PLN CMM at SocialDisNER: Improving Detection of Disease Mentions in Tweets by Using Document-Level Features

Matías Rojas[1], Jose Barros[1], Kinan Martin[3], Mauricio Araneda[2], and Jocelyn Dunstan[1,4]
[1]Center for Mathematical Modeling (CMM), University of Chile.
[2]National Center for Artificial Intelligence (CENIA), Chile.
[3]Department of Computer Sciences, Massachusetts Institute of Technology, USA.
[4]Initiative for Data & Artificial Intelligence, University of Chile.
{matias.rojas.g, jose.barros.s, mauricio.araneda}@ug.chile.cl
krmkrm@mit.edu, jdunstan@uchile.cl

Abstract

This paper describes our approaches used to solve the SocialDisNER task, which belongs to the Social Media Mining for Health Applications (SMM4H) shared task. This task aims to identify disease mentions in tweets written in Spanish. The proposed model is an architecture based on the FLERT approach. It consists of fine-tuning a language model that creates an input representation of a sentence based on its neighboring sentences, thus obtaining the document-level context. The best result was obtained using an ensemble of six language models using the FLERT approach. The system achieved an F_1 score of 0.862, significantly surpassing the average performance among competitor models of 0.680 on the test partition.

1 Introduction

Understanding the social perception of diseases is crucial for quantifying the effectiveness of public policies on medicine. In particular, analyzing social network data allows us to study this issue in depth. In this context, the SocialDisNER shared task (Gasco et al., 2022) aims to better understand diseases' social perception by using Named Entity Recognition (NER) to identify diseases mentioned in Twitter data: from highly prevalent conditions such as cancer and diabetes to rare immunological and genetic diseases. In this paper, we discuss our approaches used to solve the SocialDisNER task. Specifically, we use several language models trained on both general and domain-specific corpora, exploiting the use of document-level features. Finally, we study the use of ensembles and their impact on performance compared to individual models.

2 Task and Data Description

2.1 Task

Task 10 of SMM4H aims to identify disease mentions in tweets written in Spanish. The identification of entity mentions must be strict, meaning that an entity is considered correct when the system simultaneously identifies both entity types and boundaries.

2.2 Data

Regarding the statistics of the dataset, $10,000$ tweets were released: 50% for training, 25% for validation, and 25% for testing. There are $15,173$ disease mentions in the training partition, while $4,252$ for validation. There were also nested entities within the annotations, which we simplified using the common strategy of keeping only the outermost entity in a nesting, as pointed out in Báez et al. (2022).

3 Methodology

This section describes the models used in our experiments and the chosen baseline.

3.1 Baseline

The baseline model consists of fine-tuning the *xlm-roberta-large-ner-hrl* model for the token classification task. We decided on using this baseline due to its previous NER-specific training on high-resource multilingual corpora, including the Spanish CoNLL 2002 corpus (Tjong Kim Sang, 2002). Unlike our primary FLERT-based approach, the training of this architecture is performed at the sentence level, i.e., only the tokens of the current sentence are used as input to the model. This difference allows comparing the performance of sentence-level context against document-level context models.

3.2 FLERT Model

Our main proposal for the task is based on the FLERT approach (Schweter and Akbik, 2020). This architecture is similar to the baseline since it fine-tunes a language model but differs in that it considers document-level context rather than sentence-level context. To this purpose, we add

Proceedings of the 29th International Conference on Computational Linguistic, pages 52–54
October 12–17, 2022.

a window of 64 tokens from the previous sentence and 64 tokens from the following sentence. We consider the document-level context relevant since there are tweets containing multiple sentences, which implies that two contiguous sentences describe the same idea with a high probability.

The Spanish language models selected to test this approach are the following: the biomedical and clinical versions of RoBERTa (*bsc-bio-es* and *bsc-bio-ehr-es*) (Carrino et al., 2022), the Spanish version of BERT (BETO) (Cañete et al., 2020) and the Spanish version of RoBERTa (*roberta-base-bne*) (Gutiérrez-Fandiño et al., 2022). We also tested concatenating the representations obtained with BETO and Clinical RoBERTa. Finally, although in most of the experiments, we did not use preprocessing on the data, we tested Clinical RoBERTa but converted emojis to text, removed URLs, and unified usernames.

3.3 Ensemble

The second approach is to ensemble all FLERT-based models, regardless of their domains and data preprocessing methods. For this purpose, we used a voting system, which counts the number of times an entity mention was identified and keeps only those that exceed a defined voting threshold.

4 Experiments

The training process for the FLERT-based models was analogous to each language model. We searched for an optimal learning rate between 1e-5, 5e-5, 5e-6, and 1e-6. For the sake of brevity, we only present the best model, trained with a learning rate of 5e-6. We used the Adam optimizer with linear decay and no warm-up steps. The models were trained for 20 epochs using a batch size of 16 sentences with a maximum length of 512 tokens.

In contrast, the baseline model was trained with a learning rate of 2e-5 for 3 epochs with a weight decay of 0.01, which is the setting that gave us the best result. The rest of the hyperparameters are the same. The training of each model took approximately 3 hours using a Tesla V100 GPU.

As shown in Table 1, the baseline model obtained the lowest results, which can be explained since this model was trained on a general domain corpus, lacking both clinical language and sufficient Spanish language specialization. In contrast, the FLERT-based models obtained better results, demonstrating the importance of incorporating the document-level context. The best results were obtained using the Clinical RoBERTa model, reaching an F_1 score of 0.870, and the ensemble model reaching 0.884 points. Analyzing the ensemble approach, which obtained the best performance, the best results in the validation set were obtained using a threshold of two votes. The Clinical RoBERTa and ensemble-based models were the final systems submitted to the shared task. The source code to reproduce our experiments is freely available to the research community [1].

Model	P	R	F_1
Baseline	0.860	0.763	0.809
FLERT [BETO]	0.865	0.835	0.850
FLERT [RoBERTa]	0.863	0.833	0.848
FLERT [Bio RoBERTa]	0.887	0.854	0.867
FLERT [Clinical RoBERTa]	0.881	0.857	0.870
FLERT [Stacked]	0.880	0.837	0.858
FLERT [Ensemble]	**0.903**	**0.866**	**0.884**

Table 1: Overall results in the validation partition.

Model	P	R	F_1
Other systems (mean)	0.680	0.677	0.675
Other systems (median)	0.758	0.780	0.761
FLERT [Clinical RoBERTa]	0.867	0.831	0.848
FLERT [Ensemble]	**0.882**	**0.843**	**0.862**

Table 2: Overall results in the testing partition.

5 Results and Conclusion

The overall results of our submissions are shown in Table 2. The ensemble approach obtained the best results, achieving a strict F_1 score of 0.862, followed by the Clinical RoBERTa model with 0.848. Both models outperformed the average F_1 score of the rest of the systems submitted (0.675) by 0.187 and 0.173 points, respectively. This demonstrates that using FLERT document-level features delivers great results independent of its simplicity, especially when using domain-specific language models. In future work, we believe that incorporating a more extensive range of domain-specific models into the ensemble architecture, such as models trained on social media, can improve performance further. In addition, we would like to test character-level contextualized embeddings, such as Clinical Flair (Rojas et al., 2022), since it is mentioned that they are helpful for clinical corpora in Spanish. Finally, to measure the impact of the unstructured nature of tweets, we would like to test our models on other datasets created for disease recognition, such as the Chilean Waiting List (Báez et al., 2020).

[1] https://github.com/plncmm/socialdisner

References

Pablo Báez, Felipe Bravo-Marquez, Jocelyn Dunstan, Matías Rojas, and Fabián Villena. 2022. Automatic extraction of nested entities in clinical referrals in spanish. *ACM Trans. Comput. Healthcare*, 3(3).

Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. 2020. The Chilean waiting list corpus: a new resource for clinical named entity recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 291–300, Online. Association for Computational Linguistics.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Luis Gasco, Darryl Estrada-Zavala, Eulàlia Farrá-Maduell, Salvador Lima-López, Antonio Miranda-Escalada, and Martin Krallinger. 2022. Overview of the SocialDisNER shared task on detection of diseases mentions from healthcare related and patient generated social media content: methods, evaluation and corpora. In *Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Matías Rojas, Jocelyn Dunstan, and Fabián Villena. 2022. Clinical flair: A pre-trained language model for Spanish clinical natural language processing. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 87–92, Seattle, WA. Association for Computational Linguistics.

Stefan Schweter and Alan Akbik. 2020. FLERT: Document-level features for named entity recognition.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.