dezzai@SMM4H'22: Tasks 5 & 10 - Hybrid models everywhere

Miguel Ortega-Martín
dezzai, UCM
m.ortega@dezzai.com
m.ortega@ucm.es

Alfonso Ardoiz
dezzai, UCM
alfonso.ardoiz@dezzai.com
aardoiz@ucm.es

Jorge Álvarez
dezzai
jorge.alvarez@dezzai.com

Óscar García-Sierra
dezzai, UCM
oscar.garcia@dezzai.com
oscarg02@ucm.es

Adrián Alonso
dezzai, URJC, Data Science Lab URJC
a.alonso@dezzai.com
adrian.barriuso@urjc.es

Abstract

This paper presents our approaches to SMM4H'22 task 5 - Classification of tweets of self-reported COVID-19 symptoms in Spanish, and task 10 - Detection of disease mentions in tweets – SocialDisNER (in Spanish). We have presented hybrid systems that combine Deep Learning techniques with linguistic rules and medical ontologies, which have allowed us to achieve outstanding results in both tasks.

1 Introduction

The Social Media Mining for Health Applications (SMM4H) (Weissenbacher et al., 2022) workshop aims to promote automatic methods for mining social media data for health informatics. In order to support Spanish Natural Language Processing (NLP), we have focused on both Spanish tasks.

The SMM4H 2022 task 5 focuses on classification of tweets containing self-reports of COVID-19 symptoms in Spanish. It consists of a triple classification task in which participants have to distinguish personal symptoms from symptoms reported by others and references to news articles or other sources, giving rise to three labels: non-personal reports, news and literature mentions, and self-reports.

In the case of SMM4H task 10 - Detection of disease mentions in tweets – SocialDisNER (in Spanish) (Gasco et al., 2022), the task is focused on the recognition of disease mentions in tweets written in Spanish, with the aim of using social media to better understand societal perception of diseases. Therefore, we have tackled this task as a NER offset detection and token classification problem.

Our contributions to these tasks are the following:

- remarkable results for both tasks.
- an analysis of the tasks and the datasets used in them.
- a comparison of several approaches, finding that combining Deep Learning models with linguistic and clinical knowledge achieves the best results.

2 System description

2.1 Task 5

Our pipeline is based on a linguistic preprocessing that allows us to split the tweets in two subgroups, with each of which we use a different RoBERTa (Liu et al., 2019) classifier.

In first place we apply a filter where two subsets of the training set are created. For each tweet we analyze the number of "first person flags", that is, the number of first person verbs and pronouns, and the number of "other person flags", that is, the number of verbs which are not in first person. Then we use a threshold to compare this values: if at least a third of the total flags correspond "first person flags" the tweet is added to subset 1, and else to subset 2. This way only the tweets which the system filter as subset 1 can be labeled as self-report.

In second place, we trained 2 XLM Roberta base classifiers: one ternary classifier trained on all 3 labels from training set, and one binary classifier with only the "news and literature mentions" and the "non-personal report" labeled tweets from the training set. Subsequently, during inference, we apply the previously explained filter on the test dataset. In the end the 3-label classifier is used for subset 1 and the binary classifier for subset 2. This

Proceedings of the 29th International Conference on Computational Linguistic, pages 7–10
October 12–17, 2022.

Model	Post-processing	Precision	Recall	F1 score
RoBERTa	no	0.8434	0.8434	0.8434
RoBERTa	yes	0.8445	0.8445	0.8445
Hybrid	yes	0.8487	0.8487	0.8487

Table 1: Task 5 validation results

Model	Nº predicted	Overlap P	Overlap R	Overlap F1	Strict P	Strict R	Strict F1
RoBERTa	4270	0.951	0.948	0.949	0.852	0.856	0.854
Hybrid	4364	0.939	0.957	0.948	0.856	0.874	0.865

Table 2: Task 10 validation results

way we achieved slightly better results than using just one ternary classifier for all the tweets.

As seen in table 1, we found that cleaning emojis and hashtags slightly improved validation results.

2.2 Task 10

We face this task as a token classification problem with a hybrid system where we combine the outputs from a Transformer model and a medical rule-based NER model. However, this architecture focus heavily in the Transformer side and use the rule model just as a complement. Therefore, we have separately submitted the hybrid system approach results and Transformer only approach results.

As Transformer model we use multilingual XLM-RoBERTa-base (Conneau et al., 2019) as the base model in order to address the multilingual annotation stance of some tweets. This model is fine-tuned in order to adapt it to the token classification task using the IOB tag format strategy (Ramshaw and Marcus, 1995). To perform this strategy, a pre-processing was needed in order to convert all the tokens in the training dataset to this tag set format.

Additionally, a medical rule-based system is built using spaCy utilities and UMLS annotations (Bodenreider, 2004). In particular a pre-built Spanish spaCy model is used and a matcher using all the terms associated to the semantic type T047 "Disease or Syndrome" is built. This technique allows us to find more rare diseases that the Transformer model is not able to recognize since it has not been trained on them. However, this rule-based matcher cannot deal with orthography errors or entities in other languages.

2.3 Model hyperparameters

Models for both tasks were fine-tuned using 4 Nvidia Tesla v100 32GB. Our experiments showed that task 5 models converge after 3 epoch and task 10 model after 5 epochs. Regarding the hyperparameters, the following optimized the evaluation on the development sets for both tasks: batch size = 32, Adam epsilon = 1e-8, learning rate = 5e-5, warm up ratio = 0.1, weight decay = 0.0.

3 Dataset, results and error analysis

3.1 Task 5

Task 5 dataset contains 10,052 training tweets (1,654 labeled as self-reports, 2,413 labeled as non-personal reports and 5,985 as literature/news mentions); 3,578 validation tweets and 6,851 test tweets. Task was evaluated using precision, recall and F1 for the positive class (self-report). Evaluation on validation set showed results seen in table 1.

We consider the quality and inconsistency in the labeling of the dataset to be one of the main problems we have faced, which in many cases leads to confusion between the self-report and the non-personal-report labels. In this regard, we have identified three groups of errors:

(i) Incorrect inclusion of tweets with no symptoms in the training, validation and test sets.

Model	Post-processing	Precision	Recall	F1 score
RoBERTa	yes	0.8464	0.8464	0.8464
Hybrid	yes	0.8521	0.8521	0.8521

Table 3: Task 5 test results

Model	Nº Predict	Strict P	Strict R	Strict F1
RoBERTa	30066	0.821	0.826	0.824
Hybrid	31297	0.828	0.845	0.836

Table 4: Task 10 test results

(ii) Not a clear answer to how to handle the indirect speech annotations (self-reports and non-personal-reports are assigned unclearly in these cases).

(iii) Inconsistencies when labeling tweets describing other people COVID-19 symptoms (in this case, there are more samples labeled as non-personal reports than self-reports).

We developed a filtering system in order to overcome this annotations problem. The main idea behind it is to reinforce the weak barrier between self-reports and non-personal reports that the training data carries, and to reduce the quantity of false positives that the final model could outcome. As stated before, this way only the tweets which the system filter as group 1 can be labeled as self-report.

As seen in table 3, in the end the model got a good grasp of the task, achieving a 0.8521 test F1 score. We consider that better annotated data could lead the model to a better performance.

3.2 Task 10

Task 10 dataset contains 5000 train tweets, 2500 validation tweets and 23430 (2000) test tweets. Task was evaluated with Strict Precision, Recall and F1-score as Leaderboard Scores, and with Overlapping Precision, Recall and F1-score as Additional Scores.

We have encountered some minor issues of wrong entity labeling through the dataset. For instance, there are tweets where a term mainly related with diseases is used with another meaning and therefore should not be labeled as a disease; some tweets in both Spanish and English which contain annotated diseases in both languages; or incoherence in the annotation of non-strict diseases like "fumar" (in English, "smoke") or "enfermedad" (in English, "disease") which are only annotated in some cases.

Table 2 shows our results in the validation dataset. As it can be seen, there is a huge contrast between strict and overlapping scores. This is mainly caused by our approach as a token-based classification problem, which heavily relies on the RoBERTa tokenization process, where the first part of many entities is not classified as entity but as not-entity, therefore the strict score decreases, but the overlapping score keeps its robustness. Table 4 contains our final test results for task 10.

4 Conclusions

In this paper we have reviewed our approaches to SMM4H'22 tasks 5 and 10. For both tasks we have developed hybrid systems combining Deep Learning techniques with rule-based modules. We have analysed our results as well as both datasets, and presented some insights of each one. We also reviewed the limitations of our models and highlighted their weaknesses and strengths. As seen in table 3, in the end we achieved a score of 0.8521 in the test set for task 5. For Task 10, as seen in table 4, we achieved a Strict Precision of 0.828, a Strict Recall of 0.845 and a Strict F1 of 0.836. For both tasks these represents remarkable and top-of-the-chart results.

We consider this two tasks specially relevant due to the need to promote NLP in Spanish, and even more so when applied to the clinical domain using real data.

References

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

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. *CoRR*, abs/1911.02116.

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*.

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*.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. *CoRR*, cmp-lg/9505040.

Davy Weissenbacher, Ari Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting, 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.*