ICT-NLP at SemEval-2026 Task 3: Less Is More — Multilingual Encoder with Joint Training and Adaptive Ensemble for Dimensional Aspect Sentiment Regression

Liyuan Huang[1,2,3], Jiawei He[1,2], Wutao Shen[1,2,3], Lin Li[1,2], Jin Zhang[1,2]
[1]State Key Laboratory of AI Safety
[2]Institute of Computing Technology, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
{huangliyuan25e,shenwutao25e,lilin2020,jinzhang}@ict.ac.cn
hepisces@qq.com

Abstract

This paper describes our system to SemEval-2026 Task 3 Track A Subtask 1 on Dimensional Aspect Sentiment Regression (DimASR). We propose a lightweight and resource-efficient system built entirely on multilingual pre-trained encoders, without relying on LLMs or external corpora. We adopt joint multilingual and multi-domain training to facilitate cross-lingual transfer and alleviate data sparsity, introduce a bounded regression transformation that improves training stability while constraining predictions within the valid range, and employ an adaptive ensemble strategy via subset search to reduce prediction variance. Experimental results demonstrate that our system achieves strong and consistent performance, ranking 1st on *zho-res*, 2nd on *zho-lap*, and 3rd on *jpn-hot*, with all remaining datasets placed within the top half of participating teams[1].

1 Introduction

Aspect-based sentiment analysis (ABSA) aims to identify the sentiment expressed toward specific aspects in text (Pontiki et al., 2014; Zhang et al., 2022). Most existing ABSA studies formulate the task as coarse-grained classification (e.g., *positive*, *negative*, *neutral*), which cannot capture fine-grained affective distinctions such as those between *good* and *excellent*. Drawing on the theory from affective science (Russell, 1980), sentiment can be represented along fine-grained, real-valued dimensions of **valence** (negative–positive) and **arousal** (sluggish–excited).

The DimABSA shared task (Yu et al., 2026) introduces a multilingual and multi-domain benchmark that integrates dimensional sentiment analysis into the traditional ABSA framework. We focus on Track A Subtask 1, Dimensional Aspect Sentiment Regression (DimASR), which requires predicting continuous valence–arousal (VA) scores (1–9 scale) for a given aspect within a text (Figure 1).

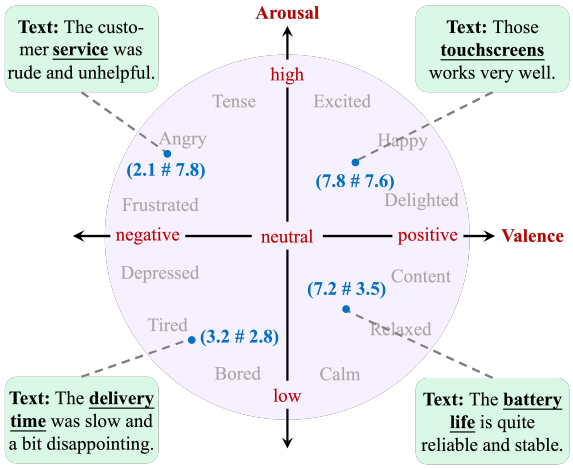

Figure 1: Illustration of Subtask 1 (DimASR).

While large language models (LLMs) have become a dominant strategy for sentiment tasks, such approaches often involve high computational cost and limited reproducibility. In contrast, we investigate how competitive a lightweight system can be without using LLMs.

In this paper, we present a lightweight and resource-efficient system built entirely on multilingual pre-trained encoders without leveraging LLMs, external corpora, or data augmentation techniques. We adopt a joint multilingual and multi-domain training strategy to facilitate cross-lingual transfer and alleviate data scarcity. We further introduce a sigmoid-based bounded output transformation to improve training stability and ensure predictions within the valid range. Finally, we apply an adaptive ensemble strategy via exhaustive search to select the optimal subset from the candidate model pool for each specific language–domain pair, significantly reducing prediction variance.

Our system achieves strong performance across all 10 language–domain datasets, ranking 1st on

[1]Our code is available at https://github.com/huangly312/SemEval2026-task3-DimASR.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 950–957

zho-res, 2nd on *zho-lap*, and 3rd on *jpn-hot*, with all remaining datasets placed within the top half of participating teams. These results suggest that, in dimensional sentiment regression, less is more: a carefully designed lightweight encoder-based system without LLMs can remain highly competitive against approaches that demand far greater computational resources.

2 Related Work

Prior work on dimensional sentiment analysis has focused on building affective resources and developing regression models across multiple granularities. On the resource side, sentiment lexicons provide word-level VA norms (Warriner et al., 2013; Mohammad, 2018), while sentence-level corpora offer broader contextual coverage (Preoţiuc-Pietro et al., 2016; Buechel and Hahn, 2017). Chinese dimensional resources have also been developed to address the gap in non-English coverage (Yu et al., 2016; Lee et al., 2022). On the modeling side, early approaches relied on LSTM-based architectures for VA prediction (Wu et al., 2017; Cheng et al., 2021; Wang et al., 2016), while pre-trained Transformers subsequently became the dominant paradigm. For instance, by employing pre-trained BERT and MLP for V-A prediction, (Xu et al., 2024), or incorporating contrastive learning approaches (Tong and Wei, 2024). More recently, LLM-based methods have shown strong performance on dimensional sentiment tasks, either via in-context learning(ICL) or supervised fine-tuning(SFT) (Xu et al., 2024), and represent the prevailing approach in recent shared task competitions.

3 System Overview

Figure 2 illustrates the overall pipeline of our system. Given a text and its associated aspect, we encode them as a sentence pair using a multilingual pre-trained encoder, and predict the VA scores via a regression head. Models are trained jointly across all language–domain pairs, and the final predictions are obtained via a development-set-guided adaptive ensemble.

3.1 Data Processing

The training data is provided in a quadruplet format (*Aspect*, *Category*, *Opinion*, *VA*) shared across all subtasks of Track A. Since we focus on Subtask 1 (DimASR), we extract only the (*Text*, *Aspect*, *VA*) fields and apply several preprocessing steps before training.

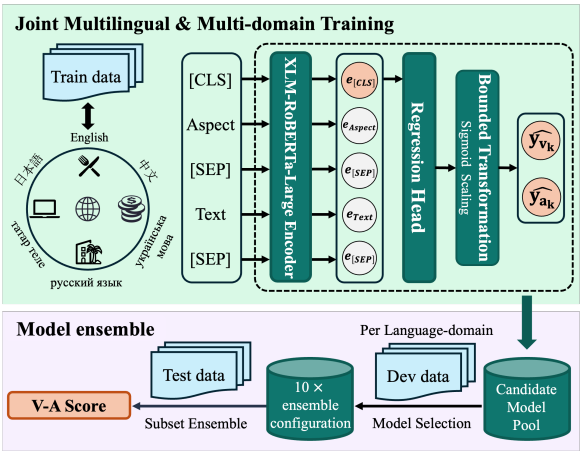

Figure 2: The architecture of our DimASR system.

Following the official dataset note that the test set excludes implicit `NULL` annotations, we remove all training instances where the aspect is `NULL` to ensure consistency between training and test distributions. We also discard 3 instances with VA values outside the valid range $[1, 9]$. For instances with multiple aspect terms, we expand them into independent samples, each paired with the full review text and a single aspect term along with its corresponding VA score. For instances where multiple opinions are associated with the same aspect, we retain only the VA score corresponding to the first opinion.

3.2 Model Architecture

Backbone Encoders. We explore three multilingual pre-trained encoders as the backbone: mBERT (Devlin et al., 2019), XLM-RoBERTa-base, and XLM-RoBERTa-large (Conneau et al., 2020). These models provide strong cross-lingual representations and are well suited to our multilingual setting. In addition, these models span a range of scales and architectures, allowing us to examine the effect of model size on cross-lingual sentiment regression.

Input Representation. The aspect term and review text are encoded as a sentence pair. Depending on the backbone, the tokenizer inserts model-specific special tokens: mBERT formats the input as `[CLS] aspect [SEP] text [SEP]`, while XLM-RoBERTa uses `<s> aspect </s></s> text </s>`. All sequences are truncated or padded to a maximum length of 128 tokens.

Regression Head. We take the hidden state of the first special token ([CLS] or <s>) as the sentence-level representation. A dropout layer is applied before a linear projection that maps the hidden representation to two raw output values corresponding to valence and arousal:

$$\mathbf{h} = \text{Dropout}(\mathbf{e}_{\texttt{[CLS]}}), \quad \hat{\mathbf{y}} = \mathbf{W}\mathbf{h} + \mathbf{b}, \quad (1)$$

where $\mathbf{W} \in \mathbb{R}^{2\times d}$ and $\mathbf{b} \in \mathbb{R}^2$ are learnable parameters, and d is the hidden size of the backbone. The model is trained by minimizing the Mean Squared Error (MSE) loss, which is consistent with the official evaluation metric RMSE_{VA} (Equation 3).

Bounded Output Transformation. The official task defines a valid VA range of $[1, 9]$. To align the model's output space with this requirement and enhance training stability, we further apply a sigmoid-based bounded transformation to the raw regression output:

$$\hat{\mathbf{y}}_{\text{bounded}} = \sigma(\hat{\mathbf{y}}) \times 8 + 1, \quad (2)$$

where $\sigma(\cdot)$ denotes the sigmoid function.

This ensures that all predictions are valid at inference time and facilitates smoother convergence during the early stages of training. During our experiments, we observe that although the transformation stabilizes the training process and yields a consistent overall improvement, its effect varies across individual settings. Therefore, we treat it as a flexible component and include both bounded and unbounded model variants in the candidate pool for the subsequent ensemble stage to leverage their complementary strengths.

3.3 Joint Multilingual and Multi-Domain Training

We adopt a joint multilingual and multi-domain training strategy, pooling data from all language–domain pairs to train a single unified model. Each instance is represented solely by its text–aspect pair, without explicit language or domain indicators.

Compared with language–domain specific training, which learns independent models for each pair, joint training mitigates overfitting in low-resource settings and encourages the model to learn more generalizable representations that facilitate cross-lingual and cross-domain knowledge transfer (Conneau et al., 2020; Thin et al., 2023). In particular, for language–domain pairs with limited training instances, the shared encoder is exposed to a substantially larger and more diverse set of sentiment expressions, allowing low-resource pairs to benefit from representations learned from richer languages and domains, thereby compensating for the lack of in-domain supervision.

3.4 Model Ensemble

To reduce prediction variance and improve robustness, we adopt a structured ensemble strategy.

Candidate Pool Construction. We evaluate all trained models on the development sets across language–domain pairs and select 7 checkpoints that consistently achieve strong overall performance while collectively covering the top-performing models for each individual pair, forming the candidate pool for ensemble. All 7 candidates are built on the XLM-RoBERTa-large backbone, but differ in batch size, learning rate, number of training epochs, and whether the sigmoid output constraint (Equation 2) is applied, providing ensemble diversity beyond hyperparameter variation alone. Detailed configurations of the candidate models are provided in Appendix A.

Ensemble Subset Selection. Rather than a naive uniform averaging, we perform an exhaustive search over all possible subsets of size 2 to 7 drawn from the candidate model pool for each language–domain pair independently. For each subset, predictions are obtained by averaging the valence and arousal outputs of the selected models element-wise. The subset that achieves the lowest RMSE_{VA} on the development set is selected as the final ensemble configuration for that specific pair. As the optimal subset varies across pairs, this process results in 10 distinct ensemble configurations, which are subsequently applied to the official test sets for final submission. The selected optimal subsets for each language–domain pair are reported in Appendix B.

This exhaustive yet controlled ensemble selection strategy effectively reduces prediction variance and consistently improves performance across all language–domain pairs.

4 Experimental Setup

Dataset. The official DimABSA Track A dataset (Lee et al., 2026) covers 6 languages (Chinese, English, Japanese, Russian, Tatar, and Ukrainian) and 4 domains (Hotel, Laptop, Restaurant, and Finance), organized into 10 language–domain pairs for Subtask 1 (DimASR).

In our experiments, we utilize the official training and development splits, and reserve 10% of the training data as an internal validation set.

Evaluation Metrics. According to the official task guidelines, Subtask 1 (DimASR) is evaluated by measuring the prediction error in the VA space using Root Mean Squared Error (RMSE). The metric is defined as:

$$\mathrm{RMSE}_{\mathrm{VA}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(V_p^{(i)} - V_g^{(i)}\right)^2 + \left(A_p^{(i)} - A_g^{(i)}\right)^2} \tag{3}$$

where N is the number of instances; V_p and A_p denote the predicted valence and arousal values for an instance; and V_g and A_g denote the corresponding gold values.

Implementation Details. Our models are implemented using PyTorch 2.3 and Hugging Face Transformers 4.46.1. All experiments are conducted on a single NVIDIA A800 GPU (40GB VRAM). The model parameters are optimized using AdamW with early stopping based on validation RMSE, where the patience is set to 2. To promote model diversity for ensemble, we train multiple models under different hyperparameter configurations. Specifically, we vary the batch size in $\{16, 32, 64\}$ and the learning rate within the range of 8×10^{-6} to 3×10^{-5}. All experiments are conducted with a fixed random seed of 42.

Baselines. We compare our method with the official baseline results reported by the task organizers (Lee et al., 2026). These baselines include LLM-based approaches evaluated in zero-shot and few-shot settings, such as **Kimi K2 Thinking** (MoonshotAI, 2025), as well as supervised fine-tuned LLMs, including **Qwen-3 14B** (Alibaba, 2025) and **GPT-OSS 120B** (OpenAI, 2025).

5 Results

5.1 Main Results

Table 1 reports the final performance of our system on the official test sets across all 10 language–domain pairs, in comparison with the official baselines. Our system outperforms the strongest supervised fine-tuning baseline, GPT-OSS 120B, on 9 out of 10 language–domain pairs. Moreover, it consistently surpasses both the one-shot prompted closed-source LLM Kimi K2 Thinking and the QLoRA fine-tuned Qwen3-14B across all pairs.

On the official leaderboard, our system ranks 1st on *zho-res*, 2nd on *zho-lap*, and 3rd on *jpn-hot*, with all remaining datasets placed within the top half of participating teams. These results demonstrate that a lightweight PLM-based system, when equipped with joint multilingual training and adaptive ensemble selection, can be highly competitive against LLM-based approaches without relying on external corpora or data augmentation.

5.2 Ablation Study

To validate the contribution of each component in our framework, we conduct a series of ablation experiments on the official development set. Results are reported in Table 2.

Joint Multilingual Training. Taking mBERT as a representative backbone, we observe that joint training significantly outperforms separate training across all pairs, yielding an overall average RMSE reduction of 10.9%. Notably, the gains are most pronounced for low-resource language-domain pairs, such as *ukr-res* (-21.3%) and *tat-res* (-19.9%), whereas improvements are comparatively modest for higher-resource pairs such as *eng-lap* (-1.2%) and *zho-lap* (-2.9%). This asymmetric pattern suggests that pooling data across languages and domains provides implicit supervision for low-resource pairs, effectively alleviating data sparsity through shared cross-lingual representations.

Backbone Model Size. Block 2 compares three backbone encoders under joint training. Scaling from mBERT to XLM-R Large results in a consistent performance boost, demonstrating that the superior cross-lingual representation capability of XLM-R Large provides a more robust foundation for capturing the subtle nuances of valence and arousal dimensions.

Bounded Output Transformation. We examine the effect of the sigmoid-based bounded transformation (Equation 2) by using a representative configuration (XLM-R Large, batch size 16, learning rate 1e-5), and observe an average RMSE reduction of 3.7%. Although the gains are not uniform across all language–domain pairs, the bounded transformation generally stabilizes training and prevents out-of-range predictions. In the final ensemble, we include both bounded and unbounded models to balance robustness and flexibility, leveraging their complementary strengths.

Ensemble Strategy. Block 4 shows that the adaptive subset ensemble consistently outperforms the best single model across all 10 datasets (avg.

Methods	eng-res	eng-lap	jpn-hot	jpn-fin	rus-res	tat-res	ukr-res	zho-res	zho-lap	zho-fin	**Avg.**
Zero-Shot Learning											
Kimi K2 Thinking†	2.3432	2.6546	2.3294	2.3379	2.0630	2.3636	2.0782	2.6230	2.0426	2.9662	2.3802
One-Shot Learning											
Kimi K2 Thinking†	2.1461	2.1893	1.7553	1.6396	1.7768	1.9380	1.7805	1.8959	1.6440	1.9652	1.8731
Supervised Fine-Tuning											
Qwen-3 14B†	2.6427	2.8089	2.2906	1.8964	2.1528	2.6367	2.2121	2.0073	1.7706	1.4707	2.1889
GPT-OSS 120B†	1.4605	1.5269	0.7188	1.0188	1.4775	**1.7153**	1.5166	1.0349	0.8032	0.6511	1.1924
Ours	**1.2676**	**1.3098**	**0.6289**	**0.8331**	**1.4420**	1.8596	**1.4750**	**0.9256**	**0.6553**	**0.4892**	**1.0886**

Table 1: Track A Subtask 1 Results. RMSE_{VA} on the official test sets across 10 language–domain pairs. The **best** results are in bold, and the second-best are underlined. Baseline results† from (Lee et al., 2026).

Variant	eng-res	eng-lap	jpn-hot	jpn-fin	rus-res	tat-res	ukr-res	zho-res	zho-lap	zho-fin	**Avg.**
1. Training Strategy (mBERT)											
Separate Training	1.3086	1.2863	1.1860	1.2455	1.8048	2.1028	1.8087	0.8985	0.8467	0.6697	1.3158
Joint Training	**1.1537**	**1.2715**	**1.1590**	**1.2090**	**1.5532**	**1.6845**	**1.4242**	**0.8048**	**0.8223**	**0.6482**	**1.1730**
Relative Change	-11.8%	-1.2%	-2.3%	-2.9%	-13.9%	-19.9%	-21.3%	-10.4%	-2.9%	-3.2%	**-10.9%**
2. Backbone Model Size (Joint Training)											
mBERT	1.1537	1.2715	1.1590	1.2090	1.5532	1.6845	1.4242	0.8048	0.8223	0.6482	1.1730
XLM-R Base	1.1498	1.1072	0.9672	0.9663	1.5050	1.8109	1.4582	0.7790	0.7951	0.6048	1.1144
XLM-R Large	**1.0892**	**1.0383**	**0.9416**	**0.9286**	**1.4000**	**1.6519**	**1.4198**	**0.6987**	**0.7550**	**0.5808**	**1.0504**
3. Bounded Output (Joint Training; XLM-R Large)											
w/o sigmoid	1.0892	1.0383	0.9416	0.9286	1.4000	1.6519	1.4198	0.6987	0.7550	0.5808	1.0504
w/ sigmoid	1.1294	0.9978	0.9157	0.8299	1.3598	1.6599	1.3258	0.7420	0.6763	0.4838	1.0120
Relative Change	+3.7%	-3.9%	-2.8%	-10.6%	-2.9%	+0.5%	-6.6%	+6.2%	-10.4%	-16.7%	**-3.7%**
4. Inference Strategy											
Best Single Model	0.9299	0.9287	0.8823	0.7898	1.2777	1.5109	1.3198	0.6614	0.6763	0.4838	0.9461
Ensemble	**0.9165**	**0.9020**	**0.8568**	**0.7601**	**1.2648**	**1.4397**	**1.2661**	**0.6492**	**0.6509**	**0.4766**	**0.9183**
Relative Change	-1.4%	-2.9%	-2.9%	-3.8%	-1.0%	-4.7%	-4.1%	-1.8%	-3.8%	-1.5%	**-2.9%**

Table 2: Ablation study on the official development set (RMSE_{VA}). Bold font indicates the best performance within each comparison block.

−2.9%), confirming that exhaustive subset search over the candidate pool effectively reduces prediction variance.

6 Conclusion

In this paper, we present a lightweight and resource-efficient system for Subtask 1 (Dimensional Aspect Sentiment Regression) of SemEval-2026 Task 3 Track A. Without relying on LLMs, external corpora, or data augmentation, our approach leverages joint multilingual training and an adaptive ensemble strategy.

Extensive experiments across 10 language–domain pairs demonstrate that, in dimensional sentiment regression, less is more: a carefully designed PLM-based system remains highly competitive against strong LLM-based baselines.

7 Ethical Considerations

All data used in this work are sourced from the official competition release and are used solely for scientific research purposes in strict accordance with the provided data usage agreements. Our system does not involve sensitive personal information. AI-assisted tools were used only for language polishing, while all research ideas, experimental design, and conclusions were independently developed and verified by the authors.

References

Alibaba. 2025. Qwen3 technical report. *arXiv preprint*, page arXiv:2505.09388.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis.

In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Yu-Ya Cheng, Yan-Ming Chen, Wen-Chao Yeh, and Yung-Chun Chang. 2021. Valence and arousal-infused bi-directional lstm for sentiment analysis of government social media management. *Applied Sciences*, 11(2):880.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–18.

Lung-Hao Lee, Liang-Chih Yu, Natalia Loukashevich, Ilseyar Alimova, Alexander Panchenko, Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, Guangmin Zheng, Jin Wang, Sharanya Awasthi, Jonas Becker, Jan Philip Wahle, Terry Ruas, Shamsuddeen Hassan Muhammad, and Saif M. Mohammad. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *Preprint*, arXiv:2601.23022.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

MoonshotAI. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint*, page arXiv:2507.20534.

OpenAI. 2025. Gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint*, page arXiv:2508.10925.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Dang Van Thin, Hung Quoc Ngo, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023. Exploring zero-shot and joint training cross-lingual strategies for aspect-based sentiment analysis based on contextualized multilingual language models. *Journal of Information and Telecommunication*, 7(2):121–143.

Zeliang Tong and Wei Wei. 2024. CCIIPLab at SIGHAN-2024 dimABSA task: Contrastive learning-enhanced span-based framework for Chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 102–111, Bangkok, Thailand. Association for Computational Linguistics.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 225–230, Berlin, Germany. Association for Computational Linguistics.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, Sixing Wu, and Zhigang Yuan. 2017. THU_NGN at IJCNLP-2017 task 2: Dimensional sentiment analysis for Chinese phrases with deep LSTM. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 47–52, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Hongling Xu, Delong Zhang, Yice Zhang, and Ruifeng 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 (SIGHAN-10)*, pages 175–185, Bangkok, Thailand. Association for Computational Linguistics.

Liang-Chih Yu, Jonas Becker, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Lung-Hao Lee, Ying-Lung Lin, Jin Wang, Jan Philip Wahle, Terry

Ruas, Alexander Panchenko, Ilseyar Alimova, Kai-Wei Chang, Lilian Wanzare, Nelson Odhiambo, Bela Gipp, and Saif M. Mohammad. 2026. SemEval-2026 task 3: Dimensional aspect-based sentiment analysis (DimABSA). In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

Appendix

A Candidate Model Hyperparameter Configurations

Model ID	Batch Size	Learning Rate	Epoch	Sigmoid
M1	16	1e-5	7	✓
M2	32	1e-5	3	
M3	32	1e-5	5	✓
M4	32	1e-5	7	✓
M5	32	2e-5	5	✓
M6	32	8e-6	3	✓
M7	32	8e-6	7	

Table 3: Hyperparameter configurations of the seven candidate models. All models utilize XLM-RoBERTa-large as the backbone and are trained under a joint multilingual and multi-domain setting.

B Ensemble Selection Results

Pair	M1	M2	M3	M4	M5	M6	M7	Number
eng-lap				✓	✓			2
eng-res				✓		✓		2
jpn-fin	✓		✓	✓			✓	4
jpn-hot			✓			✓	✓	3
rus-res	✓				✓			2
tat-res		✓		✓			✓	3
ukr-res	✓				✓			2
zho-fin	✓			✓			✓	3
zho-lap	✓				✓	✓	✓	4
zho-res		✓	✓				✓	3

Table 4: Selected ensemble configurations for each language–domain pair. ✓ indicates that the corresponding candidate model is included in the final ensemble.