Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7235–7243
Marseille, 20-25 June 2022

RoBERTuito: a pre-trained language model for social media text in Spanish

Juan Manuel Pérez[12], **Damián A. Furman**[12], **Laura Alonso Alemany**[3], **Franco Luque**[23]
[1]Universidad de Buenos Aires, [2]CONICET, [3]Universidad Nacional de Córdoba
{jmperez, dfurman}@dc.uba.ar, {francolq, alemany}@famaf.unc.edu.ar

Abstract

Since BERT appeared, Transformer language models and transfer learning have become state-of-the-art for natural language processing 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 has some cross-lingual abilities, achieving top results for English-Spanish tasks of the Linguistic Code-Switching Evaluation benchmark (LinCE) and also competitive performance against monolingual models in English Twitter tasks. To facilitate further research, we make RoBERTuito publicly available at the HuggingFace model hub together with the dataset used to pre-train it.

Keywords: Pre-trained Language Models, BERT, Spanish

1. Introduction

Pre-trained language models have become a basic building block in the area of natural language processing. In the last years, since the introduction of the transformer architecture (Vaswani et al., 2017), they have been used in many other different natural language understanding tasks, outperforming previous models based on recurrent neural networks. BERT, GPT-2, and RoBERTa are some of the most well-known such tools.

Most models are trained on large-scale corpora taken from news or Wikipedia, which are considered general enough to comprise a large part of the language. Some domains, however, are very specific and have their own vocabulary, jargon, or complex expressions. The medical or scientific domains, for example, use terms and concepts which are not found in a general corpus or occur just a few times. In some other cases, words have specific meanings within a particular domain. Colloquial language –as found in Twitter and other social networks– is more informal, with slang and other expressions which rarely occur in Wikipedia.

For these reasons, a number of pre-trained models have been created to handle these domains. SciBERT (Beltagy et al., 2019) and MedBERT (Rasmy et al., 2021) are examples of domain-specific models. For user-generated text, many models have been trained on Twitter for different languages. However, Spanish lacks pre-trained models for user-generated text, or they are not easily available in the most popular model repositories, such as the HuggingFace model hub. This hinders the development of accurate applications for user-generated text in Spanish.

In this paper, we present RoBERTuito, a large-scale transformer model for user-generated text trained on Spanish tweets. We show that RoBERTuito outperforms other Spanish pre-trained models for a number of classification tasks on Twitter. In addition to this, and due to the collection process of our pre-training data, RoBERTuito is a very competitive model in multilingual and code-switching settings including Spanish and English. Our contributions are the following:

- We publish the data used to train RoBERTuito (around 500M tweets in Spanish), 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.
- We make the weights of our model available through the HuggingFace model hub, thus sparing computation for researchers with no access to the computational power or simply sparing extra computation, while making the model transparent (albeit not interpretable).
- 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 the corresponding domain-specific tasks.
- We assess the impact of preprocessing strategies for our models: cased input vs. uncased input text vs. uncased input text without accents, showing that the uncased version of the corpus yields better performance.

- 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, showing that it achieves competitive results.

2. Previous Work

Language models based on transformers (Vaswani et al., 2017) have become a key component in state-of-the-art NLP tasks, from text classification to natural language generation. One of the most popular transformer-based tools, BERT (Devlin et al., 2019) is a neural bidirectional language model trained on the Masked-Language-Model (MLM) task and in the Next-Sentence-Prediction (NSP) task. This language model can then be fine-tuned for a downstream task or can be used to compute contextualized word representations. RoBERTa (Liu et al., 2019) is an optimized pre-training approach that differs from BERT in four aspects: it trains the model with more data; it removes the NSP objective; it uses longer batches; and it dynamically changes the masking pattern applied to the data. These models, along with GPT (Radford et al., 2018), supposed breakthroughs in the performance on benchmarks such as GLUE (Wang et al., 2018). Nozza et al. (2020) provides a good overview of the BERT-based language models.

After the explosion of language models based on transformers, some models have been trained on corpora that target more specifically a domain of interest instead of generic texts such as Wikipedia or news. For example, SciBERT (Beltagy et al., 2019) is a BERT model trained on scientific texts, and MediBERT (Rasmy et al., 2021) was crafted on medical documents. AlBERTo (Polignano et al., 2019) is one of the first models trained on tweets –particularly, in Italian. BERTweet (Nguyen et al., 2020) is a RoBERTa model trained on about 850M tweets in English, a part of them related to the COVID-19 pandemic.

Multilingual models have also been successful at many tasks comprising more than one language. Multilingual BERT (mBERT) (Devlin et al., 2019) was pre-trained on the concatenation of the top-104 languages from Wikipedia. In a parallel fashion, XLM-R (Conneau et al., 2020) was trained using RoBERTa guidelines on the concatenation of Common Crawl data containing more than 100 languages, obtaining a considerable performance boost over several multilingual tasks while keeping competitive with monolingual models.

BETO (Canete et al., 2020) was the first publicly available pre-trained model in Spanish, following mainly a BERT style of training (MLM + NSP) with some ideas taken from Liu et al. (2019). More recently, some other pre-trained models have been developed for this language, such as RoBERTa-BNE (Gutiérrez-Fandiño et al., 2021), a RoBERTa-based model trained on a database of 500GB of all the *.es* domain websites. BERTin (Rosa et al., 2022) is also a RoBERTA-based model, for whose development the authors explored sampling strategies over the Spanish portion of the *mc4* corpus (Raffel et al., 2020).

To the best of our knowledge, TwilBERT (Gonzalez et al., 2021) is the only specialized pre-trained model on Twitter data for the Spanish language. However, this model has some limitations: first, the training data is not available, making it not auditable. Second, it is not clear how long its pre-training was. Third, the authors used a variant of the NSP task adapted to Twitter (Reply Order Prediction), in spite of many works showing that the type of training based on RoBERTa (MLM only) improves performance on downstream tasks. Finally, the model is not easily available (for instance, in the HuggingFace model hub[1]), which makes its use difficult.

3. Data

In this section, we describe the tweet collection process used to build the corpus to train RoBERTuito. Twitter's free access streaming API (also known as *Spritzer*) provides a sample of around 1% of the overall published tweets, supposedly random, although some studies have shown some concerns about the possible manipulation of this sample (Pfeffer et al., 2018). Unrepresentative, biased samples may produce biased behaviours in the resulting model and systematic, possibly harmful errors in downstream tasks that use this model. That is why we make available the training dataset and the specifics of the model, so that it can be fully inspected in case biases are suspected. In following releases of this tool, an extensive audit of the training corpus will be carried out.

First, we downloaded a Spritzer collection uploaded to Archive.org dating from May 2019 [2]. From this, we only keep tweets with *language* metadata equal to Spanish, and mark the users who posted these messages. Then, the tweetline from each of these marked users was downloaded. We decided to download data from users already represented in the initial collection to facilitate user-based studies in this dataset, and also because we believe the original sample of users to be representative, and thus we hope to maintain this representativeness by sampling from the same users. In total, the collection consists of around 622M tweets from about 432K users.

Finally, we filtered tweets with less than six tokens, because language contained in those is very different from the language in longer tweets. To identify tokens we used the tokenizer trained in BETO (Canete et al., 2020), without counting character repetitions and emojis. This leaves a training corpus of around 500M tweets, which we split in many files to facilitate reading in later processes. The code for the collection process can be found at `https://github.com/finiteautomata/spritzer-tweets`.

[1] https://huggingface.co/models

[2] https://archive.org/details/archiveteam-twitter-stream-2019-05

Something to remark is that this collection process allows the data to contain code-mixed text or even tweets from other languages, as we only required the post on the original sample to be in Spanish. While other works such as Nguyen et al. (2020) required every tweet to be in English, we let other languages to be included in the pre-training data. A rough estimate of the language population using *fasttext*'s language-detection module (Joulin et al., 2016) suggests that 92% of the data is in Spanish, 4% in English, 3% in Portuguese, and the rest in other languages.

4. RoBERTuito

In this section, we describe the training process of RoBERTuito. Three versions of RoBERTuito were trained: a *cased* version which preserves the case found in the original tweets, an *uncased* version, and a *deacc* version, which lower-cases and removes accents on tweets. Normative Spanish prescribes marks for (some) accents in words, but their usage is inconsistent in user-generated text, so we want to test if removing them improves the performance on downstream tasks.

For each of the three configurations, we trained tokenizers using *SentencePiece* (Kudo and Richardson, 2018) on the collected tweets, limiting vocabularies to 30,000 tokens. We used the *tokenizers* library (Moi et al., 2019) which provides fast implementations in Rust for many tokenization algorithms.

4.1. Preprocessing

Preprocessing is key for models consuming Twitter data, which is quite noisy, have user handles (@username), hashtags, emojis, misspellings, and other non-canonical text. Nguyen et al. (2020) tried two normalization strategies: a soft one, in which only minor changes are performed to the tweet such as replacing usernames and hashtags, and a more aggressive one using the ideas of Han and Baldwin (2011). The authors found no significant improvement by using the harder normalization strategy. Having this in mind, we followed an approach similar to the one used both in this work and in Polignano et al. (2019):

- Character repetitions are limited to a max of three
- User handles are converted to a special token `@usuario`
- Hashtags are replaced by a special token `hashtag` followed by the hashtag text and split into words if this is possible
- Emojis are replaced by their text representation using *emoji* library[3], surrounded by a special token `emoji`.

4.2. Architecture and training

A RoBERTa base architecture was used in RoBERTuito, with 12 self-attention layers, 12 attention heads, and hidden size equal to 768, in the same fashion as BERTweet (Nguyen et al., 2020). We used a masked-language objective, disregarding the next-sentence prediction task used in BERT or other tweet-order tasks such as those used in Gonzalez et al. (2021).

Taking into account successful hyperparameters from RoBERTa and BERTweet, we decided to use a large batch size for our training. While an 8K batch size is recommended in RoBERTa, due to resource limitations, we decided to balance the number of updates using a 4K 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. This is roughly half the number of updates used to train BETO (and also BERTweet) but this difference is compensated by the larger batch size used to train RoBERTuito.

We trained our models for about three weeks on a v3-8 TPU and a preemptible *e2-standard-16* machine on GCP. Our codebase uses *HuggingFace's transformers* library and their RoBERTa implementation (Wolf et al., 2020). Each sentence is tokenized and masked dynamically with a probability equal to 0.15. Further details on hyperparameters and training can be found in the Appendix 11.

5. Evaluation

We evaluated RoBERTuito in two monolingual settings (Spanish and English) and also in a code-mixed benchmark for tweets containing both Spanish and English. As the collection process allowed non-Spanish tweets to be included, we assess not only the performance of our model in Spanish, but also in other environments. Table 1 summarizes the evaluation tasks.

5.1. Spanish evaluation

For the **Spanish** evaluation, we set a benchmark for this model following TwilBERT (Gonzalez et al., 2021), AlBERTo (Polignano et al., 2019) and BERTweet (Nguyen et al., 2020). Four classification tasks for Twitter data in Spanish were selected, three of them coming from the Iberian Languages Evaluation Forum (IberLEF) and one from SemEval 2019: **sentiment analysis**, **emotion analysis**, **irony detection** and **hate speech detection**. These tasks are particularly relevant in the context of social media analysis and provide a broad perspective about how our model is able to capture useful information on this specific domain.

Regarding sentiment analysis, we relied on the TASS 2020 (García-Vega et al., 2020) dataset. This dataset uses a three-class polarity model (negative, neutral, positive) and is separated according to different geographic variants of Spanish. For the purpose of benchmarking our model, we disregarded these distinctions

[3] `https://github.com/carpedm20/emoji/`

Language	Tasks	Type of task	Dataset	Num. posts
Spanish	Sentiment analysis	Text Classification	TASS 2020 Task A	$14,500$
	Emotion analysis		TASS 2020 Task B	$8,400$
	Hate speech detection		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 detection		HatEval	$13,000$
Spanish-English	Sentiment analysis	Text Classification	LinCE	$18,789$
	POS tagging	Text Labelling		$42,911$
	NER	Text Labelling		$67,233$

Table 1: Evaluation tasks for RoBERTuito. Tasks are grouped by setting: Spanish-only tasks, English-only tasks, and code-mixed Spanish-English tasks. Num. posts is the number of instances contained in the dataset of each task.

and merged all the data into a single dataset, with respective train, dev, and test splits.

For emotion analysis, we also used the dataset from the TASS 2020 workshop, *EmoEvent* (Plaza del Arco et al., 2020). This is a multilingual emotion dataset labeled with the six Ekman's basic emotions (*anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*) (Ekman, 1992) plus a neutral emotion. It was built retrieving tweets associated with eight different global events from different domains (political, entertainment, catastrophes or incidents, global commemorations, etc.), so emotions are always related to a particular phenomenon. We only kept the Spanish portion of it, containing 8,409 tweets.

Hate speech detection in social media has gained much interest in the past years, based on the need to act against the spread of hate messages that develops in parallel with an increasing amount of user-generated content. It is a difficult task that requires a deep, contextualized understanding of the semantic meaning of a tweet as a whole. For this reason, we selected the *hatEval* Task A dataset (Basile et al., 2019), which is a binary classification task for misogyny and racism, to benchmark our model. The authors collected the dataset by three combined strategies: monitoring potential victims of hate accounts, downloading the history of identified producers of hateful content, and filtering Twitter streams with keywords. This dataset distinguishes between hate speech targeted to individuals and generic hate speech, and between aggressive and non-aggressive messages. For this work, we do not consider these classifications, and we are interested in predicting only the binary label of whether the tweet is hateful or not. The Spanish subset of this dataset comprises 6,600 instances.

Irony detection has also recently gained popularity. Many works show that it has important implications in other natural language processing tasks that require semantic processing. Gupta and Yang (2017) showed that using features derived from sarcasm detection improves the performance on sentiment analysis. In addition to this, user-generated content is a rich and vast source of irony, so being able to detect it is of particular importance for the domain of social networks. IroSVa (Ortega-Bueno et al., 2019) is a recent dataset published in 2019, that has the particularity of considering the messages not as isolated texts but with a given context (a headline or a topic). It consists of 7,200 train and 1,800 test examples divided into three geographic variants from Cuba, Spain, and Mexico, each with a binary label indicating if the comment contains irony or not. Unlike the previous three tasks mentioned here, this dataset contains not only messages from Twitter but also news comments and debate forums as 4forums.com and Reddit.

We compare RoBERTuito performance for these tasks with other Spanish pre-trained models: BETO (Canete et al., 2020), RoBERTa-BNE (Gutiérrez-Fandiño et al., 2021) and BERTin (Rosa et al., 2022). All these models share a base architecture with a similar number of parameters to our model.

5.2. English evaluation

As for **English**, we tested RoBERTuito in three tasks: emotion analysis, hate speech detection, and sentiment analysis. For emotion analysis and hate speech we used the English sections from the aforementioned datasets (*EmoEvent* and *HatEval*), while for sentiment analysis *SemEval 2017 Task-4* dataset (Rosenthal et al., 2017) was used, which shares the same labels as its Spanish counterpart (negative, neutral, positive).

In this case, we compare RoBERTuito abilities in English with monolingual models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and BERTweet (Nguyen et al., 2020); and also against multilingual models such as XLM-R$_{BASE}$ (Conneau et al., 2020) and mBERT. While all these models share a base architecture, the different vocabulary sizes and number of parameters (see Appendix 11) make the comparison a bit more difficult.

5.3. Code-switching evaluation

Finally, we assessed the code-switching abilities of our model in the *Linguistic Code-Switching Evalua-*

tion Benchmark (LinCE) (Aguilar et al., 2020). LinCE comprises five tasks for code-switched data in several language pairs (Spanish-English, Hindi-English, Modern Standard Arabic-Egyptian Arabic, Arabic-English, among others), many of which were part of previous shared tasks. We evaluated RoBERTuito on three different tasks of the benchmark: part of speech (POS) tagging (AlGhamdi et al., 2016), named entity recognition (NER), and sentiment analysis (Patwa et al., 2020). As the collection process of the data centered on Spanish-speaking users, some of which also speak English and Spanglish [4], we test RoBERTuito on the Spanish-English subsection of the benchmark.

This benchmark has a centralized evaluation system, not releasing gold labels for the test split of the tasks. We evaluated our models in the dev datasets and compared our results against the ones provided by Winata et al. (2021), which achieves the best performance for the NER and POS tagging. As competing models to RoBERTuito for the Spanish-English evaluation, we have mBERT, XLM-R (both in base and large architecture) and monolingual models BERT and BETO.

5.4. Model fine-tuning

We followed fairly standard practices for fine-tuning, most of which are described in Devlin et al. (2019). For sentence classification tasks, we fine-tuned the pretrained models for 5 epochs with a triangular learning rate of 510^{-5} and a warm-up of 10% of the training steps. The best checkpoint at the end of each epoch was selected based on each task metric.

For sentence classification tasks, a linear classifier is put on top of the representation of the `[CLS]` token. For token classification tasks (NER and POS tagging), we predicted the word tag by putting a linear classifier on top of the first corresponding sub-word representation.

6. Results

Table 2 displays the results for the evaluation for the four proposed classification tasks in Spanish. Figures are expressed as the mean of 10 runs of the experiments, along with a **score** averaging the metrics for each task in a similar way as the GLUE score. We can observe that, in most cases, all RoBERTuito configurations perform above other models, in particular for hate speech detection and sentiment analysis. For most tasks, no big differences are observed between *uncased* and *deacc* models, but both perform consistently above the cased model.

Table 3 displays the results for the evaluation of the selected models for the three tasks in English. We can observe that RoBERTuito outperforms both mBERT and XLM-R, which are the other multilingual models evaluated for the tasks. Compared to monolingual English models, the results of RoBERTuito are similars to those of RoBERTa and slightly above BERT. As expected, BERTweet obtains the best results.

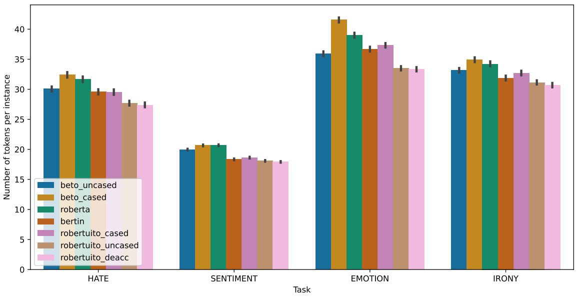

Figure 1: Distribution of the number of tokens per instance. Bars are grouped by task, and display the mean number of tokens per instance with their 95% confidence interval. Less is better

As for the code-switching evaluation, Table 4 displays the results on the dev dataset of LinCE for NER, POS Tagging, and sentiment analysis. We compare RoBERTuito against other multilingual models such as mBERT and XLM-R$_{BASE}$, and also listing the dev results reported by (Winata et al., 2021). The results for XLM-R and mBERT are consistent with the results in that work. A minor improvement is observed but this could an artifact of different choices of hyperparameters or even a slightly different preprocessing.

Finally, Table 5 displays the results from the leaderboard of the LinCE benchmark[5] for the three selected tasks: Spanish-English sentiment analysis, NER and POS tagging. For sentiment analysis, it obtains the best results in terms of Micro F1. For the other two tasks, it obtains the second position, for which an XLM-R$_{LARGE}$ model (Winata et al., 2021) has the top results. Among the compared models, RoBERTuito has 108 million parameters, while XLM-R$_{LARGE}$ sums up to around five times this number, making our model the most efficient in terms of size for this subsection of the benchmark (see Appendix 11 for further details on the size of the different models).

6.1. Vocabulary efficiency

Figure 1 shows the distribution of the number of tokens in the input text for the Spanish tasks. We can observe that RoBERTuito models have more compact representations than BETO and *RoBERTa-BNE* for this domain, and, between them, the *deacc* version has a slightly lower mean-length compared to the *uncased* version. RoBERTuito has also shorter representations for the code-mixed datasets, but longer in the case of the English tasks. Appendix 11 lists the complete figures for the three evaluation settings.

[4] The morphosyntactic mixture of Spanish and English

[5] `https://ritual.uh.edu/lince/leaderboard`

Model	Hate	Sentiment	Emotion	Irony	Score
RoBERTuito$_{uncased}$	**80.1**	**70.7**	**55.1**	73.6	**69.9**
RoBERTuito$_{deacc}$	79.8	70.2	54.3	**74.0**	69.6
RoBERTuito$_{cased}$	79.0	70.1	51.9	71.9	68.2
RoBERTa	76.6	66.9	53.3	72.3	67.3
BERTin	76.7	66.5	51.8	71.6	66.7
BETO$_{cased}$	76.8	66.5	52.1	70.6	66.5
BETO$_{uncased}$	75.7	64.9	52.1	70.2	65.7

Table 2: Evaluation results for Spanish classification tasks: hate speech detection, sentiment analysis, emotion analysis and irony detection. Results are expressed as the mean Macro F1 score of 10 runs of the classification experiments. Bold indicates best performing models

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-R$_{BASE}$*	45.7	68.0	35.7

Table 3: Evaluation results for the three English classification tasks. Results are expressed as the mean Macro F1 score of 10 runs of the classification experiments. * marks multilingual models

Model	Sentiment	NER	POS
RoBERTuito$_{uncased}$	**53.2**	67.2	97.0
RoBERTuito$_{deacc}$	52.7	**67.4**	96.8
RoBERTuito$_{cased}$	50.1	66.3	**97.3**
mBERT	51.3	64.8	96.7
XLM-R$_{BASE}$	48.8	63.7	**97.3**
mBERT†	–	63.7	**97.3**
XLM-R$_{BASE}$†	–	62.8	97.1

Table 4: Development results for the code-mixed tasks from the LinCE dataset. Sentiment Analysis task performance is measured with Macro F1, Named Entity Recognition with Micro F1, and Part-of-Speech through accuracy. $_U$, $_C$ and $_D$Each figure is the mean of 5 independent runs of the classification experiments. Results marked with † are reported in Winata et al. (2021)

7. Discussion

For this small set of Spanish tasks in the social-media domain, the results show that RoBERTuito outperforms other pre-trained models for Spanish, namely BETO, RoBERTa and BERTin. This result is in line with other works which show the effectiveness of social-media-specific language models. A limitation in the evaluation of our model for Spanish tasks is the lack of datasets for other tasks rather than text classification. As far as we know, there are no Spanish-only datasets for sequence labeling (NER and POS tagging) in the social domain.

Among the three proposed variants of RoBERTuito, the cased version is behind the others in terms of performance for most tasks –with the exception of POS tagging and NER– while the other two (uncased and uncased and deaccented) have comparable performances across all the tasks. We can read this in two ways: one, that a stronger normalization of the input text in Spanish results in no significant improvement in the performance of the model, and two, that keeping accent marks in the input text is neither beneficial nor harmful for the performance of the model.

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
C2S mBERT	59.1	64.6	96.9
mBERT	56.4	64.0	97.1
BERT	58.4	61.1	96.9
BETO	56.5	–	–

Table 5: Test results for the code-mixed tasks from Spanish-English section of the LinCE benchmark. Results are taken from the official leaderboard of this benchmark. Sentiment Analysis performance is measured by Macro F1 score, Named Entity Recognition (NER) with Micro F1 score, and Part-of-Speech (POS) with accuracy. C2S is an acronym for Char2Subword BERT, presented in (Aguilar et al., 2021)

The reasons for the differences in performance for the uncased models need further investigation. We have some working hypotheses for these differences. First, we believe that accents and non-ASCII characters –with the exception of emojis– are used in a much more inconsistent way in user-generated text than in more normative text. Therefore, no regularities can be inferred from the data concerning those marks. Second, a bigger amount of data is required to account for the possible differences in meaning for upper case or lower case forms or the lack of difference between them. Future experiments will delve into those two.

Our data collection process for the pre-training stage was centered in Spanish but it allowed other languages and other regional variants to be part of our dataset as well. This point made our model develop some multilingual features, in particular in the code-switching LinCE benchmark. The results for this benchmark highlights that RoBERTuito is suited for Spanish-English code-mixing tasks, obtaining better results than mBERT and matching those of XLM-R. This comparison, however, is not completely fair because XLM-R and RoBERTa can handle over one hundred languages but this is not the case for our model. Lastly, the results for the English tasks show that RoBERTuito keeps competitive against monolingual models for the social domain.

8. Conclusion

In this work, we presented RoBERTuito, a large-scale model trained on user-generated tweets. We set up a benchmark of classification tasks in social-media text for Spanish, and we showed that RoBERTuito outperforms other available general domain pre-trained language models. Moreover, our model features good code-switching performance in Spanish-English tasks and is competitive against monolingual English models in the social domain.

We proposed three versions of this new model: cased, uncased, and deaccented. We observed that the uncased model performs slightly better than the cased one, and similarly to the deaccented version. Further research is needed to systematize the reasons behind these results. We have made our pretrained language models public through the *HuggingFace* model hub, and our code can be found at GitHub [6]. We will also make the training corpus available, thus facilitating the development of other models for user-generated Spanish, like word embeddings or other language models. It is even feasible to extract subsets of the corpus representing subdomains of interest, like regional variants of Spanish or specific topics, to develop even more specific models.

Future work includes enhancing our benchmark to include assessment of the performance of such models in open-ended tasks, and experiments for specific subdomains of Spanish.

9. Acknowledgements

The authors would like to thank the Google TPU Research Cloud program[7] for providing free access to TPUs for our experiments. We also thank Juan Carlos Giudici for his help in setting up the GCP environment, and Huggingface's team for their support in the development of RoBERTuito.

10. Bibliographical References

Aguilar, G., McCann, B., Niu, T., Rajani, N., Keskar, N. S., and Solorio, T. (2021). Char2Subword: Extending the subword embedding space using robust character compositionality. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1640–1651, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., and Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

García-Vega, M., Díaz-Galiano, M., García-Cumbreras, M., Plaza-Del-Arco, F., Montejo-Ráez, A., Zafra, S. M., Martínez-Cámara, E., Aguilar, C., Antonio, M., Cabezudo, S., Chiruzzo, L., and Moctezuma, D. (2020). Overview of tass 2020: Introducing emotion detection. 09.

Gupta, R. K. and Yang, Y. (2017). CrystalNest at SemEval-2017 task 4: Using sarcasm detection for enhancing sentiment classification and quantification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 626–633, Vancouver, Canada, August. Association for Computational Linguistics.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., and Villegas, M. (2021). Spanish language models.

Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 368–378.

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

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Moi, A., Cistac, P., Patry, N., Walsh, E. P., Morgan, F., Pütz, S., Wolf, T., Gugger, S., Delangue, C., Chaumond, J., Debut, L., and von Platen, P. (2019). Hugging face tokenizers library. `https://github.com/huggingface/tokenizers`.

Nozza, D., Bianchi, F., and Hovy, D. (2020). What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.

Ortega-Bueno, R., Rangel, F., Hernández Farıas, D., Rosso, P., Montes-y Gómez, M., and Medina Pagola, J. E. (2019). Overview of the task on irony detection in spanish variants. In *Proceedings of the Iberian*

[6] `https://github.com/pysentimiento/robertuito`

[7] `https://sites.research.google/trc/`

languages evaluation forum (IberLEF 2019), co-located with 34th conference of the Spanish Society for natural language processing (SEPLN 2019). CEUR-WS. org, volume 2421, pages 229–256.

Pfeffer, J., Mayer, K., and Morstatter, F. (2018). Tampering with twitter's sample api. *EPJ Data Science*, 7(1):50.

Plaza del Arco, F. M., Strapparava, C., Urena Lopez, L. A., and Martin, M. (2020). EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France, May. European Language Resources Association.

Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., and Basile, V. (2019). Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Rosenthal, S., Farra, N., and Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Winata, G. I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., and Fung, P. (2021). Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online, June. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

11. Language Resource References

Aguilar, Gustavo and Kar, Sudipta and Solorio, Thamar. (2020). *LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation*. European Language Resources Association.

AlGhamdi, Fahad and Molina, Giovanni and Diab, Mona and Solorio, Thamar and Hawwari, Abdelati and Soto, Victor and Hirschberg, Julia. (2016). *Part of Speech Tagging for Code Switched Data*. Association for Computational Linguistics.

Beltagy, Iz and Lo, Kyle and Cohan, Arman. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. Association for Computational Linguistics.

Canete, José 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*.

Conneau, Alexis and Khandelwal, Kartikay and Goyal, Naman and Chaudhary, Vishrav and Wenzek, Guillaume and Guzmán, 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.

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Association for Computational Linguistics.

Gonzalez, Jose Angel and Hurtado, Lluís-F and Pla, Ferran. (2021). *TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter*. Elsevier.

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*.

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

Patwa, Parth and Aguilar, Gustavo and Kar, Sudipta and Pandey, Suraj and PYKL, Srinivas and Gamb"ack, Bj"orn and Chakraborty, Tanmoy and Solorio, Thamar and Das, Amitava. (2020). *SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets*. Association for Computational Linguistics.

Hyperparameter	Value
#Heads	12
#Layers	12
Hidden Size	768
Intermediate Size	3072
Hidden activation	GeLU
Vocab. size	$30,000$
MLM probability	0.15
Max Seq length	128
Batch Size	$4,096$
Learning Rate	$3.5 * 10^{-4}$
Decay	0.1
β_1	0.9
β_2	0.98
ϵ	10^{-6}
Warmup steps	36,000 (6%)
Training steps	$600,000$

Table 6: Hyperparameters of the RoBERTuito model pre-training.

Model	Train loss	Eval loss	Eval ppl
Cased	1.864	1.753	5.772
Uncased	1.940	1.834	6.259
Deacc	1.951	1.826	6.209

Table 7: Training results for each of the three configurations of RoBERTuito, expressed in cross-entropy loss for the Masked-Language-Modeling task and perplexity (ppl)

Radford, Alec and Narasimhan, Karthik and Salimans, Tim and Sutskever, Ilya. (2018). *Improving language understanding by generative pre-training*.

Javier De La Rosa and Eduardo G. Ponferrada and Manu Romero and Paulo Villegas and Pablo González de Prado Salas and María Grandury. (2022). *BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling*.

Appendix

Hyperparameters and training

Table 6 displays the hyperparameters of the RoBERTuito training for its three versions. For the first prototype of the model, a larger learning rate was tried but it usually diverged near its peak value, so we decided to lower it for the definitive training. Table 7 displays the results of the training in terms of cross-entropy loss and perplexity for the three versions of RoBERTuito.

Model comparison

A comparison of the models in terms of the number of parameters and vocabulary sizes is presented in Table 8. RoBERTuito is the smallest model in terms of vocabulary size, and as most models share a base architecture (see Table 6), this accounts for the difference in the number of parameters.

Model	Language	Size	Vocabulary
RoBERTuito	es	108M	30K
BETO		108M	30K
BERTin		124M	50K
RoBERTa-BNE		124M	30K
BERT	en	109M	30K
RoBERTa		124M	50K
BERTweet		134M	64K
mBERT	multi	177M	105K
C2S mBERT		136M	–
XLM-R$_{BASE}$		278M	250K
XLM-R$_{LARGE}$		565M	250K

Table 8: Comparison in terms of model size (measured in number of parameters) and vocabulary sizes for the considered models in this work. C2S mBERT is an abbreviation for Char2subword mBERT (Aguilar et al., 2021)

Spanish

Model	EMOTION	HATE	SENTIMENT	IRONY
RoBERTuito$_D$	**33.31**	**27.39**	**17.92**	**30.70**
RoBERTuito$_U$	33.50	27.70	18.08	31.12
RoBERTuito$_C$	37.33	29.51	18.64	32.71
BETO	35.94	30.06	19.95	33.14
BERTin	36.68	29.56	18.33	31.87
RoBERTa	39.02	31.67	20.68	34.21

English

Model	EMOTION	HATE	SENTIMENT
BERTweet	**33.22**	**29.24**	**25.70**
BERT	35.20	32.05	26.93
RoBERTa	37.14	31.29	27.00
RoBERTuito$_D$	38.79	36.20	28.82
RoBERTuito$_U$	39.00	36.29	29.22
RoBERTuito$_C$	44.27	39.47	31.49

Spanish-English Code-switching

Model	NER	POS	SENTIMENT
RoBERTuito$_D$	**14.27**	**8.8**	**20.65**
RoBERTuito$_U$	14.34	8.84	20.75
RoBERTuito$_C$	15.41	9.01	22.14
BETO	16.99	10.78	24.16
BERT	20.88	10.32	29.87
BERTweet	18.47	9.55	26.29
mBERT	16.47	9.63	23.82
XLM-R$_{BASE}$	18.23	9.92	25.36

Table 9: Distribution of sentence length measured by number of tokens for each evaluation setting. D, U, and C correspond to deaccented, uncased and cased versions. Results are displayed as the mean of the number of tokens per each sentence. Bold marks best results (less is better)

Vocabulary efficiency

Table 9 displays the mean sentence length for each considered model and group of tasks. For the Spanish and the code-mixed Spanish-English benchmark, RoBERTuito achieves the more compact representations in mean length in their uncased and deaccented forms. In the case of English, BERTweet achieves the shortest representations, with RoBERTuito having longer sequences of tokens than its monolingual counterparts.