NTULAW at ROCLING-2025 Shared Task: Domain-Adaptive Modeling of Implicit Emotions in Medical Reflections

Sieh-chuen Huang
College of Law,
National Taiwan University,
No. 1, Sec. 4, Roosevelt Rd.,
Taipei, Taiwan
schhuang@ntu.edu.tw

Hsuan-Lei Shao*
Graduate Institute of Health
and Biotechnology Law,
Taipei Medical University
301 Yuantong Rd. Zhonghe Dist.
New Taipei City, Taiwan, 235603
hlshao@tmu.edu.tw

Abstract

This paper describes the NTULAW team's participation in the ROCLING 2025 Dimensional Sentiment Analysis (DSA) shared task, which focuses on predicting valence and arousal ratings for Chinese doctors' self-reflection texts. Unlike previous editions of the DSA task that targeted words, phrases, or educational comments, this year's dataset consists of domain-specific multi-sentence medical narratives, posing challenges such as low-arousal writing styles, implicit emotion expressions, and discourse complexity. To address the domain shift between general affective resources (Chinese EmoBank) and medical reflections, we designed a multi-scale BERT-based architecture and explored different data selection strategies. Our final system adopted a hybrid submission: using a model trained solely on doctors' annotations for arousal prediction, and a combined model with Chinese EmoBank for valence prediction. The system achieved stable performance, ranking third among six participating teams. Error analysis shows systematic overestimation of implicit or negated expressions for valence and regression toward mid-range predictions for arousal. We conclude with limitations of relying only on BERT and outline future work involving domain adaptation, discourse-aware modeling, and large language models (LLMs).

Keywords: Dimensional Sentiment Analysis, Valence—Arousal BERT Modeling, Chinese EmoBank, Medical Self-Reflection Texts, Domain Adaptation

1 Introduction

Sentiment analysis has become one of the most widely studied topics in natural language processing (NLP), with applications ranging from social media mining to healthcare. While categorical sentiment classification maps texts into discrete classes such as positive, negative, or neutral, dimensional sentiment analysis (DSA) offers a more fine-grained representation by positioning affective states in the valence—arousal (VA) space (Russell, 1980). Valence measures the degree of pleasantness, whereas arousal reflects the level of activation from calm to excited. This continuous framework allows researchers to capture subtle affective differences beyond simple polarity.

In Chinese NLP, DSA has been advanced through several shared tasks. The first edition, organized at IALP 2016, focused on word-level prediction (Yu et al., 2016). The newer edition of Chinese EmoBank as a resource for phrase- and sentence-level prediction, educational self-evaluation comments (Lee et al., 2019, 2022). These efforts provided annotated corpora, baselines, and evaluation protocols, which laid the groundwork for subsequent research in this area.

The ROCLING 2025 shared task further extends this line of research into a new and challenging domain: Chinese doctors' self-reflection writings. Unlike short words or sentences, these multi-sentence texts combine clinical event descriptions with professional reflections(Lee et al., 2025).

In this paper, we present the NTULAW team's system and findings. We first analyze the differences between Chinese EmoBank and the doctors' corpus, showing that domain shift is a major factor affecting model performance. We then describe our multi-scale BERT-based architecture and alternative data selection strategies. Our final hybrid submission, which combines domain-specific training for arousal with EmoBank resources for valence, achieved third place among six teams.

*Corresponding author.

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 390–398

Finally, we provide quantitative and qualitative error analysis, highlighting how implicit negativity, negation, and mixed polarity remain key challenges for future DSA systems.

2 Related Work

2.1 Dimensional Sentiment Analysis

Dimensional sentiment analysis (DSA) models emotions as continuous values—typically in the valence—arousal (VA) space—providing a finer-grained representation of affective meaning beyond categorical sentiment classification (Russell, 1980). In Chinese NLP, research has explored character-level affective annotations (Peng et al., 2024), hybrid deep learning models such as CNN—BiLSTM for text classification (Liu, 2024), and valence—arousal predictors that combine knowledge-based and embedding-based methods, which ranked top in the IALP 2016 shared task (Wang and Ma, 2016). These efforts demonstrate both the feasibility and effectiveness of applying DSA in the Chinese context (Yu et al.,2016)(Lee et al., 2022).

Recently, DSA research has advanced rapidly with the rise of aspect-based sentiment analysis (ABSA) and transformer-based architectures. The SIGHAN-2024 shared task introduced Chinese Dimensional ABSA, integrating BERT and large language models (LLMs) for entity extraction, relation classification, and intensity prediction (Xu et al., 2024; Lee et al., 2024). A bibliometric review covering 2010—2025 (Gao et al., 2025) revealed a surge in DSA-related studies after 2019, driven by deep learning methods such as BiLSTM, CRF, and attention mechanisms. Further evaluations of transformer variants—BERT, RoBERTa, DistilBERT, and particularly Electra—have demonstrated superior performance in large-scale sentiment classification (Supal et al., 2025). Complementary hybrid approaches combining traditional machine learning and deep neural encoders have improved feature extraction and emotional intensity regression (Singh et al., 2025).

In applied domains, education and public health studies have shown that DSA can capture nuanced affective shifts, such as declining positivity during the post-pandemic transition to in-person learning (Tanquis et al., 2025) and optimistic sentiment toward policy relaxation (Wang and Wang, 2023). However, persistent challenges remain, including data sparsity, class imbalance, and the need for standardized datasets and emotion-expression benchmarks (Yan and Cui, 2025; Kastrati et al., 2021). These recent works collectively underscore the ongoing shift toward transformer-driven, multi-domain approaches to dimensional sentiment modeling—an evolution that also motivates our multi-scale BERT design for medical reflective texts.

2.2 Applications in Chinese NLP

Beyond research on affective resources, sentiment analysis techniques have also been applied across diverse Chinese-language domains, such as e-commerce product reviews (Lee et al., 2019). These studies highlight the practical importance of sentiment analysis in real-world applications while also demonstrating the adaptability of advanced neural architectures.

2.3 Cross-Lingual and Multilingual Perspectives

Sentiment analysis is also an active area in multilingual contexts, where training data in low-resource languages is often generated using machine translation. Comparative experiments have shown that multilingual sentiment analysis with translation-based methods can reach performance comparable to English when combined with supervised learning algorithms (Balahur and Turchi, 2014, 2012a,b). In addition, lexicon construction remains a crucial component, such as the development of sentiment lexicons for Urdu, Roman Urdu, Pashto, and Roman Pashto (Khan et al., 2024). Studies on morphologically rich languages like Arabic have further emphasized the need for comprehensive lexicons and hybrid learning approaches to handle linguistic complexity (Sabih et al., 2018; Obaidat et al., 2015).

2.4 Affective Analysis in the Medical Domain

Despite progress in general and multilingual settings, affective analysis in healthcare remains relatively underexplored. Medical reflective writings contain implicit affective ex-

pressions tied to professional experiences and wellbeing. However, few prior studies have addressed dimensional sentiment prediction in this domain, particularly for Chinese. The ROCLING 2025 shared task therefore extends previous work to multi-sentence doctors' self-reflection texts, introducing new challenges in domain-specific language and discourse-level affective modeling.

These gaps in prior work motivate our participation in the ROCLING 2025 shared task, where we specifically address dimensional sentiment prediction in doctors' reflective writings.

3 Research Design

3.1 Task Briefing

The shared task organizers provided two datasets: (1) **Chinese EmoBank**, a general-purpose affective resource for dimensional sentiment analysis, and (2) a domain-specific **validation set** consisting of doctors' self-reflection texts. These two corpora differ significantly in their distributions, sources, and linguistic styles.

Chinese EmoBank. Chinese EmoBank is a large-scale affective resource developed for dimensional sentiment analysis, where each unit (word, phrase, or sentence) is annotated with valence—arousal (VA) scores. It includes several sub-corpora: CVAW (words), CVAP (phrases), CVAT (texts such as reviews), and CVAS (social media posts such as Twitter). The corpus covers multiple genres, ranging from formal written text to colloquial and user-generated content. This diversity results in a broad coverage of affective expressions, including both high-arousal and low-arousal emotions.

Validation Set. The validation set for ROCLING 2025 consists of Chinese doctors' self-reflection writings. These texts are typically multi-sentence narratives describing clinical experiences and professional feelings. Unlike EmoBank, the domain-specific nature of the validation set makes it more homogeneous, focusing on reflective and observational content rather than overtly emotional language. This dataset better represents the task's real-world application in medical contexts.

3.2 Dataset Comparison

We conducted a preliminary comparison between Chinese EmoBank and the doctors' reflection corpus, focusing on their valence—arousal distributions, source characteristics, and stylistic properties.

1. Arousal Distribution. Figure 1–4 illustrates the differences in valence-arousal distribution. Chinese EmoBank (especially CVAW, CVAP, and CVAT) shows a wide spread across the arousal scale, with substantial samples in the mid-to-high range (5—8). In contrast, the doctors' corpus is concentrated in the low-to-mid arousal region (1—7), with very few high-arousal instances. This reflects the writing conventions of clinical texts, where doctors favor neutral and professional language, avoiding overly emotional expressions.

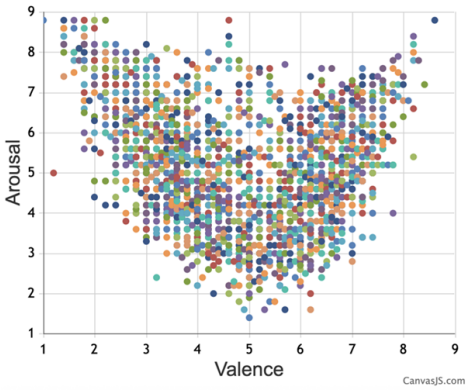

Figure 1: Valence—Arousal distribution of CVAW subset.

Figure 2: Valence—Arousal distribution of CVAP subset.

2. Source and Diversity. Chinese EmoBank draws from heterogeneous sources,

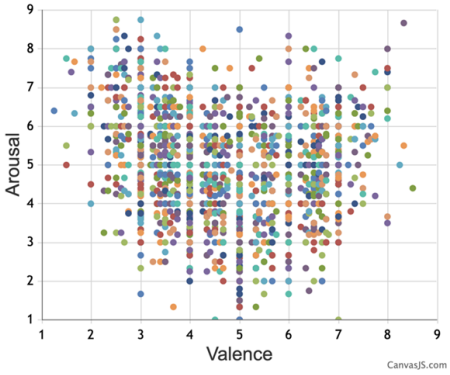

Figure 3: Valence—Arousal distribution of CVAT subset.

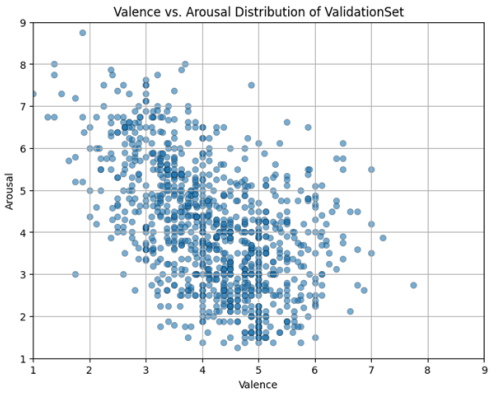

Figure 4: Valence—Arousal distribution of the doctors' validation set.

including news articles, online forums, product reviews (books, hotels, laptops), and Twitter posts. These varied genres contribute to richer emotional content and a wider valence—arousal coverage. In contrast, the doctors' corpus originates from a single domain—medical reflections—focused on clinical records and interactions. This lack of source diversity reduces the presence of explicit affective vocabulary.

3. Style and Emotional Expression. Doctors' reflections typically combine event descriptions with clinical observations, resulting in longer sentences and more complex structures, but with more implicit emotional markers. For example, a sentence like "矛盾心情全透露在他們臉上" (Their conflicting feelings were fully revealed on their faces.) conveys emotion indirectly through observation rather than direct affective terms. In contrast, Chinese EmoBank contains abundant explicit emotion words such as "快樂" (happy), "氣死" (furious), or "害怕" (afraid). Consequently, the doctors' corpus tends to cluster in the low-to-mid valence range with relatively low arousal values.

3.3 Data Selection and Model Design

Our preliminary comparison indicates that a model trained solely on Chinese EmoBank may not accurately capture the characteristics of the doctors' corpus, since medical reflective texts tend to exhibit low-arousal and indirect emotional expressions. To address this issue, we consider **multi-source data integration**, where EmoBank provides general affective knowledge and the doctors' corpus supplies domain-specific adaptation. From a data distribution perspective, domain shift is a critical factor influencing prediction performance, making domain adaptation or multi-task learning necessary for this task.

Training Data and Annotation.

- **Arousal labels**: annotated by medical professionals on the doctors' corpus.
- **Valence labels**: derived from both Chinese EmoBank and doctors' annotations.

Data Selection Strategies.

- **Only-train (Arousal-oriented)**: use only the doctors' annotated data with more reliable arousal labels, ensuring stability and quality for arousal prediction.
- **Train + ChineseEmo (Valence-oriented)**: combine doctors' data with Chinese EmoBank to improve valence prediction, leveraging the richer coverage of valence annotations.

Data Characteristics. Doctors' reflective texts are generally neutral in tone, but contain subtle lexical variations that encode fine-grained emotions. Therefore, models need to be sensitive to weak affective signals while avoiding overfitting to overtly emotional vocabulary found in EmoBank.

3.4 Model Architecture

The proposed system is designed to capture emotional cues in Chinese texts at multiple

levels of granularity. Since doctors' reflective writings often encode emotions subtly, we adopt a multi-scale architecture that processes inputs at the sentence, phrase, and word levels. This design allows the model to combine global semantics, local collocations, and character-level nuances, thereby enhancing its sensitivity to weak affective signals.

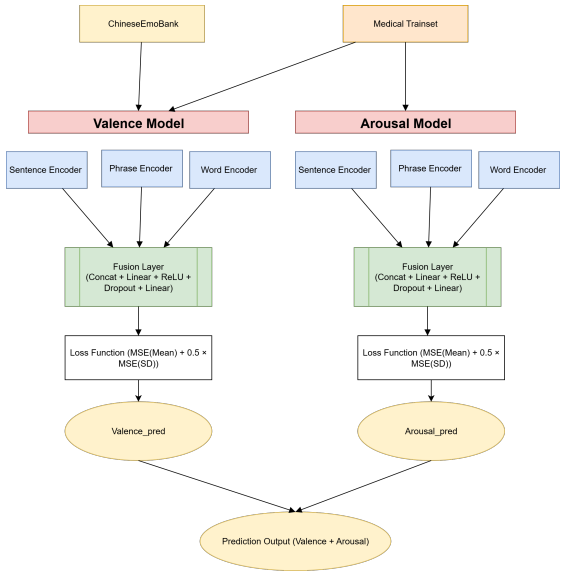

Figure 5: Modal Structure and Processing

Model Input. Three input representations are constructed from each text. At the sentence level, full sentences or paragraphs are used to model discourse and context. At the phrase level, we extract 2-gram segments to highlight collocations and local emotional patterns. Finally, at the word level, character-based sequences are included to capture fine-grained lexical information, which is particularly suitable for Chinese.

Encoder and Fusion Layers. Each input representation is processed by a dedicated BERT encoder. The sentence encoder focuses on capturing global semantics and discourse-level information. The phrase encoder emphasizes phrase-level collocations that often carry affective meaning. The word encoder specializes in character-level features, improving recognition of subtle emotion expressions. Together, the three encoders provide a layered semantic understanding of the input. The output vectors of the three encoders are concatenated and passed through a feed-forward network. This network consists of a linear layer, followed by ReLU activation, dropout for regularization, and a final linear layer that produces two regression outputs: one for arousal and one for valence.

Training Target and Loss. Each text in the dataset is annotated with both mean and standard deviation (SD) values for arousal and valence. The mean ratings ($Arousal_{Mean}$, $Valence_{Mean}$) capture the central tendency of annotators, while the SD values ($Arousal_{SD}$, $Valence_{SD}$) reflect the degree of agreement or variability across annotators. To leverage this richer annotation scheme, we design a composite loss that considers both accuracy and stability. Formally, the loss function is defined as:

$$\mathcal{L} = MSE(\hat{A}, A_{Mean}) + 0.5 \times MSE(\hat{A}, A_{SD}) + MSE(\hat{V}, V_{Mean}) + 0.5 \times MSE(\hat{V}, V_{SD}) \quad (1)$$

where $\hat{A}$ and $\hat{V}$ denote the predicted arousal and valence scores, respectively. This design encourages the model not only to approximate the average affective ratings but also to account for annotator disagreement, leading to smoother and more robust predictions.

Training Characteristics. We optimize the model using the AdamW optimizer with a linear learning rate scheduler. Early stopping is applied to prevent overfitting and ensure the best performance on the validation set. The architecture also supports cases where phrase-only data are available by inserting dummy vectors for missing sentence or word inputs, ensuring flexibility across different text granularities.

Advantages. The proposed architecture offers three key advantages. First, the multi-scale fusion enables the model to simultaneously capture global context, local collocations, and character-level nuances. Second, the joint prediction of arousal and valence reduces the need for training separate models, making the system more resource-efficient. Finally, by incorporating SD values into the loss, the model becomes more robust to annotator disagreement and produces smoother predictions that are well suited for Chinese reflective texts.

4 Results

4.1 Internal Experiments

We first attempted to build a model trained only on Chinese EmoBank (**M(ChineseE)**). However, the results were unsatisfactory, with high error and weak correlations: MAE of 1.10 and PCC of 0.44 for arousal, and MAE of 0.61 and PCC of 0.65 for valence. This indicates that the model trained purely on general-domain data struggles to adapt to the characteristics of medical reflective texts.

To address this limitation, we explored two alternative data selection strategies. The first, **M(Val)**, used only the valence-annotated subset of the doctors' corpus. This approach achieved the best performance for arousal prediction (MAE = 0.79, PCC = 0.59) and also produced strong results for valence (MAE = 0.50, PCC = 0.72). These findings suggest that restricting training to high-quality annotations enhances prediction stability, particularly for the arousal dimension.

The second approach, **M(Val+ChineseE)**, combined the doctors' valence-annotated data with Chinese EmoBank. This strategy did not improve arousal performance (MAE = 1.10, PCC = 0.44), but slightly enhanced valence prediction (MAE = 0.50, PCC = 0.75) compared to M(Val). This result highlights a trade-off: while external resources enrich valence prediction by providing broader coverage, they may introduce domain shift that harms arousal prediction.

Overall, the experiments demonstrate that domain-specific annotations are crucial for accurate arousal prediction, whereas valence prediction can benefit from multi-source integration with general-domain affective resources. Therefore, in our final submission, we adopted a hybrid approach: using **M(Val)** for arousal prediction and **M(Val+ChineseE)** for valence prediction.

Model	A (MAE,PCC)	V (MAE,PCC)
M(Val)	(0.79, 0.59)	(0.50, 0.72)
M(Val+ChiE)	(1.10, 0.44)	(0.50, 0.75)
M(ChiE)	(1.10, 0.44)	(0.61, 0.65)

Table 1: Internal experiment results with different training data strategies.

4.2 Official Evaluation Results

Table 2 shows the official evaluation results of the shared task. Our system (**ntulaw_**) achieved a balanced performance across all metrics. In particular, the model produced competitive results for both valence and arousal, although slightly behind the top two teams. Overall, our submission ranked **third place** among six participating teams, demonstrating stable and reliable performance.

Team (ID)	V-MAE	V-PCC	A-MAE	A-PCC
CYUT-NLP	0.46	0.78	0.74	0.63
TCU	0.46	0.81	0.76	0.61
ntulaw	0.50	0.75	0.79	0.59
SCU-NLP	0.51	0.76	0.87	0.59
Monokeros	0.53	0.76	0.82	0.58
Hey Vergil	0.63	0.62	1.01	0.21

Table 2: Official evaluation results of the task

5 Discussion

5.1 Error Analysis

To further evaluate the distributional properties of our predictions, we examined Q–Q and P–P plots for valence and arousal.

Valence. As shown in Figure 6, the plots demonstrate a reasonably good alignment along the 45-degree line, though with slight deviations in the mid-to-high quantile range. This suggests that the model captures the central tendency of valence effectively, but tends to underestimate extreme positive values. The P–P plot confirms this observation, showing strong overall agreement between the cumulative distributions of predictions and ground truth.

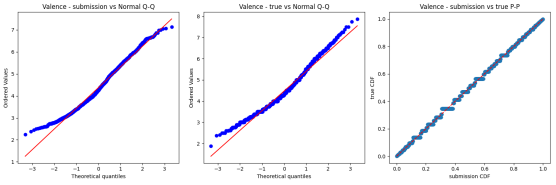

Figure 6: Diagnostic Q–Q and P–P plots for valence predictions.

Arousal. Figure 7 presents similar analyses for arousal. Compared to valence, the predicted arousal values deviate more at the tails, indicating that the model underestimates variance and struggles with extreme arousal lev-

els. Nevertheless, the P–P plot shows that the predicted distribution still closely follows the true cumulative distribution, confirming that the model is reliable in the mid-range but less accurate for highly activated emotional expressions.

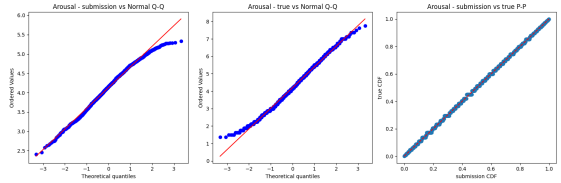

Figure 7: Diagnostic Q–Q and P–P plots for arousal predictions.

5.2 Qualitative Error Analysis (Valence)

We further investigated the sentences with the largest prediction errors in valence. In all cases, the model predicted values much higher than the ground truth, revealing systematic overestimation when emotional expressions are implicit, negated, or embedded in mixed contexts.

For example, in ID V0256 the true valence was 2.63, but the model predicted 5.86. The text contains words like ” 理解”(understanding) and ” 同理”(empathy), which are usually positive in general-domain corpora. Yet here the doctor is empathizing with distressed family members, and the overall emotional tone should remain negative. The lexical cues thus misled the model into an overly positive prediction.

In ID V0013 (true = 1.63, pred = 4.67), the doctor narrates a case in procedural detail, describing responsibility for a critically ill patient. Although emotionally heavy, the text lacks overt negative emotion words, causing the model to assign a mid-range valence.

ID V0185 shows a similar issue: the true rating was 1.25, but the prediction reached 4.18. The sentence reflects on ” 活著”(being alive) while also describing unrelieved pain and swelling. Abstract reflective terms (e.g., 思考, thinking) appear neutral or even positive to the model, diluting the strong negative context of suffering.

Finally, in ID V0064 (true = 2.38, pred = 4.93), the sentence includes the word ” 悲傷”(sadness), but it is negated by ” 不能夠顯露出悲傷的情緒”(Unable to display sorrowful emotions.). The model likely ignored the scope of negation and misinterpreted ” 悲傷”(sadness) as a direct negative marker, again resulting in overestimation.

In summary, these error cases reveal that the model often fails to handle (1) lexical—context mismatches (empathetic words in tragic situations), (2) implicit negativity in reflective writing, (3) mixed polarity within single sentences, and (4) negation scope. Future work should therefore integrate domain-adaptive fine-tuning, negation-aware processing, and discourse-level segmentation to better capture subtle emotional signals in medical reflections.

5.3 Qualitative Error Analysis (Arousal)

We also examined the cases with the largest prediction errors for arousal. In these examples, the model systematically underestimated high-arousal texts and overestimated low-arousal ones, reflecting its tendency to regress towards the mid-range values (around 4—5).

For instance, in ID V0574 the true arousal was 7.88, but the model predicted only 4.45. The sentence describes the final day of ICU training, with urgency and emotional weight. However, the reflective and narrative style downplayed explicit high-arousal cues, leading the model to underestimate the intensity.

A similar pattern appears in ID V0034 (true = 7.50, pred = 4.38). The doctor urgently describes controlling seizures with medications and intubation due to respiratory acidosis. Although the clinical situation is clearly intense, the text contains mostly procedural terms (BZD, Keppra, intubation) that the model may associate with neutral reporting, resulting in lowered arousal prediction.

On the opposite end, ID V0314 (true = 1.38, pred = 4.47) was substantially overestimated. The sentence emphasizes acceptance of illness and appreciation of care, which should indicate calmness. Yet phrases like ” 病人剛好跳短暫 VT”(The patient suddenly went into a short episode of ventricular tachycardia.) introduce suddenness that may have been misinterpreted as high arousal.

Finally, ID V0151 (true = 1.88, pred = 4.93) illustrates reflective calmness: "我們是不是可以更加專注做自己想做的事情了.(Can we now focus more on doing what we truly want to do?)" This expresses philosophical contemplation rather than excitement. Nevertheless, the rhetorical framing and modal verb usage might have been interpreted as emotionally charged, causing overestimation.

In summary, the arousal errors reveal two major tendencies: (1) underestimation of truly high-arousal emergency contexts, when described with technical or reflective wording; and (2) overestimation of calm or philosophical passages that include interrogatives, sudden events, or modal expressions. Future work should incorporate domain-adaptive embeddings that better distinguish between clinical urgency and rhetorical style, as well as discourse-level modeling to capture shifts between calm reflection and acute events.

6 Conclusion

6.1 Overall Task Review

In this paper, we presented our system for the ROCLING 2025 Dimensional Sentiment Analysis (DSA) shared task, focusing on doctors' reflective texts. Our study highlighted the challenges of applying general-domain affective resources, such as Chinese EmoBank, to the medical domain. Through systematic experiments, we found that domain-specific annotations are crucial for arousal prediction, whereas valence prediction benefits from multi-source integration. Based on these findings, we adopted a hybrid submission strategy: using the M(Val) model for arousal and the M(Val+ChineseE) model for valence. This approach achieved stable performance, ranking third among six participating teams.

Beyond quantitative results, our qualitative error analysis revealed important insights into model limitations. For valence, errors often stemmed from empathetic words used in tragic contexts, implicit negativity, mixed polarity, and negation scope. For arousal, the model underestimated high-arousal emergency descriptions that were written in technical terms, and overestimated calm, reflective passages that contained interrogatives or rhetorical devices.

6.2 Limitations

We outline the main limitations of our work and discuss directions for future improvements.

Looking forward, we plan to incorporate domain-adaptive fine-tuning, negation- and discourse-aware modeling, and clause-level segmentation to better capture subtle emotional signals. In addition, variance-aware training objectives may help the system better model extreme values on both valence and arousal scales.

Another limitation of this study is that our experiments relied mainly on BERT-based encoders. While effective, such models may lack the capacity to fully capture nuanced discourse and implicit affective cues. Future work should therefore explore larger pre-trained language models (LLMs) and hybrid architectures, combining domain adaptation, variance-aware objectives, and sentence-level reasoning to better capture subtle emotional signals.

By participating in ROCLING 2025, we aimed to bridge computational linguistics, healthcare, and law—demonstrating how interdisciplinary collaboration can contribute to affective computing research in Chinese NLP.

Ethical Considerations

The dataset used in this study was provided by the ROCLING 2025 shared task organizers and consists of anonymized Chinese doctors' self-reflection texts. No personally identifiable information (PII) was included, and we did not conduct any additional data collection. Our models are intended solely for research purposes. They should not be applied directly in clinical decision-making, as misinterpretation of affective predictions in sensitive medical contexts may pose ethical risks.

References

A. Balahur and M. Turchi. 2012a. Comparative experiments for multilingual sentiment analysis using machine translation. In *CEUR Workshop Proceedings*.

A. Balahur and M. Turchi. 2012b. Multilingual sentiment analysis using machine translation? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

A. Balahur and M. Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech and Language*, 28(1):56–75.

R. Gao, Y. Liu, and Y. Qiao. 2025. Trends and developments in aspect-based sentiment analysis: A bibliometric study using citespace and the web of science database (2010–2025). In *Proceedings of the 4th International Symposium on Computer Applications and Information Technology (ISCAIT 2025)*. IEEE.

Z. Kastrati, F. Dalipi, A. S. Imran, and M. A. Wani. 2021. Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. *Applied Sciences (Switzerland)*, 11(3):987–1002.

Z. A. Khan, Y. Xia, A. Khan, and E. A. A. Ismail. 2024. Developing lexicons for enhanced sentiment analysis in software engineering: An innovative multilingual approach for social media reviews. *Computers, Materials and Continua*.

J. S. Lee, D. Zuba, and Y. Pang. 2019. Sentiment analysis of chinese product reviews using gated recurrent unit. In *Proceedings of the 5th IEEE International Conference on Big Data Service and Applications (BigDataService)*.

Lung-Hao Lee, 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 (ROCLING 2025)*.

Lung-Hao Lee, Liang-Chih Yu, S. Wang, and J. Liao. 2024. Overview of the sighan-2024 shared task for chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics.

Y. Liu. 2024. Role of natural language processing in document understanding and semantic analysis: A chinese perspective. *Profesional de la Informacion*.

I. Obaidat, R. Mohawesh, M. Al-Ayyoub, and Y. Jararweh. 2015. Enhancing the determination of aspect categories and their polarities in arabic reviews using lexicon-based approaches. In *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*.

C. Peng, X. Xu, and Z. Bao. 2024. Sentiment annotations for 3827 simplified chinese characters. *Behavior Research Methods*.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

S. Sabih, A. Sallam, and G. S. El-Taweel. 2018. Manipulating sentiment analysis challenges in morphological rich languages. In *Advances in Intelligent Systems and Computing*.

B. Singh, K. Kaur, and G. Kaur. 2025. Optimizing emotion detection: An nlp-driven deep learning approach to sentiment encoding. In *Proceedings of the International Conference on Data Science and Business Systems (ICDSBS 2025)*. IEEE.

S. Supal, S. M. Anzar, C. Jacob, and D. Aji. 2025. Deep learning transformers for sentiment classification: A performance evaluation. In *Proceedings of the 6th International Conference on Control Communication and Computing (ICCC 2025)*. IEEE.

J. Q. Tanquis, L. Feliscuzo, and C. L. S. Romana. 2025. Data collection tools in faculty evaluation sentiment analysis. In *Proceedings of the 9th International Symposium on Innovative Approaches in Smart Technologies (ISAS 2025)*. IEEE.

H. Wang and X. Wang. 2023. Sentiment analysis of tweets and government translations: Assessing china's post-covid-19 landscape for signs of withering or booming. *Global Media and China*.

H.-Y. Wang and W.-Y. Ma. 2016. Ckip valence-arousal predictor for ialp 2016 shared task. In *Proceedings of the 2016 International Conference on Asian Language Processing (IALP)*.

H. Xu, D. Zhang, Y. Zhang, and R. Xu. 2024. Hitsz-hlt at sighan-2024 dimabsa task: Integrating bert and llm for chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics.

S. Yan and S. Cui. 2025. Fine-grained sentiment analysis of movie reviews based on machine learning and deep learning models. In *Proceedings of the 5th Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS 2025)*. IEEE.

Liang-Chih Yu, Lung-Hao Lee, and Kam-Fai Wong. 2016. Overview of the ialp 2016 shared task on dimensional sentiment analysis for chinese words. In *Proceedings of the 2016 International Conference on Asian Language Processing (IALP)*, pages 156–160. IEEE.