ITAINNOVA at SocialDisNER: A Transformers cocktail for disease identification in social media in Spanish

Rosa M. Montañés-Salas, Irene López-Bosque
Luis García-Garcés and **Rafael del-Hoyo-Alonso**
Technological Institute of Aragón (ITAINNOVA), Zaragoza (Spain)
{rmontanes, ilopez, lggarcia, rdelhoyo}@itainnova.es

Abstract

The Social Media Mining for Health Applications (#SMM4H) Shared Task aims to promote the state of the art of health informatics challenges for social media. The 10th track of the task, SocialDisNER, focuses on the identification of disease mentions in tweets written in Spanish. ITAINNOVA presents a hybrid system based on Transformer Language Models in combination with Natural Language Processing techniques such as the development and use of a diseases gazetteer for approximate string matching. A comprehensive exploration of the components contributions is presented as well as the final test results obtained, which outperform the mean overall performance in the task.

1 Motivation

The detection of disease mentions in tweets (SocialDisNER, Gasco et al., 2022b) is part of the SMM4H Shared Task (Weissenbacher et al., 2022). SocialDisNER proposes a challenging task: NER (Named-Entity Recognition) offset detection of diseases by finding the span of its mentions in tweets published in the Spanish language. Using social media data for health research involves facing multiple Natural Language Processing (NLP) challenges: multilingualism, usage of formal and informal expressions, misspellings, ambiguity and so on, which may be better tackled unifying state of the art approaches and more conventional methods.

In this context, ITAINNOVA participates with a hybrid system which combines Transformer-based Language Models (LMs) with a custom-built gazetteer for Approximate String Matching (ASM) and dedicated text processing techniques for the social media domain. Additionally zero-shot classification capabilities (Pushp and Srivastava, 2017) have been explored in order to support different parts of the system. An extensive analysis on the interactions of these components has been accomplished, making the system stand out above the mean performance of all the participating teams.

2 System description

The overall architecture is illustrated in figure 1.

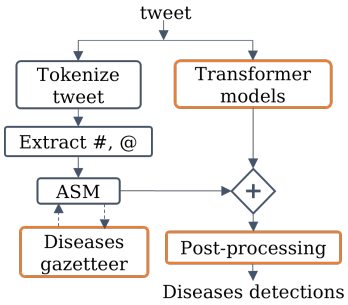

Figure 1: System architecture.

2.1 Transformer-based Language models

The core of the system is a parallel aggregation ensemble of fine-tuned Transformer-based LMs. Nine publicly available pretrained models from the HuggingFace hub, were selected following these requirements: the model has support for Spanish or multiple languages, it has been pretrained with a health related or social media corpus, and its architecture may be fine-tuned for token classification.

Each model is subjected to a hyper-parameter tuning using both gold and silver standard corpus (Gasco et al., 2022a), which allows ranking the models by performance: (1)*wikineural-multilingual-ner* (Tedeschi et al., 2021)[1], (2)*bsc-bio-ehr-es-cantemist* (Miranda-Escalada et al., 2020a), (3)*bertin-base-ner-conll2002-es* (de la Rosa et al., 2022), (4)*bsc-bio-es* (Carrino et al., 2022), (5)*twitter-xlm-roberta-base* (Barbieri et al., 2022)[2], (6)*bsc-bio-ehr-es* and (7)*bsc-bio-ehr-es* (Carrino et al., 2022), (8)*roberta-large-bne* and (9) *roberta-large-bne-capitel-ner* (Gutiérrez-Fandiño et al., 2022)[3]. Tables 3 and 4 in the appendix show the values and metrics of fine-tuned models.

[1]Based on mBERT(Devlin et al., 2018) + Bi-LSTM + CRF
[2]XLM-Roberta (Conneau et al., 2019)
[3]Rest of models are built on RoBERTa (Liu et al., 2019)

Proceedings of the 29th International Conference on Computational Linguistic, pages 71–74
October 12–17, 2022.

2.2 Diseases gazetteer string matching

In conjunction with the neural models, an approximate string matching is performed with an in-domain gazetteer. The gazetteer has been built with diseases-related concepts from publicly available corpora: DisTEMIST (Gasco et al., 2022c), AbreMES-DB (Intxaurrondo, 2018), CodiEsp (Miranda-Escalada et al., 2020b), SNOMED-CT (International Health Terminology Standards Development Organisation - IHTSDO, 2014) and ICD-10-CM (CodeBooks, 2016).

An iterative curating process has been performed over the 122620 entries gathered. Firstly, normalization and duplicates removal is needed. Thereafter a filter on the number of tokens is applied: 5 and 3 tokens-length are considered relating to the average length of tweets. An analysis of n-gram frequencies enables to extract sets of general common-used terms and "stop-terms", which are included if not previously present or removed, respectively. Then a zero-shot classifier built on BETO (Cañete et al., 2020)[4] is used to filter no disease-related terms. Finally, two versions of the gazetteer are consolidated: *Final*, which compile up to 5-token length entries, containing 69655 health-related terms; a 3-token length version of the former called *Reduced* having 32852 terms.

2.3 Text processing and filtering

Various text processing steps are performed within the prediction flow. After ignoring breaklines, hashtags (#) and mentions (@) are extracted as plain tokens, and analyzed applying morpho-lexical rules to gather meaningful words (i.e. *#cancerdemama* gets transformed into "cancer", "de", "mama"). Then, once the disease mentions are extracted, punctuation marks, special characters and emojis at the beginning and the end of each one are removed and offsets are adjusted. After that, the zero-shot model is applied to filter generic mentions. Finally, duplicates and overlapped entities are excluded.

The code of the system is available in GitHub[5].

3 Results and discussion

A thorough study on different state of the art Transformer Language Models for NER in Spanish, their aggregation and integration with other NLP techniques was conducted using the datasets of SocialDisNER. The top three configurations are shown in Table 1, while the whole set of results can be reached at table 5 of the appendix.

Conf.	Gaz.	E.	st.F	st.P	st.R
B2	X	F	**0,817**	**0,840**	0,795
B5	X	T	**0,817**	0,833	**0,802**
B5	X	F	0,815	0,831	0,800

Table 1: Validation results ranked by strict F1.
"B"s refer to the ensemble of the best N top models.

The obtained results demonstrate that ensembling approaches provide the best performance, since standalone models enriched with specific rules work reasonably well on the task. The mBERT with LSTM-CRF model outperforms other architectures, but in terms of model sizes no significant differences have been found. Despite the overall strong performance, many false positives are extracted, so that further research on the effect of the pretraining corpus would be needed.

Using gazetteers alongside LMs, when their size and quality are extensive enough, have a positive impact on the recall. However, large gazetteers are time-consuming in building and predicting phases comparing with their performance contribution. In regard to zero-shot classification, it has contributed to build gazetteers by filtering out of domain terms. Nevertheless, when applied to the prediction pipeline it filters true positives and thus worsens performance, so that it has not been used on test. Domain specific text processing, such as hashtag segmentation and filtering rules to remove false positives, is needed due to the linguistic complexity and orthographic diversity of social media.

Final results in the test set obtained with two different configurations focusing on comparativeness over gazetteer usage, are depicted in table 2. In comparison to the official results of SocialDisNER, the developed system outperforms average values of the participants' submissions.

Model	Gaz./E.	st.F	st.P	st.R
B5	X/T	**0,774**	**0,779**	0,769
B5	Reduce/T	0,752	0,691	**0,826**
Mean task	-	0.675	0.680	0.677
Median task	-	0.761	0.758	0.780

Table 2: Test results.

[4] https://huggingface.co/Recognai/bert-base-spanish-wwm-cased-xnli

[5] https://github.com/ITAINNOVA/SocialDisNER

Acknowledgements

This work has been partially funded by the Department of Big Data and Cognitive Systems at the Technological Institute of Aragon, by IODIDE group of the Government of Aragon, grant number T1720R and by the European Regional Development Fund (ERDF).

References

Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of LREC*.

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*.

Medical CodeBooks. 2016. *ICD-10-CM Complete Code Set 2016*, volume 1. Medical Code Books.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Javier de la Rosa, Eduardo G. Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Maria Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Luis Gasco, Darryl Estrada, Eulàlia Farré, and Martin Krallinger. 2022a. SocialDisNER corpus: gold standard annotations for detection of disease mentions in Spanish tweets. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Luis Gasco, Darryl Estrada-Zavala, Eulàlia Farré-Maduell, Salvador Lima-López, Antonio Miranda-Escalada, and Martin Krallinger. 2022b. 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*.

Luis Gasco, Eulàlia Farré, Antonio Miranda-Escalada, Salvador Lima, and Martin Krallinger. 2022c. DisTEMIST corpus: detection and normalization of disease mentions in spanish clinical cases. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

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.

International Health Terminology Standards Development Organisation - IHTSDO. 2014. SNOMED CT. Http://www.ihtsdo.org/snomed-ct/.

Ander Intxaurrondo. 2018. Abremes-db. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

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.

A Miranda-Escalada, E Farré, and M Krallinger. 2020a. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, and Martin Krallinger. 2020b. CodiEsp corpus: gold standard Spanish clinical cases coded in ICD10 (CIE10) - eHealth CLEF2020. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train once, test anywhere: Zero-shot learning for text classification. *CoRR*, abs/1712.05972.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

A Complete result tables

Model	Learning rate	Epochs	Batch size	Warmup ratio
M1	1e-05	6	4	0.05
M2	2e-06	5	4	0.1
M3	1e-05	10	6	0.2
M4	2e-06	5	4	0.1
M5	1e-05	8	4	0.01
M6	1e-05	6	4	0.05
M7	5e-06	10	16	0.005
M8	3.4e-05	10	8	0.1
M9	5e-05	10	10	0.01

Table 3: Hyperparameter tuning.

R.	Model	ov.F	ov.P	ov.R	st.F	st.P	st.R
1	Babelscape/wikineural-multilingual-ner	0,882	0,937	0,833	0,769	0,813	0,729
2	PlanTL-GOB-ES/bsc-bio-ehr-es-cantemist	0,775	0,891	0,686	0,715	0,822	0,632
3	bertin-project/bertin-base-ner-conll2002-es	0,713	0,835	0,622	0,674	0,758	0,565
4	PlanTL-GOB-ES/bsc-bio-es	0,708	0,832	0,616	0,655	0,769	0,570
5	cardiffnlp/twitter-xlm-roberta-base	0,704	0,836	0,607	0,650	0,772	0,562
6	PlanTL-GOB-ES/bsc-bio-ehr-es	0,706	0,841	0,608	0,648	0,773	0,557
7	PlanTL-GOB-ES/roberta-base-biomedical-clinical-es	0,699	0,814	0,612	0,630	0,733	0,552
8	PlanTL-GOB-ES/roberta-large-bne	0,694	0,822	0,601	0,626	0,741	0,541
9	PlanTL-GOB-ES/roberta-large-bne-capitel-ner	0,675	0,811	0,578	0,596	0,716	0,511

Table 4: Ranking of fine-tuned pretrained models according to strict F1 metric. Columns "R.", "ov." and "st." refer to *Ranking*, *overlap* and *strict* respectively, and *F*, *P*, *R* to F1, Precision and Recall.

Config	Gaz.	Gaz v.	Zero-shot	Emoji	ov.F	ov.P	ov.R	st.F	st.P	st.R
Best 2	F	X	F	F	0,899	0,928	0,872	0,817	0,840	0,795
Best 5	F	X	F	T	0,899	0,919	0,879	0,817	0,833	0,802
Best 5	F	X	F	F	0,899	0,919	0,879	0,815	0,831	0,800
Model 1	T	Final	F	F	0,894	0,859	0,932	0,807	0,768	0,649
Best 2	T	Final	F	F	0,898	0,854	0,947	0,807	0,761	0,859
Best 5	T	Final	F	F	0,897	0,849	0,952	0,805	0,754	0,862
Model 2	T	Final	F	F	0,887	0,873	0,901	0,802	0,784	0,820
All 9	F	X	F	F	0,887	0,889	0,885	0,796	0,794	0,798
Best 5	T	Reduced	F	T	0,886	0,827	0,954	0,793	0,732	0,864
Best 2	T	Reduced	F	F	0,886	0,831	0,950	0,792	0,735	0,858
Best 5	T	Reduced	F	F	0,886	0,827	0,954	0,791	0,731	0,861
Model 4	T	Reduced	F	F	0,877	0,848	0,907	0,790	0,758	0,825
Model 2	T	Reduced	F	F	0,879	0,847	0,914	0,788	0,754	0,825
Model 1	F	X	F	F	0,882	0,937	0,833	0,769	0,813	0,729
All 9	F	X	T	T	0,850	0,857	0,842	0,760	0,763	0,757
Model 1	T	Reduced	F	F	0,882	0,833	0,938	0,759	0,709	0,815
Best 5	T	Reduced	T	T	0,853	0,802	0,911	0,758	0,706	0,819
All 9	T	Reduced	T	T	0,845	0,784	0,916	0,744	0,682	0,817
Model 2	F	X	F	F	0,775	0,891	0,686	0,715	0,822	0,632
Model 1	T	Reduced	T	T	0,837	0,798	0,881	0,713	0,673	0,757

Table 5: System configurations ranked by strict F1.
Gaz and *Gaz v.* refers to the use of the gazetteer and its corresponding version, *Zero-shot* and *Emoji* indicates whether zero-shot filter and emoji cleaning is performed.