READ-BioMed@SocialDisNER: Adaptation of an Annotation System to Spanish Tweets

Antonio Jimeno Yepes[1,2] and Karin Verspoor[1,2]

[1]School of Computing Technologies, RMIT University, Melbourne, Australia
[2]School of Computing and Information systems
The University of Melbourne, Melbourne, Australia
[1]{antonio.jose.jimeno.yepes,karin.verspoor}@rmit.edu.au

Abstract

We describe the work of the READ-BioMed team for the preparation of a submission to the SocialDisNER Disease Named Entity Recognition (NER) Task (Task 10) in 2022. We had developed a system for named entity recognition for identifying biomedical concepts in English MEDLINE citations and Spanish clinical text for the LivingNER 2022 challenge (Miranda-Escalada et al., 2022). Minimal adaptation of our system was required to perform named entity recognition in the Spanish tweets in the SocialDisNER task, given the availability of Spanish pre-trained language models and the SocialDisNER training data. Minor additions included treatment of emojis and entities in hashtags and Twitter account names.

1 Motivation

In this paper, we describe the READ-BioMed (Reading, Extraction, and Annotation of Documents in BioMedicine) approach to the 2022 SocialDisNER Task 10 (Weissenbacher et al., 2022). The documents in this task are Spanish-language tweets and the task involves the annotation of disease entities in tweet text.

The READ-BioMed team has extensive experience in the processing of Twitter to identify medical entities (Jimeno-Yepes et al., 2015; Jimeno Yepes and MacKinlay, 2016), the analysis of large sets of tweets for trend analysis (MacKinlay et al., 2015; Jimeno Yepes et al., 2015; Huang et al., 2016), pharmacovigilance using social media (MacKinlay et al., 2017; Li et al., 2020), and for syndromic surveillance (Ofoghi et al., 2016).

Our approach was to adapt a system previously developed for annotation of MEDLINE citations in the English language and clinical text in Spanish. Our system relies on a pre-trained transformer based language model, which was fine-tuned for biomedical concept recognition for the LitCOIN challenge earlier this year[1], where we ranked in the top 5 submissions[2]. More recently, we adapted this system for processing Spanish clinical texts (Jimeno Yepes and Verspoor, 2022) in the LivingNER challenge (Miranda-Escalada et al., 2022).

2 Methods

In this section, we describe the methods that we have used in the challenge. We describe the data and the pre-processing step, the training and annotation of the data and the post-processing steps that we have followed to prepare our submissions using the validation and testing sets.

2.1 Data

We used the data set provided by the task organisers (Gasco et al., 2022). This data set contains 5,000 tweets in the training set, 2,500 tweets in the validation set and 23,430 tweets in the testing set.

2.2 From tweets to BIO

Tweets are messages of up to 280 characters. So, we did not split the tweets into sentences as done in previous challenges (Jimeno Yepes and Verspoor, 2022) to support analysis with BERT-like models (Devlin et al., 2019), instead processing a complete tweet as a single unit.

We define the following list of characters used to identify token boundaries, some of these tokens were defined as Unicode characters, e.g. ’. . . ’. All characters except for space (’ ’) and the new line (’\n’) were included as tokens.

```
split_tokens = [’¡’,’\xa0’,’_’,
’¿’,’=’,’+’,’*’,’;’,’&’,’ ’,
’–’,’.’,’[’,’]’,’...’,’(’,’)’,
’,’,’/’,’%’,’:’,’#’,’@’,’”’,
’“’,’-’,"\n",’?’,"’",’!’,’"’]
```

[1] https://ncats.nih.gov/funding/challenges/litcoin

[2] https://ncats.nih.gov/funding/challenges/litcoin/winners

In addition to the token list defined above, we also used the python library `emoji`[3], set up for the Spanish language, to identify emoji tokens. Tweets in all the data sets (training, validation and testing) were tokenised using the same process.

The training and validation sets include information about the disease entities annotated per tweet, which are provided in a tab separated values file. For each entity, the tweet id and the offset of the entities are available and are used to identity the tokens and label each one of them using the IOB labeling (Ramshaw and Marcus, 1995). The B label is used to denote the first token of a disease entity, the I label is used to denote any additional token within the entity and the O label is to denote any token that is not part of a disease entity. The IOB labels were used to train our system.

2.3 Training

In our previous work in biomedical Spanish named entity recognition as part of the LivingNER challenge (Miranda-Escalada et al., 2022) and for English language concept recognition in the LitCoin challenge,[4] we evaluated pre-trained language models such as BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019) and PubMedBERT (Gu et al., 2021).

We utilised the NERDA[5] framework for named entity recognition. It consists of a BERT-based system followed by a fully connected layer with as many outputs as labels need to be predicted. Our motivation was to be able to make changes and adapt the software according to (Jimeno Yepes, 2022), to prevent overfitting.

Since the challenge data was in Spanish, we reused the pre-trained language model for Spanish biomedical data (Carrino et al., 2021) that we used in the LivingNER challenge. More specifically, we have used the model *PlanTL-GOB-ES/roberta-base-biomedical-es* available from Huggingface[6]. This model is based on a RoBERTa (Liu et al., 2019) and trained trained on a biomedical-clinical corpus in Spanish collected from several sources.

We used the training data set provided by the organisers (5,000 tweets) for learning the model and the validation set (2,500 documents) was used to control the training process. The IOB labels that our system was trained for included the O label for out entity tokens, representing a total of 3 token labels. We did not consider any of the extended data sets provided by the organisers.

2.4 Annotation

The trained model obtained using the process described above has been used to annotate the tokens in the validation and testing sets with IOB labels. The IOB labels were matched to the tweet text and the positions of these tokens were used to generate the tab separated values submission files. A pipeline of this process is shown in figure 1.

We submitted the annotations on the validation set to the submission system provided by the organisers, which showed that precision was relatively high, while recall could be improved. The analysis of errors in the validation set showed that disease entities in hashtags, identified by the hashtag symbol # and account names @, were not identified. The main reason for this problem is that such terms were not tokenised to allow disease identification.

To solve this problem, the predicted disease entities in the tweet were identified and these entities were matched against these special terms by lower casing the terms and doing an exact match. This method boosted the recall from 0.856 in the strict evaluation to 0.887 and from 0.914 in the relaxed evaluation to 0.947, while suffering only a small decrease in precision. The validation set contains 2,500 tweets and it took over 7 minutes to annotate them using the trained model.

The testing set was processed using the methodology presented above. The testing set contained 23,430 tweets and it took just over one hour to annotate them using the trained model.

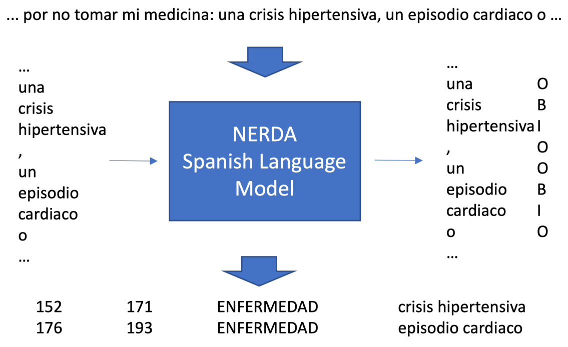
... por no tomar mi medicina: una crisis hipertensiva, un episodio cardiaco o ...
...
una
crisis
hipertensiva
,
un
episodio
cardiaco
o
...
NERDA
Spanish Language
Model
...
una O
crisis B
hipertensiva I
, O
un O
episodio B
cardiaco I
o O
...
152 171 ENFERMEDAD crisis hipertensiva
176 193 ENFERMEDAD episodio cardiaco

Figure 1: Annotation pipeline. A tweet is tokenised, processed by the trained model in NERDA and the IOB output is converted into the submission format.

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

[4] https://bitgrit.net/competition/13#private-leaderboard

[5] https://github.com/ebanalyse/NERDA

[6] https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es

3 Results

We generated only one submission for the testing set using the procedure presented above. Results of our submission (READ-BioMed) using strict entity matching are presented in table 1, which also includes the results for the mean and median values of all participants. Our submission achieved a substantially higher performance compared to the mean and median of the challenge participants.

	Precision	Recall	F1
READ-BioMed	0.868	0.875	0.871
Mean	0.680	0.677	0.675
Median	0.758	0.780	0.761

Table 1: Official results on the testing set using strict matching. These results contain out submission (READ-BioMed) and the mean and median of all participants.

Table 2 shows the results of the fine tuned model of the validation set. Both results are reported, the precision, recall and F1 of the B and I labels and the results by the organisers system on a submission built on the validation set. Comparing the results of our submission in table 1 and the results on the validation set using strict matching, we can see that despite the lower result of our submission, the decrease in performance is limited.

Evaluation	Precision	Recall	F1
B label	0.936	0.951	0.944
I label	0.904	0.877	0.890
Strict	0.881	0.887	0.884
Relaxed	0.944	0.947	0.946

Table 2: Evaluation of the fine tuned models on the validation data set for B and I labels and results for the strict and relaxed evaluation.

The B label is predicted with high precision and recall, which might explain the performance on the relaxed evaluation, e.g. being able to find the beginning of an entity (B label) but not being so successful with the following tokens (I label).

Acknowledgements

This work was supported by Commonwealth Innovation Connections Grant ICG001884 from the Australian Department of Industry, Science, Energy and Resources to RMIT. This research used the HPC-GPGPU Facility at the University of Melbourne, supported by LIEF Grant LE170100200.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for Spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv preprint arXiv:2109.03570*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luis Gasco, Darryl Estrada-Zavala, Eulàlia Farré-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*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Pin Huang, Andrew MacKinlay, and Antonio Jimeno Yepes. 2016. Syndromic surveillance using generic medical entities on twitter. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 35–44.

Antonio Jimeno Yepes. 2022. Hyperplane bounds for neural feature mappings. *arXiv preprint arXiv:2201.05799*.

Antonio Jimeno Yepes and Andrew MacKinlay. 2016. Ner for medical entities in twitter using sequence to sequence neural networks. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 138–142.

Antonio Jimeno Yepes, Andrew MacKinlay, and Bo Han. 2015. Investigating public health surveillance using twitter. In *Proceedings of BioNLP 15*, pages 164–170.

Antonio Jimeno-Yepes, Andrew MacKinlay, Bo Han, and Qiang Chen. 2015. Identifying diseases, drugs, and symptoms in twitter. In *MEDINFO 2015: eHealth-enabled Health*, pages 643–647. IOS Press.

Antonio Jimeno Yepes and Karin Verspoor. 2022. The read-biomed team in livingner task 1 (2022): Adaptation of an english annotation system to spanish. In *LiverNER challenge, IberLEF 2022*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ying Li, Antonio Jimeno Yepes, and Cao Xiao. 2020. Combining social media and fda adverse event reporting system to detect adverse drug reactions. *Drug safety*, 43(9):893–903.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andrew MacKinlay, Hafsah Aamer, and Antonio Jimeno Yepes. 2017. Detection of adverse drug reactions using medical named entities on twitter. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1215. American Medical Informatics Association.

Andrew MacKinlay, Antonio Jimeno Yepes, and Bo Han. 2015. Identification and analysis of medical entity co-occurrences in twitter. In *Proceedings of the ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics*, pages 22–22.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima-López, Darryl Estrada, Luis Gascó, and Martin Krallinger. 2022. Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of livingner shared task and resources. *Procesamiento del Lenguaje Natural*.

Bahadorreza Ofoghi, Meghan Mann, and Karin Verspoor. 2016. Towards early discovery of salient health threats: A social media emotion classification technique. In *Proceedings of the Pacific Symposium on Biocomputing 2016*, pages 504–515. World Scientific.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications smm4h shared tasks at coling 2022. In *Proceedings of the Seventh Social Media Mining for Health (SMM4H) Workshop and Shared Task*.