CASIA@SMM4H'22: A Uniform Health Information Mining System for Multilingual Social Media Texts

Jia Fu[1,2,†], **Sirui Li**[1,3,†], **Minghui Yuan**[1,4], **Zhucong Li**[1,2], **Zhen Gan**[1,4], , **Yubo Chen**[1,2], **Kang Liu**[1,2], **Jun Zhao**[1,2], **Shengping Liu**[5]
[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Department of Computer Science, Emory University, Altlanta
[4] Beijing University of Chemical Technology
[5] Beijing Unisound Information Technology Co., Ltd
{yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn
sirui.li@emory.edu, ganzhen@mail.buct.edu.cn
{fujia2021, lizhucong2019}@ia.ac.cn, liushengping@unisound.com

Abstract

This paper presents a description of our system in SMM4H-2022, where we participated in task 1a, task 4, and task 6 to task 10. There are three main challenges in SMM4H-2022, namely the domain shift problem, the prediction bias due to category imbalance, and the noise in informal text. In this paper, we propose a unified framework for the classification and named entity recognition tasks to solve the challenges, and it can be applied to both English and Spanish scenarios. The results of our system are higher than the median F1-scores for 7 tasks and significantly exceed the F1-scores for 5 tasks. The experimental results demonstrate the effectiveness of our system.

1 Introduction

A large amount of health-related data exists on social media, and this data has attracted a lot of attention in medical applications. The Social Media Mining for Health Applications 2022 (SMM4H-2022) Shared Task(Davy Weissenbacher, 2022) involves natural language processing (NLP) challenges of using social media data for health research, including informal, colloquial expressions and misspellings of clinical concepts, noise, data sparsity, ambiguity, and multilingual posts. We participated in seven English classification tasks (i.e., task 1a, 4, 6, 7, 8, 9) and one Spanish named entity recognition task (i.e., task 10). The classification tasks focus on the classification of tweets mentioning drug changes, tweets indicating self-reported COVID-19 vaccination status, Reddit posts self-reporting exact age, etc. The named entity recognition task is designed to detect the diseases mentioned in the Spanish tweets.

There are three challenges in the SMM4H 2022 shared tasks. The first one is the domain shift problem. When fine-tuning specific downstream tasks using large-scale pre-trained models, the increasing size of the pre-trained models leads to the domain bias problem. To address this problem, we apply continue pre-training (Gururangan et al., 2020) to enhance the adaptive ability of the model in the corresponding domain. Also, we incorporate FGM adversarial training and Child-Tuning (Xu et al., 2021) to improve the robustness of the model. The second challenge is the prediction bias due to category imbalance. The category imbalance in the annotated data affects the focus of the model learning, which eventually leads to prediction bias and also hinders the robustness and generalization of the model. Therefore, we apply oversampling to generate samples from the minority class to achieve data balance. The third challenge is the noise in informal text. When learning from informal social media text data, the text contains a large number of emojis, mentioned usernames, and hyperlinks, which can affect the model's ability to learn the true semantic information in the text. Thus, we apply data cleaning to remove the noise from the data.

[†]First Author and Second Author contribute equally to this work.

Proceedings of the 29th International Conference on Computational Linguistic, pages 143–147
October 12–17, 2022.

Therefore, addressing the three challenges, the three approaches of this paper can be summarized as below:

- Applying continue pre-training and adding FGM adversarial training and Child-Tuning in fine-tuning to address the domain shift problem.
- Applying oversampling to generate more data to address the category imbalance problem.
- Applying data cleaning to solve the noise problem in the corpus.

2 Task Description

2.1 Classification Tasks

We participated in seven English classification tasks including task 1a: Classification of adverse events mentions in tweets;Task 4: Classification of tweets self-reporting exact age; Task 6: Classification of tweets which indicate self-reported COVID-19 vaccination status; Task 7: Classification of self-reported intimate partner violence on Twitter (Al-Garadi et al., 2022); Task 8: Classification of self-reported chronic stress on Twitter; and Task 9: Classification of Reddit posts self-reporting exact age. All the tasks are binary classification and use F1-score for the positive class as the evaluation metric. The training set and evaluation set are released by the organizers, while the testing set is provided afterward.

2.2 Named Entity Recognition Task

Task 10 focuses on detecting drug mentions and entity spans in Spanish tweets(Gasco et al., 2022b). The task aims to use social media as a proxy to better understand the societal perception of disease, from rare immunological and genetic diseases such as cystic fibrosis, highly prevalent conditions such as cancer and diabetes, to often controversial diagnoses such as fibromyalgia and even mental health disorders. Only one class of entities is included in the tweets. The dataset consists of a training set, a validation set and a test set with 5,000, 2,500 and 23,430 data respectively (Gasco et al., 2022a).

3 Methods

3.1 Overall Framework

A uniform framework is used for all the classification and named entity recognition tasks in both English and Spanish. The framework includes three major components: pre-processing social media data, continuing pre-training transformer-based models, and fine-tuning with three strategies.

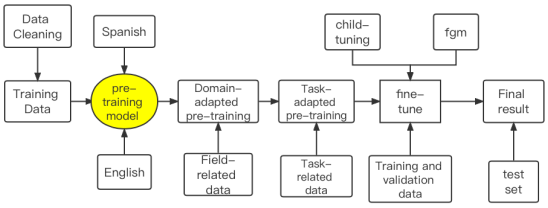

Figure 1: Overall system structure.

First, we performed data cleaning to reduce the noise of the tweets and Reddit posts in both languages.

Then, we continued domain-adaptive pre-training using unlabeled data from all the related tasks, aiming to tailor the pre-trained model to the domain of the tasks. Specifically, we continued pre-training RoBERTa_base with data from all the English tasks for 20 rounds, using a smaller learning rate (e.g., 6e-5). We continued pre-training bert_spanish_cased_finetuned_ner (Cañete et al., 2020) using data from all the Spanish tasks, followed by 20 rounds of continued pre-training using large-scale (85k) unlabeled Spanish tweets with mentions of disease.

Finally, for each task, we fine-tuned the model using its training set and validation set. We adopted FGM adversarial training, Child-Tuning, and five-fold cross-validation in this stage to improve the performance. For certain classification tasks, we oversampled the minority class to overcome class imbalance and fine-tuned the model for 24 rounds. For the named entity recognition task, we fuse 15 models with different learning rates and different training epochs to obtain the final results. Our overall system structure is shown in Figure 1.

4 Our Method

4.1 Pre-processing

Data Cleaning: There are a large number of emojis, usernames and hyperlinks in tweets and Reddit posts. For the model to better learn the semantic information in the text, we perform three cleaning operations on the data: replacing all emojis with *emoji*, usernames with @user, and hyperlinks with http.

Oversampling: The annotated data for task 1a,task 6, task 7 are significantly imbalanced where only around 10% of the data are from the positive class

(task 1a: 7%,task 6: 11%, task 7: 11%). To overcome the class imbalance problem, we performed oversampling on the training set. Specifically, random samples are duplicated from the minority class until the selected ratio between the positive class and negative class is reached.

4.2 Training Method

Five-fold Cross-voting: Utilizing five-fold cross-validation, we divide the training set into five different datasets, and the inconsistencies of entity labeling in each dataset varies. With the same model structure, we train five models on five training sets and integrate their prediction results on the same test set through majority voting.

Adversarial Training: In order to obtain a model with better robustness, we employ adversarial training to improve the model's stability. Referring to the FGM (Miyato et al., 2016) adversarial training mechanism, we directly impose a small disturbance on the embedding representation of the model and assume the embedding representation of the input text sequence $[v_1, v_2, \ldots, v_T]$ as x. Then the small disturbance r_{adv} applied each time is:

$$r_{adv} = \epsilon \cdot g/\|g\|_2 \quad (1)$$

$$g = \nabla_x L(\theta, x, y) \quad (2)$$

The formulas' meaning is to move the input one step further in the direction of rising loss, which causes the model loss to rise in the fastest direction, generating an attack. In this case, the model must identify more robust parameters in the optimization phase to deal with the attacks against the samples.

Child-Tuning: Nowadays, pre-trained models are becoming increasingly large, and when using pre-trained models to fine-tune on a smaller amount of downstream data, overfitting often occurs, resulting in poor model generalization on downstream tasks. To address the problem, Child-Tuning proposes a new fine-tuning method.in backward propagation, instead of adjusting all the parameters, only a part of them is updated, which is the child network.

Child-Tuning has two algorithms: Child-tuning-F and Child-Tuning-D. The Child-Tuning-F approach is to randomly discard a portion of the gradient when the network parameters are updated in the fine-tuning process.

$$w_{t+1} = w_t - \eta \frac{\partial \zeta(w_t)}{\partial w_t} \odot M_t \quad (3)$$

$$M_t \sim Bernoulli(P_F) \quad (4)$$

Task	AdamW	ChildTuning_F
Task 1a	0.7833	**0.8095**
Task 4	**0.9229**	0.9169
Task 6	**0.7089**	0.6953
Task 7	0.7379	**0.7500**
Task 8	**0.7742**	0.7662
Task 9	**0.8989**	0.8964

Table 1: F1 scores of the classification tasks using different optimizing strategies. Within the same task, all other parameters remain the same.

The Child-Tuning-D approach is to select a child network whose policy can be adaptively adjusted for different downstream tasks, selecting the most relevant and important parameters for the downstream tasks to act as the child network. The Fisher Information Matrix (FIM) (Tu et al., 2016) is introduced to estimate the importance of each parameter to the downstream task.

$$\mathbf{F}^{(i)}(\mathbf{w}) = \frac{1}{|D|} \sum_{j=1}^{|D|} \left(\frac{\partial \log p(y_j|\mathbf{x}_j; \mathbf{w})}{\partial \mathbf{w}^{(i)}} \right)^2 \quad (5)$$

4.3 Continue Pre-training

Due to the increasing size of pre-trained models, domain bias is a problem when fine-tuning for specific downstream tasks using large-scale pre-trained models. This can be well addressed by continuing pre-training. Continue pre-training can be interpreted as the continuation of domain-adaptive or task-adaptive training of a model based on a large-scale pre-trained language model for a specific corpus of downstream NLP tasks. Continued pre-training can be classified into two types:

- Domain-adapted pre-training (DAP), which improves the performance of the model for the corresponding domain-specific task in both low- and high-resource cases
- Task-adapted pre-training (TAP), which is very efficient in improving the performance of the model on specific tasks despite the lack of a small corpus.

5 Experiment

5.1 Classification Tasks

We trained each model with a batch size of 16 for 24 epochs. For RoBERTa_base parameters, we set the learning rate to 3e-5 and L2 normalization's weight to 0.01; for other parameters, we set

Task	TRAIN	TRAIN_1	TRAIN_05
Task 1a	0.7183	0.7343	**0.7368**
Task 6	0.7500	0.7538	**0.7608**
Task 7	0.7015	0.7379	**0.7619**

Table 2: F1 scores of the classification tasks with imbalanced data using oversampling. TRAIN is the training set with the original class distribution. TRAIN_1 is the oversampled set in which the ratio between the positive and negative class is 1 to 1. TRAIN_05 is the other oversampled set in which the ratio between the positive and negative class is 0.5 to 1.

Task	RoBERTa_base	cp-RoBERTa_base
Task 1a	**0.8095**	0.7595
Task 4	**0.9194**	0.7206
Task 6	0.7127	**0.7249**
Task 7	**0.7379**	0.7170
Task 9	0.8973	**0.8989**

Table 3: F1 scores of the classification tasks using continue pre-training. The continue pre-trained RoBERTa_base model is represented by cp-RoBERTa_base.

the learning rate to 3e-4 and L2 normalization's weight to 0. We fine-tuned the all models using five-fold cross-validation on the training set. We experimented on two optimizers, namely AdamW (Loshchilov and Hutter, 2018) and Child-Tuning. The experiments' results are shown in Table 1.

For tasks with significantly imbalanced data (task 1a, 6, 7), we oversampled the minority class with ratios of 1 to 1 and 0.5 to 1 (i.e., the ratio between the number of samples in the minority class and majority class is 0.5 to 1). For other tasks (task 4, 8, 9), we kept the original data and oversampled the minority class with a ratio of 1 to 1. We experimented on these ratios, and the results are shown in Table 2.

Furthermore, we experimented with using the continue pre-trained RoBERTa_base model (Liu et al., 2019). The model is pre-trained on the data from all the English tasks to acquire domain knowledge. The results are shown in Table 3.

For each task, we submitted the predictions of models with the best performances on the validation set. The performances of our framework are above the median and mean for all tasks except task 9. Significantly, our model the mean of task 1a by 7%, exceeds the median of task 4 by 4.7%. exceeds the median of task 7 by 7%, and exceeds the median of task 8 by 3%. The final results are shown in Table 4.

Task	Validation F1	Test F1	Mean F1	Median F1
Task 1a	0.809	0.638	0.562	-
Task 4	0.929	0.916	0.847	0.869
Task 6	0.729	0.78	-	0.77
Task 7	0.780	0.833	-	0.763
Task 8	0.774	0.781	-	0.750
Task 9	0.897	0.885	0.901	0.891

Table 4: Our results, mean results, and median results of each classification task.

Model	Validation F1	Test F1
60	0.9104±0.11	-
20+60	0.9105±0.13	-
20+60+Child-Tuning-F	0.9121±0.21	-
20+60+Child-Tuning-D	0.9064±0.15	-
Final Result	-	0.891

Table 5: Results on the SMM4H-task10 validation set and test set.

5.2 Named Entity Identification Task

For task 10, the BERT model was fine-tuned to 60 epochs on the training set with a learning rate of 1e-5 and a batch size of 12, and we continued to pre-train 10 or 20 epochs on large-scale unlabeled data with a learning rate of 6e-5, followed by adversarial training and Child-Tuning training. The experimental results are shown in the following table. Here, "60" means the BERT model is trained for 60 epochs, "20+60" means the pre-training is continued for 20 epochs and then fine-tuned for 60 epochs, and "20+60+Child-Tuning-F" means adding Child-Tuning-F. The results are the mean value of the training process plus the variation range. The final result is the result of the fusion of 15 models. The results are shown in Table 5.

6 Conclusion and Future Work

The above tasks in the SMM4H 2022 explore social perceptions of health and diseases using social media data, including classification and information extraction tasks. In this paper, we propose a unified system to solve the classification and named entity recognition tasks, which can be applied to both English and Spanish scenarios. Our framework consists of data cleaning, oversampling, continued pre-training, and applying adversarial training and Child-Tuning for fine-tuning. We evaluated different variants of our framework and demonstrated the effectiveness of the framework. For future work, we will introduce knowledge graphs from the medical field to further improve our system.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0106400), the National Natural Science Foundation of China (No.62176257， 61976211). This work is also supported by the CCF-Tencent Open Research Fund and the Youth Innovation Promotion Association CAS.

References

Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2022. Natural language model for automatic identification of intimate partner violence reports from twitter. *Array*, page 100217.

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*.

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 Graciela Gonzalez-Hernandez Davy Weissenbacher, Ari Z. Klein. 2022. Overview of the seventh social media mining for health applications (#smm4h) shared tasks at coling 2022. In *In Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop & Shared Task.*

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*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

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*.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Ming Tu, Visar Berisha, Martin Woolf, Jae-sun Seo, and Yu Cao. 2016. Ranking the parameters of deep neural networks using the fisher information. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2647–2651. IEEE.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. *arXiv preprint arXiv:2109.05687*.