Evaluation of Methods to Improve Hate Speech Detection in Social Media for Low-Resource Languages with a Focus on Spanish: XLM-RoBERTa-CNN

Xiyana V. Figuera

Department of Computer Science and Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea

xyin@unist.ac.kr

***Abstract*—**Toxic comments or hate speech are a widely spread problem in social media. Numerous platforms and service companies are seeking efficient artificial-intelligence systems that can help human moderators on this task. Spanish is one of the languages that faces the most difficulties on this topic, given the scarcity of publicly available data. This work proposes to implement the multilingual model XLM-RoBERTa-CNN for this task and offers a comparative study on fine-tuning approaches of the model. The findings of this study suggest that pure-language datasets can obtain better performance with a longer fine-tuning, while translated datasets can achieve a good performance when doing a short fine-tuning.

1. INTRODUCTION

Toxic language refers to all those words that seek to denigrate, humiliate, belittle and/or ridicule an individual or group of people, harassing them both in daily life and in social networks, with the purpose of generating hatred towards them, as well as instilling fear in those who are victims of these attacks. It often causes irreversible damage to their mental, social, physical, or emotional health. Among the most harmful examples of toxic language that exist today are the so-called hate speeches that abound on all social networks. According to the Council of Europe (October 20, 1997), hate speech comprises "all those expressions that propagate, incite, promote or justify hatred," since studies have determined that verbal acts of hate are dispersive, contagious and emotionally effective, and they tend to be dehumanizing because the "other" becomes a thing.

These hate speeches and incitement to physical and verbal violence have become a threat to society, and are categorized today as a cyber-crime in many countries. Currently, social networks have positioned themselves as the main channel of communication and dissemination of information. The possibility of maintaining a hidden profile and avoiding a physical confrontation has sometimes increased the level of aggression in the way in which netizens communicate. Recent studies have shown that with the application of modern Natural Language Processing (NLP) techniques, it is possible to effectively detect digital violence in social networks. Thus, the analysis of hate speech is a current and relevant problem for the field of NLP, with many techniques and approaches (Alrehili, 2019); among them, the use of transformer-based models has proved to be useful in the detection of hate comments (Luu et al., 2021).

Nevertheless, research on this topic faces a major problem for low-resource languages (e.g., scarcity of data for some languages). This work addresses this issue, more specifically for the Spanish language which, to the best of the author's knowledge, has few available datasets and only one publicly available dataset for toxic-comment detection. This work proposes to use XLM-RoBERTa-CNN to address the Spanish toxic-comment detection task. The use of multilingual models and data in other languages can benefit low-resource languages (Calizzano et al., 2021), and CNN models have shown outstanding performance on this task (Georgakopoulos et al., 2018). There are also different ways in which the fine-tuning of multilingual models can be carried out; this work focuses on the use of pure-language datasets

(i.e., no translation in the data) and translated datasets, as well as a different number of epochs for the fine-tuning.

In summary, the main contributions of this work are the following:

- Comparing the performance of the proposed model with the performance of BETO-CNN (a BERT model trained on a large Spanish corpus).
- An analysis of methods to fine-tune XLM-RoBERTa-CNN focused on Spanish, specifically a comparison of the performance of translated and pure-language data with a different number of epochs.

2. RELATED WORKS

There exist several approaches to the detection of hate speech in comments made on social media. The use of toxic-span detection is among one of the ways to detect these comments. One of the publicly available datasets for the task of toxic-span detection is provided by SemEval-2021 Task 5: Toxic Spans Detection (Pavlopoulos et al., 2021). It includes training and test sets, both consisting of two parts: the content of posts and the spans denoting the toxic words in the posts, where spans represent toxic words as a set of character indexes (Luu and Nguyen, 2021). One of the proposed models for this task combines a detection and a classification model: the detection model (BiLSTM-CRF) returns the toxic spans from the post, while the classification model (ToxicBERT) classifies whether a post is toxic or non-toxic. If a post is non-toxic, the classification model returns an empty span; otherwise, it reserves the spans of the detection model. Although this efficiency is limited to the case of a single-word span, the authors suggest that the use of character-level representations and attention mechanisms may improve the performance for the task of toxic/hate comment detection.

For multilingual data, the work by Calizzano et al. (2021) compares the performance of two multilingual models (XLM-RoBERTa and mT5) against a German-specific language model (GBERT) on a German binary classification task, with and without task-specific pre-training for the multilingual models. The task-specific pre-training was toxicity classification, carried out by taking 12 toxicity or hate-speech classification datasets and training the language models on these datasets before fine-tuning them on the shared-task dataset (German only). The task-specific pre-training dataset was composed of a total of 105,142 examples in nine different languages. The study shows that models can benefit from hate-speech detection datasets in other languages to improve the performance of multilingual models through task-specific pre-training. This is especially important for the training of models whose target language has a relatively small dataset available for the specific task.

For text classification, the use of Convolutional Neural Networks (CNNs) has recently been proposed, approaching text analytics in a modern manner that emphasizes the structure of words in a document (Georgakopoulos et al., 2018). In their work, the authors employ the dataset provided by the Kaggle challenge on toxic-comment classification in an attempt to investigate whether the emergence of CNNs in text mining, using word embeddings to encode texts, provides any significant benefit over traditional approaches accompanied by Bag-of-Words (BoW) representations. That work concludes that CNNs can outperform well-established methodologies, providing enough evidence that their use is appropriate for toxic-comment classification.

3. PROPOSED APPROACH

In the multilingual case, there are not many human-annotated datasets publicly available for the task of toxic-word span detection, and thus a different approach is necessary. The datasets for toxic-comment detection in languages other than English are usually annotated in a binary or hierarchical way for the

whole post or comment, rather than providing span information; hence, in this work the dataset is of the latter type. In the work by Calizzano et al. (2021), datasets for task-specific training are mixed between pure-language datasets (i.e., not translated) and datasets augmented via translation, such as the CONAN dataset (Chung et al., 2019). One of the aims of this work is to offer a comparison of the performance when using both pure-language datasets and translated datasets. More specifically, we analyze whether using a diversified dataset with natural sentences of different languages — that might or might not include the same toxic words and contexts — performs better than a translated dataset. A translated dataset contains the same toxic words and contexts, which might not be natural for the translated language, but would yet allow the model to learn more about the specific toxic words and contexts of the language of translation.

This work proposes combining the pre-trained multilingual model XLM-RoBERTa with a CNN for the classification of hate speech. XLM-RoBERTa is particularly useful as it is pre-trained on the languages of interest, which makes it possible to simply fine-tune for the task, and the CNN is useful to find features that allow the construction of representations of the text.

4. EXPERIMENTS

4.1. Dataset

There are not many datasets available for the toxic-comment detection task in Spanish, which is one of the motivations for the focus on this language. The only publicly available data for this language is the Jigsaw Multilingual Toxic Comment Classification challenge dataset from Kaggle, and thus the data of this work is from this source. This dataset is composed of six different languages and, for languages other than English, is labelled in a binary way. The languages used for the multilingual task are Spanish, Turkish and French. Spanish and French are Indo-European languages; Turkish comes from Anatolia. While Turkish has become quite different, Spanish and French retain a 75% lexical similarity.

4.2. Models

This work proposes to use XLM-RoBERTa-CNN for the detection of hate speech in Spanish, Turkish and French, with a focus on Spanish. The performance of XLM-RoBERTa-CNN is compared to that of BETO-CNN. While BETO is a model based on BERT, XLM-RoBERTa is a multilingual version of RoBERTa.

BERT stands for Bidirectional Encoder Representations from Transformers and is pre-trained on a large corpus from Wikipedia and Book Corpus for English. The pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks (Devlin et al., 2019). *RoBERTa* is a Robustly Optimized BERT Pretraining Approach; one of the main differences with BERT is that, contrary to the static masking of BERT (performed once during data preprocessing), for RoBERTa the masking pattern is generated every time a sequence is fed to the model (Liu et al., 2019). *BETO* is a BERT model trained on a large Spanish corpus, of a size similar to BERT-Base and trained with the Whole Word Masking technique (Cañete et al., 2020). *XLM-RoBERTa* (Conneau et al., 2020) is the multilingual version of RoBERTa, trained on the Common Crawl corpus in 100 languages using masked language modeling.

Convolutional Neural Networks for text classification are multi-stage trainable neural-network architectures developed for classification tasks (Georgakopoulos et al., 2018). Each stage consists of layers such as: (1) *Convolutional layers*, the major components of CNNs, consisting of kernel matrices that perform convolution on their input and produce an output matrix of features to which a bias value is

added; (2) *Pooling layers*, which perform dimensionality reduction by subsampling the output of the convolutional layers — the most common being the max-pooling function; (3) the *Embedding layer*, a special component for text-classification problems that transforms text inputs into a suitable form for the CNN, where each word is transformed into a dense vector of fixed size; and (4) the *Fully-Connected layer*, a classic feed-forward hidden layer used in the final stages of CNNs.

In this work, the embedding layer is the implementation of the XLM-RoBERTa model, which works as self-attention layers that output contextualized vector representations of the multilingual dataset, and the CNN is used as the classifier. This model combines the strengths of the attention mechanisms of the pre-trained transformers of XLM-RoBERTa with the outstanding pattern-detection capabilities of CNNs. The CNN model consists of a 2D convolution layer, fed with the hidden states of the XLM-RoBERTa transformers (the embedding layer), followed by the commonly used ReLU activation and 2D max-pooling; the outputs of the CNN are passed through a fully-connected layer and, lastly, a sigmoid activation function. The loss function of this model is the Binary Cross-Entropy loss, given the binary nature of the data.

4.3. Evaluation Method

The evaluation metrics for the performance are recall and F1 score, as they are widely used metrics for classification tasks. F1 ensures balance in the evaluation, and recall evaluates how many hate comments are detected. All reported values correspond to the toxic (positive) class on the test set.

4.4. Experimental Details

The model is evaluated on three different approaches to multilingual fine-tuning, for 1 and 4 epochs, using a learning rate of 1×10^{-4}, a batch size of 16, and a dropout probability of 0.1, following the standard recommended methods of fine-tuning for the BERT model (as XLM-RoBERTa and BETO are both based on BERT). These three approaches to multilingual fine-tuning use datasets that are: (1) *Pure-language* — no translated datasets; (2) *Translated to Spanish* — from French and Turkish to Spanish; and (3) *Translated from Spanish* — from Spanish to Turkish and French.

4.5. Results

The first experiment is a comparison of the performance of XLM-RoBERTa-CNN and BETO-CNN using only Spanish data. From Table I, it is possible to observe that while both models' performance on the F1 score remains stable, the recall is higher for XLM-RoBERTa-CNN.

Table I. XLM-RoBERTa-CNN vs. BETO-CNN on Spanish data

Model	F1	Recall
XLM-RoBERTa + CNN	0.80	0.93
BETO + CNN	0.80	0.77

From Table II, it is noticeable that the F1 score of XLM-RoBERTa-CNN for the fine-tuning of a single epoch shows a clear tendency, in both Spanish and French, to improve the detection of toxic language when translation is applied, while Turkish remains relatively stable for both pure and translated datasets.

Table II. F1 score of XLM-RoBERTa-CNN for the three fine-tuning approaches — 1 epoch

Approach	Spanish	Turkish	French
Pure	0.61	0.88	0.69
Translated to es	0.74	0.83	0.73
Translated from es	0.80	0.86	0.76

In Table III (recall, single epoch), Turkish maintains the same behavior as for the F1-score metric, but both Spanish and French show a notorious difference between a pure-language dataset and the third approach, in which the recall improves significantly to 0.93 and 0.95 respectively.

Table III. Recall of XLM-RoBERTa-CNN for the three fine-tuning approaches — 1 epoch

Approach	Spanish	Turkish	French
Pure	0.47	0.89	0.63
Translated to es	0.66	0.82	0.81
Translated from es	0.93	0.91	0.95

Contrary to what happens with single-epoch fine-tuning, the pure-language datasets show the best performance at four epochs for both the F1 score (Table IV) and the recall (Table V).

Table IV. F1 score of XLM-RoBERTa-CNN for the fine-tuning approaches — 4 epochs

Approach	Spanish	Turkish	French
Pure	0.83	0.87	0.78
Translated from es	0.76	0.79	0.73

The "Translated to es" run at 4 epochs was not available and is therefore omitted from Tables IV–V.

Table V. Recall of XLM-RoBERTa-CNN for the fine-tuning approaches — 4 epochs

Approach	Spanish	Turkish	French
Pure	0.85	0.95	0.95
Translated from es	0.71	0.69	0.71

5. DISCUSSION AND ANALYSIS

The use of a multilingual model improves hate-speech detection, as shown in the experiment comparing XLM-RoBERTa-CNN and BETO-CNN, since recall is an important metric for this task. Spanish, French and Turkish share the same birthplace, but it is noticeable that the similarity of Spanish and French is high, while the Turkish language has become more independent. This helps to explain why Turkish was not considerably affected by any of the three methods, though for the majority of cases the use of pure-language data for Turkish yielded the highest performance for both the F1 score and the recall. On the other hand, French and Spanish were significantly affected by the choice of a pure versus a translated dataset, which corroborates their well-known similarity.

Interestingly, the fine-tuning for a single epoch and for four epochs revealed opposite tendencies, which could occur because of the nature of the data. In the case of the pure-language datasets, more epochs

showed improved performance, possibly due to the fact that the data is more diversified and the model needs more time to learn the words and contexts particular to each language. For the translated datasets, fewer epochs obtained the best performance, while more epochs had a negative effect. This is likely to result from a bias that is produced due to translation: when translating either from French and Turkish to Spanish, or from Spanish to French and Turkish, the learning is faster because it harnesses the similarities of the languages, but the more epochs are used, the more translation bias is introduced.

6. CONCLUSION

NLP models are a promising solution for the fast and efficient detection of hate speech in social media. This work proposed the implementation of XLM-RoBERTa-CNN, which combines the attention mechanisms of the pre-trained multilingual model XLM-RoBERTa with the outstanding pattern-detection capabilities of CNNs, for the detection of toxic comments in Spanish — a low-resource language for this task. XLM-RoBERTa-CNN has a better performance on recall than BETO-CNN, which suggests that, for the particular task of toxic-comment detection, this multilingual model is better suited than the Spanish version of the BERT model. This work also compared and analyzed three different approaches to fine-tune XLM-RoBERTa-CNN for Spanish, Turkish and French, using F1 score and recall for a single epoch and four epochs. The results showed that pure-language datasets can benefit from a longer fine-tuning, while translated datasets can offer a better performance when the fine-tuning is short; and for the particular case of Spanish and French, translating from Spanish can obtain better performance.

REFERENCES

Calizzano, R., Ostendorff, M., & Rehm, G. (2021). *DFKI SLT at GermEval 2021: Multilingual Pre-training and Data Augmentation for the Classification of Toxicity in Social Media Comments*. Papers with Code.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). *Spanish Pre-Trained BERT Model and Evaluation Data (BETO)*. PML4DC at ICLR 2020.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (2019). *CONAN – COunter NArratives through Nichesourcing*. ACL 2019.

Conneau, A., et al. (2020). *Unsupervised Cross-lingual Representation Learning at Scale (XLM-RoBERTa)*. ACL 2020.

Council of Europe (1997, October 20). *Recommendation on hate speech*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL 2019. arXiv:1810.04805.

Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., & Plagianakos, V. P. (2018). *Convolutional Neural Networks for Toxic Comment Classification*. Papers with Code.

Liu, Y., et al. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692.

Luu, S. T., & Nguyen, N. L. (2021). *UIT-ISE-NLP at SemEval-2021 Task 5: Toxic Spans Detection with BiLSTM-CRF and ToxicBERT Comment Classification*. Papers with Code.

Pavlopoulos, J., Sorensen, J., Laugier, L., & Androutsopoulos, I. (2021). *SemEval-2021 Task 5: Toxic Spans Detection*.

Scikit-learn (2020). *sklearn.metrics.f1_score*.