Identification of profession & occupation in Health-related Social Media using tweets in Spanish

Victoria Pachón Álvarez
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
vpachon@uhu.es

Jacinto Mata Vázquez
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
mata@uhu.es

Juan Luis Domínguez Olmedo
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
juan.dominguez@dti.uhu.es

Abstract

In this paper we present our approach and system description on Task 7a in ProfNer-ST: Identification of profession & occupation in Health related Social Media. Our main contribution is to show the effectiveness of using BETO-Spanish BERT for classification tasks in Spanish. In our experiments we compared several architectures based on transformers with others based on classical machine learning algorithms. With this approach, we achieved an F1-score of 0.92 for the positive class in the evaluation process.

1. Introduction

The battle against COVID-19 is present in practically every country in the world. Confinement, curfews and restrictions on the movement of personnel and cargo, are part of the strategy to stop the transmission of the virus. Some workers are at the forefront of the battle against the COVID-19 pandemic, and they are more exposed to the virus and also more likely to suffer from mental health problems because of the stress caused by the pandemic. The detection of vulnerable occupations is essential to prepare preventive measures. In ProfNer-ST: Task 7, Track A (Tweet binary classification) (Miranda-Escalada et al. 2021) participants must determine whether a tweet contains a mention of occupation, or not.

Despite Spanish being the 4th most spoken language, finding resources to train and evaluate for Spanish text is not an easy task. We hypothesized that automatically translating a text from Spanish to English to use and model based on this language would not be as good as working straight away with a model pre trained with a Spanish corpus. The idea behind all our experiments was to compare models pre-trained in Spanish with models pretrained in English and using automatic translations. In this context of work we have been heavily using BETO-Spanish BERT (Cañete et al. 2020) , BERT-Multilingual (Devlin et al. 2018) and RuPERTa: the Spanish RoBERTa (GitHub - mrm8488/RuPERTa-base: Spanish RoBERTa), and we compared them with the results obtained by BERT (Devlin et al. 2018) using the official translation of the given datasets in English.

2. Data Description and preprocessing

The corpus provided to perform Task 7a (classification) is described in (Magge et al. 2021). Since tweets have a very specific language, for this task we have not performed a very exhaustive data preprocessing. The only text processing performed was to convert all characters to lowercase. In order to carry out different experiments to evaluate the performance of our systems, we have used 3 files:

- **Original**. The original text of the tweets was preserved.
- **URLs_removed.** The URLs of the tweets were removed.
- **Hashtags_URLs_removed.** Both URLs and hashtags were removed from the tweets.

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 105–107

3. Methods

Our methodology is based on working directly with texts in Spanish and applying multilingual or Spanish pre-trained models, instead of translated and using pre-trained English models. The system consists of fine-tuning a BETO model for classification tasks. Before starting to study the performance of our systems, we design a baseline and use its results as a starting point to improve our approaches. We trained a Bidirectional Long-Short Term Memory RNN with one dense layer and the *skipgram uncased* Spanish COVID-19 Twitter Embeddings (Miranda-Escalada et al. 2021b) for the word embedding layer. BERT variants gave good results in #SMM4H 2020 (Klein et al. 2020) so we decided to focus on them to develop our proposal. We have carried out several experiments to compare the results of BETO with multilingual Bert (mBert) and RuPERTa as well as English pretrained BERT (cased and uncased). For all the experiments, we used the training and test dataset supplied by the organizers. The training dataset was split up in two parts to get a validation dataset (30%). We used a batch size of 32 instances, and we trained with 4 epochs and max length of 256. In our experiments with the BERT English pretrained model we have used the translation to English provided by the organizers. In all experiments with uncased models, we have transformed each tweet to lowercased. BERT (Bidirectional Encoder Representations for Transformers) also offers a multilingual model (mBERT) pretrained on concatenated Wikipedia data for languages without any cross-lingual alignment. BETO (Spanish Pre-Trained BERT Model and Evaluation Data) is a model similar in size to a BERT-Base model with 12 self-attention layers, 16 attention-heads each (Vaswani et al. 2017) and 1024 as hidden size. The total size of the corpora gathered was comparable with the corpora used in the original BERT. RuPERTa-base (uncased) is a RoBERTa model trained on an uncased version of big Spanish corpus and its architecture is the same as Roberta-base (Liu et al. 2019).

4. Experiments and Results

We fine-tuned mBert, BETO, RuPERTa and BERT with the training dataset provided by the organizers and we test the model obtained using the test dataset described before. The measure was F1-score for the positive class, according to the one used for the ranking of the systems in the competition. Table 1 shows a summarization of the experimental results obtained. BETO obtained the best results for all the measures. We fine-tuned BETO-cased using all the tweet from the training and test datasets with 5 epoch and we made predictions on the unseen evaluation examples as our first and only submission. We achieved an F1-score of 0.92 in the evaluation process.

	Original Dataset			**URL_removed Dataset**			**Hashtags_URLs_removed Dataset**		
	F1-score (class 1)	**macro-F1**	**AUC**	**F1-score (class 1)**	**macro-F1**	**AUC**	**F1-score (class 1)**	**macro-F1**	**AUC**
baseline	0.77	0.86	0.830	0.75	0.84	0.816	0.78	0.86	0.844
mBert -uncased-	0.88	0.92	0.915	0.88	0.92	0.918	0.87	0.92	0.910
mBert -cased-	0.88	0.92	0.919	0.88	0.92	0.915	0.87	0.91	0.911
BETO -uncased-	0.91	0.94	0.934	0.90	0.93	0.930	0.89	0.93	0.918
BETO -cased- (our propousal)	**0.91**	**0.94**	**0.939**	0.90	0.94	0.936	0.89	0.93	0.923
RuPERTa-spanish	0.75	0.83	0.834	0.76	0.84	0.844	0.76	0.84	0.848
BERT -uncased-	0.88	0.92	0.915	0.88	0.92	0.919	0.88	0.92	0.909
BERT -cased-	0.88	0.92	0.916	0.87	0.91	0.904	0.87	0.91	0.904

Table 1. Results on test dataset

5. Conclusions

In this paper we present our approach and system description on Task 7a in ProfNer-ST: Identification of profession & occupation in Health related Social Media. The main idea was checking the use of models trained with a Spanish corpus. Our model was based on fine tuning a pretrained model in Spanish: BETO for classification tasks. In our experiments we also tested and compared several architectures based on transformers with others based on classical machine learning algorithms. In the future we want to keep testing BETO in other contests. With this approach, we achieved an F1-score of 0.92 in the evaluation process for class "1". In this way, we proved the accuracy and usability of pretrained models with a Spanish Corpus.

References

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. *Spanish Pre-Trained BERT Model and Evaluation Data*. PML4DC at ICLR 2020

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. DOI: 10.18653/v1/N19-1423

Ari Klein, Ilseyar Alimova, Ivan Flores, et al. 2020. *Overview of the Fifth Social Media Mining for Health Applications (\#SMM4H) Shared Tasks at COLING 2020.* Proc. of the Fifth Social Media Mining for Health Applications Workshop & Shared Task

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, et al. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arxiv.org/abs/1907.11692

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, et al. 2021. *Overview of the Sixth Social Media Mining for Health Applications (\#SMM4H) Shared Tasks at NAACL 2021*

Antonio Miranda-Escalada, Eulália Farré-Maduell, Salvador Lima, Vicent Briva-Iglesias, et al. 2021. *The ProfNER Shared Task on Automatic Recognition of Professions and Occupation Mentions in Social Media: Systems, Evaluation, Guidelines, Embeddings and Corpora*

Antonio Miranda-Escalada, Marvin Aguero, and Martin Krallinger,. 2021. *Spanish COVID-19 Twitter Embeddings in FastText*. DOI: http://doi.org/10.5281/zenodo.4449930

Jörg Tiedemann. 2012. *Parallel Data, Tools and Interfaces in OPUS*. Proc. of the Eighth International Conference on Language Resources and Evaluation (LREC'12)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, et al. 2017. *Attention is all You Need*. arXiv:1706.03762v5