RoBERTuito: a pretrained language model for social media text in Spanish

Juan Manuel Pérez[12] Damián Ariel Furman[12] Laura Alonso Alemany[3] Franco Luque[23]

[1] Universidad de Buenos Aires [2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) [3] Universidad Nacional de Córdoba

Abstract

Since BERT appeared, Transformer language models and transfer learning have become state-of-the-art for Natural Language Understanding tasks. Recently, some works geared towards pre-training specially-crafted models for particular domains, such as scientific papers, medical documents, user-generated texts, among others. These domain-specific models have been shown to improve performance significantly in most tasks. However, for languages other than English such models are not widely available.

In this work, we present RoBERTuito, a pre-trained language model for user-generated text in Spanish, trained on over 500 million tweets. Experiments on a benchmark of tasks involving user-generated text showed that RoBERTuito outperformed other pre-trained language models in Spanish. In addition to this, our model achieves top results for some English-Spanish tasks of the Linguistic Code-Switching Evaluation benchmark (LinCE) and has also competitive performance against monolingual models in English tasks. To facilitate further research, we make RoBERTuito publicly available at the HuggingFace model hub together with the dataset used to pre-train it.

Introduction

- RoBERTuito is a pretrained language model trained on 500M tweets in Spanish. **We publish the data used for training** to facilitate the development of other language models or embeddings, also using subsets of the corpus that model specific subdomains, like regional or thematic variants.
- The **weights of our model are available** through the HuggingFace model hub. Search for pysentimiento/robertuito-* at https://huggingface.co/models
- We set up a benchmark for classification tasks involving user-generated text in Spanish
- We assess the performance of **domain-specific models with respect to general-language models**, showing that the first **outperform** the latter in four classification tasks: **Sentiment Analysis, Emotion Analysis, Irony Detection and Hate Speech**.
- We assess the impact of different preprocessing strategies for our models
- We also evaluate our model in a **code-switching benchmark** for Spanish-English and in a small number of English tasks, both for user-generated text.

Dataset

We downloaded a multilingual database from Archive.org, filtered tweets in Spanish and downloaded all those tweets user's timetlines using Spritzer

- 622M Tweets from 432K users
- Filter tweets with less than 6 tokens. 500M tweets left for training

Available to facilitate the development of other Language Models of Embeddings

Corpus can be fully inspected in case bias is suspected

https://github.com/pysentimiento/robertuito

Collected tweets were not required to be in Spanish (only the ones on the original sample). Language population estimated using fasttext (Joulin et al) is 92% Spanish, 4% English, 3% Portuguese and 1% Others.

Preprocessing

@PerezJotaEme	⇒	@usuario
Faaaaaaa!!!!!!!	⇒	Faaa!!!
#LanguageModels	⇒	<hashtag> Language Models
👍	⇒	emoji pulgar_hacia_arriba emoji

Models

- RoBERTa base architecture with 12 self-attention layers, 12 attention head and hidden size 768
- Three versions: **cased, uncased and deacc** (lowercase and remove accents)
- 4K batch size. To check convergence, we first trained an uncased model for 200K steps. After this, we then proceeded to run it for 600K steps for the three models
- Models were trained for three weeks on a v3-8 TPU and a preemptive e2-standar-16 machine on GCP
- Codebase used Huggingface transformers library.
- Model is uploaded to Huggingface and available to download

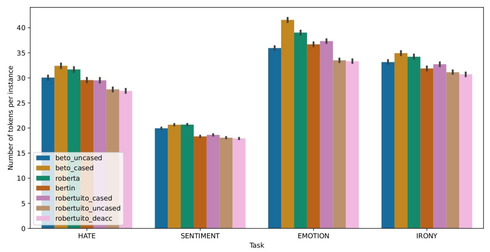
Parameter	Value
#Heads	12
#Layers	12
HiddenSize	768
Intermed Size	3072
Activation	GeLU
Vocab Size	30,000
MLM Prob	0.15
Max Seq	128
Batch Size	4096
Learning Rate	$3.5 * 10^{-4}$
Training Steps	600,000

Evaluation

Language	Tasks	Type of Tasks	Dataset	Num. posts
Spanish	Sentiment Analysis	Text Classification	TASS 2020 Task A	14.500
	Emotion Analysis		TASS 2020 Task B	8.400
	Hate Speech		HatEval	6.600
	Irony Detection		IrosVA 2019	9.000
English	Sentiment Analysis	Text Classification	SemEval 2017 Task 4	61.900
	Emotion Analysis		TASS 2020 Task B	7.303
	Hate Speech		HatEval	13.000
Spanish-English	Sentiment Analysis	Text Classification	LinCE	18.789
	POS Tagging	Text Labelling		42.911
	NER	Text Labelling		67.233

- **Sentiment**: Positive, Negative, Neutral
- **Hate**: Binary classification task for misogyny and racism
- **Emotion**: anger, disgust, fear, joy, sadness
- **Irony**: Binary classification with contextual information

Distribution of the number of tokens per instance: We can observe that RoBERTuito models have more compact representations



SPANISH

Model	Emotion	Hate	Sentiment	Irony
RoBERTuito$_{UNCASED}$	**80.1**	**70.7**	**55.1**	73.6
RoBERTuito$_{DEACC}$	79.8	70.2	54.3	**74.0**
RoBERTuito$_{CASED}$	79.0	70.1	51.9	71.9
RoBERTa	76.6	66.9	53.3	72.3
BERTin	76.7	66.5	51.8	71.6
BETO$_{CASED}$	76.8	66.5	52.1	70.6
BETO$_{UNCASED}$	75.7	64.9	52.1	70.2

Results are expressed as the mean Macro F1 score of 10 runs of the classification experiments. Bold indicates best performing models.

ENGLISH

Model	Hate	Sentiment	Emotion
BERTweet	**55.3**	**70.3**	42.8
RoBERTuito	54.2	68.4	44.1
RoBERTa	45.8	69.5	**46.3**
BERT	48.9	68.9	42.8
mBERT	43.3	66.6	40.4
XLM-RoBERTA$_{BASE}$	45.7	68.0	35.7

Results from the English evaluation setting. Results are the mean Macro F1 of ten runs of the experiments.

CODE-SWITCHING

Model	Sentiment	NER	POS
RoBERTuito	**60.6**	68.5	**97.2**
XLM-R$_{LARGE}$	-	**69.5**	**97.2**
XLM-R$_{BASE}$	-	64.9	97.0
BERT	58.4	61.1	96.9
BETO	56.5	-	-
mBERT*	59.1	64.6	96.9

Results from the code-mixed tasks, taken from the official leaderboard of the LinCE benchmark. Sentiment Analysis is measured by Macro F1 Score, NER with Micro F1 score, and POS Tagging by accuracy. mBERT is a Char2SubWord mBERT, from Aguilar et al (2021

Discussion and conclusions

- This results are in line with other studies that show the effectiveness of Social-Media specific language models
- A limitation on the evaluation is the lack of datasets in Spanish for other tasks rather than text classification
- While uncased and deaccented versions have similar performance, the cased version is behind these two. This can be interpreted in two ways: a strong normalization of the input text in Spanish doesn't yield a significant improvement in the performance and keeping accent marks in the input text is neither beneficial nor harmful
- The data collection process allowing other languages and regional variants made our models develop some multilingual features. Results show that RoBERTuito is suited for code-mixing tasks, obtaining better results than mBERT and similar to XLM-R (although comparison is unfair since XLM-R can handle over one hundred languages)

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, January). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT (1).

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Cañete, Jose and Chaperon, Gabriel and Fuentes, Rodrigo and Ho, Jou-Hui and Kang, Hojin and Pérez, Jorge. (2020). Spanish pre-trained bert model and evaluation data.

Liu, Yinhan and Ott, Myle and Goyal, Naman and Du, Jingfei and Joshi, Mandar and Chen, Danqi and Levy, Omer and Lewis, Mike and Zettlemoyer, Luke and Stoyanov, Veselin. (2019). Roberta: A robustly optimized bert pretraining approach

Conneau, Alexis and Khandelwal, Kartikay and Goyal, Naman and Chaudhary, Vishrav and Wenzek, Guillaume and Guzm´an, Francisco and Grave, Edouard and Ott, Myle and Zettlemoyer, Luke and Stoyanov, Veselin. (2020). Unsupervised Cross-lingual Representation Learning at Scale. Association for Computational Linguistics

Dat Quoc Nguyen and Thanh Vu and Anh Tuan Nguyen. (2020). BERTweet: A pre-trained language model for English Tweets.