NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001296

IMNTPU at the NTCIR-17 Real-MedNLP Task: Multi-Model Approach to Adverse Drug Event Detection from Social Media

Hsiao-Chuan Liu
Information Management,
National Taipei University
New Taipei City, Taiwan
s711136108@gm.ntpu.edu.tw

Vidhya Nataraj
Smart Healthcare Management,
National Taipei University
New Taipei City, Taiwan
vidhyanataraj99@gmail.com

Chia-Tung Tsai
Information Management,
National Taipei University
New Taipei City, Taiwan
s711136106@gm.ntpu.edu.tw

Wen-Hsuan Liao
Information Management,
National Taipei University
New Taipei City, Taiwan
s711136109@gm.ntpu.edu.tw

Tzu-Yu Liu
Business Administration,
National Taipei University
New Taipei City, Taiwan
s411084008@gm.ntpu.edu.tw

Mike Tian-Jian Jiang
Zeal Co., Ltd
Tokyo, Japan
tmjiang@gmail.com

Min-Yuh Day*
Information Management,
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

ABSTRACT

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.

KEYWORDS

Medical Natural Language Processing, Adverse Drug Event, Large Language Models, Social Media, Multi-class Classification.

TEAM NAME

IMNTPU

SUBTASKS

SM-ADE-EN

1 INTRODUCTION

As technologies are growing over the years, medical histories are updated to electronic format from the level of paper records, along with the increasing significance of information processing technology in the medical field. Among the shared task workshop NTCIR-17 Real-MedNLP task is one such task employing actual medical documents in the case of medical language processing (case reports and radiology reports).[1] This task is widely utilized to endorse the development of practical systems of practical medical NLP tools applicable in the hospital for the medical industry to support medical decisions and help physicians and paramedical staff support various medical services. This remains the ultimate goal for the organizers since they are involved in restructuring the scheme (so-called medical AI task). This task delivers core resources into two; (1) Social media corpus (shortly MedTxt-SM) and this task is divided into two parts:(a) Classification of messages into “contains ADE (22 symptoms)” vs. “does not contain ADE (22 symptoms)”, (b)multi-labeling of symptoms for all documents containing ADEs.2) Radiology-Report corpus (shortly MedTxt-RR).

It's fascinating to learn about IMNTPU's involvement in the Real-MedNLP track this year, particularly in the subtask focusing on the social aspect of Adverse Drug Event (ADE) detection from English social media texts. Their approach involved constructing pre-trained language models (PLM). The paper will likely provide detailed insights into their methodologies and will discuss the official results. In the context of ADE, the primary objective of

this task is to extract ADE information from case reports and determine whether a particular drug has elicited a positive or negative response. In the realm of pre-trained models, there have been various domain-specific models established previously, including Biobert[2], Roberta large, and the Generative pre-trained model (GPT) 3.5. These models have shown promise in tasks related to biomedical and healthcare text processing. IMNTPU's work in this area is contributing to advancements in understanding and handling adverse drug events, particularly in the context of social media data. This research is instrumental in enhancing the capabilities of natural language processing models in healthcare-related applications.

These domain-specific Bidirectional Encoder Representations from Transformers (BERT) based models are predicted to produce an improved result in the tasks. We also compared 3 different pre-trained models and evaluated the results.

The following section of this paper is organized as follows Section 2, defines the related studies whereas Section 3, introduces our approaches for each of the subtasks. Section 4 provides a comprehensive account of the official experiment results and a detailed analysis. Finally, in Section 5, we present our conclusions obtained from this study.

2 RELATED WORKS

2.1 BioBERT-Base

BioBERT becomes more adept at handling biomedical terminology, scientific jargon, biomedical question answering, and context-specific information [2,3]. It is constructed on the BERT architecture, which is pre-trained on a large corpus of general text data.

2.2 Roberta base and large

We assessed the RoBERTa transformer model (Liu et al., 2019)[4] in both of its base and large versions for our evaluation. Employing the pre-trained Roberta model, a study carried out a comprehensive investigation to categorize English tweets discussing COVID-19 vaccination. The central goal was to differentiate between tweets that mentioned one or more ADE and those with no ADE [5] and evaluated as the best model for classification and showed a good performance.

2.3 GPT 3.5 and GPT 4.0

A remarkable capability in both general and health domains, large language models (LLMs) like GPT-3.5 and GPT-4.0 are capable of processing image and text inputs to generate text outputs. They have impressed researchers with their impressive zero-shot and few-shot abilities, as evidenced in recent studies (OpenAI, 2023; Bubeck et al., 2023) [6,7], and have also been proven in health-related contexts. (Nori et al., 2023) [8]. Significant, Gu et al., conducted a case study on the extraction of adverse drug events (ADE), which is an important domain for enhancing healthcare [9] and this model achieved comparable accurateness as administered state-of-the-art models without the need for labeled data.

2.4 Prompt Engineering for LLMs

Prompt Engineering is increasingly recognized for its crucial role in ensuring effective communication with Large Language Models (LLMs). A "prompt" serves as a specific command to LLMs, aiming to enforce particular rules, automate processes, and maintain the quality and scope of outputs. In essence, prompts can be viewed as a programming approach, allowing users to tailor their interactions and desired results with LLMs. The significance of efficient prompt design in LLMs is widely acknowledged in academia [10,11] Past studies have delved into the precise effects of prompts on AI generative models. For instance, Wang, et al.[12] assessed the efficiency of prompts in literature searches, and in a more specialized domain, Xia and Zhang [13]concentrated on software prompt design and its role in error correction.

3 PROPOSED METHODS

In this section, we will first introduce the PLMs used to compare this work in Section 3.1.1, and Section 3.1.2 presents tokenization, Section 3.1.3 is the prompt engineer for GPT.

3.1.1 Finetuned Pretrained Language Models (PLMs)

The application of PLMs like BERT transformers-based large-scale language models has witnessed remarkable achievements across various domains in the field of NLP as well as in the domain specific[14]. Additionally, recent research revealed a significant and increasing interest in large language models, primarily focused on NLP applications[15].

In this task, we adopted 3 different types of PLM models:

Table 1 Model Summary

Models	Pre-trained Data	Backbone	Features
BioBERT-Base	PubMed abstracts and PubMed Central full-text articles	BERT	Mixed Domain
Roberta-Base	Wikipedia, books, and publicly available data	BERT	Mixed Domain
Roberta-Large	Wikipedia, books, and publicly available data	BERT	Mixed Domain
GPT 3.5	Wikipedia, books, and publicly available text data	GPT 3	Mixed Domain
GPT 4.0	Wikipedia, books, and publicly available text data	GPT 3.5	Mixed Domain

Table 2 The Main Hyperparameters for Subtask 1-ADE- EN

Hyperparameters	BioBERT-Base	Roberta-Base	Roberta-Large	GPT 3.5
Batch Size	16	16	16	Prompt Engineering for LLMs
Optimizer	Adam	Adam	Adam	
Learning Rate	2e-5	5e-5	5e-5	
Epoch	10	3	3	

Table 3 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	-

```
You are a medical expert analyzing tweets to check whether the user suffers adverse drug events.
**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, fatigue, vomiting, loss of appetite, headache, fever, interstitial lung disease, liver damage, dizziness, pain, alopecia, analgesic asthma syndrome, renal impairment, hypersensitivity, insomnia, constipation, bone marrow dysfunction, abdominal pain, hemorrhagic cystitis, 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
Text: Why do I feel so dizzy after taking an antibiotic called minocycline!!!! I have to take it again tomorrow, but what if I don't take it?
Label: dizziness
Text: <user_name> Ah, I knew it, you have aspirin asthma! I also coughed and threw up when I took Loxonin. I also had difficulty breathing just because my throat was swollen (sweat).
Label: vomiting
Text: Day 2 of cisplatin administration. The side effect of nausea has subsided, but the diarrhea and loss of appetite was so bad that I've been eating all yesterday and now this morning I'm having an upset stomach again...
```Other 15 samples for GPT to know.```
```

Figure 1 Prompt used for GPT3.5 and 4.0

BioBERT[2,3], Roberta-large, and GPT 3.5 (Table 1). Their model structure is the same as BERT, but the differences are that a pre-trained model undergoes initial training to learn general language features, while a large language model is characterized by its substantial parameter count.

We treat the ADE task as text classification, aiming to predict whether each text corresponds to an ADE. We focus on to identify the two categories: "NO ADE" (0) or "ADE" (1). Our approach involves fine-tuning a pre-trained model using ADE-related data from tweets containing ADE information. This fine-tuning helps the model better distinguish ADE-related text. A list of each ADE entity (case number, case report text, and ADE entity label) was established as training data. The development dataset is 7,964 records. The training data were 90 % case, 10% cases were utilized for data validation. The hyperparameters of the models were max epochs 10, learning rate 1e-5, max length 128, batch size 32, and optimizer Adam. (Table 2)

Table 4 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**

3.1.2 Tokenization

We used a tokenizer to convert text to numerical inputs with a maximum length of 512 tokens and applied truncation and padding to ensure uniformity. This preprocessing step is used to bridge the gap between raw text and machine learning models. For prediction, the logits (raw model outputs) are passed through the sigmoid function to calculate the predicted probabilities ('batch probs') for each class. This is done using the sigmoid function from PyTorch's ('nn. functional') module.

3.1.3 Prompt Engineer for Large Language Models (LLMs)

We used a one-shot prompt approach to guide both GPT-3.5 and GPT-4.0 regarding the definition and expected results from each dataset. Below is the prompt given to GPT. (Refer to Figure 1) Upon comparing the accuracy of the development dataset for GPT-3.5 and GPT-4.0, it's evident that GPT-3.5 performs better under the same prompt in Table 3. Therefore, we chose GPT-3.5 for our submissions.

4 EXPERIMENTAL RESULTS

4.1 Experiment Structure

For the evaluation on the MedNLP SM-ADE-EN test dataset, we presented three distinct runs for comparative analysis. However, during our developmental phase, we assessed five different models to determine the most suitable for our final submission.

Figure 2 shows the experiment structure of this study by using the development dataset. We split the development data into three parts, 70% (5,580) for training, 15% (1,192) for validation. 15% (1,192) for testing. The foundation of our pre-training model is the BERT-based transformer. The variations among these models are primarily attributed to their structures and hyperparameters. Among the LLMs we evaluated, GPT 3.5 and GPT 4.0 were our selections. Below is a detailed description of the five models considered in our research:
```

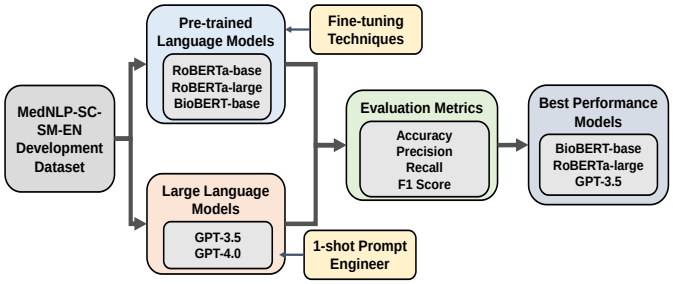

Figure 2 Experiment Structure of MedNLP SM-ADE-EN Task

1) BioBERT-Base: This model did not employ prompt learning and was designed with 22 labels complemented by corresponding text sentences.

2) Roberta-Base: This model was fine-tuned with specific hyperparameters tailored to cater to the multi-label text classification task.

3) Roberta-Large: Similar to Roberta-Based, this model was adapted with fine-tuned hyperparameters suitable for the multi-label text classification objective.

4) GPT3.5: With the application of prompt engineering combined with one-shot learning, this model utilized LLM to detect ADEs.

5) GPT4.0: Employing strategies akin to GPT3.5, this model also leveraged prompt engineering and one-shot learning, harnessing the capabilities of LLM to pinpoint ADEs.

4.2 Submitted Runs

In our experiment, we designated Roberta-Base as the baseline model. We then conducted evaluations on the development dataset using five distinct models: BioBERT-Base, Roberta-Large, Roberta-Base, GPT3.5, and GPT 4.0. The exact accuracies for both the Test and Development Datasets are presented in Table 3.

From Table 3, it's evident that BioBERT-Base achieved the highest accuracy on the development dataset with a score of 0.92 and maintained commendable performance on the test dataset with an accuracy of 0.82. This makes it the top performer among the models evaluated for Subtask1-SM-EN. On the other hand, Roberta-Base, which served as our baseline, showed a marginal improvement from the development dataset (0.76) to the Test Dataset (0.81).

We chose to submit the results from three models: Roberta-Large, BioBERT-Base, and GPT3.5. This decision was made by the comparative performance, as the remaining two models, particularly GPT 4.0, lagged in terms of accuracy on the development dataset.

For a comprehensive evaluation, Table 4 presents the performance metrics on the test dataset, while Table 5 emphasizes individual model performance across different classes.

4.3 Error Analysis

Table 6 shows the detail binary and per label performance metrics in development dataset. Following session is the error analysis of different evaluation metrics.

A. Precision Analysis

Table 5 Subtask1-SM-ADE-EN Per label (Individual) Submitted Per Label Score

Symptoms	Precision			Recall			F1 Score		
	BioBERT	**Roberta-L**	**GPT3.5**	**BioBERT**	**Roberta-L**	**GPT3.5**	**BioBERT**	**Roberta-L**	**GPT3.5**
Diarrhea	**0.75**	0.73	0.64	0.83	**0.84**	0.17	**0.79**	0.78	0.27
Fatigue	**0.76**	0.73	0.64	0.89	**0.93**	0.16	**0.82**	0.82	0.26
Vomiting	0.83	**0.84**	0.57	**0.86**	0.73	0.18	**0.84**	0.78	0.28
Appetite	**0.75**	0.69	0.50	0.87	**0.96**	0.15	0.80	**0.81**	0.24
Headache	0.77	0.77	**0.87**	0.95	**0.96**	0.23	0.85	**0.86**	0.36
Fever	**0.63**	0.61	0.40	**0.58**	0.47	0.23	**0.61**	0.53	0.29
Interstitial lung disease	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Liver damage	0.00	0.00	**0.17**	0.00	0.00	**1.00**	0.00	0.00	**0.29**
Dizziness	0.53	0.50	**0.67**	**0.69**	0.62	0.31	**0.60**	0.55	0.42
Pain	0.53	**0.67**	0.15	0.44	**0.54**	0.15	0.48	**0.60**	0.15
Alopecia	**0.70**	0.67	0.38	**0.88**	0.75	0.38	**0.78**	0.71	0.38
Analgesic asthma syndrome	0.81	0.94	**1.00**	0.94	**0.94**	0.17	0.87	**0.94**	0.29
Renal impairment	0.83	**1.00**	0.00	**1.00**	0.40	0.00	**0.91**	0.57	0.00
Hypersensitivity	0.76	**0.68**	0.38	**0.68**	0.61	0.11	**0.72**	0.64	0.17
Insomnia	**0.84**	0.55	0.60	**0.47**	0.18	0.18	**0.60**	0.27	0.27
Constipation	**0.74**	0.66	0.56	0.84	**0.87**	0.16	**0.79**	0.75	0.25
Bone marrow dysfunction	**0.33**	0.00	0.14	**0.50**	0.00	0.50	**0.40**	0.00	0.22
Abdominal pain	**0.72**	0.64	0.54	0.75	**0.86**	0.23	0.73	**0.74**	0.32
Hemorrhagic cystitis	**0.67**	0.00	0.50	**0.50**	0.00	0.25	**0.57**	0.00	0.33
Rash	**0.75**	0.69	0.14	**0.64**	0.61	0.09	**0.69**	0.65	0.11
Stomatitis	**0.67**	0.62	0.38	0.64	**0.68**	0.14	**0.65**	0.65	0.20

Table 6 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

1) ADE Precision

BioBERT-Base exhibits the highest precision in detecting ADEs for both binary and per-label scores, closely followed by Roberta-Large. However, GPT4.0 has the lowest precision for ADE detection, indicating a higher number of false positives. (Refer to Figure 3)

2) NO ADE Precision

GPT4.0 and Roberta-Base in the per-label score, as well as Roberta-Large, demonstrate almost perfect precision. But Roberta-Base in the binary score exhibits the lowest precision in detecting NO ADEs.

B. Recall Analysis

1) ADE Recall

GPT4.0 has the highest recall for detecting ADEs, especially in the binary score, implying that it correctly identifies most of the ADE instances. Roberta-Base, in its per-label score, exhibits the lowest recall, suggesting it misses a significant number of actual ADE instances. (Refer to Figure 3)

2) NO ADE Recall

Roberta-Base and Roberta-Large, in their per-label scores,

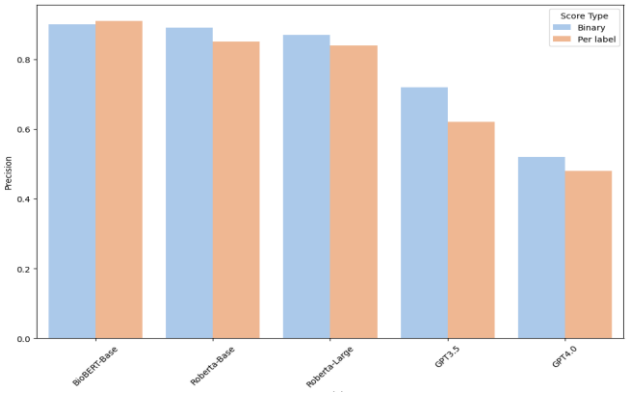

Figure 3 Precision of ADE in Development Dataset of MedNLP SM-ADE-EN Task

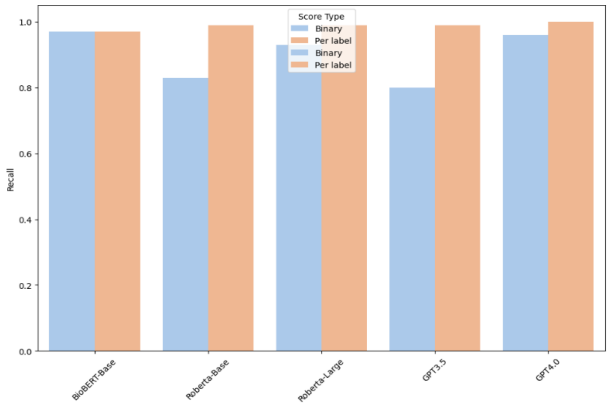

Figure 3 Recall of ADE in Development Dataset of MedNLP SM-ADE-EN Task

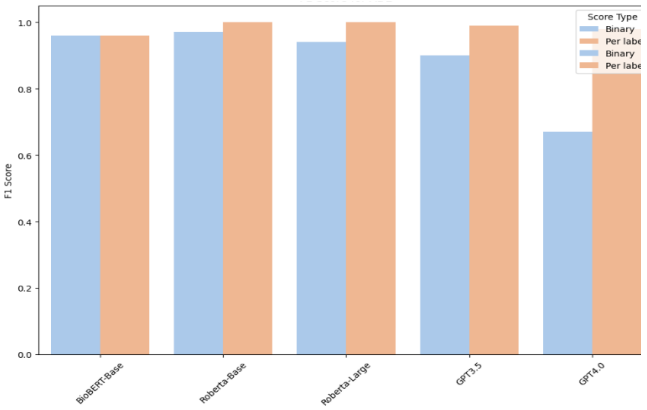

Figure 5 F1 Score for ADE in Development Dataset of MedNLP SM-ADE-EN Task

and GPT4.0 in its per-label score, show near-perfect recall for NO ADE detection. GPT4.0 in its binary score has the lowest recall for NO ADEs.

C. F1 Score Analysis

1) ADE F1 Score

BioBERT-Base and Roberta-Large showcase the highest F1 scores, indicating a balanced performance between precision and recall. GPT4.0's per-label score is the lowest, indicating a disparity between precision and recall for ADE detection. (Refer to Figure 5)

2) NO ADE F1 Score

BioBERT-Base and Roberta-Large showcase the highest F1 scores, indicating a balanced performance between precision and recall. GPT4.0's per-label score is the lowest, indicating a disparity between precision and recall for ADE detection.

To sum up the error analysis, choosing the best model depends on the specific requirements of an application. If avoiding false negatives (missing actual ADE instances) is crucial, a model like GPT4.0 with high recall might be preferred. For balanced performance, BioBERT-Base stands out.

Table 7 Exact Match Accuracy Results of Multi-label in MedNLP SM-ADE-EN Development Dataset

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

5 FURTHER EXPERIMENTS

After our initial submissions, we carried out a more detailed analysis on the MedNLP SM-ADE-EN Development Dataset. The preliminary results in Table 7, utilizing a 5-fold cross-validation coupled with a majority voting ensemble on the original development dataset, were not as promising as we had hoped for the PLMs.

To enhance our model's performance, we opted for data augmentation using GPT-3.5. This strategic decision expanded our training dataset from 7,167 entries to 8,959. As a result, the exact accuracy of RoBERTa-Base witnessed a significant improvement, moving from 0.76 to 0.86.

In the context of the dataset, which primarily consists of succinct tweets, we refined our prompts to offer a more explicit context to the GPT models. Our detailed error analysis hinted at an intriguing observation: models like GPT4.0, which inherently have a high recall, might be more suitable for this task. Given the concise nature of tweets, there's a plausible risk of ambiguity, which could mislead the LLMs. To counteract this, we crafted our prompts to emphasize the specific scenario of the dataset. This tailored approach played a pivotal role in enhancing the accuracy of GPT4.0 from 0.61 to 0.70.

Drawing from these results, we can infer that the combined efforts of data augmentation and 5-fold cross-validation provided us with models that outperformed the GPT variants in terms of exact accuracy. Interestingly, the GPT3.5 model, after fine-tuning, showcased commendable performance, registering an exact accuracy score close to the PLMs. This underscores the potential of GPT models in multi-label text classification tasks, especially when optimized with fine-tuning.

6 CONCLUSIONS

In our research for the NTCIR 17 RealMedNLP task, our team, IMNTPU, focused on two primary tasks: Adverse Drug Events (ADE) extraction and identifying radiology reports from the same patient (CI). For the ADE extraction, we deployed five diverse models to discern their efficacy in this specific domain. Our comparative analysis revealed that BioBERT-Base stands out, exhibiting robust performance in contrast to its counterparts.

Our exploration brought forth a noteworthy contribution by delving into one-shot learning within the realm of multiclass labeling, utilizing the GPT framework. While the potential of one-shot learning is immense, our findings indicated it might not be the most optimal choice for scenarios where an input can be linked to multiple labels. As a future trajectory, researchers might consider more tailored strategies for multiclass labeling, such as refining the GPT model or integrating specialized deep learning structures designed for multilabel outputs. Such endeavors could not only elevate the model's efficiency but also offer solutions to computational intricacies often associated with using GPT for multiclass labeling tasks.

ACKNOWLEDGEMENT

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.

REFERENCES

[1] S. Wakamiya, L.K. Pereira, L. Raithel, H. Yeh, P. Han, T. Nishiyama, G. Herman, B. Andrade, N. Nishida, N. Tokunaga, P. Thomas, R. Roller, P. Zweigenbaum, A. Aizawa, S. Möller, C. Grouin, T. Lavergne, P. Paroubek, S. Yada, E. Aramaki, D. Berlin, NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection : Subtask Overview, 2 (n.d.).

[2] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: A pre-trained biomedical language representation model for biomedical text mining, Bioinformatics. 36 (2020) 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

[3] X. Yu, W. Hu, S. Lu, X. Sun, Z. Yuan, BioBERT based named entity recognition in electronic medical record, Proc. - 10th Int. Conf. Inf. Technol. Med. Educ. ITME 2019. (2019) 49–52. https://doi.org/10.1109/ITME.2019.00022.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A Robustly Optimized BERT Pretraining Approach, (2019). http://arxiv.org/abs/1907.11692.

[5] C. Palmer, S. Khademi, M. Javed, G.L. Dimaguila, J. Buttery, Task 6 : Classification of tweets which indicate self-reported COVID-19 vaccination status, (2022) 81–84.

[6] OpenAI, GPT-4 Technical Report, 4 (2023) 1–100. http://arxiv.org/abs/2303.08774.

[7] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y.T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M.T. Ribeiro, Y. Zhang, Sparks of Artificial General Intelligence: Early experiments with GPT-4, (2023). http://arxiv.org/abs/2303.12712.

[8] H. Nori, N. King, S.M. McKinney, D. Carignan, E.

Horvitz, Capabilities of GPT-4 on Medical Challenge Problems, (2023) 1–35. http://arxiv.org/abs/2303.13375.

[9] Y. Gu, S. Zhang, N. Usuyama, Y. Woldesenbet, C. Wong, P. Sanapathi, M. Wei, N. Valluri, E. Strandberg, T. Naumann, H. Poon, Distilling Large Language Models for Biomedical Knowledge Extraction: A Case Study on Adverse Drug Events, (2023) 1–15. http://arxiv.org/abs/2307.06439.

[10] T. Webb, K.J. Holyoak, H. Lu, Emergent Analogical Reasoning in Large Language Models, (2022). https://doi.org/10.1038/s41562-023-01659-w.

[11] L. Reynolds, K. McDonell, Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm, Conf. Hum. Factors Comput. Syst. - Proc. (2021). https://doi.org/10.1145/3411763.3451760.

[12] S. Wang, H. Scells, B. Koopman, G. Zuccon, Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?, 1 (2023). http://arxiv.org/abs/2302.03495.

[13] C.S. Xia, L. Zhang, Conversational Automated Program Repair, (2023) 1–11. http://arxiv.org/abs/2301.13246.

[14] S. Yada, Y. Nakamura, S. Wakamiya, E. Aramaki, Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task, Proc. 16th NTCIR Conf. Eval. Inf. Access Technol. (2022) 285–296.

[15] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, B. Ge, Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models, (2023) 1–35. http://arxiv.org/abs/2304.01852.