ROCLING-2026 Shared Task: Chinese Dimensional Sentiment Analysis for New Immigrants' Feeling Texts

Tzu-Mi Lin[1] **Wen-Ni Liu**[1] **Yen-Chen Tang**[2] **Lung-Hao Lee**[1,*]

[1] Institute of Artificial Intelligence Innovation,
National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2] Department of Public Administration,
National Dong Hwa University, Hualien, Taiwan
[*] lhlee@nycu.edu.tw

Abstract

This paper presents the ROCLING-2026 shared task on Dimensional Sentiment Analysis for New Immigrants' Feeling Texts (DSA-NIFT), the fifth edition of the Chinese Dimensional Sentiment Analysis shared task series. The task requires systems to predict continuous valence and arousal scores for Chinese narratives written by new immigrants. Chinese EmoBank serves as the source-domain training data, while the development and test sets contain 1,300 new-immigrant narratives, introducing a challenging domain shift to linguistically diverse non-native Chinese texts. Eight teams participated, with five submitting system description papers. Their approaches include domain adaptation, transfer learning, LLM-assisted augmentation, affective lexicons, synthetic data, and model combination. The best system achieved MAE/PCC scores of 0.5366/0.8971 for valence and 0.8658/0.4376 for arousal. Overall, valence prediction consistently outperformed arousal prediction, highlighting arousal as a key challenge. DSA-NIFT provides a benchmark for advancing robust fine-grained affective modeling of linguistically diverse Chinese texts.[1]

Keywords: Sentiment intensity, valence-arousal prediction, affective computing, Chinese language processing, language resources

1 Introduction

Dimensional sentiment analysis (DSA) represents emotions using continuous valence–arousal (VA) scores, where valence reflects the degree of emotional positivity or negativity and arousal indicates the level of emotional activation, as illustrated in Fig. 1. Compared with categorical sentiment representations, the VA framework provides a fine-grained characterization of affective states and can capture subtle variations in emotional intensity. It has therefore been widely studied in affective computing and natural language processing.

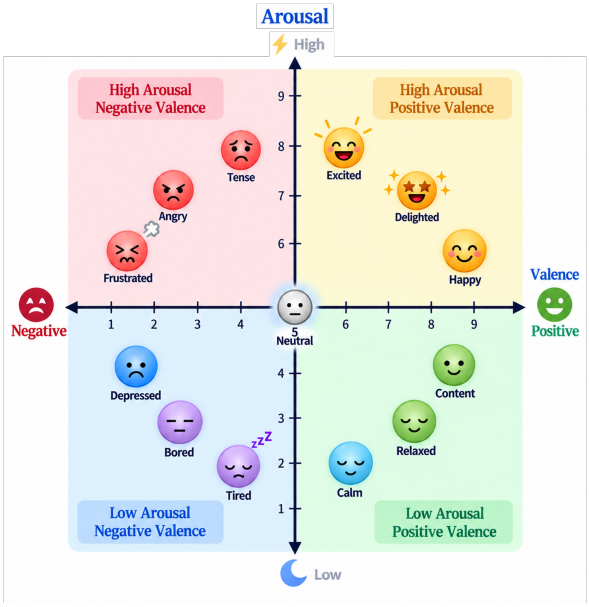

Figure 1: Two-dimensional valence-arousal space.

Existing approaches to VA prediction at different levels of granularity, ranging from words and phrases to sentences and longer texts, can be broadly categorized into lexicon-based methods (Taboada et al., 2011; Thelwall et al., 2012; Paltoglou and Thelwall, 2013), regression-based models (Wei et al., 2011; Malandrakis et al., 2013; Wang et al., 2016), neural-network-based methods (Kulshreshtha et al., 2018; Yu et al., 2018; Zhu et al., 2019; Yu et al., 2020; Deng et al., 2022), and transformer-based approaches (Hung et al., 2021; Mukherjee et al., 2021; Park et al., 2021; Deng et al., 2023; Lin et al., 2023; Mendes and Martins, 2023). More recently, large language models (LLMs) have demonstrated promising capabilities for fine-grained sentiment prediction by leveraging their strong semantic understanding and reasoning

[1] https://github.com/NYCU-NLP/ROCLING-2026-ST-DSA-NIFT

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 446–455
November 6-7, 2026 ©2026 Association for Computational Linguistics

abilities, further extending dimensional sentiment analysis to longer and more complex texts (Liu et al., 2024; Xu et al., 2025).

Natural language processing has also been increasingly applied to understand migrants' experiences through textual data, including identifying migrant workers' concerns (Dou et al., 2021), assessing emotional well-being (Knudson et al., 2016), and examining immigrants' adaptation and sense of belonging (Samanta and Chattopadhyay, 2025). However, most existing studies have focused on sentiment polarity classification, topic discovery, or qualitative analysis, whereas fine-grained dimensional sentiment analysis of migrants' narratives remains largely unexplored. Moreover, publicly available benchmark datasets for dimensional sentiment analysis of Chinese texts in this context remain scarce.

To address this gap, the Chinese DSA shared task series has been organized since 2016 to promote research on fine-grained affective prediction. Previous editions have successively investigated word-, phrase-, sentence-, and multi-sentence-level prediction across different application domains (Yu et al., 2016, 2017, 2021; Lee et al., 2025). Building on these efforts, the fifth edition introduces the **Dimensional Sentiment Analysis for New Immigrants' Feeling Texts (DSA-NIFT)** shared task, providing a new benchmark for dimensional sentiment analysis of Chinese narratives written by new immigrants.

Participants are required to predict continuous valence and arousal scores for narratives describing the writers' recent daily lives and emotional experiences. As these narratives are written by non-native Chinese speakers, they exhibit diverse lexical choices, grammatical variations, and expressions influenced by the writers' native languages. Such linguistic diversity poses additional challenges to semantic understanding and fine-grained affective prediction, making DSA-NIFT a challenging benchmark for evaluating the robustness of dimensional sentiment analysis models in real-world multilingual and cross-linguistic settings.

2 Task Description

The goal of the DSA-NIFT shared task is to predict the valence and arousal values expressed in Traditional Chinese texts written by new immigrants. The input consists of free-text responses describing the participants' emotional experiences during the previous two months, including messages they would like to convey to a specific person.

Given an input text, participating systems are required to predict its corresponding valence and arousal intensity scores on a nine-point scale. For the valence dimension, a score of 1 indicates an extremely negative emotional state, whereas a score of 9 indicates an extremely positive emotional state. For the arousal dimension, a score of 1 represents a very calm or low-activation emotional state, while a score of 9 represents a highly activated or excited emotional state. The resulting VA scores therefore provide a two-dimensional representation that jointly captures emotional polarity and activation intensity.

Table 1 presents two examples from the DSA-NIFT task, including the original Chinese texts, their English translations, and the corresponding gold-standard VA scores. In Example 1, the writer expresses missing her mother and her mother's cooking while living in Taiwan. The text receives a valence score of 4.500 and an arousal score of 7.125, reflecting relatively neutral-to-negative valence accompanied by a comparatively high level of emotional activation. In Example 2, the writer describes the joy of helping others and sharing experiences with her daughter. This text receives a high valence score of 8.750 but a relatively low arousal score of 2.750, representing a strongly positive yet comparatively calm emotional state.

These examples illustrate that valence and arousal capture complementary aspects of affective expression. Texts with positive or negative emotional content do not necessarily exhibit corresponding levels of emotional activation. Consequently, the DSA-NIFT task requires participating systems to move beyond conventional sentiment polarity classification and estimate fine-grained affective intensity along both dimensions.

3 Data Construction

The training data for the DSA-NIFT shared task is derived from Chinese EmoBank (Lee et al., 2022), a dimensional sentiment resource annotated with continuous valence and arousal (VA) scores. Chinese EmoBank provides affective annotations at multiple levels of granularity, ranging from individual words and phrases to sentences and multi-sentence texts.

Specifically, Chinese EmoBank consists of four resources: Chinese Valence-Arousal Words

Example		Input & Output
Example 1	**Input**	想對媽媽說，我想妳了，還想吃妳做的飯，在台灣吃到的菜雖然好吃，但沒有妳的味道，在家裡吃多了還覺得膩了，現在卻很想吃了。 *English translation:* I want to tell my mom that I miss her. I also miss the meals she used to cook for me. Although the food in Taiwan is delicious, it does not taste like her cooking. When I was at home, I sometimes got tired of eating her food, but now I miss it very much.
	Output	**Valence:** 4.500 **Arousal:** 7.125
Example 2	**Input**	我可以把我的在大陸幫助別人的喜悅，分享給我最親愛的人，目前唯一的親人是女兒，我們無話不說，彼此分享。 *English translation:* I can share the joy I feel from helping others in mainland China with the person dearest to me. At present, my daughter is my only family member. We can talk about anything and share everything with each other.
	Output	**Valence:** 8.750 **Arousal:** 2.750

Table 1: Examples of input texts and corresponding valence–arousal scores in the DSA-NIFT task.

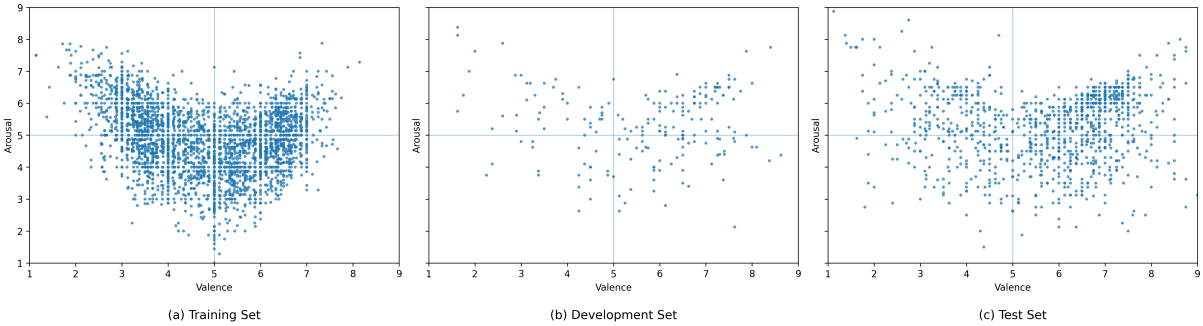

Figure 2: Scatter plots of valence-arousal distributions

(CVAW), containing 5,512 single-word entries; Chinese Valence-Arousal Phrases (CVAP), containing 2,998 multi-word expressions; Chinese Valence-Arousal Sentences (CVAS), comprising 2,582 sentences; and Chinese Valence-Arousal Texts (CVAT), consisting of 2,969 multi-sentence documents. Among these resources, CVAT was designated as the official training set because its multi-sentence format is most consistent with the narrative texts used in the DSA-NIFT task.

The development and test sets were constructed from free-text responses collected from new immigrants. Participants were asked to describe their emotional experiences and life situations during the previous two months. Based on these experiences, they were also invited to write a message to a particular person, specifying whom they wished to address and what they wanted to convey. These responses form naturally occurring narratives that reflect the participants' recent experiences and affective states.

Each text was independently rated by five annotators who were native speakers of Chinese. Following the annotation procedure of Chinese EmoBank (Lee et al., 2022), ratings that deviated by more than 1.5 standard deviations from the mean were treated as outliers and removed. The gold-standard VA score for each instance was then calculated by averaging the remaining ratings separately for the valence and arousal dimensions.

The annotated narratives were randomly divided into non-overlapping development and test sets. The development set contains 200 instances, with an average length of 73.73 tokens, and was released with gold-standard VA annotations for system development and validation. The test set contains 1,100 instances, with an average length of 70.99 tokens, and its gold-standard annotations were withheld for the official evaluation of submitted systems.

Figure 2 illustrates the distributions of valence and arousal scores across the training, development, and test sets. While CVAT serves as the official training set, the development and test sets

consist exclusively of narratives written by new immigrants, introducing a domain shift between the training and evaluation data. This setting allows the shared task to evaluate how effectively dimensional sentiment models trained on general Chinese affective texts can generalize to the linguistically diverse narratives of non-native Chinese speakers.

Participants were also permitted to use additional publicly available resources for model development. Any external datasets or resources used for training were required to be clearly documented in the corresponding system description papers.

4 Performance Metrics

The DSA-NIFT shared task evaluates system performance using two complementary metrics: Mean Absolute Error (MAE) and the Pearson Correlation Coefficient (PCC). MAE measures prediction accuracy in terms of the absolute differences between the predicted and gold-standard scores, whereas PCC measures the strength of the linear correlation between them.

The Mean Absolute Error (MAE) is defined as

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

where N denotes the number of test instances, a_i is the gold-standard score of the i-th instance, and p_i is the corresponding predicted score. Since both valence and arousal are measured on a nine-point scale ranging from 1 to 9, MAE ranges from 0 to 8, with a lower value indicating better prediction accuracy.

The Pearson Correlation Coefficient (PCC) is defined as

$$PCC = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{a_i - \mu_A}{\sigma_A}\right)\left(\frac{p_i - \mu_P}{\sigma_P}\right) \tag{2}$$

where μ_A and σ_A denote the mean and standard deviation of the gold-standard scores, respectively, while μ_P and σ_P denote those of the predicted scores. PCC ranges from -1 to 1, where values closer to 1 indicate a stronger positive linear correlation between the predicted and gold-standard scores.

MAE and PCC are computed separately for the valence and arousal dimensions, resulting in four evaluation measures: valence MAE, arousal MAE, valence PCC, and arousal PCC. Participating systems are ranked independently on each measure, with lower MAE and higher PCC indicating better performance. The overall ranking is determined by averaging the four individual ranks to obtain an average rank for each system. Systems are then ordered according to their average ranks, with a lower value indicating better overall performance.

5 Participating Systems

A total of eight teams participated in the DSA-NIFT shared task, of which seven completed their submissions before the official deadline and were included in the final rankings. This section summarizes and compares the approaches presented by CYUT (Jian and Wu, 2026), SCUNLP (Chen et al., 2026), aoff123 (Cheng et al., 2026), NCKU_NLP (Chen and Tan, 2026), and YCW-NTUST (You et al., 2026). Although the participating systems differ considerably in model architecture and adaptation strategy, most of them address a common challenge: transferring affective knowledge learned from existing Chinese VA resources to the new-immigrant narrative domain.

Table 2 summarizes the main characteristics of the five participating systems with submitted system description papers. The approaches exhibit substantial methodological diversity, ranging from target-domain adaptation and staged transfer learning to LLM-assisted semantic augmentation, affective-lexicon fusion, and heterogeneous ensembling. Despite these differences, a common objective across the systems is to mitigate the distributional gap between existing Chinese affective resources and the new-immigrant narratives in DSA-NIFT.

5.1 Overview of Participating Approaches

CYUT (Jian and Wu, 2026) proposed a perplexity-guided target-domain adaptation approach based on Qwen2-7B-Instruct and Ministral-3-8B-Instruct-2512. The models were first fine-tuned on annotated source-domain data from CVAS and CVAT. Perplexity was then used to identify target-domain texts that differed substantially from the source distribution, and selected unlabeled texts were used for continued language-model training. To mitigate the loss of task-specific capabilities during domain adaptation, the models were subsequently realigned with the source-domain VA prediction task. Finally, predictions from the two backbones were

Team	Main Backbone	Main Strategy	Additional Resources	LLM Usage
CYUT	Qwen2-7B, Ministral-3-8B	PPL-guided target-domain adaptation	-	Primary prediction models
SCUNLP	BERT-base-Chinese	LLM-generated affective-text augmentation	DSA-MST	Semantic augmentation
aoff123	MacBERT	Staged domain adaptation and model selection	DSA-ET DSA-MST	–
NCKU_NLP	MacBERT	Lexicon fusion and source-aware supervision	DSA-ET DSA-MST CVAW/CVAP lexicons	–
YCW-NTUST	RoBERTa, MacBERT, multilingual-e5	Heterogeneous prediction fusion	DSA-MST CVAW lexicon; synthetic	Direct VA scoring

Table 2: Comparison of the five participating systems with submitted system description papers. All systems used the CVAS and CVAT portions of Chinese EmoBank as their primary supervised resources.

combined using a dimension-specific weighted ensemble. The submitted system used the top 30% of target-domain texts selected by perplexity for adaptation.

SCUNLP (Chen et al., 2026) adopted an LLM-assisted feature augmentation strategy. Instead of directly using an LLM to generate the final VA scores, the system employed GPT-5.4-mini to analyze affective information in each input text, including emotional polarity, intensity, negation, and contrast. The generated descriptions were concatenated with the original text and subsequently encoded by a Chinese BERT model for joint valence–arousal regression. The training data combined Chinese EmoBank with labeled data from the ROCLING-2025 shared task DSA-MST data (Lee et al., 2025). This approach treats LLM-generated affective descriptions as auxiliary semantic information while leaving the final numerical prediction to a supervised regression model.

The aoff123 (Cheng et al., 2026) system focused on staged domain adaptation and dimension-specific model selection. A bounded MacBERT valence regressor was initially trained using the CVAT and CVAS portions of Chinese EmoBank and subsequently adapted using educational DSA-ET (Yu et al., 2021) and medical DSA-MST data (Lee et al., 2025). Candidate models were selected according to valence MAE subject to a PCC constraint, and fixed interpolation was used to combine candidate predictions. In contrast to the adaptive valence stream, arousal predictions were obtained from a frozen reference model. This design explicitly treats valence and arousal as distinct prediction problems and applies different adaptation strategies to the two dimensions.

NCKU_NLP (Chen and Tan, 2026) developed a cross-domain regression framework combining MacBERT representations with affective-lexicon features and source-aware supervision. Its formal submission incorporated Chinese affective lexicons into the neural representation and applied source-aware weighting to arousal supervision. The study additionally investigated several diagnostic alternatives, including pseudo-labeling, pairwise supervision, different encoders, model ensembles, and graph-guided LLM-generated synthetic data. These experiments were designed to examine whether additional weak supervision and synthetic target-style data could improve cross-domain VA prediction, particularly for the more challenging arousal dimension.

The YCW-NTUST (You et al., 2026) system adopted a heterogeneous ensemble strategy that integrates multiple sources of affective information. Its components included fine-tuned Chinese RoBERTa and MacBERT regressors, frozen multilingual-e5 embeddings with support vector regression, a model trained on target-style synthetic data, CVAW lexicon-based predictions, and LLMs used as direct VA scorers. The system further employed dimension-wise hybridization and affine rescaling for arousal. Unlike the other systems, its ensemble procedure exploited leaderboard PCC feedback together with locally computable prediction statistics to derive weights for combining previously submitted prediction vectors.

5.2 Comparison of System Strategies

Several common methodological trends can be observed across the five system description papers.

First, **domain adaptation is a central concern** in DSA-NIFT. Because the official training data and the new-immigrant narratives differ in topic, writing style, and linguistic characteristics, participating teams explored various mechanisms for reducing the source–target gap. CYUT (Jian and Wu, 2026) directly adapted its language models using perplexity-selected unlabeled target texts, whereas aoff123 performed staged transfer from general affective data to educational and medical reflection domains. NCKU_NLP (Chen and Tan, 2026) instead incorporated source-aware supervision, while SCUNLP (Chen et al., 2026) used LLM-generated affective descriptions to enrich the semantic representation of target texts.

Second, the systems demonstrate an increasing use of **LLMs as auxiliary components rather than relying exclusively on direct prompting**. CYUT (Jian and Wu, 2026) fine-tuned instruction-tuned LLMs as the primary prediction models, whereas SCUNLP (Chen et al., 2026) used an LLM to generate intermediate affective descriptions that were subsequently processed by BERT. NCKU_NLP (Chen and Tan, 2026) explored LLM-generated synthetic texts as an auxiliary supervision strategy, while YCW-NTUST (You et al., 2026) incorporated direct LLM scoring as one component of a heterogeneous ensemble. These approaches illustrate several distinct roles for LLMs in dimensional sentiment analysis, including backbone modeling, semantic augmentation, synthetic-data generation, and direct affective scoring.

Third, several systems explicitly recognized that **valence and arousal may require different modeling strategies**. CYUT (Jian and Wu, 2026) employed dimension-specific ensemble weights, aoff123 (Cheng et al., 2026)s eparated adaptive valence prediction from a frozen arousal stream, and YCW-NTUST (You et al., 2026) independently combined prediction sources for the two dimensions and applied additional calibration to arousal. This design trend is consistent with the official evaluation results, in which arousal prediction exhibits substantially lower PCC values and higher MAE values than valence prediction.

Finally, **model combination and complementary supervision** are common across the participating systems. Rather than relying solely on a single pretrained encoder, the systems explored cross-backbone ensembling, prediction interpolation, affective lexicons, synthetic data, LLM-generated information, and multiple source-domain datasets. Overall, the submitted approaches suggest that successful performance on DSA-NIFT depends not only on the choice of pretrained language model but also on how effectively source-domain affective knowledge, target-domain characteristics, and dimension-specific information are integrated.

6 Evaluation Results

Table 3 presents the official evaluation results of the eight participating teams in the DSA-NIFT shared task. Performance was evaluated using MAE and PCC for both valence and arousal, and the overall ranking was determined by the average rank across the four evaluation measures. Among the eight teams, five submitted system description papers, whose main approaches are summarized in Table 2. Accordingly, the following method-oriented analysis focuses on these five systems, while the results of BEDO, wuhongyi and TBD are reported only in terms of their official evaluation scores.

CYUT (Jian and Wu, 2026) achieved the best performance across all four evaluation measures, with a valence MAE of 0.5366, valence PCC of 0.8971, arousal MAE of 0.8658, and arousal PCC of 0.4376. It therefore obtained an average rank of 1.00 and ranked first overall. As described in its system paper, CYUT employed perplexity-guided target-domain adaptation with Qwen2-7B-Instruct and Ministral-3-8B-Instruct-2512, followed by source-task realignment and dimension-specific weighted ensembling. Its strong performance across both affective dimensions suggests that explicitly adapting language models to the linguistic characteristics of the unlabeled target-domain texts was effective in this cross-domain setting.

SCUNLP (Chen et al., 2026) ranked second overall with an average rank of 2.25. It achieved the second-best valence MAE (0.5793), valence PCC (0.8815), and arousal MAE (0.8951), together with an arousal PCC of 0.4050. Rather than directly relying on LLM-generated VA scores, SCUNLP used GPT-5.4-mini to generate descriptions of affective information and concatenated them with the original texts before BERT-based regression. Its system paper reports that this augmentation reduced valence MAE compared with its BERT baseline, indicating that LLM-generated affective descriptions can provide useful auxiliary semantic information for a supervised regression model.

The third-ranked system, aoff123 (Cheng et al.,

Team (Submission)	Evaluation Metric				Avg. Rank	Overall Rank
	V-MAE (rank)	V-PCC (rank)	A-MAE (rank)	A-PCC (rank)		
CYUT (#848249)	**0.5366 (1)**	**0.8971 (1)**	**0.8658 (1)**	**0.4376 (1)**	**1.00**	**1**
SCUNLP (#851040)	0.5793 (2)	0.8815 (2)	0.8951 (2)	0.4050 (3)	2.25	2
aoff123 (#849556)	0.6018 (3)	0.8688 (3)	0.9403 (7)	0.4052 (2)	3.75	3
NCKU_NLP (#850173)	0.6200 (4)	0.8663 (4)	0.9259 (4)	0.3566 (5)	4.25	4
BEDO (#851016)	0.6418 (5)	0.8619 (5)	0.8992 (3)	0.3204 (7)	5.00	5
wuhongyi (#845768)	0.6495 (6)	0.8586 (6)	0.9294 (5)	0.3366 (6)	5.75	6
TBD (#845310)	0.6684 (7)	0.8542 (7)	0.9341 (6)	0.3581 (4)	6.00	7
YCW-NTUST (#849542)	0.5555	0.8951	0.8159	0.4438	–	–

Table 3: Evaluation results and overall rankings of the participating systems in the DSA-NIFT task. YCW-NTUST did not complete its submission before the official deadline and was therefore excluded from the official rankings. Its evaluation scores are reported for reference only.

2026), obtained an average rank of 3.75. It achieved competitive valence performance, with an MAE of 0.6018 and a PCC of 0.8688, and obtained the second-highest arousal PCC of 0.4052. However, its arousal MAE of 0.9403 ranked seventh among all participating systems. The system adopted a dimension-specific strategy in which valence was optimized through staged domain adaptation and validation-guided model selection, whereas arousal predictions were preserved from a frozen reference model. The contrasting arousal ranks illustrate that a system can capture relative variation in affective intensity, as measured by PCC, while still exhibiting comparatively large absolute prediction errors.

NCKU_NLP (Chen and Tan, 2026) ranked fourth overall with an average rank of 4.25, obtaining a valence MAE/PCC of 0.6200/0.8663 and an arousal MAE/PCC of 0.9259/0.3566. Its formal submission combined MacBERT representations, affective-lexicon features, and source-aware arousal supervision. The accompanying study also explored pseudo-labeling, pairwise supervision, ensembling, and graph-guided LLM augmentation as diagnostic alternatives. These experiments further highlight the difficulty of transferring auxiliary supervision to arousal prediction: the system paper reports trade-offs between improvements in arousal correlation and increases in absolute error under some augmentation settings.

BEDO, wuhongyi, and TBD ranked fifth, sixth, and seventh overall, respectively. No system description papers were submitted; therefore, their underlying modeling strategies are not analyzed here. Among these three systems, BEDO achieved the third-best arousal MAE of 0.8992 but the lowest arousal PCC of 0.3204, whereas wuhongyi obtained an arousal MAE/PCC of 0.9294/0.3366.

The YCW-NTUST (You et al., 2026) system used a heterogeneous approach combining fine-tuned Chinese encoders, embedding-based regression, synthetic target-style data, affective lexicon signals, and LLM direct scoring, together with dimension-wise prediction fusion. This represents a substantially different strategy from the more unified adaptation pipelines adopted by several other participating systems and illustrates the methodological diversity explored in the shared task.

A notable overall pattern is the substantial performance gap between valence and arousal prediction. Across the eight participating teams, valence PCC ranges from 0.8542 to 0.8971, showing consistently strong positive correlations with the gold-standard scores. In contrast, arousal PCC ranges from only

0.3204 to 0.4376. The same tendency is observed for absolute prediction error: the best valence MAE is 0.5366, compared with 0.8658 for arousal. These results indicate that arousal is considerably more difficult to predict than valence in the DSA-NIFT dataset.

The submitted system papers also reflect this difficulty. Several teams explicitly adopted dimension-specific strategies, including separate ensemble weights, frozen prediction streams, source-aware supervision, and arousal-specific calibration or model combination. Rather than assuming that valence and arousal can be modeled identically, these approaches suggest that the two dimensions may benefit from different sources of supervision and adaptation strategies.

Overall, the results reveal two major trends. First, reducing the mismatch between source-domain affective resources and target-domain new-immigrant narratives is an important consideration across the participating systems. The strongest systems explored target-domain adaptation, LLM-assisted semantic augmentation, staged transfer, affective knowledge integration, or combinations of these techniques. Second, MAE and PCC provide complementary perspectives on prediction quality, particularly for arousal, where several systems exhibit noticeably different rankings under the two metrics. These findings support the use of multiple evaluation criteria and dimension-specific modeling when developing robust dimensional sentiment analysis systems for linguistically diverse narrative texts.

7 Conclusions

This paper presented the ROCLING 2026 shared task on Dimensional Sentiment Analysis for New Immigrants' Feeling Texts (DSA-NIFT), the fifth edition of the Chinese DSA shared task series. DSA-NIFT extends previous dimensional sentiment analysis research to Chinese narratives written by new immigrants, requiring systems to predict continuous valence and arousal scores. By using Chinese EmoBank as the source-domain training resource and new-immigrant narratives for development and evaluation, the task provides a challenging setting for evaluating cross-domain generalization in fine-grained affective prediction.

A total of eight teams participated in the shared task, and five submitted system description papers. The participating systems explored a diverse range of approaches, including target-domain adaptation, staged transfer learning, LLM-assisted semantic augmentation, affective-lexicon integration, synthetic-data augmentation, and heterogeneous model combination. Among the submitted systems, CYUT (Jian and Wu, 2026) achieved the best performance across all four evaluation measures and ranked first overall.

The evaluation results reveal a clear performance gap between the two affective dimensions. While all participating systems achieved relatively strong correlations for valence prediction, arousal prediction consistently resulted in higher MAE and substantially lower PCC. The system papers also reflect this challenge, with several teams adopting dimension-specific supervision, adaptation, calibration, or model combination strategies. These findings suggest that valence and arousal should not necessarily be treated as equally difficult or modeled using identical strategies.

Overall, DSA-NIFT provides a benchmark for investigating dimensional sentiment analysis under both domain and linguistic variation. Future editions may expand the scale and diversity of new-immigrant narratives, investigate more effective domain adaptation and data augmentation strategies, and further explore dimension-specific modeling for arousal. The shared task is expected to encourage continued research on robust and fine-grained affective computing for linguistically diverse Chinese texts.

Acknowledgments

This study was partially supported by the National Science and Technology Council, Taiwan, under the grant NSTC 114-2221-E-A49-059-MY3. This work was also financially supported by the Co-creation Platform of the Industry Academia Innovation School, NYCU.

References

Chih-Hsi Chen and Hang Tan. 2026. NCKU_NLP at ROCLING-2026 Shared Task: Source-aware lexicon-augmented regression for valence–arousal prediction. In *Proceedings of the 38th Conference on Computational Linguistics and Speech Processing*.

Hsiang-Wei Chen, He-Chang Chan, and Jheng-Long Wu. 2026. SCUNLP at ROCLING-2026 Shared Task: Augmenting BERT with LLM-generated affective text for valence-arousal prediction of new immigrants' texts. In *Proceedings of the 38th Con-*

ference on Computational Linguistics and Speech Processing.

Chih-Yuan Cheng, Yi-Hsuan Huang, and Bor-Jiunn Hwang. 2026. aoff123 at ROCLING-2026 Shared Task: A validation-guided protocol for staged domain adaptation in Chinese dimensional sentiment analysis. In *Proceedings of the 38th 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.

Yu-Chih Deng, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2023. Toward transformer fusions for Chinese sentiment intensity prediction in valence-arousal dimensions. *IEEE Access*, 11:109974–109982.

Zhijie Dou, Zixuan Cheng, and Dongmei Huang. 2021. Research on migrant works' concern recognition and emotion analysis based on web text data. *Frontiers in Psychology*, 12.

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 and Shih-Hung Wu. 2026. CYUT at ROCLING-2026 Shared Task: Perplexity-guided target-domain adaptation for Chinese dimensional sentiment regression. In *Proceedings of the 38th Conference on Computational Linguistics and Speech Processing*.

Sarah Knudson, Srijita Sarkar, and Abhik Ray. 2016. Connecting data science and qualitative interview insights through sentiment analysis to assess migrants' emotion states post-settlement. *arXiv:1609.08776*.

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(3):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*, pages 375–380.

Tzu-Mi Lin, Jung-Ying Chang, and Lung-Hao Lee. 2023. NCUEE-NLP at WASSA 2023 Shared Task 1: Empathy and emotion 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.

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.

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.

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*, pages 548–552.

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.

Georgios Paltoglou and Michael Thelwall. 2013. Seeing stars of valence and arousal in blog posts. *IEEE Transactions on Affective Computing*, 4(1):116–123.

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.

Purbita Samanta and Subrata Chattopadhyay. 2025. Exploring the perception of ethnic migrants in Kolkata, India: A comparative study using sentiment analysis. *Journal of Urban Affairs*, 47(10):3634–3653.

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

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):163–174.

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.

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.

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.

Chuan-Jhih You, Yu-Cheng Chao, and Kuan-Yu Chen. 2026. YCW-NTUST at ROCLING-2026 Shared Task: Board-feedback closed-form ensembling with LLM direct scoring for Chinese valence–arousal regression in new-immigrant texts. In *Proceedings of the 38th Conference on Computational Linguistics and Speech Processing*.

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, 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 20th International Conference on Asian Language Processing*, pages 156–160.

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.

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.

Liang-Chih Yu, Jin Wang, Bo Peng, and 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.