KU_ED at SocialDisNER: Extracting Disease Mentions in Tweets Written in Spanish

Antoine Lain[1] **Wonjin Yoon**[2] **Hyunjae Kim**[2] **Jaewoo Kang**[2,3] **T Ian Simpson**[1]
[1] Institute for Adaptive and Neural Computation, The University of Edinburgh,
10 Crichton Street, Edinburgh EH8 9AB, UK [2]Korea University [2]AIGEN Sciences
{Antoine.Lain,ian.simpson}@ed.ac.uk
{wjyoon,hyunjae-kim,kangj}@korea.ac.kr

Abstract

This paper describes our system developed for the Social Media Mining for Health (SMM4H) 2022 SocialDisNER task. We used several types of pre-trained language models, which are trained on Spanish biomedical literature or Spanish Tweets. We showed the difference in performance depending on the quality of the tokenization as well as introducing silver standard annotations when training the model. Our model obtained a strict F1 of 80.3% on the test set, which is an improvement of +12.8% F1 (24.6 std) over the average results across all submissions to the SocialDisNER challenge.

1 Introduction

In this system description paper, we aim to detect disease mentions from tweets written in Spanish as part of the Social Media Mining for Health (SMM4H) 2022 SocialDisNER task (Gasco et al., 2022). The organizers provided the participants with various sets of data, either labelled by healthcare experts or machine-generated by the organizers themselves. All the sets contain anonymous tweets written in Spanish, each saved as a txt file, with a corresponding tsv file that contains the label for each tweet. Since the data is user-generated text limited to 280 characters as per the limitation from Twitter, it contains misspelled words, abbreviations, emojis, links, hashtags, and mentions to other users of the platform, making the task challenging.

Acknowledging the outstanding performance for named entity recognition presented in Devlin et al. (2019) and from its adaptation to the biomedical domain presented in Lee et al. (2020), we decided to use pre-trained biomedical language models for detection of disease mentions in the SocialDisNER task. We selected four publicly available pre-trained language models on Spanish corpora (Carrino et al., 2022; Cañete et al., 2020; Chizhikova et al.; Huertas-Tato et al., 2022), one pre-trained on English corpora (Sanh et al., 2019) to retrain one of the models for named entity recognition. We reported the results of our experiments that range from 62.4% to 82.9% strict F1 score on the validation set.

Model	Precision	Recall	Tag-F1
Carrino et al. (2022)	0.94	0.95	0.94
Chizhikova et al.	0.95	0.94	0.94
Cañete et al. (2020)	0.91	0.92	0.92
Sanh et al. (2019)	0.88	0.90	0.89
Huertas-Tato et al. (2022)	0.92	0.92	0.92

Table 1: Token-level performance of pre-trained language models on the SocialDisNER validation set.

2 Data Description

The first set released by the organizers was the gold standard (GS), 7,500 tweets, it was accompanied by a tab-separated file with healthcare experts' annotations containing the unique tweet ID, beginning position, end position, type and extraction. This set was divided into 2 subsets, a training set of 5,000 tweets and a development set of 2,500 tweets. The second set, 85,077 tweets, is similar as it offers the same amount of information but is generated by a machine at the discretion of the organizers, making this set a silver standard ultimately less reliable than human expert annotations. Finally, the third set is the test set. The organizers used 2,000 of 23,430 tweets from the test set to evaluate the performance of each team but the labels were not disclosed by the time of submission. For our final system, we only used the training dataset annotated from healthcare professionals (GS). There is an average of 3,3 labels per tweet in the training set when the average is 1,7 labels per tweet in the development set supplied by the organizers.

3 System Description

BERT type models have demonstrated improvement compared to previous state of the art meth-

Proceedings of the 29th International Conference on Computational Linguistic, pages 78–80
October 12–17, 2022.

ods in NER (Devlin et al., 2019; Lee et al., 2020). We decided to select a few pre-trained language models from Hugging Face[1] to train each of these models for the NER task using the transformer architecture (Vaswani et al., 2017) and the seqeval library implemented in Python (Nakayama, 2018). We first selected four models trained on Spanish corpora (Carrino et al., 2022; Cañete et al., 2020; Chizhikova et al.; Huertas-Tato et al., 2022) and one trained on English corpora (Sanh et al., 2019). We also needed to convert the data to the correct format. We decided to use the BIO tagging format (Begin, Interior, Outside) (Ramshaw and Marcus, 1999), where each token within a tweet was coupled with one of the three tags.

3.1 Experiments and System Selection

First, we ran all five models using the same input parameters in order to select the best performing model for the task, the results are reported in Table 1. The model from Carrino et al. (2022) had the best F1-score.

Due to the challenging style of writing that tweets present, we worked on improving the quality of the tokenization as it has shown improvement in Kim et al. (2021). We split each tweet by space, then we split every token to separate the punctuation from the rest. Hashtags (#) and at (@) were separated from the original token. We also cut the words where only parts of the word were annotated. We ran three experiments: simple tokenization, improved tokenization and simple tokenization + silver standard data. Each experiment uses the same pre-trained language model 'PlanTL-GOB-ES/bsc-bio-es'[2], with 3 epochs, learning rate = 1e^{-4} and weight decay = 1e^{-5}. Since we needed to give the span for each extraction, we replaced every emoji with '@' since it was changing the position of the text. As a post-processing rule, if the length of the prediction of the model was shorter than 3 characters the prediction was ignored.

4 Results and analysis of the error

We reported the results of our experiments in Table 2. The best performing model on the validation set was when we improved the quality of the tokenization. We gained 6.1% F1-score compared to a tokenization where the punctuation was kept. The difference in performance between the overlap and strict measures showed that the model was able to identify a part of the mentions but still resulted in around 10% difference between both measures. Looking at the mentions from the model compared to the mentions from the healthcare expert it seems that extra rules in the post-processing step could have been implemented to reduce the gap between both performances. For example when the identified token is part of a word, which is not a hashtag, extract the word instead of the token.

Model	Set	Measure	Precision	Recall	F1
Simple	Validation	Strict	71.9	82.4	76.8
Improved	Validation	Strict	83.1	82.7	82.9
Simple + Silver	Validation	Strict	66.0	59.2	62.4
Simple	Validation	Overlap	84.5	95.1	89.5
Improved	Validation	Overlap	94.6	92.7	93.7
Simple + Silver	Validation	Overlap	89.0	78.6	83.5
All participants	Test	Strict Mean	68.0	67.7	67.5
All participants	Test	Strict Median	75.8	78.0	76.1
Our Model	Test	Strict	80.9	79.8	80.3

Table 2: Summary of our results based on the official overlap and strict evaluation.

5 Conclusion

In this work, we studied the difference in performance of pre-trained language models depending on the quality of the labels and tokenization for detection of disease mentions in tweets. At the end, the model achieved 80.3% F1 score on the test set. For future work, one direction would be to use ensemble models for token classification, where we combine a Twitter pre-trained language model with a biomedical pre-trained language model.

Acknowledgments

This research was supported by (1) the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2022-2020-0-01819) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), (2) a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0021), (3) National Research Foundation of Korea (NRF-2014M3C9A3063541, NRF-2020R1A2C3010638) and (4) School of Informatics at UoE.

[1] https://huggingface.co/

[2] PlanTL-GOB-ES/bsc-bio-es/

References

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estap'e, Joaqu'in Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pre-trained biomedical language models for clinical nlp in spanish. In *BIONLP*.

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*.

Mariia Chizhikova, Jaime Collado-Montañéz, Pilar López-Úbeda, Manuel C. Díaz-Galiano, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. Sinai at clef 2022: Leveraging biomedical transformers to detect and normalize disease mentions. pages 265–273.

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.

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*.

Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2022. Bertuit: Understanding spanish language in twitter through a native transformer. volume abs/2204.03465.

Hyunjae Kim, Mujeen Sung, Wonjin Yoon, Sungjoon Park, and Jaewoo Kang. 2021. Improving tagging consistency and entity coverage for chemical identification in full-text articles. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

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*, 36:1234 – 1240.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.