EXPLORING PORTUGUESE HATE SPEECH DETECTION WITH TRANSFORMERS

Gabriel Assis (1), Annie Amorim (1), Jonnathan Carvalho (2),
Daniel de Oliveira (1), Daniela Vianna (3), Aline Paes (1)

(1) Institute of Computing, Universidade Federal Fluminense, Niterói, RJ, Brazil
(2) Department of Informatics, Instituto Federal Fluminense, Itaperuna, RJ, Brazil
(3) JusBrasil, Brazil

Email: assisgabriel@id.uff.br, annieamorim@id.uff.br, alinepaes@ic.uff.br

Machine Learning and Learning from Language (MeLLL IC-UFF)

PROBLEM

- Social media allows for broad communication but can also foster hateful content;
- Linguistic and cultural aspects of PT-BR are not properly framed in multilingual analyses;
- Binary classification does not acknowledge the difference between hate speech (mostly a crime) and offenses (bad behavior).

Given a social media post P written in Portuguese, pre-process it returning X, and classify it as belonging to one of the three classes in Y = {"hate speech", "offensive" or "neutral"}

OBJECTIVE

- Investigate the performance of Transformers variants in tackling hate speech in PT-BR;
- Comparing the potential of PT-BR and multilingual encoders with LLMs trained in PT-BR and LLMs with the emerging multilingual ability.

Group 1:	Group 2:	Group 3:
• BERTimbau • AIBERTina • BERTweet.BR • Bernice	• Sabiá-7B • Gervásio-7B	• Gemini-pro • GPT-3.5

METHOD

Datasets

- **HateBR** (Instagram) • **OLID-BR** (YouTube, Twitter) • **ToLD-BR** (Twitter)

Encoder-Classifier Models

The encoder-based models pre-trained with PT-BR corpora include
- **BERTimbau-large**
- **AIBERTina-PTBR-100m**
- **BERTweet.BR** (tweets corpus)

A model pre-trained with Multilingual corpus is also selected:
- **Bernice** (tweets corpus)

Our strategy consists of:
- **Fine-tuning** by stacking a classifier layer onto the language model and adjusting the entire model.

Decoder-Classifier Models

The decoder-based models pre-trained with PT-BR corpora include
- **Sabiá-7B-1** (built on LLaMA-1 architecture)
- **Gervásio-7B-PTBR** (built on LLaMA-2 architecture)

Our strategy consists of:
- Also **fine-tuning** as the encoders but here we (unusually) put the classifier on the top of the decoder.

General Purpose Generative LLMs

LLMs that are activated with prompts include
- **Gemini-pro 1.0**
- **GPT-3.5-turbo**

Our strategy consist of:
- **In-context learning** with prompts
- Some variations include demonstration instances

The instruction used is as follows:
CLASSIFY THE SOCIAL MEDIA TEXT AS "HATE SPEECH", "OFFENSIVE", OR "NEUTRAL". \N TEXT: *target* \N CLASS:

Demonstration selection strategies were based on the number of examples:
- **zero-shot**, with no examples;
- **one-shot**, with a single example;
- **one-class-shot**, with one example per class;
- **few-shot**, with *N* examples per class.

We select demonstration options considering
- **random choice**
- **semantic similarity**
- **the number of tokens**

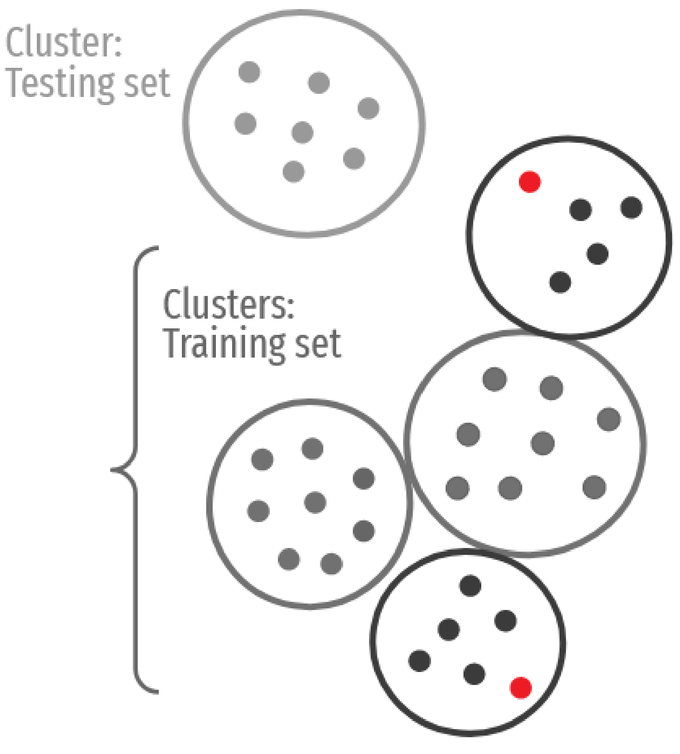

Figure 1

Figure 1 illustrates the process used to select examples based on clustering.

RESULTS

- Decoder-based 7-B models underperformed compared to encoder-based models, despite their higher number of parameters,
  - Training deficiency regarding hateful content?
- Selection strategy for prompts, particularly the GPT and Gemini models, showed promise: they outperformed models specifically adapted and fine-tuned for Portuguese, such as Sabiá and Gervásio
- Targeted nature of encoder language models pre-trained on specific domains like social media and Portuguese, and explicitly adjusted for the task, proves crucial for effectively identifying hate speech

Table 1 summarizes the results.

	HateBR						OLID-BR						ToLD-BR					
	Strategy	**prec.**	**rec.**	**acc.**	**f1**	**f1$_{hs}$**	**Strategy**	**prec.**	**rec.**	**acc.**	**f1**	**f1$_{hs}$**	**Strategy**	**prec.**	**rec.**	**acc.**	**f1**	**f1$_{hs}$**
BERTimbau	Fine Tuning	0.803	**0.822**	0.862	**0.811**	**0.667**	Fine Tuning	0.637	0.664	0.663	0.623	0.596	Fine Tuning	0.529	0.598	0.599	0.474	0.065
AIBERTina	Fine Tuning	0.793	0.707	0.800	0.734	0.569	Fine Tuning	0.599	0.589	0.613	0.563	0.544	Fine Tuning	0.417	0.453	0.538	0.399	0.059
BERTweet.BR	Fine Tuning	0.768	0.793	0.846	0.779	0.583	Fine Tuning	0.625	**0.665**	**0.672**	**0.635**	**0.606**	Fine Tuning	**0.543**	**0.671**	**0.708**	**0.548**	**0.178**
Bernice	Fine Tuning	**0.830**	0.788	**0.863**	0.805	0.656	Fine Tuning	**0.640**	0.660	0.666	0.620	0.579	Fine Tuning	0.534	0.640	0.704	0.536	0.141
Sabiá-7b-1	Fine Tuning	0.465	0.379	0.526	0.322	0.129	Fine Tuning	0.422	0.434	0.532	0.437	0.412	Fine Tuning	0.383	0.381	0.531	0.355	0.000
Gervásio-7b	Fine Tuning	0.595	0.619	0.672	0.595	0.345	Fine Tuning	0.464	0.480	0.495	0.457	0.475	Fine Tuning	0.361	0.388	0.446	0.336	0.042
GPT-3.5-turbo	size-based one-class-shot	0.654	0.696	0.697	0.621	0.408	sim-based one-class-shot	0.526	0.567	0.553	0.528	0.564	sim-based few-shot	0.486	0.543	0.621	0.447	0.081
Gemini-pro 1.0	size-based one shot	0.602*	0.609*	0.601*	0.562*	0.407*	sim-based one shot	0.592*	0.476*	0.607*	0.460*	0.554*	rand-based few shot	0.475*	0.526*	0.609*	0.455*	0.100*

Table1. Macro results of precision, recall, accuracy, f1-score, and also the hate speech class f1-score for each model in its best configuration. Gemini* results may slightly fluctuate due to the rate of responses blocked by Google API filters. This rate was 0.15%, 0.79% and 0.12% for each dataset, respectively. Best results in **bold**.

KEY FINDINGS

- Small tuned models still play a crucial role in addressing this task;
- Adding context and carefully selected examples benefits prompt-activated generative models;
- Specialized training generates better results than the in-context learning abilities of LLMs;
- Our results illustrate AI limitations in this critical domain.

FUTURE DIRECTIONS

- Analysing the impact of different layers, pre-training corpora and variations in the architectures;
- Provide some explanation on how the models achieve different results;
- Other prompt strategies;
- Investigating the prompt strategies in other sensitive classification scenarios.

Acknowledgments

uff CAPES CNPq FAPERJ