RACAI@SMM4H'22: Tweets Disease Mention Detection Using a Neural Lateral Inhibitory Mechanism

Andrei-Marius Avram, Vasile Păis and **Maria Mitrofan**
Research Institute for Artificial Intelligence "Mihai Drăgănescu"
{andrei.avram,vasile,maria}@racai.ro

Abstract

This paper presents our system employed for the Social Media Mining for Health (SMM4H) 2022 competition Task 10 - SocialDisNER. The goal of the task was to improve the detection of diseases in tweets. Because the tweets were in Spanish, we approached this problem using a system that relies on a pre-trained multilingual model and is fine-tuned using the recently introduced lateral inhibition layer. We further experimented on this task by employing a conditional random field on top of the system and using a voting-based ensemble that contains various architectures. The evaluation results outlined that our best performing model obtained 83.7% F1-strict on the validation set and 82.1% F1-strict on the test set.

1 Motivation

Social media data (e.g. Twitter, Facebook) analysis for health informatics is an active area of research that aims to understand public opinion of health-related topics. The Social Media Mining for Health Applications (#SMM4H) workshop is a venue that brings together researchers that want to contribute to this area, including, but not limited to subjects such as: deriving health trends from social media, health-related message classification or disease monitoring from messages on social media.

We chose to participate at the tenth task of the #SMM4H competition - SocialDisNER - whose aim was to improve the disease detection in Spanish tweets by identifying `ENFERMEDAD` entities in the given text (Gasco et al., 2022). We approached this problem using the XLM-RoBERTa multilingual pre-trained model (Conneau et al., 2020) on which we incorporated the recently introduced lateral inhibitory (LI) layer (Păiș, 2022) in the fine-tuning process, together with a conditional random field (CRF) (Lafferty et al., 2001) layer on top of the model and an ensemble system.

The interest in this task came from our involvement within the CURLICAT project for the CEF AT action (Váradi et al., 2022), where named entity recognition systems need to be developed for various domains (health-related included), in Romanian language.

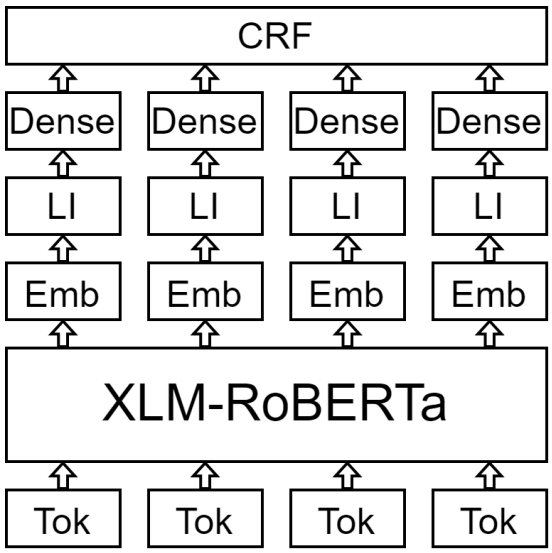

Figure 1: Model architecture that contains XLM-RoBERTa, the LI layer and the CRF layer (in this figure, "Tok" stands for "Token" and "Emb" stands for "Embedding").

2 System Description

We treat the SocialDisNER task as a sequence tagging problem, so we try to predict a label for each token given as input. The architecture of our system consists of using the XLM-RoBERTa pre-trained model that produces an embedding for each token in the input sequence. Then, we employ the LI layer on each token embedding, followed by a dense layer to produce the output probabilities for each of the three possible classes: `O`, `B-ENFERMEDAD` or `I-ENFERMEDAD`[1]. We further use a CRF layer on top of the predicted logits to

[1]We transform the input data to match the Inside–Outside–Beginning format (Ramshaw and Marcus, 1999).

Proceedings of the 29th International Conference on Computational Linguistic, pages 1–3
October 12–17, 2022.

Model	F1-Overlap			F1-Strict		
	P	R	F1	P	R	F1
XLM-RoBERTa	**95.7**	88.3	**91.8**	86.2	75.4	80.4
XLM-RoBERTa+LI*	94.7	88.4	91.4	87.5	**80.3**	**83.7**
XLM-RoBERTa+CRF	95.3	87.8	91.4	**87.8**	79.6	83.5
XLM-RoBERTa+LI+CRF	95.0	88.4	91.6	87.4	80.0	83.5
Ensemble*	94.9	**88.6**	91.6	87.5	80.2	**83.7**

Table 1: The F1-overlap and F1-strict evaluation results obtained by our models on the SocialDisNER task validation set. *These models were sent for the final evaluation.

learn to better predict the most probable sequence of classes for a given input. The overall system architecture is depicted in Figure 1.

It must be noted that we also experimented with models that do not include the LI and/or the CRF layers. In addition, we created a voting-based ensemble system that incorporates the predictions of all four architectures developed in this work: XLM-RoBERTa, XLM-RoBERTa+LI, XLM-RoBERTa+CRF and XLM-RoBERTa+LI+CRF.

We fine-tune each model for 80 epochs using a batch size of 16 and a gradient accumulation of 8, resulting in a total batch size of 128. We use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 2e-5 and the default values for the weight decay, epsilon and the betas in PyTorch (Paszke et al., 2019). During training, we save only the checkpoint that obtained the highest performance on the validation set.

3 Evaluation

We evaluate and present the results of our models on the validation and test sets. The results of the XLM-RoBERTa with different layers and of the ensemble system are listed in Table 1. Here we compute both the F1-overlap and F1-strict, of which the latter is the main metric used in the competition. The highest F1-overlap was obtained by the XLM-RoBERTa model that contains no special layers with 91.8%. However, when considering the F1-strict as the evaluation metric, this model underperformed all the other architectures by more than 3%, obtaining a F1-strict score of 80.4%. This result shows that, without any additional layers, the model is not capable of predicting exact boundaries for the entities. The highest F1-strict score of 83.7% was obtained by both XLM-RoBERTa+LI and the ensemble system, these two being the ones that we sent for the final evaluation.

Model	P	R	F1
XLM-RoBERTa+LI	**86.8**	**77.9**	**82.1**
Ensemble	86.7	77.9	82.0
Mean	68.0	67.7	67.5
Median	75.8	78.0	76.1

Table 2: The F1-strict evaluation results obtained by our models on the SocialDisNER test set, together with the mean and median results of the task.

The results on the test set are outlined in Table 2. The two models we sent for evaluation, XLM-RoBERTa+LI and the ensemble, obtained similar results, with a 82.1% F1-strict for the former and 82.0% for the later. This puts our solution 14.6% F1-strict above the mean and 6% F1-strict above the median.

4 Conclusion

Analysing social media data remains an important area of research for health informatics, offering a better understanding of public opinion in health-related topics. This work describes our system used to participate at the SMM4H competition SocialDisNER shared task. We treated the problem as a sequence tagging task and employed the XLM-RoBERTa multilingual pre-trained model to process the Spanish tweets. In addition, we analysed the capabilities of the LI layer on this task in combination with a CRF module. We sent to evaluation the XLM-RoBERTa model with the LI layer, as well as an ensemble system. The two systems obtained 82.1% and 82.0% F1-strict scores on the evaluation set, respectively. No additional resources were employed. In the future we plan to apply this experience also on Romanian microblogging text (Păiș et al., 2022).

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Vasile Păiș. 2022. RACAI at SemEval-2022 Task 11: Complex named entity recognition using a lateral inhibition mechanism. In *Proceedings of the 3International Workshop on Semantic Evaluation (SemEval)*.

Vasile Păiș, Maria Mitrofan, Verginica Barbu Mititelu, Elena Irimia, Roxana Micu, and Carol Luca Gasan. 2022. Challenges in creating a representative corpus of romanian micro-blogging text. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-10)*, pages 1–7, Marseille, France. European Language Resources Association.

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.

Tamás Váradi, Marko Tadić, Svetla Koeva, Maciej Ogrodniczuk, Dan Tufiş, Radovan Garabík, Simon Krek, and Andraž Repar. 2022. Curated multilingual language resources for cef at (curlicat): Overall view. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 339–340.