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 Spanish. The development of other language models or embeddings, also using subsets of the corpus that model specific subdomains, like regional or thematic variants.

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 through the HuggingFace model hub.

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.

Pre-processing

We download a multilingual database from Archivo.org. We only consider tweets in Spanish and download all those tweets user’s timelines using the @username.

Dataset

622M Tweets from 422K users
Filters tweets with less than 6 tokens.
500M tweets left for training

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.

Evaluation

RoBERTuito 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 c2-standard-16 machine on GCP
Codebase used Huggingface transformers library
Model is uploaded to Huggingface and available to download

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

Discussion and conclusions

This is 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 make 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

Acknowledgements

Juan Manuel Pérez¹ Damián Ariel Furman¹ Laura Alonso Alemany² Franco Luque³

¹ Universidad de Buenos Aires ² Tecnología de la Información y la Comunicación (TINyCO) ³ Universidad de Córdoba

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

Juan Manuel Pérez¹ Damián Ariel Furman¹ Laura Alonso Alemany² Franco Luque³

¹ Universidad de Buenos Aires ² Tecnología de la Información y la Comunicación (TINyCO) ³ Universidad 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 Spanish. The development of other language models or embeddings, also using subsets of the corpus that model specific subdomains, like regional or thematic variants.

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 through the HuggingFace model hub.

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 pysentiment/robertuito on 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 Archivo.org. We only consider tweets in Spanish and download all those tweets user’s timelines using the @username.

622M Tweets from 422K users
Filters tweets with less than 6 tokens.
500M tweets left for training

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.

Evaluation

RoBERTuito 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 c2-standard-16 machine on GCP
Codebase used Huggingface transformers library
Model is uploaded to Huggingface and available to download

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

Discussion and conclusions

This is 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 make 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


