CTYUN-AI@SMM4H-2024: Knowledge Extension Makes Expert Models

Yuming Fan* and **Dongming Yang** * † and **Lina Cao**
{fanym, yangdm1, caoln}@chinatelecom.cn
China Telecom Cloud Technology Co., Ltd

Abstract

This paper explores the potential of social media as a rich source of data for understanding public health trends and behaviors, particularly focusing on emotional well-being and the impact of environmental factors. We employed large language models (LLMs) and developed a suite of knowledge extension techniques to analyze social media content related to mental health issues, specifically examining 1) effects of outdoor spaces on social anxiety symptoms in Reddit, 2) tweets reporting children's medical disorders, and 3) self-reported ages in posts of Twitter and Reddit. Our knowledge extension approach encompasses both supervised data (i.e., sample augmentation and cross-task fine-tuning) and unsupervised data (i.e., knowledge distillation and cross-task pretraining), tackling the inherent challenges of sample imbalance and informality of social media language. The effectiveness of our approach is demonstrated by the superior performance across multiple tasks (i.e., Task 3, 5 and 6) at the SMM4H-2024. Notably, we achieved the best performance in all three tasks, underscoring the utility of our models in real-world applications.

1 Introduction

In recent years, the surge in social media usage has transformed these platforms into valuable repositories of public health attitudes and behaviors. Users not only share snippets of their daily lives but also discuss a variety of health issues, drug reactions, and treatment outcomes, providing a wealth of real-time data for medical and health research. Social media plays an especially crucial role in monitoring adverse drug reactions, tracking diseases, and facilitating public discussions on health conditions. This data aids healthcare organizations and researchers in understanding disease trends and patient needs, enhancing drug safety monitoring and optimizing treatment plans.

Against this backdrop, the significance of the 9th Social Media Mining for Health Research and Applications Workshop (SMM4H-2024)(Xu et al., 2024) is particularly pronounced. This workshop brings together researchers, developers, and medical professionals from around the world to address the challenges of automating the extraction and analysis of health information from social media. The conference not only serves as a platform for sharing the latest research findings and cutting-edge technologies but also organizes multiple shared tasks targeting specific practical application problems, attracting numerous teams. These tasks are designed to enhance data processing capabilities across languages and cultural contexts, aiming to more accurately parse and utilize health-related information from social media, thereby supporting global health research and public health surveillance.

In this workshop, our goal is to construct and enhance the given limited data, including both unsupervised and supervised data, to achieve maximum knowledge extension for training large language models. This will enable us to turn the models into specialized experts for each individual task. The core points are summarized as follows:

- Sample Augmentation: by performing self-augmentation on long-tail samples, we aim to address the issue of sample imbalance in the tasks.
- Knowledge Distillation: we utilized ensemble learning with multiple models to process unsupervised data, thereby generating supervised samples that are beneficial for model training.
- Cross-Task Training: we applied both unsupervised and supervised data from one task to train the model for another task, in order to

*Equal Contribution.
†Corresponding Author.

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 5–9

expand the model's background knowledge in that task.

These strategies not only resolved issues related to uneven class distribution and data scarcity but also refined the sentiment analysis process. Finally, experimental results confirm the effectiveness of our strategies, as our team (i.e., CTYUN-AI) achieved **best performance** in tasks 3, 5, and 6 among all participating teams.

2 Related Work

Recent advances in natural language processing (NLP) have significantly enhanced the ability to analyze health-related discussions on social media. Zanwar et al. utilized advanced NLP techniques alongside psycholinguistic features to effectively detect chronic stress expressions on social media, addressing data imbalance issues in the process (Zanwar et al., 2022). Liu et al. demonstrated the use of multiple pre-trained models to detect adverse drug reactions on Twitter, showcasing methods that tackle the complexities of social media data (Liu et al., 2022). Additionally, Tamayo et al. developed a transfer learning approach with post-processing enhancements to accurately extract disease mentions from Spanish tweets, improving the robustness of disease monitoring across different languages (Tamayo et al., 2022). These contributions highlight the evolving capabilities of NLP to provide valuable insights into public health from social media content.

Concurrently, generative models in the realm of NLP, exemplified by the GPT (Brown et al., 2020) series, have exhibited remarkable abilities in comprehending and producing natural language. Bai et al. (Bai et al., 2023) developed the Qwen models, which excel at multiple tasks. Consequently, we employ the Qwen models as our base model to cultivate further specialized experts.

3 System Overview

In this section, we systematically explicate the knowledge extension strategies employed by our team for the respective sub-tasks. We commence with a descriptive analysis of the datasets pertinent to each sub-task, followed by an exposition of our methodologies, specifically devised and optimized in accordance with the task-specific data characteristics.

3.1 Task 3: Classification of reported effects of outdoor spaces on social anxiety symptoms

Task 3 is centered on categorizing Reddit posts by individuals aged 12 to 25 discussing the effects of green or blue spaces on symptoms of Social Anxiety Disorder (SAD). The dataset comprises 3,000 annotated posts, divided into 1,800 for training, 600 for validation, and 600 for testing.

Considering the long-tail distribution of the task data, which poses a challenge in achieving satisfactory performance for certain classes on the test set and detrimentally impacts the overall F1 score, we created a knowledge extension approach relying on random shuffling to alleviate this concern, as illustrated in Figure 1.

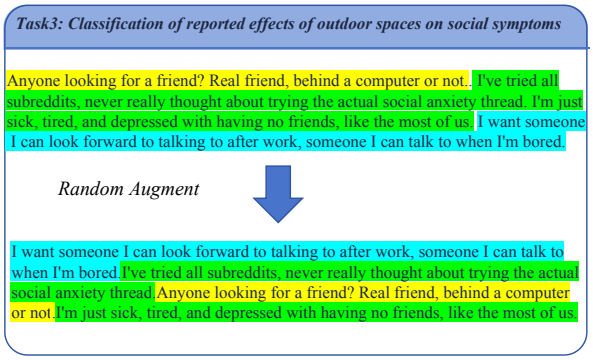

Figure 1: Example of random shuffling.

Unlike formal articles, Reddit posts are less structured, and some level of sequence disorder can still reflect the sentiment analysis inherent in social media text. Therefore, we utilized commas and periods as delimiters to randomly shuffle and augment the split data. We then balanced the dataset by augmenting the less frequent classes to match the quantity of the most populous 'positive effect' category.

3.2 Task 6: Self-reported exact age classification with cross-platform evaluation in English

Task 6 focuses on identifying precise self-reported ages from social media posts on Twitter and Reddit. This enables the analysis of health-related observational studies by determining the age of users directly from their posts. The dataset comprises 8,800 labeled tweets and 100,000 unlabeled Reddit posts, with the F1-score for the positive class serving as the evaluation metric.

To effectively extends professional knowledge

of the LLM, we employed the unlabeled data for pre-training, thereby strengthening the model's capacity to learn domain-specific features pertinent to social media text. Using the unlabeled data for pre-training involves treating each post sample directly as a training example.

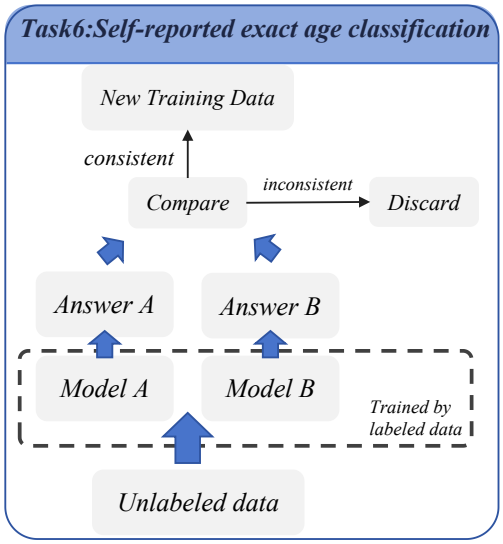

Figure 2: Example of random shuffling.

Furthermore, we introduced an ensemble model voting strategy to further mine professional knowledge from unsupervised data. Specifically, we first trained two additional models, qwen72b and qwen1.5 72b(Bai et al., 2023), using train set of the task and inferred the unlabeled data employing the trained LLMs. Then, the inference results from the two models are compared to see if they are consistent. Next, the samples with consistent inference results are retained as extended fine-tuning data, as shown in Figure 2. Finally, the participating model was fine-tuned using a combination of the original training data and augmented training data. In summary, by integrating responses from multiple models, we significantly expanded the training dataset and enhanced the robustness and generalization capability of our model across different textual contexts in social media.

3.3 Task 5: Binary classification of tweets reporting children's medical disorders

Task 5 focuses on binary classification of tweets to determine whether they report a child's medical condition, such as Attention Deficit/Hyperactivity Disorder (ADHD). The dataset consists of 7,398 training tweets, 389 validation tweets, and 1,947 test tweets, differentiating between tweets that directly report children's disorders from those that merely mention such conditions. The performance is assessed using the F1-score for tweets that substantively report on a child's medical disorder.

As task 6 is also a binary classification task focused on social media content analysis, task 5 and 6 demonstrated promising potential for transferability due to the similarities in task nature and data characteristics. Thus, our cross-task training strategy utilized the unlabeled and labeled data from task 6, which involved analyzing Twitter and other social media content, to improve the model at task 5. Building on this, we adopted the supervised fine-tuned model from Task 6 as the base model and further fine-tune the model using the train set of task 5. The experiments have proven that cross-task training not only deepens the model's comprehension of the domain-specific data but also improves its generalization capabilities in practical applications.

It is important to note that while cross-task training strategy have achieved notable success, there are still some limitations. The effectiveness of this method largely depends on the correlation between the source task and the target task. If there is a significant difference in data characteristics or objectives between the two tasks, the performance of the pre-trained model may be substantially compromised. This means that selecting tasks with high relevance for pre-training is a critical issue in practical applications. Additionally, the decline in model performance may be more pronounced when there are significant differences in the nature of the tasks, data distribution, or language style.

4 Experiment

4.1 Implement Detail

We employed the Qwen-72B-Chat(Bai et al., 2023) as our base model, upon which post pre-traing and fine-tuning of all parameters was carried out. The computational experiments were executed on an Nvidia A800 GPU, equipped with 80GB of VRAM. In the training phase, we configured the model to handle sequences with a maximum of 2048 tokens, a batch size of 8, and accumulated the gradient after every training step. The training initiated with a learning rate of 5e-6, adopting a cosine decay schedule, and spanned across three complete epochs. For the inference process, the model's built-in default parameters were utilized.

4.2 Result

In this section, we present the evaluation results of our participation in Tasks 3, 5, and 6 at the SMM4H-2024, comparing our system's performance against the mean and median scores of all participating teams. According to the organizers' assessment, our CTYUN-AI team achieved the best performance across these tasks.

Table 1: Evaluation Result on SMM4H Task 3.

Task 3 Result	F1-score	P	R	Acc
CTYUN-AI	0.692	0.704	0.686	0.726
Mean	0.5186	0.5649	0.5379	0.5746
Median	0.5795	0.63	0.5885	0.627

For Task 3, which involved classifying social media posts about the impact of social anxiety disorder, we achieved an F1 score of 0.692, as shown in Table 1. This performance significantly surpassed the mean F1 score of 0.5186 and the median of 0.5795, demonstrating our model's robust capability in accurately classifying relevant posts. The precision and recall were 0.704 and 0.686 respectively, with an accuracy of 0.726.

Table 2: Evaluation Result on SMM4H Task 5.

Task 5 Result	F1-score	P	R
CTYUN-AI	0.956	0.954	0.959
Mean	0.822	0.818	0.838
Median	0.901	0.885	0.917

In Task 5, aimed at identifying tweets reporting on children's medical conditions, our model demonstrated exceptional effectiveness with an F1 score of 0.956, considerably outperforming the mean score of 0.822 and the median score of 0.901, as shown in Table 2. This indicates that our model was highly precise (P=0.954) and sensitive (R=0.959) in identifying relevant tweets.

Table 3: Evaluation Result on SMM4H Task 6.

Task 6 Result	F1-score	P	R
CTYUN-AI	0.970	0.976	0.963
Mean	0.924	0.924	0.926
Median	0.936	0.934	0.949

Finally, in Task 6, our approach achieved an F1 score of 0.970, which is notably higher than the average F1 score of 0.924 and the median of 0.936 reported by other teams, as shown in Table 3. Our model exhibited a precision of 0.976 and a recall of 0.963, indicating superior performance in accurately identifying and classifying age-related information from the posts.

These results across different tasks highlight the efficacy of our approaches and underline the potential of our model configurations in effectively handling diverse and complex social media datasets.

5 Conclusion

In this work, we employed large language models and crafted a comprehensive set of knowledge extension techniques for the purpose of analyzing social media content pertaining to mental health concerns. We have provided a detailed account of how to leverage knowledge extension techniques to maximize the utilization of limited data, enabling us to train a general large language model into a domain-specific expert model. Specifically, our approach encompasses both supervised data and unsupervised data, including sample augmentation, knowledge distillation and cross-task training. We achieved the best performance at multiple SMM4H-2024 tasks (i.e., Task 3, 5 and 6), validating the effectiveness of our approach.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, and et al. 2023. Qwen technical report. *arXiv:2309.16609*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and et al. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Xi Liu, Han Zhou, and Chang Su. 2022. Pingantech at smm4h task1: Multiple pre-trained model approaches for adverse drug reactions. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 4–6.

Antonio Tamayo, Alexander Gelbukh, and Diego A Burgos. 2022. Nlp-cic-wfu at socialdisner: Disease mention extraction in spanish tweets using transfer learning and search by propagation. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 19–22.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Sai Tharuni Samineni, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th social media

mining for health applications (smm4h) shared tasks at acl 2024. *arXiv preprint arXiv:2405.02994*.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2022. Mantis at smm4h'2022: pre-trained language models meet a suite of psycholinguistic features for the detection of self-reported chronic stress. In *Proceedings of the Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 16–18.

A Appendix

Additionally, we will report the improvements in model performance on the validation set. It is important to note that the results shown in the previous tables are from the final test sets, which differ from the results presented here. Specifically, we initially defined the baseline method by directly feeding the labeled training data into the qwen-72b-chat model without applying any augmentation strategies.

In Task-6, starting from the initial baseline model, we achieved the highest score of 94.78 by incorporating unsupervised data and semantic alignment processing. In contrast, the baseline model using only labeled training data scored 92.29. This indicates that the strategies of using unsupervised data and semantic processing significantly improved the model's performance, especially when the accuracy was already at a high level.

In Task-5, we experimented with different learning rates and model bases. The initial baseline model scored 93.92. By adjusting the learning rate and employing a pre-trained model with unsupervised data, we improved the score to 95.88. This result was achieved using the model from Task-6 as the base, demonstrating that selecting the appropriate pre-trained model and fine-tuning parameters can significantly enhance classification performance.

In Task-3, we improved the model's classification performance through a mix of data augmentation strategies. The initial baseline model scored 57. After applying class-wise random exchange augmentation, the score increased to 61. Further, by enhancing the smaller classes, the model score rose to 64. This shows that appropriate training strategies and data augmentation techniques play a crucial role in improving multi-class classification task performance.

In summary, by integrating unsupervised data, using pretrained model from other tasks, and employing data augmentation strategies, we achieved significant performance improvements across these three tasks, validating the effectiveness of these methods in enhancing model performance.