Preliminary Symptom Classification with Natural Language Processing

Han-Wei Wang, Ting-Wei Wu, Wei-Jen Lan, Ruei-En Liang, Kai-Hsiang Tu, Kuan-Chung Wang
{b10508018, r11922008, r12922120, r11922201, b11401016, b09402007}@ntu.edu.tw
Advisor: Chung-Wei Lin

I. Abstract

The contemporary healthcare landscape faces significant challenges, primarily stemming from the scarcity of medical resources and the need for efficient utilization in patient care. In response to this, **our study introduces an innovative approach using natural language processing (NLP) and machine learning to conduct preliminary patient symptom classification.** The motivation arises from the pressing need to assist doctors in classifying symptoms and optimizing the allocation of limited resources.

Our proposed solution involves the development of a distillation-based language model DistilBERT, for the filtering and categorization of patients based on their chief complaints. This model is trained on diverse datasets encompassing cancers, heart diseases, and kidney diseases. For creating datasets for training, special prompts for ChatGPT are created to generate patients' descriptions while ensuring a balance between simplicity and medical accuracy. The dataset includes lung, breast, and colorectal cancers, heart diseases with complications, and various kidney diseases, demonstrating the model's adaptability.

In conclusion, our study introduces a practical application of NLP and machine learning in healthcare, offering a tool to expedite the diagnostic process and optimize resource allocation. The findings underscore the potential of our approach, emphasizing the need for ongoing refinement and exploration in real-world conditions before fully functioning in actual clinical situations.

II. Motivation and Problem Formulation

In contemporary healthcare settings, the scarcity of medical resources, such as doctors and hospitalization facilities, poses a significant challenge. The constraints on these resources become even more pronounced when doctors are tasked with the general classification of symptoms through diagnosis. The allocation of already limited resources becomes a critical concern as medical professionals strive to meet the diverse needs of patients. With these challenges in mind, there is an urgent need for innovative solutions to streamline the initial stages of patient care and optimize the utilization of medical resources.

One promising solution to addressing the proposed issue involves the development of a language model designed to assist in the preliminary filtering and separation of patients based on their chief complaints. By utilizing the power of natural language processing and machine learning algorithms, the model can analyze the chief complaint acquired from the patient, and then categorize them into different sectors to provide doctors with a valuable tool to speed up the diagnostic process and avoid patients going to incorrect department when seeking medical support. Such a system has the potential to enhance the efficiency of healthcare delivery by allowing medical professionals to focus their attention on cases that require immediate and specialized attention.

III. Theory

BERT is a general-purpose language representation model based on the "pretrain-finetune" framework, which involves designing a self-supervised "pretext" task that can be learned without human labeling. Examples of such tasks include the Masked Language Model (MLM) and Next Sentence Prediction (NSP). After pre-training the model using a large amount of unlabeled data, such as Wikipedia, the set of parameters is initialized in a good manner. The model is then fine-tuned with a few labeled data of the downstream task, such as sentence classification in our case. Although using less labeled data, it outperforms most of the baselines' training from scratch with a lot of labeled data.

DistillBERT is a smaller, faster, cheaper, and lighter version of BERT that was introduced by Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf in 2019. It is trained using a technique called knowledge distillation, which involves transferring knowledge from a complex "Teacher Model" to a lightweight "Student Model". The Teacher Model is larger and more complex, while the Student Model is a smaller and more lightweight version of the Teacher Model. During pre-training, the Student Model learns from the pseudo labels generated by the Teacher Model, which helps to improve its learning efficiency.

One of the benefits of using knowledge distillation is that it can help the Student Model learn the potential relationship between each class. Often, labels are too "sharp" for a model to learn this relationship. For example, a class label [1, 0, 0] only tells the model what the correct answer is, but a "smoother" pseudo label [0.7, 0.25, 0.05] generated by the Teacher Model, e.g., BERT, can provide a hint to the model that the first two classes are more similar.

DistillBERT is 40% smaller than BERT, 60% faster, and cheaper to pre-train. To leverage the inductive biases learned by larger models during pre-training, DistillBERT uses a triple loss that combines language modeling, distillation, and cosine-distance losses. This allows the model to retain 97% of BERT's language understanding capabilities while being significantly smaller and faster, and that's why it is a popular model for practical use cases.

IV. Method

1. Dataset

1.1. Cancers

Cancers are known for their threat to personal health, sometimes proving fatal. Early detection of any type of cancer can benefit the future prognosis. According to the 2018 statistics from the World Health Organization (WHO), lung cancer had the highest prevalence globally, with the highest mortality rate as well. Following lung cancer, breast cancer, and colorectal cancer are also among the most notorious cancers of which people should be aware. We chose the three cancers mentioned above as our models for training and provided ChatGPT with information on symptoms that patients may notice at home and the risk factors associated with these cancers. We specifically instructed ChatGPT to avoid using complicated medical terms, aiming to mimic the tone of an actual patient. For example, in the case of lung cancer, the initial prompt was as follows:

"Here are some general early symptoms and risk factors of lung cancer:
- Symptoms: (Provide symptoms)
- Risk factors: (Provide risk factors)

Please play the role of 20 patients who were eventually diagnosed with lung cancer and describe your chief complaints to the doctor when you first entered the clinic, etc."

During the data-generating process, we identified additional diagnostic risk factors that should be addressed in the clinic. For instance, we considered habits like smoking and occasions for lung cancer, as well as whether the breast cancer patients had children or not. Taking these factors into consideration, we modified our prompts by adding specific requirements, such as

-*"Modify the examples by including their pack years of smoking."*

-*"If there are occupationally related risk factors for lung cancer, please add some descriptions about the job of the specific patients."*

In the end, our dataset for cancers consisted of 302 descriptions for lung cancer, 305 descriptions for breast cancer, and 393 descriptions for colorectal cancer, respectively.

1.2. Heart diseases

The dataset of heart diseases consists of 2 parts: the first part includes about 700 data based on medical records written in five medical papers about “acute decompensated heart failure with hepatic congestion”, “Rheumatic heart disease, chest obstruction, heart, and kidney deficiency”, “massive pericardial effusion caused by tuberculosis, infective endocarditis, mild tricuspid regurgitation, and abscess at femur sinistra”, “Patent Ductus Arteriosus”, “Endocarditis, neonatal systemic lupus erythematosus, complete AV block”, “Congenital heart disease associated with a novel variant in the USP9X gene” and “Mechanical Aortic Valve Thrombosis with Heart Failure” published in 2023, and generated by ChatGPT with prompt *“Assume you are the man/woman in the medical record "...", what will you tell your doctor? Please give me 20 examples. You need to rearrange the sentences or use synonyms to rewrite your expression so every example will not be similar to each other. you should express more details, including symptoms, and how long the symptoms last. Start with “Hello/Hi doctor, …” or “I feel…”.”.*

In this part, ChatGPT generated the expressions of patients with not only heart disease but also other complications.

The second part includes about 300 data, which are based on the common heart diseases in Taiwan, such as arteriosclerosis, hypertension, myocardial infarction, arrhythmia, and heart failure, and generated by ChatGPT with the prompt

“What will happen to a patient with ... disease?” followed by *“Based on your answer, assume that you are a patient with ... disease and you go see a doctor, what will you describe your symptoms to your doctor? Give me 20 examples. You need to rearrange the sentences or use synonyms to rewrite your expression so every example will not be similar to each other. you should express more details, including symptoms, and how long the symptoms last. Start with “Hello/Hi doctor, …” or “I feel…”.”.*

In this part, ChatGPT generated the expressions of heart disease patients with fewer complications.

To increase data variability, we use new prompts such as:

“Okay, now you can use more professional words, assuming that you are a patient with some medical knowledge. Give me 20 more examples. Remember to rearrange the sentences to make every expression different from each other.”, *“Now change your expression more orally, like*

a patient with little medical knowledge. Also, you should change your expression order to make it not similar to your previous answer. Give me 20 examples." or *"Now, make a prediction about your disease with "I guess it may be ... disease"* or *"Doctor, is this ... disease", don't make exactly the correct prediction. Also, the expression of symptoms and lasting time are still needed. Give me 20 examples, do not only rewrite from your previous answers, or it may be too similar. You need to change the order of your expression and use synonyms to further rewrite the expressions."*

1.3. Kidney

The dataset of kidney diseases is composed of six subclasses and generated by ChatGPT respectively. It includes acute kidney injury, chronic kidney disease, glomerulonephritis, interstitial nephritis, kidney stones, and kidney cysts, as these are more common kidney diseases in Taiwan. An example of a prompt:
"Please play the role of 5 Interstitial nephritis patients and describe their condition to the doctor in English in the clinic. These patients do not know what disease they have. The description includes symptoms, and how long these symptoms last. For variety, the number of characters can be more or less, the content may not be complete. Don't exceed 500 characters for each patient."

Since symptoms and their duration are important parts of patients' descriptions, we especially ask ChatGPT to meet this requirement. The limitation of characters is due to the constraint of input of our model. It can only input 512 tokens per data sample. In addition, one prompt usually fails to generate data with high variety, even though we ask it to do so in the prompt. Thus, for every 20 samples, we asked ChatGPT to generate data with more details or more like spoken language or asked to idle chat, inform their age, guess their diseases, or guess what leads to their symptoms. With this strategy, our dataset can meet a higher variety and be close to real patients' descriptions.

2. Training

In this study, we employed the DistilBERT Classification pretrain-model for a multi-label classification task with five distinct labels. The model was trained using a training set comprising 2,241 sentences, each associated with one of the five predefined labels. The training process utilized a batch size of 16 and employed a learning rate of $2e^{-5}$, with a weight decay of 0.01. The total duration of training spanned 100 epochs, during which the Mean-Square Error (MSE) was utilized as the loss function. Additionally, a separate test set of 961 sentences was employed to evaluate the model's performance.

3. Environment

This hardware used in this study features the Nvidia GeForce RTX 4090 with CUDA 12.1 for efficient parallel processing. PyTorch 2.1.1, a dynamic deep learning framework, was employed on Ubuntu Linux 22.04, utilizing Python 3.11 for seamless integration.

V. Results

The model training consisted of 14,100 total steps, achieving a training loss of 0.0068. The training metrics indicate a runtime of 417.56 seconds, with a processing speed of 536.69 samples per second and 33.77 steps per second. The final training loss for the model was 0.0068, reflecting the convergence of the model during the 100th epoch.

The confusion matrix follows:

	precision	recall	f1-score	support
lung cancer	1.00	1.00	1.00	162
kidney disease	0.98	1.00	0.99	301
heart disease	1.00	0.98	0.99	279
breast cancer	1.00	1.00	1.00	88
colorectal cancer	1.00	1.00	1.00	130
accuracy			0.99	960
macro avg	1.00	1.00	1.00	960
weighted avg	0.99	0.99	0.99	960

For the case study, only 6 out of 960 test samples were misclassified, and all of them mistook heart disease to be kidney disease. Some of them are weird, for example:
"I've been bloated in my stomach a lot, even when I haven't eaten much. It's uncomfortable, and I don't know why my stomach feels so full."
Also, some of them are highly similar:
"Lately, I've been experiencing abdominal bloating and fullness, even when I haven't eaten much. It's uncomfortable, and I'm uncertain about the cause."

VI. Discussion

1. High similarity of generated data

   Even though a high variety of generated data was required in our prompts and we further increased variety by extra prompts about per 20 samples, the similarity of generated data is still high and could make it too easy to classify for the model. Thus, we cannot conclude that the prediction accuracy of our model is close to 100% in real conditions, unless there are metrics for evaluating the similarity of sentences, or there are enough real patients' descriptions as our test data.
2. Inappropriate division of generated data

   Data for each class of diseases was generated by different prompts. This further makes prediction easier. When generating data, focusing on various kinds of patients instead of different classes of diseases is more reasonable.
3. Well-designed prompts are needed

   Since ChatGPT does not always follow prompts to generate data, it is crucial to design prompts that are followed well. If ChatGPT can generate data with higher variety according to the real distribution of statistics, the generated data reflects reality well, and it is more efficient to generate data.

VII. Work Distribution

Han-Wei Wang

1. Propose project title
2. Presentation Sides
3. I. Abstract and II. Motivation, report wrap up

Ting-Wei Wu

1. Model Structure & Analysis
2. Model Slides & Analysis Slides
3. III. Theory in the report

Wei-Jen Lan

1. Providing medical advice for model application
2. Prepare the dataset of kidney disease
3. Kidney diseases in IV. Method of the report
4. VI. Discussion of the report

Ruei-En Liang

1. Model training
2. Model results analysis

Kai-Hsiang Tu

1. Prepare the dataset of heart disease with medical papers for training
2. Heart disease in IV. Method of the report

Kuan-Chung Wang

1. Prepare the dataset of lung cancer, breast cancer, and colorectal cancer by ChatGPT
2. Cancer in IV. method of the report

References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[2] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[3] M. L. Pfieffer, and J. Hannah, “Liver Disease and Heart Failure”, Managing Heart Failure in Primary Care: A Case Study Approach, pp. 237–255, Mar. 2023, https://link.springer.com/chapter/10.1007/978-3-031-20193-6_15

[4] W. H. Chen, Y. Tan, Y. L. Wang, X. Wang, and Z. H. Liu, “Rheumatic valvular heart disease treated with traditional Chinese medicine: A case report”, World J Clin Cases, vol. 11, no. 7, pp. 1600–1606, Mar. 2023, doi: 10.12998/wjcc.v11.i7.1600

[5] S. E. Rahayuningsih, R. B. Kuswiyanto, P. Apandi, D. Setiabudi, B. J. Manurung, and M. Hasna, “Unique Clinical Manifestation of Infective Endocarditis in Children: A Case Series”, Open Access Maced J Med Sci, vol. 11, no. C, pp. 57–61, Feb. 2023, https://oamjms.eu/index.php/mjms/article/view/11223

[6] C. Agazzi, M. Magliozzi, O. Iacoviello, S. Palladino, M. Delvecchio, M. Masciopinto, A. Galati, A. Novelli, F. A. Causio, G. Zampino, C. Ruggiero, and R. Fischetto, “Novel Variant in the USP9X Gene Is Associated with Congenital Heart Disease in a Male Patient: A Case Report and Literature Review”, Molecular Syndromology, vol. 14, no. 2, pp. 158–163, Apr. 2023, https://doi.org/10.1159/000527424

[7] M. Ababneh, A. Al-Kasasbeh, and E. Algorani, “Mechanical Aortic Valve Thrombosis with Heart Failure Successfully Treated with Oral Anticoagulation: A Case Report”, Vasc Health Risk Manag, vol. 19, pp. 617–620, Jun. 2023, https://doi.org/10.2147/VHRM.S425525