus into different topics and then fit a model for each topic. Better results were unfortunately not achieved in this was, with accuracy dropping to around 70%. We assume this is due to the way the dataset was created, with specific words being filtered on Twitter, to be later classified.

TABLE II
MOST IMPORTANT WORDS

Rank	Word	Importance
1	cu	0.029665
2	porr	0.018472
3	caralh	0.014739
4	put	0.011070
5	fud	0.009615
6	pqp	0.008925
7	pau	0.006563
8	fod	0.006276
9	fdp	0.006075
10	tom cu	0.005695
11	rt	0.005219
12	vagabund	0.005196
13	burr	0.004829
14	filh put	0.004745
15	car	0.004546
16	tom	0.003632
17	babac	0.003629
18	piranh	0.003590
19	fei	0.003434
20	filh	0.003397

REFERENCES

[1] S. Roberts, *Commercial Content Moderation: Digital Laborers' Dirty Work*, 01 2016.

[2] J. A. Leite, D. Silva, K. Bontcheva, and C. Scarton, "Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, K.-F. Wong, K. Knight, and H. Wu, Eds. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 914–924. [Online]. Available: https://aclanthology.org/2020.aacl-main.91

[3] G. D. Saraiva, R. Anchiêta, F. A. R. Neto, and R. Moura, "A semi-supervised approach to detect toxic comments," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, R. Mitkov and G. Angelova, Eds. Held Online: INCOMA Ltd., Sep. 2021, pp. 1261–1267. [Online]. Available: https://aclanthology.org/2021.ranlp-1.142

[4] J. D. Garrett, "garrettj403/SciencePlots," Sep. 2021. [Online]. Available: http://doi.org/10.5281/zenodo.4106649