TMUNLPG2 at the NTCIR-18 MedNLP-CHAT Task

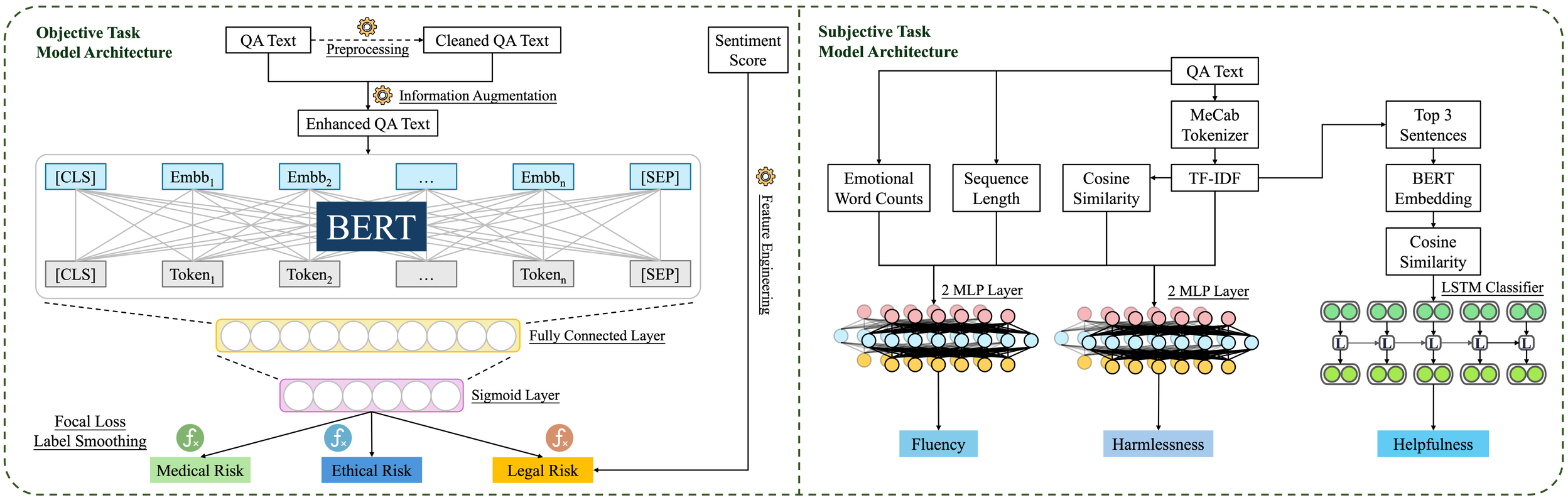
Pei-Ying, Yang[1], Tzu-Cheng Peng[2], Wen-Chao Yeh[3], Chien Chin Chen[2], Yung-Chun Chang[1,*]
[1]Graduate Institute of Data Science, Taipei Medical University, Taiwan
[2]Department of Information Management, National Taiwan University
[3]Department of Information Systems and Applications, National Tsing Hua University

Introduction

In the NTCIR-18 MedNLP-CHAT, we participated in the Japanese subtask, focusing on both the Expert Evaluation (Objective Task) and the Public Evaluation (Subjective Task). Our approach centered on leveraging advanced natural language processing (NLP) techniques and feature engineering methods to detect implicit semantic relationships within medical chatbot interactions. The goal was to develop a model that achieves both high accuracy and generalizability, capable of identifying semantic consistency and detecting potential misinformation or harmful risks in medical dialogues. Experimental results confirmed the effectiveness of our method, demonstrating its potential applications in the field of medical NLP.

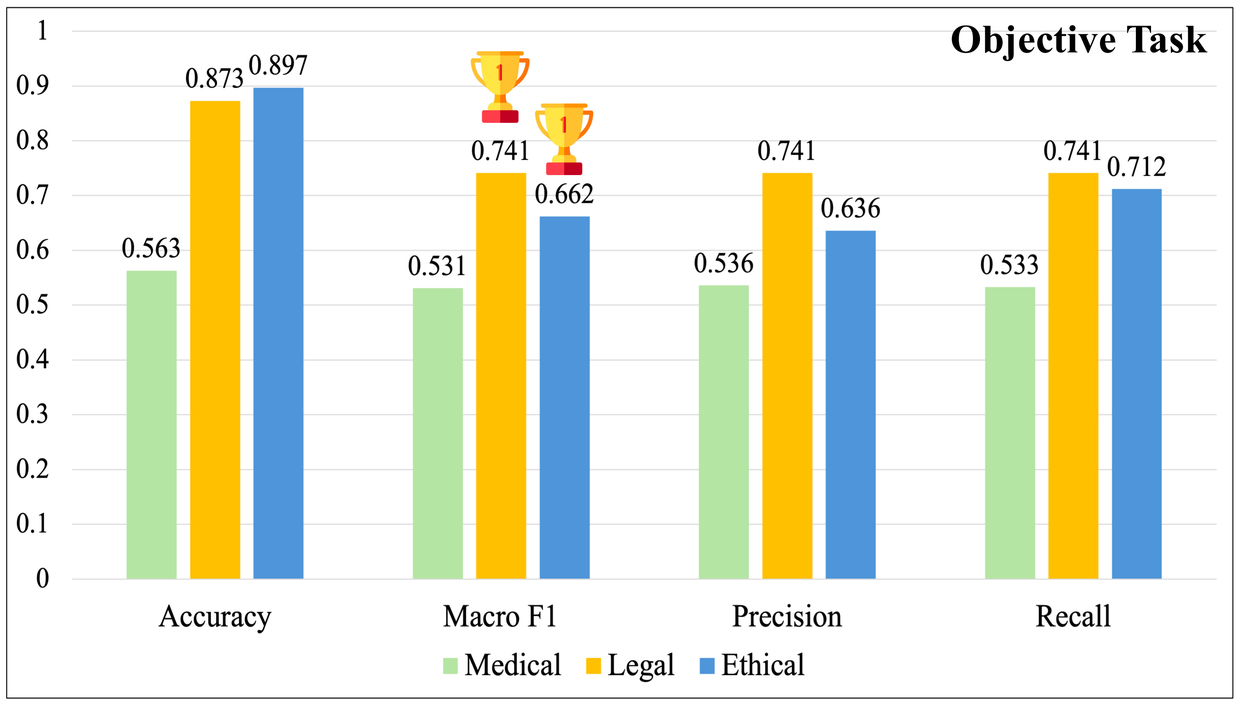
Methodology

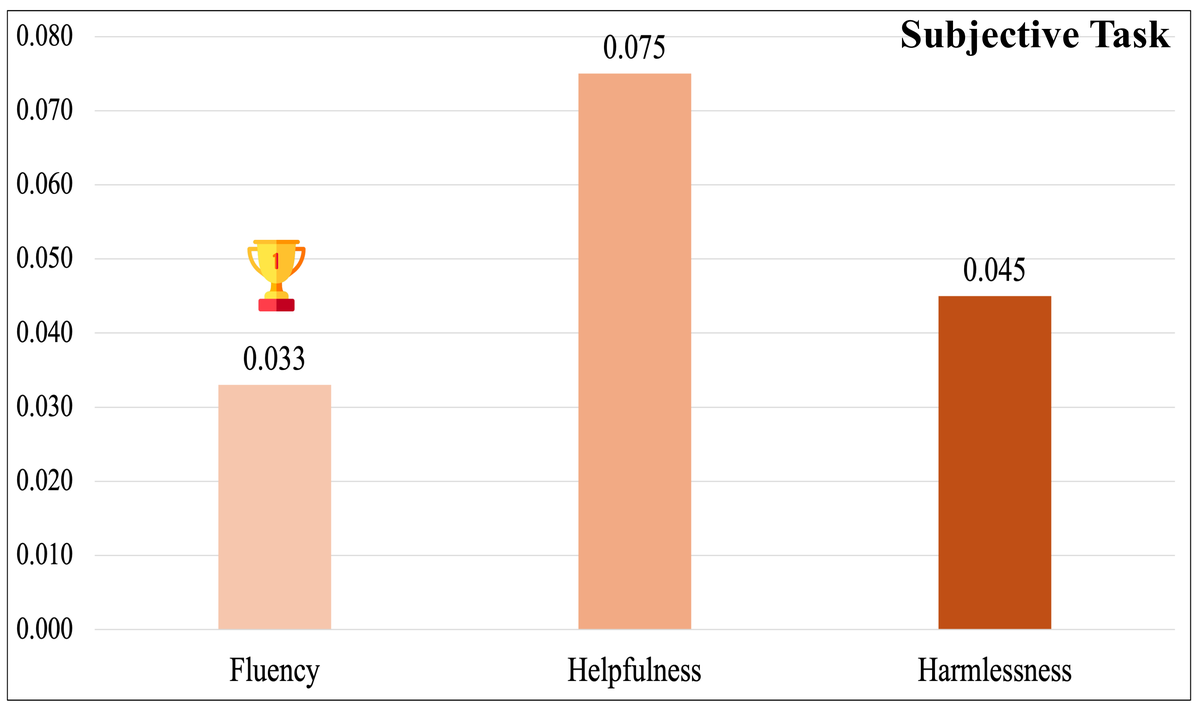
In the Objective Task, we enhanced the input representation through information augmentation by concatenating QA pairs and integrating sentiment scores specifically for the Legal Risk category. To address class imbalance, we applied Focal Loss combined with Label Smoothing. In the Subjective Task, Fluency and Harmlessness were modeled using two-layer MLPs, with features including TF-IDF vectors, cosine similarity, sequence length, and emotion word counts. For Helpfulness, we incorporated BERT embeddings of key sentences and their semantic similarity, and employed an LSTM model to improve the detection of content.



Result

In the Objective Task, our system demonstrated outstanding performance on the test set. Compared to the official baseline method, our Macro F_1-score improved by 0.1777 in the Ethical Risk category and by 0.295 in the Legal Risk category, achieving the highest scores among all participating systems in both. Although performance in the Medical Risk category was not the best, our system ranked first overall across all three risk types, showcasing its robustness and comprehensive strength in multi-faceted risk classification tasks. In the Subjective Task, our system outperformed the baseline across all three EMD Loss metrics, with the greatest improvement reaching a reduction of 0.121. This indicates a more precise prediction of semantic distributions. Furthermore, in the Japanese subtask, our system achieved first place in the Fluency evaluation, highlighting its exceptional ability to generate fluent and natural language.





Conclusion

Our study proposes a comprehensive approach for both the objective and subjective labeling tasks in MedNLP-CHAT. We utilized pretrained Transformer-based language models and enhanced performance through customized feature engineering and sentiment score integration. For the objective task, information augmentation improved the model's understanding of emotional tone and medical terminology, while label smoothing and Focal Loss effectively addressed class imbalance. In the subjective task, we constructed a multi-dimensional feature system and a two-layer MLP classifier or LSTM classifier to balance performance and generalization. Experimental results demonstrate the effectiveness of both approaches in analyzing complex medical dialogues. Future work will explore cross-lingual adaptation and domain generalization.

Acknowledgements

This work was supported by the National Science and Technology Council of Taiwan under grants NSTC 112-2622-E-038-001, NSTC 113-2221-E-038-019, NSTC 113-2627-M-A49-002, NSTC 113-2321-B-038-012, and NSTC 113-2321-B-038-006.