SINAI@SMM4H'22: Transformers for biomedical social media text mining in Spanish

Mariia Chizhikova[1], **Pilar López-Úbeda**[2], **Manuel C. Díaz-Galiano**[1], **L. Alfonso Ureña-López**[1], **M. Teresa Martín-Valdivia**[1]
[1]Department of Computer Science. University of Jaén. Jaén, Spain
[2]R+D+I department. HT medica. Jaén, Spain
mc000051@red.ujaen.es, p.lopez@htmedica.com,
{mcdiaz, laurena, maite}@ujaen.es

Abstract

This paper covers participation of the SINAI team in Tasks 5 and 10 of the Social Media Mining for Health (#SSM4H) workshop at COLING-2022. These tasks focus on leveraging Twitter posts written in Spanish for healthcare research. The objective of Task 5 was to classify tweets reporting COVID-19 symptoms, while Task 10 required identifying disease mentions in Twitter posts. The presented systems explore large RoBERTa language models pre-trained on Twitter data in the case of tweet classification task and general-domain data for the disease recognition task. We also present a text pre-processing methodology implemented in both systems and describe an initial weakly-supervised fine-tuning phase alongside with a submission post-processing procedure designed for Task 10. The systems obtained 0.84 F1-score on the Task 5 and 0.77 F1-score on Task 10.

1 Introduction

Social media networks are widely used as a base for public health related research. Among other websites, Twitter constitutes a useful data source for researches due to the real time nature of the content and the ease to accessing publicly available information (Sinnenberg et al., 2017).

However, extraction of information from tweets remains challenging due to some particularities of the language variety used in the social network. Spelling mistakes, shortenings, abbreviations, agrammatical word truncation alongside with frequent use of emoji, hashtags, emoticons and even usage of digits and mathematical symbols to represent phonetic sounds are some characteristics that distinguish linguistic register typically adopted by Twitter users (Jalbuena, 2012). This fact restricts the efficiency of rule-based information retrieval approaches, such as exact matching, used to extract data in many researches (Chew and Eysenbach, 2010; McNeil et al., 2012; Lyles, 2013).

Named entity recognition (NER) is a task in Natural Language Processing (NLP) which has as objective structuring free-text data by identifying relevant terms in it. Text classification is another way of structuring data which improves information retrieval by assigning standardized labels to texts(Du et al., 2019).

In this paper we present the systems developed by the SINAI team for Shared Tasks 5 and 10 at the Social Media Mining for Health (#SMM4H) workshop at COLING 2022. On the one hand, task 5 brings the community effort to design systems that would automatically classify Spanish tweets reporting COVID-19 in 3 categories: self reports, non-personal reports and literature/news mentions. The presented tweet classification systems relies on RoBERTiuto (Pérez et al., 2021), a language model pre-trained on social media text. On the other hand, the purpose of Task 10 (SocialDisNER) is to design systems that would recognize all kinds of disease mentions in tweets written in Spanish. With this objective, we developed a system based on a model pre-trained on general-domain text (Gutiérrez-Fandiño et al., 2021). In order to leverage large scale additional Silver Standard data with automatically generated labels provided by task's organizers we designed a two-stage fine-tuning framework.

Concerning data processing, we describe a submission post-processing procedure aimed to improve system's performance and we enhanced our systems with additional text pre-processing.

2 Data

The organizers of both tasks made available corpora of tweets in Spanish that represent first-hand experience of diseases and other health-related content.

Proceedings of the 29th International Conference on Computational Linguistic, pages 27–30
October 12–17, 2022.

2.1 Task 5

The annotated dataset for this classification task is a collection of 10,052 tweets characterized with class imbalance, as its major part is labeled as literature/news mentions (5,985). We randomly split this set in order to use 30% of labelled data as development set.

An unannotated collection of 3,578 was provided as validation data and the test set consisted of 6,851 tweets.

2.2 Task 10

The Gold Standard SocialDisNER corpus is a collection of 9,500 manually annotated tweets (Gasco et al., 2022a). The corpus was randomly split in train, validation and test sets by the organizers. To prevent participants from manually annotating tweets, final predictions were made on a set of 23,430 tweets (test+background data), while only 2,000 of these were used to evaluate the presented systems.

As an additional resource, a set of large-scale Silver Standard data which contained mentions automatically extracted from 85,000 tweets was released. Its subset labelled with disease mentions consists of 84,988 tweets with 116,260 annotations 16,034 from which are unique (Gasco et al., 2022b).

3 Pre-processing

The peculiarities of the linguistic register adopted by Twitter users make text pre-processing an essential step in order to get a language model to perform robustly. For both tasks, we followed the same rule-based pre-processing procedure which consists of the following steps:

1. Newline characters are replaced with spaces.
2. Character repetitions are limited to max of 3.
3. Hashtag symbol (#) is removed and its text is split into words when possible.
4. Emoji are replaced by their text representations between two 'emoji' tokens.

The steps 2-4 were carried out by using the `pysentimiento` Python package (Pérez et al., 2021).

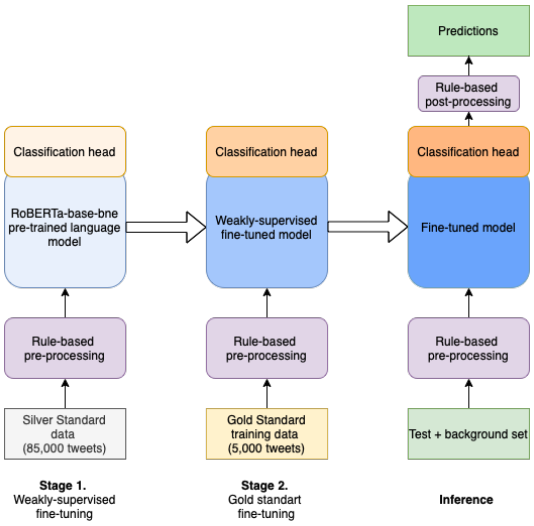

Figure 1: Task 10 model fine-tuning process overview

4 Task 5. Classification of tweets containing self-reported COVID-19 symptoms

To address the tweet classification task we opted for fine-tuning two variants (*cased* and *uncased-deacc*, which lowercases and removes accents) of RoBERTuito - a RoBERTa architecture language model (Liu et al., 2019) pre-trained on social media text (Pérez et al., 2021).

5 Task 10- Detection of disease mentions in tweets – SocialDisNER

The core element of the system presented for SocialDisNER task is the RoBERTa-base-bne: RoBERTa pre-trained on Spanish general-domain corpus (Gutiérrez-Fandiño et al., 2021).

Weak supervision In order to leverage the additional large-scale data provided by SocialDisNER organizers, we designed a two-step model fine-tuning process in which we first feed the model automatically generated (*weak*) Silver Standard data and fine-tune the model on it using as validation data the corresponding split of the Gold Standard dataset. Secondly, we take the resulting model to the second fine-tuning stage using the Gold Standard training set. Figure 1 provides a summary of the described process.

The aim of weakly supervised modelling is to bootstrap system performance without the need of manually annotating more data (Lison et al., 2020).

Submission post-processing To improve overall quality of NER performed by our system, we incor-

porated a rule-based post-processing which is used in both single-stage and two-stage fine-tuning. At this step, the entities detected by the transformer model are subjected to the following procedures:

- Strip all punctuation marks and whitespaces from detected mentions.
- Retrieve a list of relevant terms from train, validation and Silver Standard and perform an exact match.
- Select the longest mention if two entities occur within the same offset.

6 Experimental setup

All models were fine-tuned on a single NVIDIA Tesla V100 GPU by adding a linear classifier layer preceded by a 0.1 dropout layer on top of the original architecture using the Hugging Face Transformers Python library (Wolf et al., 2019). In addition, to take the maximum advantage of the pre-trained model, we performed hyperparameter optimization that relied on the Optuna framework (Akiba et al., 2019).

In the two-stage learning process, we performed optimization at each step within the same hyperparameter space which was also used for a single-stage fine-tuning.

7 Results

Our team submitted a total of two runs per task. In Task 5 these corresponded to predictions made with *cased* model (Run 1) and *uncased-deacc* model (Run 2). Only the best run was kept after the official evaluation.

As for Task 10, Run 1 corresponds to the system based on the fine-tuned RoBERTa model and Run 2 to the one based on the same pre-trained model, but subjected to the-two stage fine-tuning procedure described in Section 5.

The metrics selected to evaluate performance of the participant systems are precision, recall and F1-score for the self-report class or for each mention extracted where the spans overlap exactly in Task 5 and Task 10 respectively. Table 1 displays the results for each of presented systems.

8 Error analysis

8.1 Task 5

The best performing system reached 0.77 F-1 score during in-house evaluation. While carrying out the evaluation and analysis of the results, we were able to determine that many miss-classified tweets contained an large number of spelling errors or incorporated direct speech citations.

8.2 Task 10

During the development process, the best performing model reached 0.81 strict F1-score on the validation set. Most common false negatives were related to detection of mentions that formed part of complex hashtags, for example, *Depresión* in *#VivirConDepresión* (*eng. #LiveWithDepression*). Also were observed some cases of incorrect extraction of complex disease mentions, where the system tended to shorten complex entities: *reacciones dermatológicas* (*eng. dermatologic reactions*) instead of *reacciones dermatológicas tras la vacuna* (*eng. dermatological reactions after the vaccine*).

9 Conclusions and future work

This paper presents the participation of the SINAI team in #SMM4H workshop at COLING 2022. Firstly, we describe rule-based tweet preprocessing which was used in both tasks and aimed to normalize the peculiarities of the linguistic variety used in Twitter,.

For tweet classification task we compared transfer learning potential of two versions of RoBERTa model pre-trained on social media text: *cased* and *uncased-deacc*. The latter performed better achieving 0.84 F1-score.

To address the automatic disease mention recognition in tweets we employed a large pre-trained model and compared a state-of-the art fine-tuning approach with a weak-supervision enhanced two-stage fine-tuning. For the tweet classification task, we compared transfer learning performance of two versions of language models pre-trained on social media text.

The official evaluation for Task 10 reveals that the model subjected to weak-supervision fine-tuning performs slightly better in terms of precision (0.739 vs 0.756) and F1-score (0.769 vs 0.775), which, however, cannot be considered a significant performance boost.[1]

[1]You can access to the best performing model trained for this task at https://huggingface.co/chizhikchi/spanish-SM-disease-finder

		Precision	Recall	F1-score
Task 5	Run 2	0.84	0.84	0.84
	Median	0.84	0.84	0.84
Task10	Run 1	0.739	**0.802**	0.769
	Run 2	**0.756**	0.795	**0.77**
	Mean (STD)	0.680 (0.245)	0.677 (0.254)	0.675 (0.246)

Table 1: Official evaluation results for systems presented by the SINAI team

Acknowledgements

This work has been partially supported by Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, LIVING-LANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Cynthia Chew and Gunther Eysenbach. 2010. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118.

Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. Ml-net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, 26(11):1279–1285.

Luis Gasco, Darryl Estrada, Eulàlia Farré, and Martin Krallinger. 2022a. SocialDisNER corpus: gold standard annotations for detection of disease mentions in Spanish tweets. Type: dataset.

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*.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. MarIA: Spanish Language Models. page 22.

Maria Cecilia M. Jalbuena. 2012. Linguistic Features of English in Twitter. *MSEUF Research Studies*, 14(1):1–1.

Pierre Lison, Aliaksandr Hubin, Jeremy Barnes, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. *arXiv preprint arXiv:2004.14723*.

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*.

CR Lyles. 2013. Ló pez a, pasick r, sarkar u."5 mins of uncomfyness is better than dealing with cancer 4 a lifetime": an exploratory qualitative analysis of cervical and breast cancer screening dialogue on twitter. *Journal of Cancer Education*, 28(1):127–33.

Karen McNeil, Paula M Brna, and Kevin E Gordon. 2012. Epilepsy in the twitter era: a need to re-tweet the way we think about seizures. *Epilepsy & behavior*, 23(2):127–130.

Juan Manuel Pérez, Damián A Furman, Laura Alonso Alemany, and Franco Luque. 2021. Robertuito: a pre-trained language model for social media text in spanish. *arXiv preprint arXiv:2111.09453*.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. Technical Report arXiv:2106.09462, arXiv. ArXiv:2106.09462 [cs] type: article.

Lauren Sinnenberg, Alison M. Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina M. Merchant. 2017. Twitter as a Tool for Health Research: A Systematic Review. *American Journal of Public Health*, 107(1):e1–e8.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.