ROCLING-2025 Shared Task: Chinese Dimensional Sentiment Analysis for Medical Self-Reflection Texts

Lung-Hao Lee[1,*], **Tzu-Mi Lin**[1], **Hsiu-Min Shih**[2],
Kuo-Kai Shyu[2], **Anna S. Hsu**[3], **and Peih-Ying Lu**[3]
[1]Institute of Artificial Intelligence Innovation, National Yang Ming Chiao Tung University
[2]Department of Electrical Engineering, National Central University
[3]Department of Medical Humanities and Education, Kaohsiung Medical University
*lhlee@nycu.edu.tw

Abstract

This paper describes the ROCLING-2025 shared task aimed at Chinese dimensional sentiment analysis for medical self-refection texts, including task organization, data preparation, performance metrics, and evaluation results. A total of six participating teams submitted results for techniques developed for valence-arousal intensity prediction. All datasets with gold standards and evaluation scripts used in this shared task are publicly available online for further research.

Keywords: dimensional sentiment analysis, valence-arousal intensity prediction, medical education, domain adaption, Chinese language processing

1 Introduction

In dimensional sentiment analysis, affective states are generally represented as continuous numerical values on multiple dimensions, such as valence-arousal (VA) space, as shown in Fig. 1 (Yu et al., 2016b). Based on this two-dimensional representation, any affective state can be represented as a point in the VA coordinate plane by determining the degrees of valence and arousal of given texts.

The existing methods for sentiment valence-arousal intensity prediction at different granularities from the word, phrase to text levels can be categorized as lexicon-based (Taboada et al., 2011; Thelwall et al., 2012; Paltoglou and Thelwall, 2013), regression-based (Wei et al., 2011; Malandrakis et al., 2013; Wang et al., 2016), neural-network-based (Kulshreshtha et al., 2018; Yu et al., 2018; Zhu et al., 2019; Yu et al., 2020; Deng et al., 2022), or transformer-based (Hung et al., 2021; Mukherjee et al., 2021; Park et al., 2021; Deng et al., 2023; Lin et al., 2023; Mendes and Martins, 2023). Recently, large language models (Liu et al., 2024; Xu et al., 2025) have also been used for sentiment intensity prediction with promising results.

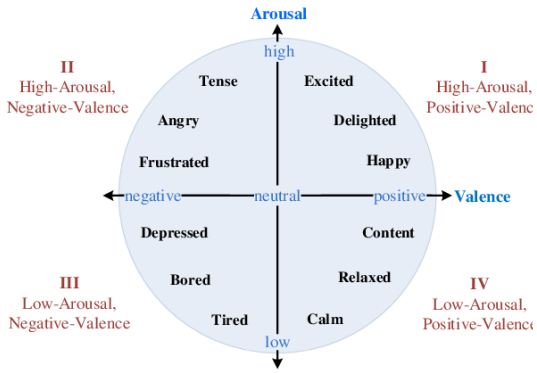

Figure 1: Two-dimensional valence-arousal space

The first dimensional sentiment analysis (DSA) task for Chinese words (Yu et al., 2016a) was organized at the IALP-2016 conference. The second edition was organized at the IJCNLP-2017 conference and included both Chinese words and phrases (Yu et al., 2017). The third edition was organized at the ROCLING-2021 conference to explore the sentence-level educational texts from students' self-evaluated comments (Yu et al., 2021). This year, we organized the fourth edition of the DSA task to analyze medical multi-sentence texts to describe doctors' self-reflection feelings.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 375–380
November 20-22, 2025

Examples	Input & Output
Example 1	*Input*: ex01, 主治醫師曾經多次強調血液透析和輸血，以病人的狀況就是不建議，已經在加護病房積極治療了兩個禮拜，家屬却遲遲無法達到共識。 (The attending physician has repeatedly emphasized that, given the patient's condition, he/she does not recommend hemodialysis or blood transfusion. The patient has already been receiving intensive care in the ICU for two weeks, yet the family has been unable to reach a consensus.) *Output*: ex01, 4.750, 2.750
Example 2	*Input*: ex02, 視病如親，這個成語一直是一個難以達成的理想，但在 ICU 我感受到醫療端與病人和家屬站在同一陣線、共同努力對抗病魔，完成病人的願望的努力，讓我十分的動容。 (The saying 'treat patients as if they were your own family' has long been an admirable yet challenging ideal to realize. However, during my time in the ICU, I was deeply moved by the dedication of the medical team, who stood in solidarity with the patient and their family, working tirelessly together to combat illness and fulfill the patient's final wishes.) *Output*: ex02, 6.900, 5.600

Table 1: Examples of the DSA-MST task.

The rest of this article is organized as follows. Section 2 provides a description of the Chinese Dimensional Sentiment Analysis for Medical Self-reflection Texts (DAS-MST) shared task. Section 3 introduces the constructed data sets. Section 4 describes the evaluation metrics. Section 5 compares evaluation results from the various participating teams. Finally, Section 6 provides conclusion and proposes future research directions.

2 Task Description

The goal of the DSA-MST shared task is to develop and evaluate the performance of Chinese sentiment analysis systems for multi-sentence texts written by doctors. The input is a self-reflective text describing a doctor's feelings and opinions regarding his/her medical internship in Intensive Care Unit (ICU) rotation. The system should predict the real-valued valence-arousal (VA) intensity ratings using a nine-degree scale. A value of 1 on the valence and arousal dimensions respectively denotes extremely high-negative and most-calm sentiment, while a 9 denotes extremely high-positive and most-excited sentiment, and 5 denotes a neutral-valence and medium-arousal sentiment.

Example instances are presented in Table 1. The input format is the instance ID followed by given texts and the output format is the same ID, followed by valence and arousal ratings. In Example 1, the valence intensity is slightly negative at 4.75 and the arousal sentiment tends to be calm at 2.75. Example 2 shows a positive sentiment of 6.9 and medium-arousal of 5.6.

3 Data Preparation

The training set for this DSA-MST shared task is the Chinese EmoBank (Lee et al., 2022), a dimensional sentiment resource annotated with real-valued scores for both valence and arousal dimensions. The valence represents the degree of positive and negative sentiment, and arousal represents the degree of calm and excitement. Both dimensions range from 1 (highly negative or calm) to 9 (highly positive or excited). The Chinese EmoBank features various levels of text granularity including two lexicons called Chinese valence-arousal words (CVAW with 5,512 single

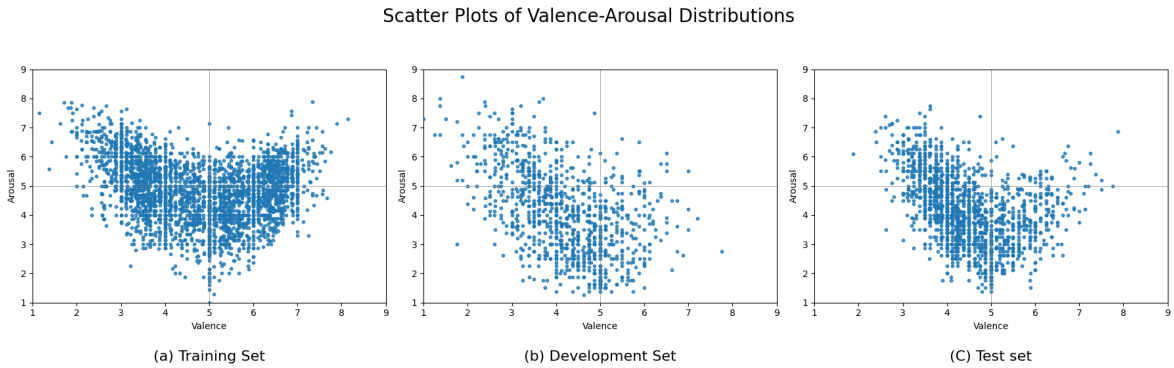

Figure 2: Scatter plots of valence-arousal distributions

words) and Chinese valence-arousal phrases (CVAP with 2,998 multi-word phrases), and two corpora called Chinese valence-arousal sentences (CVAS with 2,582 single sentences) and Chinese valence-arousal texts (CVAT with 2,969 multi-sentence texts).

The development and test sets consist of self-reflection texts written by doctors in their ICU rotation during their medical internship. The content covers the doctors' feelings and opinions towards patients and the patients' families. First, self-reflection texts were segmented into sentences and those containing sentiment words in the CVAW of Chinese EmoBank (Lee et al., 2022) were selected for manual annotation. Each sentence was presented to five Chinese native speakers for VA rating. Once the annotation process was finished, a cleanup procedure (Lee et al., 2022) was performed to remove outlier values which did not fall within 1.5 standard deviations (SD) of the mean. These outliers were then excluded from calculating the average VA values for each instance.

The annotated instances were randomly included in two mutually exclusive datasets. The development set contains 994 self-reflection texts (average 76.51 tokens) with VA ratings for system development, while the remaining 1,541 instances (average 76.81 tokens) were retained in the test set for system performance evaluation.

Figure 2 shows scatter plots of valence-arousal distributions, where the CVAT was used as the training set. Although they presented similar results, participating systems were allowed to use other publicly available data for prediction model learning, but such training data must be specified in the final system description.

4 Performance Metrics

System performance is evaluated by examining the difference between machine-predicted ratings and human-annotated ratings based on evaluation metrics including Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC), defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i|$$

$$PCC = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{a_i - \mu_A}{\sigma_A}\right)\left(\frac{p_i - \mu_P}{\sigma_P}\right)$$

where $a_i \in A$ and $p_i \in P$ respectively denote the i-th actual value and predicted value, n is the number of test samples, and μ_A and σ_A respectively represent the mean value and the standard deviation of A, while μ_P and σ_P respectively represent the mean value and the standard deviation of P.

The actual and predicted real values range from 1 to 9, so MAE measures the error rate in a range where the lowest value is 0 and the highest value is 8. The PCC is a value between −1 and 1 that measures the linear correlation between the actual value and the predicated value. A lower MAE and a higher PCC indicate more accurate prediction performance.

Each metric for the valence and arousal dimensions is ranked independently. A model's overall ranking is computed based on the mean rank across the four metrics. The lower the mean rank, the better the system performance.

Team (Submission)	Evaluation Metric				Overall Rank
	V-MAE (rank)	**V-PCC** (rank)	**A-MAE** (rank)	**A-PCC** (rank)	
CYUT-NLP (#356721)	**0.46** (1)	0.78 (2)	**0.74** (1)	**0.63** (1)	1
TCU (#356930)	**0.46** (1)	**0.81** (1)	0.76 (2)	0.61(2)	2
NTULAW (#357770)	0.50 (3)	0.75 (5)	0.79 (3)	0.59 (3)	3
SCUNLP (#357007)	0.51 (4)	0.76 (3)	0.87 (5)	0.59 (3)	4
KOLab (#358133)	0.53 (5)	0.76 (3)	0.82 (4)	0.58 (5)	5
HeyVergil (#356794)	0.63 (6)	0.62 (6)	1.01(6)	0.21 (6)	6

Table 2: Evaluation results of the DSA-MST task.

5 Evaluation Results

A total of six teams provided submissions to the leaderboard and submitted their technical papers. CYUT-NLP (Jian et al., 2025) applied the retrieval-augmented generation (RAG) and pseudo-labeling techniques to generate augmented data, and then used fine-tuned transformer-based models to predict VA ratings. TCU (Li and Lin, 2025) used several large language models (LLM) to extract contextual embedding representations and then fed semantic vectors into a regression model for VA rating prediction. The averaging ensemble technique was applied to assemble multiple prediction models for performance enhancement. NTULAW (Huang and Shao, 2025) fused encoders trained at different levels of granularity including word, phrase, and sentence to independently predict valence and arousal intensity. SCU-NLP (Pan and Wu, 2025) presented a dual-layer agent-executor framework for dimensional sentiment analysis. KOLab (Chan et al., 2025) and HeyVergil (Lin et al., 2025) systems were mainly based on BERT (Devlin et al., 2019) transformer fine-tuning for VA score prediction.

Table 2 shows the evaluation results. For the valence dimension, the best MAE of 0.46 and PCC of 0.81 was achieved by the TCU team (Li and Lin, 2025). For the arousal dimension, the best MAE of 0.74 and PCC of 0.63 was achieved by the CYUT-NLP system (Jian et al., 2025). In summary, the overall best results were provided by CYUT-NLP, followed by TCU and NTULAW (Huang and Shao, 2025).

6 Conclusion and Future Work

This paper provides an overview of the ROCLING-2025 shared DSA-MST task for Chinese dimensional sentiment analysis for medical self-reflection texts, including task descriptions, data preparation, performance metrics and evaluation results. Regardless of actual performance, all submissions contribute to the development of effective DSA systems in the medical domain, and each system description paper for this shared task also provides useful insights for further research.

We hope the data sets collected and annotated for this shared task can facilitate and expedite future development of Chinese DSA. The gold standard and evaluation scripts are made publicly available in a GitHub repository at https://github.com/NYCU-NLP/ROCLING-2025-ST-DSA-MST

Future directions will focus on the development of a Chinese domain-specific DSA. We plan to build new resources to develop techniques for the future enrichment of this research topic, especially for valence-arousal datasets in new domains.

Acknowledgments

This work was partially supported by the National Science and Technology Council, Taiwan under grant NSTC 111-2628-E-A49-029-MY3 and NSTC 114-2221-E-A49-059-MY3. This work was financially supported by the Co-creation Platform of the Industry-Ademia Innovation School, National Yang Ming Chiao Tung University.

References

Chia-Yu Chan, Chia-Wen Wang, and Jui-Feng Yeh. 2025. KOLab at ROCLING-2025 shared task: Research on emotional dimensions in Chinese medical self-reflection texts. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Yu-Chih Deng, Cheng-Yu Tsai, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2022. Predicting Chinese phrase-level sentiment intensity in valence-arousal dimensions with linguistic dependency features. *IEEE Access*, 10:126612-126620. https://doi.org/10.1109/ACCESS.2022.3226243

Yu-Chih Deng, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2023. Towards transformer fusions for Chinese sentiment intensity prediction in valence-arousal dimensions. *IEEE Access*, 11:109974-109982. https://doi.org/10.1109/ACCESS.2023.3322436

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171-4186. https://doi.org/10.18653/v1/N19-1423

Sieh-Chuen Huang, and Hsuan-Lei Shao. 2025. NTULAW at ROCLING-2025 shared task: Domain-adaptive modeling of implicit emotions in medical reflection. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Man-Chen Hung, Chao-Yi Chen, Pin-Jung Chen, and Lung-Hao Lee. 2021. NCU-NLP at ROCLING-2021 shared task: Using MacBERT transformers for dimensional sentiment analysis. *In Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing*, pages 380–384.

Yi-Min Jian, An Yu Hsiao and Shih-Hung Wu. 2025. CYUT-NLP at ROCLING-2025 shared task: Valence-arousal prediction in physicians' texts using BERT, RAG, and multi-teacher pseudo-labeling. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Devang Kulshreshtha, Pranav Goel, and Anil Kumar Singh. 2018. How emotional are you? Neural architectures for emotion intensity prediction in microblogs. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2914–2926.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions of Asian and Low-Resource Language Information Processing*, 21(4), article 65: 1-18. https://doi.org/10.1145/3489141

Hsin-Chieh Li, and Wen-Cheng Lin. 2025. TCU at ROCLING-2025 shared task: Leveraging LLM embeddings and ensemble regression for Chinese dimensional sentiment analysis. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Tzu-Mi Lin, Jung-Ying Chang, and Lung-Hao Lee. 2023. NCUEE-NLP at WASSA 2023 Empathy, Emotion, and Personality Shared Task: Perceived intensity prediction using sentiment-enhanced RoBERTa transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 548-552. https://doi.org/10.18653/v1/2023.wassa-1.49

Ting-Yi Lin, Cong-Ying Lin, and Jui-Feng Yeh. 2025. HeyVergil at ROCLING-2025 shared task: Emotion-space-based system for doctors' self-reflection sentiment analysis. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487-5496. https://doi.org/10.1145/3637528.3671552

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, Language Processing*, 21(11):2379–2392. https://doi.org/10.1109/TASL.2013.2277931

Goncalo Azevedo Mendes, and Bruno Martins. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *Proceedings of the 45th European Conference on Information Retrieval*. https://doi.org/10.1007/978-3-031-28244-7_6

Rajdeep Mukherjee, Atharva Naik, Sriyash Poddar, Soham Dasgupta, and Niloy. Ganguly. 2021. Understanding the role of affect dimensions in detecting emotions from tweets: A multi-task approach. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2303–2307. https://doi.org/10.1145/3404835.3463080

Georgios Paltoglou, and Michael Thelwall. 2013. Seeing stars of valence and arousal in blog posts. *IEEE Transactions on Affective Computing*, 4 (1): 116–123. https://doi.org/10.1109/T-AFFC.2012.36

Hong Rui Pan, and Jheng-Long Wu. 2025. SCUNLP at ROCLING-2025 shared task: Systematic guideline refinement for continuous value prediction with outlier-driven LLM feedback. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing*.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4367–4380. https://doi.org/10.18653/v1/2021.emnlp-main.358

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307. https://doi.org/10.1162/COLI_a_00049

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173. https://doi.org/10.1002/asi.21662

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Community-based weighted graph model for valence-arousal prediction of affective words. *IEEE/ACM Transactions on Audio, Speech, Language Processing*, 24(11): 1957–1968. https://doi.org/10.1109/TASLP.2016.2594287

Wen-Li Wei, Chung-Hsien Wu, and Jen-Chun Lin. 2011. A regression approach to affective rating of Chinese words from ANEW. In *Proceedings of the International Conference on Affective Computing and Intelligent Systems*, pages 121-131. https://doi.org/10.1007/978-3-642-24571-8_13

Zhe-Yu Xu, Yu-Hsin Wu, and Lung-Hao Lee. 2025. NYCU-NLP at SemEval-2025 Task 11: Assembling small language models for multilabel emotion detection and intensity prediction. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, pages 1129–1135.

Liang-Chih Yu, Lung-Hao Lee, and Kam-Fai Wong. 2016a. Overview of the IALP 2016 shared task on dimensional sentiment analysis for Chinese words. In *Proceedings of the 20th International Conference on Asian Language Processing*, pages 156-160.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016b. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540-545. https://doi.org/10.18653/v1/N16-1066

Liang-Chih Yu, Lung-Hao Lee, Jin Wang, and Kam-Fai Wong. 2017. IJCNLP-2017 Task 2: Dimensional sentiment analysis for Chinese phrases. In *Proceedings of the 8th International Joint Conference on Natural Language Processing: Shared Tasks*, pages 9-16.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2018. Refining word embeddings using intensity scores for sentiment analysis. *IEEE/ACM Transaction on Audio, Speech, Language Processing*, 26(3):671–681. https://doi.org/10.1109/TASLP.2017.2788182

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2020. Pipelined neural networks for phrase-level sentiment intensity prediction. *IEEE Transactions on Affective Computing*, 11(3): 447–458. https://doi.org/10.1109/TAFFC.2018.2807819

Liang-Chih Yu, Jin Wang, Bo Peng, Chu-Ren Huang. 2021. ROCLING-2021 shared task: Dimensional sentiment analysis for educational texts. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing*, pages 385-388.

Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial attention modeling for multi-dimensional emotion regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 471–480. https://doi.org/10.18653/v1/P19-1045