

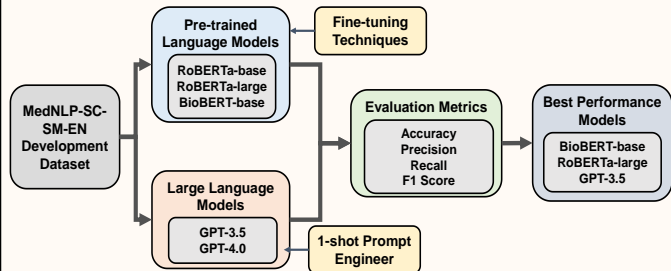


Hsiao-Chuan Liu¹, Vidhya Nataraj², Chia-Tung Tsai¹, Wen-Hsuan Liao¹, Tzu-Yu Liu³, Mike Tian-Jian⁴ Jiang and Min-Yuh Day¹, *
*myday@gm.ntpu.edu.tw

¹Information Management, ²Smart Healthcare Management, ³Business Administration, National Taipei University, New Taipei City, Taiwan
⁴Zeals Co., Ltd. Tokyo, Japan

The IMNTPU team engaged in the NTCIR-17 RealMedNLP task, specifically focusing on Subtask1: Adverse Drug Event detection (ADE) and the challenge of identifying related radiology reports. This task is centered on harnessing methodologies that offer significant aid in real-world medical services, especially when training resources are limited. In our approach, we harnessed the power of pre-trained language models (PLMs), particularly leveraging models like the BERT transformer, to understand both sentence and document structures. Our experimentation with diverse network designs based on PLMs paved the way for an enlightening comparative analysis. Notably, BioBERT-Base emerged as a superior contender, showcasing commendable accuracy relative to its peers. Furthermore, our investigation made strides in the realm of one-shot learning for multiclass labeling, specifically with the GPT framework. The insights gathered emphasized the necessity for more specialized strategies, suggesting avenues for future research in multiclass labeling tasks.

Research Architecture



Performance

Exact Accuracy of Test Dataset and Development Dataset		
Models	Development Dataset (#1,192)	Test Dataset (#1,993)
BioBERT-Base (Submission Run 1)	0.92	0.82
Roberta-Base	0.76	-
Roberta-Large (Submission Run 2)	0.85	0.81
GPT3.5 (Submission Run 3)	0.72	0.69
GPT 4.0	0.62	-

Subtask 1-SM-ADE-EN Binary and Per Label Performance Metrics							
Models	Score	Precision		Recall		F1 Score	
		ADE	NO ADE	ADE	NO ADE	ADE	NO ADE
BioBERT-Base	Binary	0.74	0.91	0.78	0.89	0.76	0.90
	Per label	0.72	1.00	0.76	0.99	0.74	0.99
RoBERTa-Large	Binary	0.73	0.93	0.83	0.88	0.78	0.90
	Per label	0.71	1.00	0.77	0.99	0.74	0.99
GPT3.5	Binary	0.47	0.47	0.20	0.91	0.28	0.82
	Per label	0.42	0.98	0.18	1.00	0.25	0.99

Subtask 1-SM-ADE-EN Binary and Per Label Performance Metrics in Development Dataset							
Models	Score	Precision		Recall		F1 Score	
		ADE	NO ADE	ADE	NO ADE	ADE	NO ADE
BioBERT-Base	Binary	0.90	0.97	0.93	0.96	0.92	0.97
	Per label	0.91	0.97	0.93	0.96	0.92	0.97
RoBERTa-Base	Binary	0.89	0.83	0.59	0.97	0.71	0.90
	Per label	0.85	0.99	0.49	1.00	0.62	0.99
RoBERTa-Large	Binary	0.87	0.93	0.85	0.94	0.86	0.93
	Per label	0.84	0.99	0.77	1.00	0.81	1.00
GPT3.5	Binary	0.72	0.80	0.53	0.90	0.61	0.85
	Per label	0.62	0.99	0.57	0.99	0.60	0.99
GPT4.0	Binary	0.52	0.96	0.95	0.61	0.67	0.75
	Per label	0.48	1.00	0.92	0.98	0.63	0.99

Exact Match Accuracy Results in Development Dataset			
Models	Accuracy	Models	Accuracy
RoBERTa-Base	0.86	GPT3.5-1-shot	0.54
RoBERTa-Large	0.87	GPT3.5-scenario	0.69
BioBERT-Base	0.85	GPT4.0-1-shot	0.61
BioBERT-Large	0.86	GPT4.0-scenario	0.70
		GPT3.5-fintune	0.85

Prompt Engineer

One-shot Learning

- Analysis showed reduced accuracy in insight extraction from short, ambiguous tweets.
- GPT models often over-labeled: GPT-3.5 labeled 929 instances, GPT-4.0 labeled 789, while the actual ground truth was 400.

Prompt

You are a medical expert analyzing tweets to check whether the user suffers adverse drug events.

****Scenario****: Because every text is from Twitter, the texts are short. Please consider this situation and annotate the text with proper labels to check whether the user suffers adverse drug events. For instance, users list the adverse drug effects rather than express personal experiences of adverse reactions.

****Your annotating steps are as follows: ****

1. Check whether the user lists the adverse drug effects rather than expressing personal experiences of adverse reactions.

2. Check whether this tweet's user suffers from adverse drug events.

3. Check the symptoms in these 22 symptoms listed below.

****Your annotation should be in the following format:****

1. If the user suffers from the tweet's symptom instead of listing the adverse drug events, output with the corresponding label.

2. If the user doesn't suffer from the symptom in the tweet, output with \"None\".

****Symptom Labels****:

nausea, diarrhea, ..., rash, stomatitis

****Here is some annotate example for you to base on.****

Text: I finished C due to side effects of the contrast dye. I was feeling kind of sick and nausea was getting worse, so I thought it would be tough, but this morning my chest hurts... I'm going to go to the hospital tomorrow, though I'm anxious because I have 2 hours left until the test results...
Label: nausea, pain

Other 22 samples for GPT to know.````

Fine-tuning Techniques

Hyperparameters

- Fine-tuned for multi-label text classification
- Max Epochs: 10
- Max Sequence Length: 512
- Learning Rate: 5e-5
- Batch Size: 16
- Loss Func.: BCEWithLogitsLoss

Conclusions and Contributions

- Expanded dataset with GPT-3.5 boosts RoBERTa accuracy from 0.76 to 0.86.
- Refined prompts for tweets, increasing GPT4.0 accuracy to 0.70.
- BioBERT excels in drug event extraction; GPT one-shot learning shows limits.
- We proposed two critical elements, Prompt Engineer and Fine-Tuning Techniques. GPT-4.0 showing enhanced performance in ambiguous datasets.

ACKNOWLEDGMENT

This research was supported in part by the National Science and Technology Council (NSTC), Taiwan, under grants MOST 110-2410-H-305-013-MY2, NSTC 112- 2425-H-305-002-, and NSTC 112-2627-M-038-001-, and National Taipei University (NTPU), Taiwan under grants 112-NTPU-ORDA-F-003, 112- NTPU-ORDA-F-004, USTP-NTPU-TMU-112-01, NTPU-112A413E01, and NTPU-112A513E01.