CYUT-NLP at ROCLING-2025 Shared Task: Valence–Arousal Prediction in Physicians' Texts Using BERT, RAG, and Multi-Teacher Pseudo-Labeling

JIAN, YI-MIN
Department of CSIE
Chaoyang University of Technology Taichung, Taiwan
s11427601@gm.cyut.edu.tw

HSIAO An Yu
Department of CSIE
Chaoyang University of Technology Taichung, Taiwan
s11227617@gm.cut.edu.tw

Shih-Hung Wu*
Department of CSIE
Chaoyang University of Technology Taichung, Taiwan
shwu@cyut.edu.tw

Abstract

Accurately modeling physicians' emotional states from self-reflection texts remains challenging due to the low-resource, domain-specific nature of medical corpora. The proposed workflow performs Retrieval-Augmented Generation (RAG) and multi-teacher pseudo-labeling to generate high-quality augmented data. This workflow enables effective cross-domain adaptation from general text corpora to professional medical texts. Evaluations on the ROCLING 2025 test set demonstrate improvements over the best-performing baseline in Valence–Arousal prediction accuracy and model stability. Importantly, the workflow is domain-agnostic and provides a generalizable methodology for systematically transferring models to new, low-resource domains, making it applicable beyond medical text analysis.

Keywords: RAG, BERT, pseudo-labeling

1 Introduction

In clinical healthcare settings, physicians are often exposed to high-pressure and high-risk working environments. Their emotional states not only influence the quality of clinical decision-making and patient care outcomes but are also closely related to their psychological well-being and professional development. Physicians' self-reflection texts provide an authentic record of their psychological experiences and emotional fluctuations. Automated analysis of these texts can facilitate emotional awareness among physicians, support hospital management decisions, and even enhance the quality of medical education and assessment.

Among various sentiment analysis approaches, traditional binary classification (e.g., positive/negative) or unidimensional scales are insufficient to capture the complex emotions commonly observed in clinical contexts, such as a "sense of heaviness in professional practice" or "perseverance amidst exhaustion." In contrast, the Valence–Arousal (V-A) two-dimensional model (Russell, 1980) can simultaneously measure the pleasantness and activation levels of emotions, providing a more nuanced representation of affective content.

Previous studies have confirmed the effectiveness of the V-A model in lexical and textual sentiment analysis (Wei et al., 2011; Wang et al., 2016a; Wu et al., 2017). In the field of Chinese sentiment analysis, Dimensional Sentiment Analysis (DSA) was first introduced at IALP 2016 (Wang et al., 2016b) and later extended to words and phrases at IJCNLP 2017 (Yu et al., 2017), followed by applications on student self-assessment texts ROCLING-2021 shared Task(Yu et al., 2021). Although these studies performed well on educational or general-domain corpora, models typically exhibit poor generalization in professional domains due to vocabulary differences, stylistic variations, and the complexity of domain-specific emotions.

The ROCLING 2025 shared task (Lee et al., 2025) applied DSA to physicians' self-reflection texts for the first time, requiring models to predict Valence and Arousal scores using Chinese EmoBank (Lee et al., 2022) as the primary data source. Compared with previous datasets, physicians' texts contain richer, multi-layered

* corresponding author

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 381–389

emotions, such as uncertainty in clinical decision-making, professional responsibility, and emotional tension in doctor–patient interactions. This domain shift poses challenges to existing models and highlights the importance of cross-domain adaptation and low-resource learning strategies.

To address this challenge, this study proposes a three-stage data augmentation framework combining Retrieval-Augmented Generation (RAG, Lewis et al., 2020) and Multi-Teacher Pseudo-Labeling (Nguyen et al., 2024). In the first stage, large language models (LLMs) combined with RAG and few-shot learning generate medical texts consistent in style and context. In the second stage, BERT-based teacher models annotate the generated texts with V-A pseudo-labels. In the third stage, high-quality augmented datasets are constructed through consistency verification and outlier removal and are then used to train downstream models.

The core contribution of this study lies not only in generating high-quality augmented data but also in demonstrating how models can be effectively transferred from general corpora to professional medical domains, providing an empirical example of cross-domain adaptation in dimensional sentiment analysis. Experimental results show that models trained with augmented data perform comparably—or even better—on physicians' texts compared with models trained solely on original data, demonstrating successful domain adaptation. Additionally, the two high-quality augmented datasets produced in this study provide valuable resources for future research on Chinese medical text sentiment analysis and low-resource cross-domain applications.

The main contributions of this study are as follows:

We propose and implement a three-stage data augmentation framework combining RAG and Multi-Teacher Pseudo-Labeling, specifically designed for V-A sentiment analysis of physicians' self-reflection texts.

We demonstrate strategies and empirical results for effectively transferring models from general corpora to the professional medical domain, providing a reference for cross-domain adaptation.

We produce two high-quality augmented datasets that can serve as valuable resources for future Chinese medical text sentiment analysis and low-resource research.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 introduces the proposed methodology; Section 4 presents experimental design and results; Section 5 concludes; and Section 6 discusses future directions.

2 Research Background

2.1 Comparison Between ROCLING 2021 Shared Task and ROCLING 2025 Shared Task

The ROCLING 2021 shared task focused on students' self-evaluation texts, in which emotional expressions were mostly related to course progress, learning new knowledge, or difficulties in comprehension. These expressions were relatively straightforward in meaning. In contrast, the ROCLING 2025 shared task (Lee et al., 2025) is the first to apply dimensional sentiment analysis to doctors' self-reflection texts, which often convey more complex emotions, such as uncertainty in clinical decision-making, a strong sense of professional responsibility, and the emotional tension inherent in doctor–patient interactions.

Both tasks adopt the same Valence–Arousal annotation scheme (real-valued scores ranging from 1 to 9, with the same input/output format). However, the shift in domain substantially increases the level of difficulty. While students' texts generally express emotions more directly, making valence easier to identify, doctors' texts frequently exhibit multi-layered or mixed emotions, which makes arousal prediction considerably more challenging.

2.2 Dimensional Model of Emotion: Valence–Arousal

Emotion, as a complex psychological and social phenomenon, has long been a central topic in both psychology and natural language processing (NLP). Traditional emotion classification models, such as Ekman's six basic emotions—happiness, anger, sadness, pleasure, surprise, and fear (Ekman, 1992)—can categorize emotional states effectively, but they are limited in capturing multidimensional and continuous affective experiences.

To overcome these limitations, Russell (1980) proposed the Circumplex Model of Affect, which maps emotions onto a two-dimensional continuous space. The valence dimension reflects the pleasantness of an emotion, ranging from negative

(unpleasant) to positive (pleasant), while the arousal dimension indicates the level of emotional activation or energy, ranging from low (calm) to high (excited). This model can capture nuanced emotional states, such as "calm joy" (high valence, low arousal) or "agitated anger" (low valence, high arousal), offering a more precise representation than traditional unidimensional sentiment classification for NLP tasks requiring fine-grained affective understanding (Schouten and Frasincar, 2015).

2.3 Retrieval-Augmented Generation (RAG) and Data Augmentation

In low-resource settings, data augmentation is a key strategy for improving model performance. Traditional techniques, such as synonym replacement (Wei and Zou, 2019), can expand the size of the training corpus but often suffer from contextual inconsistency or unnatural outputs, limiting their effectiveness in downstream tasks. The recent emergence of Large Language Models (LLMs) provides new opportunities for data augmentation, as these models can generate fluent and semantically diverse synthetic texts.

However, relying solely on LLMs may result in hallucinations, producing outputs that deviate from domain-specific contexts or contain factually incorrect information (Ji et al., 2023). To mitigate this problem, Retrieval-Augmented Generation (RAG) has been proposed (Lewis et al., 2020). RAG first retrieves relevant content from an external knowledge base or task-specific dataset and provides these retrieved examples as context to guide the LLM's generation. By anchoring outputs to the target domain, RAG enhances contextual consistency, stylistic alignment, and factual accuracy, while preserving the linguistic diversity and fluency of LLM-generated text.

In this study, we adopt RAG using the DSAMST-Validation Set as the retrieval corpus, guiding the LLM to generate synthetic texts that more closely resemble the style and context of physicians' self-reflections.

2.4 Pseudo-Labeling and Multi-Teacher Strategy

Pseudo-labeling is a widely used semi-supervised learning technique that leverages large amounts of unlabeled data to enhance model training. The typical procedure involves first training a teacher model on a small labeled dataset, then using it to predict labels for unlabeled data. High-confidence predictions are treated as pseudo-labels to expand the training set. While effective in increasing data utilization, relying on a single teacher model can introduce bias or errors, potentially misleading downstream student models.

To address this issue, a multi-teacher strategy is employed, which aggregates predictions from multiple teacher models to improve the robustness and reliability of pseudo-labeling (Nguyen et al., 2024). This approach often incorporates consistency checks and outlier removal, such as anomaly detection based on mean and standard deviation (Lee et al., 2022), to filter inconsistent or unreliable pseudo-labels. By applying these strategies, the quality of augmented datasets is enhanced, which in turn improves the generalization capability of downstream models.

3 Methods

3.1 Methodological Framework

This study focuses on predicting continuous valence and arousal values from physicians' self-reflection texts and designs a strategy combining data augmentation and multi-teacher pseudo-labeling to enhance model generalization in low-resource scenarios. The overall methodology is divided into three main stages: data augmentation, teacher model training, and annotation and corpus cleanup. Ultimately, we construct two high-quality augmented datasets, which are then applied to downstream model training and performance comparison. We illustrate the overall workflow in Figure 1.

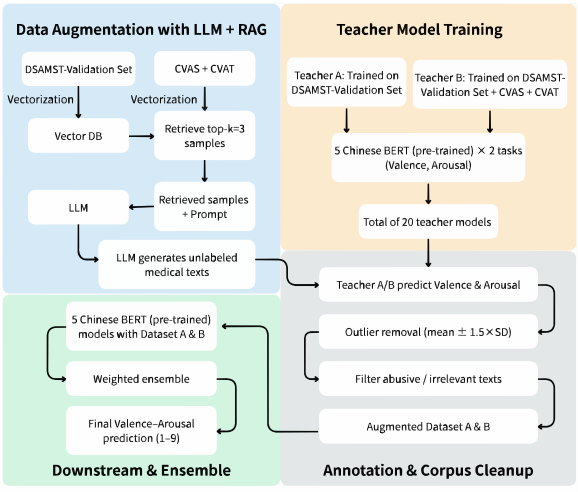

Figure 1. Methodological Framework for Data Augmentation, Teacher Annotation, and Ensemble-Based V-A Prediction.

3.2 Data Augmentation with LLM + RAG

In the data augmentation stage, we first vectorize the DSAMST-Validation Set provided by the ROCLING 2025 shared task and construct a vector database using the BAAI/bge-base-zh-v1.5 Chinese embedding model.

We retrieve the top three most similar samples from the database for each text in the Chinese EmoBank (Lee et al., 2022) derived from CVAS and CVAT and embed them as few-shot examples in the prompt to guide the large language model (LLM) in generating unlabeled medical texts that are consistent in style and context. The generated texts are guided by the following principles:

- Professionalism: Maintain domain-specific medical expression and avoid excessive colloquialism.
- Diversity: Introduce moderate variation in semantics and expression compared to retrieved examples to enhance corpus diversity.
- Authenticity: Avoid content that deviates from clinical context or is irrelevant to the task.

3.3 Teacher Model Training

We design two types of training datasets to annotate the augmented data for teacher models:

- Teacher A: DSAMST-Validation Set only.
- Teacher B: A combination of DSAMST-Validation Set, CVAS, and CVAT.

We fine-tune five pre-trained Chinese BERT models for each dataset:

1. bert-base-chinese
2. hfl/chinese-bert-wwm
3. hfl/chinese-roberta-wwm-ext
4. hfl/chinese-macbert-base
5. freedomking/mc-bert

We train two sub-models for each model and dataset because valence and arousal are two independent continuous prediction tasks. The final number of teacher models is as follows:

- Teacher A: 5 models × 2 tasks = 10 models
- Teacher B: 5 models × 2 tasks = 10 models
- Total: 20 teacher models

We fine-tune all models with an 80:20 training/validation split and use mean absolute error (MAE) as the loss function.

3.4 Pseudo-labeling and Corpus Cleanup

During annotation, we input the unlabeled augmented texts into both Teacher A and Teacher B models to obtain predicted valence and arousal values. To ensure reliability, we apply an outlier removal procedure similar to Chinese EmoBank (Lee et al., 2022):

Calculate the mean and standard deviation (SD) for each data point's predictions, remove outliers outside the range of mean ± 1.5 × SD, recalculate the mean of the remaining predictions as the final label.

Additionally, we remove any generated text that contains abusive, discriminatory, or clearly task-irrelevant content. Valence and arousal cleanup and calculations are performed independently to ensure annotation precision. After this process, two augmented datasets corresponding to Teacher A and Teacher B are generated for downstream model training and performance comparison.

3.5 Ensemble Strategy

To further improve predictive performance, this study employs a weighted ensemble approach to combine the outputs of multiple models. We first evaluate each model's performance on the validation set using mean absolute error (MAE) and Pearson Correlation Coefficient (PCC).

The initial weight of each model is defined as where W_i denotes the weight assigned to model i.

To ensure comparability across models, the weights are normalized, where M is the number of models in the ensemble and W_i' represents the normalized weight for model i.

Finally, the ensemble prediction is obtained via a weighted average, where $\hat{y}_i$ is the prediction of model i.

We integrate the predictions in this manner, effectively leveraging complementary information across individual models.

As a result, it enhances the accuracy and stability of the final predictions.

- Weighted Ensemble Initial Weight:

$$W_i = \frac{PCC_i}{MAE_i} \quad (1)$$

- Normalized Weight:

$$W_i' = \frac{W_i}{\sum_{j=1}^{M} W_j} \quad (2)$$

- Ensemble Prediction:

$$\hat{y}_{ensemble} = \sum_{i=1}^{M} W_i' \hat{y}_i \quad (3)$$

4 Experiments and Results

4.1 Training Data

This study utilizes three types of Chinese emotion datasets for model training and performance evaluation:

DSAMST-Validation Set: Derived from physicians' self-reflection texts, containing precisely annotated valence and arousal values (range 1–9). This dataset is primarily used for fine-tuning teacher models and serves as the basis for generating augmented data.

CVAS and CVAT (Lee et al., 2022): Contain valence and arousal annotations for single-sentence (CVAS) and short-text (CVAT) samples, respectively. These datasets are used to expand the training data for teacher models, enhancing their generalization to different text lengths and expression styles.

RAG-Augmented Dataset: Generated using the Retrieval-Augmented Generation (RAG) approach combined with few-shot LLMs, and annotated and cleaned by teacher models to form high-quality augmented data. The augmented datasets are categorized based on the source teacher models:

- Teacher A Augmented Data: Annotated by teacher models trained only on the DSAMST-Validation Set, focusing on the specific emotional distribution of physicians' self-reflection texts.
- Teacher B Augmented Data: Annotated by models trained on DSAMST-Validation Set combined with CVAS and CVAT, enhancing diversity and cross-style adaptability.

Dataset	Samples
DSAMST-Validation Set	994
Chinese Emobank-CVAS	2583
Chinese Emobank-CVAT	2926
RAG-CVAS-A Teacher	2583
RAG-CVAT-A Teacher	2926
RAG-CVAS-B Teacher	2583
RAG-CVAT-B Teacher	2926
ROCLING 2025 Test set	1541

Table 1. Dataset and Number of Samples

4.2 Evaluation Metrics

We assess model performance using Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC).

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |A_i - P_i| \quad (4)$$

- Pearson Correlation Coefficient (PCC):

$$PCC = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{A_i-\bar{A}}{\sigma_A}\right)\left(\frac{P_i-\bar{P}}{\sigma_P}\right) \quad (5)$$

where A_i and P_i denote the ground-truth and predicted values for sample i, and $\bar{A}$, $\bar{P}$, σ_A, σ_P are the corresponding means and standard deviations.

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.6550	0.6942	**1.0690**	0.4711
freedomking/mc-bert	**0.6093**	0.7180	1.0931	0.4843
hfl/chinese-bert-wwm	0.6289	0.6932	1.1268	0.4450
hfl/chinese-macbert-base	0.5979	**0.7237**	1.0761	0.4675
hfl/chinese-roberta-wwm-ext	0.6286	0.7048	1.0804	**0.5332**

Table 2. Performance of BERT models trained on CVAT+CVAS and evaluated on DSAMST-Validation Set

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.5182	0.7887	0.9499	**0.6258**
freedomking/mc-bert	0.5155	0.7839	0.9660	0.6173
hfl/chinese-bert-wwm	**0.5085**	**0.7909**	**0.9440**	0.6237
hfl/chinese-macbert-base	0.5154	0.7817	0.9533	0.6171
hfl/chinese-roberta-wwm-ext	0.5119	0.7898	0.9487	0.6178

Table 3. Performance of five BERT models trained on RAG-augmented data annotated by Teacher A, evaluated on DSAMST-Validation Set

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	**0.5099**	0.7843	0.9905	0.5967
freedomking/mc-bert	0.5213	0.7847	**0.9782**	0.5925
hfl/chinese-bert-wwm	0.5207	0.7812	0.9971	0.5949
hfl/chinese-macbert-base	0.5150	**0.7875**	0.9984	0.6016
hfl/chinese-roberta-wwm-ext	0.5205	0.7840	0.9850	**0.6021**

Table 4. Performance of five BERT models trained on RAG-augmented data annotated by Teacher B, evaluated on DSAMST-Validation Set

4.3 Cross-Domain Performance: Original vs. RAG-Augmented Data

From the results presented in Tables 2 to 4, we observe that when we train models on the original datasets (CVAT + CVAS) and directly evaluate them on physicians' self-reflection texts (DSAMST-Validation Set, Table 2), both Valence and Arousal MAE remain notably high, while PCC values are relatively low. This indicates that our models perform poorly under cross-domain conditions. In other words, training solely on general medical texts makes it difficult for the models to adequately capture the multi-layered and mixed emotional features present in physicians' texts.

In contrast, Tables 3 and 4 present the results of models we train on the RAG-augmented datasets proposed in this study, which combine LLM + RAG generation with multi-teacher pseudo-labeling. We observe that Valence MAE significantly decreases and PCC markedly improves, demonstrating that our approach enhances predictive performance for professional medical texts. Although the improvement in Arousal MAE is relatively modest, we find that the overall trend still surpasses the performance of models trained solely on the original datasets, indicating that our augmented data effectively provide samples similar in style and emotional distribution to the target domain.

Furthermore, when we compare Table 3 (Teacher A pseudo-labeled data) and Table 4 (Teacher B pseudo-labeled data), we observe that the models show similar performance on Valence prediction, while Teacher B achieves slightly better results on some Arousal metrics, reflecting the contribution of the multi-teacher strategy to the quality of augmented data.

Overall, these results demonstrate that our RAG-generated professional medical texts effectively mitigate the limitations of cross-domain data scarcity and substantially enhance model generalization in the professional medical domain, highlighting the innovative contribution of our study to cross-domain adaptation.

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.5	**0.77**	0.79	0.57
freedomking/mc-bert	0.5	**0.77**	**0.78**	**0.59**
hfl/chinese-bert-wwm	**0.48**	0.74	0.8	0.56
hfl/chinese-macbert-base	0.49	0.76	0.8	0.57
hfl/chinese-roberta-wwm-ext	0.49	0.76	0.81	0.55

Table 5. Performance of BERT models trained on the original combined dataset (CVAT + CVAS + DSAMST) and evaluated on ROCLING 2025 Test set

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.53	0.7	0.84	0.57
freedomking/mc-bert	0.53	**0.75**	**0.77**	**0.59**
hfl/chinese-bert-wwm	0.54	0.72	0.84	0.57
hfl/chinese-macbert-base	0.55	0.73	0.78	**0.59**
hfl/chinese-roberta-wwm-ext	**0.51**	0.72	0.8	0.56

Table 6. Performance of BERT models trained on the original DSAMST dataset and evaluated on ROCLING 2025 Test set

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.53	0.74	**0.77**	**0.59**
freedomking/mc-bert	0.55	0.74	0.81	0.58
hfl/chinese-bert-wwm	0.53	0.74	**0.77**	**0.59**
hfl/chinese-macbert-base	0.53	0.74	**0.77**	**0.59**
hfl/chinese-roberta-wwm-ext	**0.52**	**0.75**	**0.77**	0.58

Table 7. Performance of BERT models trained on Teacher A augmented dataset and evaluated on ROCLING 2025 Test set

Pre-trained BERT Model	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
bert-base-chinese	0.51	0.77	0.79	0.57
freedomking/mc-bert	0.51	0.77	**0.78**	0.57
hfl/chinese-bert-wwm	0.5	0.77	0.79	**0.58**
hfl/chinese-macbert-base	**0.48**	0.78	0.79	0.57
hfl/chinese-roberta-wwm-ext	0.51	**0.79**	**0.78**	**0.58**

Table 8. Performance of BERT models trained on Teacher B augmented dataset and evaluated on ROCLING 2025 Test set

Ensemble strategy	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑
Chinese Bank Ensemble	**0.46**	0.78	**0.74**	**0.63**
Teacher Ensemble	0.49	0.78	0.76	0.61
Combined Ensemble	0.47	**0.79**	**0.74**	0.62
Top-6 Ensemble	**0.46**	**0.79**	0.75	0.62

Table 9. Performance of various ensemble strategies on ROCLING 2025 test set

Team	Valence MAE↓	Valence PCC↑	Arousal MAE↓	Arousal PCC↑	Overall Rank
CYUT-NLP	**0.46**	0.78	**0.74**	**0.63**	1
TCU	0.46	**0.81**	0.76	0.61	2
ntulaw_	0.5	0.75	0.79	0.59	3
SCU-NLP	0.51	0.76	0.87	0.59	4
Monokeros	0.53	0.76	0.82	0.58	5
Hey Vergil	0.63	0.62	1.01	0.21	6

Table 10. Comparison of results with other groups

4.4 Performance Comparison Across Training Data on ROCLING 2025 Share Task Test Set

We observe from Table 5 and Table 6 that including general texts (CVAT + CVAS) positively impacts model performance. Models trained on CVAT + CVAS + DSAMST (Table 5) achieved lower Valence MAE and higher PCC, with slightly better Arousal metrics, compared to models trained solely on DSAMST (Table 6). These results indicate that even non-medical general texts provide additional linguistic diversity, which improves model generalization on the ROCLING 2025 Test set (Lee et al., 2025), particularly for Valence prediction.

In contrast, we find that models trained with RAG-augmented data (Teacher A and Teacher B pseudo-labeled datasets, Table 7 and Table 8) show only marginal improvements in Valence and Arousal metrics compared to DSAMST-only training. While the performance gains are limited, we note that our proposed data generation and multi-teacher labeling process maintains model performance when transferring to professional medical texts, preserving the target domain's emotional distribution and language style. These results demonstrate that our augmentation pipeline reliably produces high-quality medical texts, ensuring stable cross-domain adaptation.

Table 9 summarizes the results of four ensemble strategies on the ROCLING 2025 Test set:

- Chinese Bank Ensemble: integrates predictions from all models trained on general texts (Table 5 and Table 6) via weighted averaging to enhance stability.

- Teacher Ensemble: integrates predictions from all models trained on Teacher A/B RAG-augmented datasets (Table 7 and Table 8) via weighted averaging.

- Combined Ensemble: merges predictions from all models in Table 5–8, leveraging both general and RAG-augmented data to improve stability.

- Top-6 Ensemble: selects the six best-performing predictions from Table 5–6 (general text models) and six from Table 7–8 (RAG-augmented models), combining these 12 sets via weighted averaging to maximize complementary information and overall performance.

We analyze the trends and find that the Top-6 Ensemble and Chinese Bank Ensemble achieve the best performance, effectively improving stability and emotion prediction. We observe that the Combined Ensemble performs moderately, slightly affected by weaker predictions, while the Teacher Ensemble shows the lowest performance among the four but still outperforms single models.

Compared with the best-performing baseline (Table 5, CVAT+CVAS+DSAMST), our Top-6 Ensemble (Table 9) achieves an absolute improvement of 0.02 in Valence PCC (0.77 vs. 0.79) and a reduction of 0.03 in Valence MAE (0.49 vs. 0.46). Similarly, for Arousal, PCC improves by 0.03 (0.59 vs. 0.62) while MAE decreases by 0.04 (0.78 vs. 0.74).

Overall, we conclude that the selective ensemble of high-performing predictions is the most effective strategy for enhancing emotion prediction stability and performance.

Moreover, according to the results in Table 10, our system achieved the top-ranked performance in the ROCLING 2025 Shared Task, further validating the effectiveness of the proposed framework.

5 Conclusion

In this study, we propose a three-stage data augmentation framework combining Retrieval-Augmented Generation (RAG) and Multi-Teacher Pseudo-Labeling to enhance Valence–Arousal prediction on physicians' self-reflection texts.

We observe that integrating general texts (CVAT + CVAS) improves model performance, particularly for Valence. Meanwhile, models trained on RAG-augmented datasets maintain stable predictions when transferring to professional medical domains. In our framework, we systematically generate high-quality, domain-consistent synthetic data, leverage multiple teacher models to reduce labeling bias, and filter unreliable samples to ensure dataset quality.

We find that ensemble strategies, especially the Top-6 and Chinese Bank ensembles, further enhance stability and accuracy.

Importantly, this framework is theoretically applicable to other domains, offering a generalizable approach for cross-domain adaptation in low-resource dimensional sentiment analysis.

6 Future Work

In future work, we plan to integrate reinforcement learning (RL) into our framework to optimize the teacher-model architecture and data augmentation pipeline, to guide sample selection, teacher prediction weighting, and to identify reliable augmented data.

We also aim to further enhance our framework through advanced teacher aggregation strategies, improved retrieval methods, and semi-supervised learning, which may improve the quality of augmented datasets and downstream model performance.

Acknowledgments

This study was supported by the National Science and Technology Council under the grant number NSTC 114-2221-E-324-006.

References

Cheng, Yu-Ya, Yan-Ming Chen, Wen-Chao Yeh, and Yung-Chun Chang. 2021. Valence and Arousal-Infused Bi-Directional LSTM for Sentiment Analysis of Government Social Media Management. Applied Sciences, 11(2):880.

Ekman, Paul. 1992. An argument for basic emotions. Cognition & Emotion, 6(3-4):169–200. https://doi.org/10.1080/02699939208411068

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, et al. 2023. Survey of hallucination in natural language generation. ACM Computing Surveys (CSUR), 55(12):1–38. https://doi.org/10.1145/3571730

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented

generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33:9459–9474.

Lee, Lung-Hao, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: Building Valence-Arousal Resources for Dimensional Sentiment Analysis. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(4), article 65.

Lung-Hao Lee, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. 2025. ROCLING-2025 Shared Task: Chinese Dimensional Sentiment Analysis for Medical Self-Reflection Texts. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing.

Nguyen, Huy Thong, En-Hung Chu, Lenord Melvix, Jazon Jiao, Chunglin Wen, and Benjamin Louie. 2024. Heuristic-Free Multi-Teacher Learning. arXiv preprint arXiv:2411.12724.

Russell, James A. 1980. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161. https://doi.org/10.1037/h0077714

Schouten, Kim, and Flavius Frasincar. 2015. Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 28(3):813–830. https://doi.org/10.1109/TKDE.2015.2485209

Wang, Jin, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016a. Community-based weighted graph model for valence-arousal prediction of affective words. In Proc. of ROCLING 2021, IEEE/ACM Trans. Audio, Speech and Language Processing, 24(11):1957-1968. https://rocling2021.github.io/

Wang, Jin, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016b. Locally weighted linear regression for cross-lingual valence-arousal prediction of affective words. Neurocomputing, 194:271–278. https://doi.org/10.1016/j.neucom.2016.02.057

Wang, Jin, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2019. Investigating Dynamic Routing in Tree-Structured LSTM for Sentiment Analysis. In Proc. of EMNLP/IJCNLP-19, pages 3423–3428.

Wang, Jin, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2020. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28:581–591. https://doi.org/10.1109/TASLP.2019.2959251

Wei, Wen-Li, Chung-Hsien Wu, and Jen-Chun Lin. 2011. A regression approach to affective rating of Chinese words from ANEW. In Proc. of ACII-11, pages 121–131.

Wei, Jason, and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6383–6389. https://doi.org/10.48550/arXiv.1901.11196

Wu, Chuhan, Fangzhao Wu, Yongfeng Huang, Sixing Wu, and Zhigang Yuan. 2017. THU_NGN at IJCNLP-2017 Task 2: Dimensional Sentiment Analysis for Chinese Phrases with Deep LSTM. In Proc. of IJCNLP-17, Shared Tasks, pages 47–52.

Yu, Liang-Chih, Lung-Hao Lee, Jin Wang, and KamFai Wong. 2017. IJCNLP-2017 Task 2: Dimensional Sentiment Analysis for Chinese Phrases. In Proc. of IJCNLP-17, Shared Tasks, pages 9–16.

Yu, Liang-Chih; Wang, Jin; Peng, Bo; Huang, Chu-Ren. 2021. *ROCLING-2021 Shared Task: Dimensional Sentiment Analysis for Educational Texts*. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 385-388. Taoyuan, Taiwan. ACLCLP. URL: https://aclanthology.org/2021.rocling-1.51