Using Cross Lingual Learning for Detecting Hate Speech in Portuguese

Anderson Almeida Firmino[1(✉)], Cláudio Souza de Baptista[1], and Anselmo Cardoso de Paiva[2]

[1] Information Systems Laboratory, Federal University of Campina Grande, Campina Grande, Paraiba, Brazil
andersonalmeida@copin.ufcg.edu.br, baptista@computacao.ufcg.edu.br
[2] Applied Computing Center, Federal University of Maranhão, Maranhão, Brazil
paiva@nca.ufma.br

Abstract. Social media growth all around the world brought benefits and challenges to society. One problem that must be highlighted is hate speech proliferation on the Internet. This article proposes a technique to hate speech detection in texts, which uses a Cross-Lingual Learning classifier. In our experiments, we used a public dataset in Portuguese and achieved one of the greatest F1 Scores within our state-of-the-art. Besides, this work is the first one to perform a cross-lingual learning task for hate speech detection using a corpus in Portuguese.

Keywords: Hate speech detection · Cross-Lingual Learning · Deep learning · Social media

1 Introduction

Mobile technology growth has impacted social media. According to a recent survey[1], people prefer to use their smartphones and social media for news consumption instead of printed newspapers and television. People usually prefer platforms such as Facebook and Twitter for gathering information and news.

Besides the benefits and convenience that social media has provided, the anonymity provided by such means may be harmful to society, keeping in mind that people tend to have a more aggressive behavior while using their social networks [1]. An example of this is the growing proliferation of hate speech on the Internet. Fortuna and Nunes [1] define hate speech as a language that attacks and incites violence against certain groups of people based on their specific characteristics such as physical appearance, religion, lineage, nationality or ethnic origin and gender [2].

[1] https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020.

This research was partially funded by Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) and the Brazilian Research and Development Council (CNPq).

C. Strauss et al. (Eds.): DEXA 2021, LNCS 12924, pp. 170–175, 2021.
https://doi.org/10.1007/978-3-030-86475-0_17

As discussed by Pikuliak et al. [3], most languages do not have enough data available to create state-of-the-art models. Therefore, the ability to create intelligent systems for these languages is restricted. The importance of Natural Language Processing (NLP) tasks for languages with fewer resources has emerged recently during various crises in regions of the World where people speak languages that are not commonly dealt with in the NLP community, such as the Ebola outbreaks in West Africa (e.g. Niger-Congo languages).

This work proposes to perform hate speech detection in texts using a multilingual approach. The languages used in this work are Italian and Portuguese since both originate from the same mother language - Latin. Pelle and Moreira [4] published the Portuguese dataset we used, and the Italian dataset was made available in Evalita 2018, in the task Hate Speech Detection - Bosco et al. [5].

The contributions of this work lie because this is the first work to use CLL with hate speech data in Portuguese, besides having achieved one of the best results in the literature with the OffComBr2 database - provided by Pelle and Moreira [4].

The remainder of this paper is structured as follows. Section 2 focuses on reviewing briefly the state-of-the-art on hate speech detection in texts using CLL. Section 3 addresses the methodology and the dataset used by this research work. Section 4 presents the results and evaluation method. Finally, Sect. 5 highlights the conclusion and further work to be undertaken.

2 Related Work

Silva et al. [6] developed a novel approach to detect hate speech in Portuguese, which comprises a CNN model and a psycho-linguistic dictionary, the Linguistic Inquiry and Word Count (LIWC), with Logistic Regression (LR+LIWC). They used three Brazilian datasets; OffComBr2 and OffComBr3 [4] and HSD [7], and compared the baseline results with theirs. Using a CNN along with a 300-dimensional word embedding achieved the best F1 score.

The idea of Pagmunkas and Pati [9] was to develop a multilingual hate speech detection approach, in which the concept of transfer learning from a more resource-rich language to another with fewer resources is used. The languages studied were English, Spanish, Italian, and German. The best results were achieved using Hurtlex and multilingual embeddings as features and an LSTM architecture, and the best approach used Joint Learning and Multilingual Embeddings.

Ranasinghe and Zampieri [10] used multilingual word embeddings to detect hate speech. Besides carrying out experiments with different languages, different domains were also tested. Data were obtained in English, Spanish, Hindi, and Bengali. The XLM-R framework was used to perform the classification. The idea of using a multilingual approach is to train the model in a richer language and test it in another language with fewer resources. Ranasinghe and Zampieri trained the model in English and tested it in the other three languages researched. The results showed went beyond the state-of-the art of each dataset and language.

3 Methodology

This section describes the Portuguese and Italian datasets used. Then, we show the methodology used and how we apply cross-lingual learning to detect hate speech in Portuguese texts.

We used two datasets in this research: one comprising data in the Portuguese language (OffComBr-2 - [4]), with comments containing hate speech collected from the Brazilian news site g1.globo.com. Pelle and Moreira [4] gathered comments from pages on politics and sports. They selected a sample of 1,250 random comments and two judges noted each comment.

The other dataset comprises Facebook posts in Italian, made publicly available at the Evalita 2018 conference [5], in the Hate Speech Detection task. The dataset was developed by a research group in Pisa, created in 2016 [11], and contains about 17,000 Facebook comments, extracted from 99 posts from selected pages.

The main idea of using Cross Lingual Learning is to use one language with more resources to train a model and then use this model in a language with fewer resources. So, we chose Italian as the source language (the language with more resources) and Portuguese as the target language. We used the XLM-RoBERTa framework - called XLM-R [12] because it has presented good results in Cross-Lingual Learning tasks, achieving an accuracy 23% higher than that of BERT in using low-resource languages [13].

According to Pikuliak et al. [3], in Cross-Lingual Learning tasks, the transfer learning from the language with the most resources to the one with the least resources can be done in three ways: Zero Shot Transfer, Joint Learning and Cascade Learning. The first occurs when no data from the target language is used in the training. The joint learning approach consists of using both languages at the same time in the training; and cascade learning occurs when there is a pre-training with the source language and then the model is fine tuned with the target language.

Thus, we have conducted experiments following these three approaches. In all cases, the source language was Italian and the target language was Portuguese. We used the base and the large versions of XLM-R. The base version contains approximately 270M parameters, with 12 layers, 768 hidden states, 3072 feed-forward hidden states, and 8-heads; and the large version contains 550M parameters, with 24 layers, 1024 hidden states, 4096 feed-forward hidden-state and 16-heads [12].

We have performed some experiments initially using the first two approaches listed above (zero-shot transfer and joint learning). In these experiments, we only trained the XLM-R model using the Italian data (adding some Portuguese data sometimes) and we tested it with the Portuguese data. In the second round of experiments, we performed a fine-tuning on the model. We trained it with Italian data; after that, we did another training - this time with Portuguese data - and then we tested it with Portuguese data. This process is shown in Fig. 1.

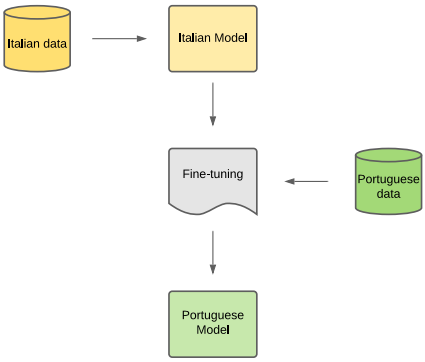

Fig. 1. The fine-tuning process on cross-lingual learning.

4 Evaluation and Results

This section presents the experiments used to evaluate our proposed model to identify hate speech in Portuguese in tweets. Moreover, we discuss the results of the performed experiments.

In our experiments, we used a Nvidia Tesla K80 GPU to train the models used. We obtained $1e^{-5}$ as the best value for the learning rate and we adopted 3 as the best value for the number of epochs. First of all, we have performed some experiments using the zero-shot transfer and joint learning approaches.

In these experiments, we performed the training with the Italian dataset with 90% used for training and the 10% remaining used for validation. In the test step, we used the Portuguese dataset (zero-shot) or splitted the Portuguese dataset in two parts (Table 1 shows the amount of Portuguese data used in training; the remaining data was used for testing). The experiment results are displayed in Table 1. As we can see, the results from the large model are greater than the base one. Also, using the joint learning approach, when we increased the amount of target language in the training process, the F1 score was increased. But we observed if we added data from the target language above a threshold, the F1 score decreased. This threshold was about 70%, and this result is listed in Table 1.

We also conducted experiments using the cascade learning approach. Here, we also trained the model with the Italian dataset, with 90% used for training and the 10% remaining for validation. In zero-shot experiments, we used 70% of data for training, 10% for validation, and 20% for test. In joint learning experiments, we also set apart some Portuguese data for the training with Italian data (as we can see in Table 1).

Table 1. Experiment results.

Model	Approach	Fine-Tuning	F1 score
XLM-R (base)	zero-shot	No	58%
XLM-R (base)	joint learning (20%)	No	69%
XLM-R (base)	joint learning (40%)	No	73%
XLM-R (large)	zero-shot	No	71%
XLM-R (large)	joint learning (20%)	No	75%
XLM-R (large)	joint learning (40%)	No	77%
XLM-R (large)	joint learning (70%)	No	74%
XLM-R (large)	zero-shot	Yes	80%
XLM-R (large)	joint learning (30%)	Yes	86%
XLM-R (large)	joint learning (50%)	Yes	74%

We trained the XLM-R with the Italian dataset and then fine-tuned it with the Portuguese dataset. We also used the zero-shot and joint learning within this approach (adding or no data from Portuguese in the training step). The experiments are listed in Table 1. We notice that when we used the joint learning approach; we obtained the best result in our experiments. But we observed we cannot add too much data from Portuguese in the training step. The F1 score decreased sharply from 86% to 74% - when the amount of Portuguese data was increased from 30% to 50% in the training step.

Silva et al. [6], Lima and Bianco [14], and Soto et al. [8] also used the OffComBr2 database [4]. In Table 2, a comparison with the F1-Score of these works is displayed. It is worth noticing that our work is the only one that used cross-lingual learning and it achieved one of the best results using the OffComBr2 database.

Table 2. Comparison of related work results.

Work	F1 score
Baseline (Pelle e Moreira, [4])	77%
Silva et al. [6]	89%
Lima and Bianco [14]	72%
Soto et al. [8]	86%
Our approach	86%

5 Final Remarks

In this paper, we presented an approach for hate speech detection in texts using a cross-lingual learning approach. Among other works that used the same dataset, we had one of the best F1 scores so far.

As further work, we point to the usage of other languages to perform the training in the model (such as English) to verify an improvement when classifying the texts. Besides, we suggest the use of more iterations on the fine-tuning step.

References

1. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. (CSUR) **51**(4), 1–30 (2018)
2. Bourgonje, P., Moreno-Schneider, J., Srivastava, A., Rehm, G.: Automatic classification of abusive language and personal attacks in various forms of online communication. In: Rehm, G., Declerck, T. (eds.) GSCL 2017. LNCS (LNAI), vol. 10713, pp. 180–191. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73706-5_15
3. Pikuliak, M., Šimko, M., Bielikova, M.: Cross-lingual learning for text processing: a survey. Expert Syst. Appl. **165**, 113765 (2021)
4. Pelle, R.P., Moreira, V.P.: Offensive comments in the Brazilian Web: a dataset and baseline results. In: Anais do VI Brazilian Workshop on Social Network Analysis and Mining. SBC, July 2017
5. Bosco, C., Felice, D. O., Poletto, F., Sanguinetti, M., Maurizio, T.: Overview of the EVALITA 2018 hate speech detection task. In: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, vol. 2263, pp. 1–9. CEUR (2018)
6. Silva, S.C., Serapião, A.B., Paraboni, I.: Hate-speech detection in Portuguese using CNN and psycho-linguistic dictionary. J. Inf. Data Manage. **5**, 1–12 (2019)
7. Fortuna, P.C.T.: Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes (2017)
8. Soto, C., Nunes, G., Gomes, J.: Avaliação de técnicas de word embedding na tarefa de detecção de discurso de ódio. In: Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional (pp. 1020–1031). SBC, October 2019
9. Pamungkas, E.W., Patti, V.: Cross-domain and cross-lingual abusive language detection: a hybrid approach with deep learning and a multilingual lexicon. In: Proceedings of the 57th Annual Meeting of the Association For Computational Linguistics: Student Research Workshop, pp. 363–370, July 2019
10. Ranasinghe, T., Zampieri, M.: Multilingual offensive language identification with cross-lingual embeddings. arXiv preprint arXiv:2010.05324 (2020)
11. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), pp. 86–95 (2017)
12. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116 (2019)
13. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
14. Lima, C., Dal Bianco, G.: Extração de característica para identificação de discurso de ódio em documentos. In: Anais da XV Escola Regional de Banco de Dados, pp. 61–70. SBC, April 2019