NCL-BU at SemEval-2026 Task 3: Fine-tuning XLM-RoBERTa for Multilingual Dimensional Sentiment Regression

Tong Wu[1], Nicolay Rusnachenko[2], and Huizhi Liang[3]
[1]Independent Researcher
[2]Centre for Applied Creative Technologies (CFACT+), Bournemouth University, UK
[3]School of Computing, Newcastle University, Newcastle upon Tyne, UK
tongwuwhitney@gmail.com, nrusnachenko@bournemouth.ac.uk,
huizhi.liang@newcastle.ac.uk

Abstract

Dimensional Aspect-Based Sentiment Analysis (DimABSA) extends traditional ABSA from categorical polarity labels to continuous valence–arousal (VA) regression. This paper describes a system developed for Track A, Subtask 1 (Dimensional Aspect Sentiment Regression), aiming to predict real-valued VA scores in the $[1, 9]$ range for each given aspect in a text. A fine-tuning approach based on XLM-RoBERTa-base is adopted, with dual regression heads with sigmoid-scaled outputs for valence and arousal prediction. Separate models are trained for each language–domain pair (English and Chinese across restaurant, laptop, and finance domains), and training and development sets are merged for final test predictions. In development experiments, the fine-tuning approach is compared against several large language models under a few-shot prompting setting, demonstrating that task-specific fine-tuning outperforms these LLM-based methods across all evaluation datasets.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) aims to identify sentiment expressed toward specific aspects in text and has been extensively studied in the NLP community (Zhang et al., 2023). Traditional ABSA assigns categorical sentiment labels (e.g., positive, negative, neutral) to each aspect, but such coarse-grained representations fail to capture the nuanced spectrum of human emotion. SemEval-2026 Task 3 - Dimensional ABSA (DimABSA) (Yu et al., 2026; Lee et al., 2026) bridges this gap by extending ABSA from categorical polarity to continuous valence–arousal (VA) regression based on the circumplex model of affect (Russell, 1980).

This paper focuses on Track A, Subtask 1 (Dimensional Aspect Sentiment Regression, DimASR), which requires predicting a pair of real-valued VA scores in the range [1, 9] for each given aspect in a text. This work addresses English and Chinese across multiple domains (restaurant, laptop, and finance). The approach fine-tunes XLM-RoBERTa-base (Conneau et al., 2020), a multilingual pretrained language model, with dual regression heads for valence and arousal prediction. The model takes the concatenation of the input text and the target aspect as input and outputs two real-valued scores via sigmoid-scaled linear heads.

LLM-based approaches have shown strong performance in ABSA classification tasks (Zhang et al., 2024), raising the question of whether they can similarly excel at continuous VA regression. To investigate this, a comparative study is conducted between task-specific fine-tuning and few-shot prompting with several state-of-the-art large language models (LLMs). Under the evaluated few-shot prompting conditions, fine-tuning XLM-RoBERTa on modest in-domain annotated data outperforms the evaluated LLM-based methods across all language–domain pairs. The system also outperforms both organizer-provided baselines (Kimi-K2 Thinking and QLoRA-fine-tuned Qwen-3 14B) on all five datasets, with relative improvements of 31–63%. Error analysis on the test predictions further reveals that the model struggles most with strongly negative sentiment, where it tends to predict values closer to the positive training mean, and that valence is harder to predict than arousal across datasets. The code is available[1].

2 Background

Task Description. This work participates in SemEval-2026 Task 3, Track A, Subtask 1 (Lee et al., 2026). Given a sentence and a target aspect, the system must predict a valence score and an arousal score, both in the range $[1, 9]$. For example, given the sentence "the food was absolutely amazing!" and the aspect "food", the expected out-

[1]https://github.com/tongwu17/SemEval-2026-Task3-Track-A

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1911–1918

put is $V = 8.50$, $A = 8.25$, indicating positive and excited sentiment. The task provides training data in two languages (English and Chinese) and three domains (restaurant, laptop, and finance).

Aspect-Based Sentiment Analysis. ABSA has been the focus of a series of SemEval shared tasks (Pontiki et al., 2014, 2015, 2016), progressing from aspect sentiment classification to structured extraction tasks such as Aspect Sentiment Triplet Extraction (ASTE) (Peng et al., 2020) and Aspect Sentiment Quad Prediction (ASQP) (Cai et al., 2021; Zhang et al., 2021). These tasks, however, represent sentiment as discrete categorical labels.

Dimensional Sentiment Analysis. In prior work, sentence-level dimensional sentiment resources such as EmoBank (Buechel and Hahn, 2017) and Chinese EmoBank (Lee et al., 2022) have enabled valence–arousal prediction. At the aspect level, the SIGHAN 2024 shared task (Lee et al., 2024) introduced dimensional ABSA for Chinese, and SemEval-2026 Task 3 (Lee et al., 2026) further extends it to multilingual and multidomain settings.

Pretrained Language Models for Regression. Transformer-based pretrained models such as BERT (Devlin et al., 2019) produce a [CLS] token representation as a sentence-level encoding that can be augmented with task-specific layers for regression. XLM-RoBERTa (Conneau et al., 2020) extends this paradigm to multilingual settings through large-scale cross-lingual pretraining on 100 languages, outperforming multilingual BERT on cross-lingual benchmarks. This capability makes it well-suited for multilingual dimensional sentiment regression across languages and domains.

3 Methodology

3.1 Task Definition

Given a text T and a set of aspects $\{a_1, a_2, \ldots, a_k\}$, the task is to predict a valence–arousal pair (V_i, A_i) for each aspect a_i, where $V_i, A_i \in [1, 9]$.

3.2 Model Architecture

XLM-RoBERTa-base (Conneau et al., 2020) is used as the backbone encoder. For each (text, aspect) pair, the input is constructed as:

$$\text{[CLS] } T \text{ [SEP] } a_i \text{ [SEP]}$$

where [SEP] denotes the separator token of the pretrained tokenizer. The [CLS] token representation $\mathbf{h} \in \mathbb{R}^d$, from the encoder is passed through a dropout layer and then fed into two independent regression heads:

$$\hat{V}_i = \sigma(\text{MLP}_V(\mathbf{h})) \times 8 + 1$$
$$\hat{A}_i = \sigma(\text{MLP}_A(\mathbf{h})) \times 8 + 1$$

where σ is the sigmoid function and each MLP is a two-layer feedforward network ($d \to d/2 \to 1$) with tanh activation and dropout between layers. Since σ maps to $[0, 1]$, the affine transformation $\sigma(\cdot) \times 8 + 1$ constrains predictions to $[1, 9]$. Figure 1 illustrates the architecture.

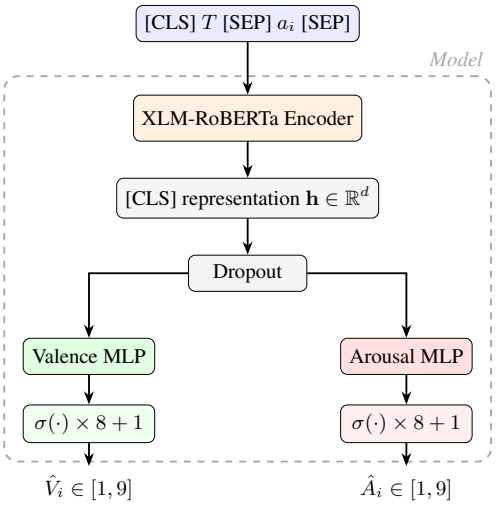

Figure 1: Model architecture.

4 Experimental Setup

4.1 Dataset and Training Strategy

For each language–domain pair, the training data comes from the DimABSA dataset (Lee et al., 2026), constructed by annotating VA values on existing ABSA resources. For example, the English restaurant and laptop data are sourced from the ACOS dataset (Cai et al., 2021). Since Subtask 1 only requires VA regression for given aspects, (text, aspect, VA) triples are extracted from the *Quadruplet* or *Aspect_VA* annotations. When a sentence contains multiple aspects, each (text, aspect) pair is treated as an independent sample with its own VA target. Multi-token aspects are kept as raw text spans and tokenized jointly with the sentence by the XLM-RoBERTa tokenizer; no manual subword aggregation is applied to aspect tokens. Training instances are shuffled at the instance level each epoch to reduce any batching bias from repeated sentence contexts shared across multiple aspects. Since each (text, aspect) pair produces a distinct input sequence of the form [CLS] T [SEP] a_i [SEP],

Language	Domain	Train	Dev	Test
eng	Restaurant	2,284	200	1,000
eng	Laptop	4,076	200	1,000
zho	Restaurant	6,050	300	1,000
zho	Laptop	3,490	300	1,000
zho	Finance	1,000	200	842

Table 1: Dataset statistics (number of sentences).

Hyperparameter	Value
Pretrained model	XLM-RoBERTa-base
Max sequence length	256
Batch size	16
Learning rate	2e-5
Optimizer	AdamW
Warmup ratio	10%
LR scheduler	Linear decay
Dropout	0.1
Max epochs	10
Early stopping patience	3
Gradient clipping	1.0
Random seed	42

Table 2: Hyperparameters for fine-tuning.

the encoder representations are not identical across aspects of the same sentence; instance-level shuffling further ensures that aspects from the same sentence are unlikely to co-occur within the same mini-batch.

Table 1 summarizes the dataset statistics. For the final test submission, the training and development sets are merged to maximize the available training data, while a 10% random split is reserved for internal validation during training.

4.2 Training Configuration

The training loss is the sum of mean squared error (MSE) for valence and arousal:

$$\mathcal{L} = \text{MSE}(\hat{V}, V^*) + \text{MSE}(\hat{A}, A^*)$$

where V^* and A^* are the gold-standard values.

Table 2 lists the hyperparameters. Separate models are trained for each language–domain pair: English restaurant, English laptop, Chinese restaurant, Chinese laptop, and Chinese finance.

The AdamW optimizer (Loshchilov and Hutter, 2019) is used with a linear learning rate schedule including 10% warmup steps. Early stopping with a patience of 3 epochs is applied based on the RMSE_{VA} metric on a held-out validation split (10% of training data). All results reflect a single training run per language–domain pair with a fixed random seed (42).

Method	eng_lap	eng_res	zho_lap	zho_res	zho_fin
Kimi-K2 Thinking	2.1893	2.1461	1.6440	1.8959	1.9652
Qwen-3 14B (QLoRA)	2.8089	2.6427	1.7706	2.0073	1.4707
XLM-RoBERTa	**1.4562**	**1.4861**	**0.7510**	**0.9553**	**0.5391**

Table 3: Test set RMSE_{VA} compared with baselines (lower is better).

4.3 Development Evaluation Protocol

To compare methods during development, the original training set is split into 80% for training and 20% for evaluation. This setup allows comparison of the fine-tuning approach against LLM-based methods on a held-out portion of annotated data.

4.4 Evaluation Metric

The evaluation metric for Subtask 1 is RMSE_{VA}, defined as:

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

where N is the total number of (text, aspect) instances, and subscripts p and g denote predicted and gold values, respectively. Lower values indicate better performance.

5 Results

5.1 Main Results

Models trained on the training set alone, without using any development gold labels for training, achieve RMSE_{VA} of 1.1045, 1.0630, 0.7207, 0.7983, and 0.5389 on the development set for eng_lap, eng_res, zho_lap, zho_res, and zho_fin, respectively. For the final test predictions, the training and development sets are merged to maximize available data, with an internal 10% validation split reserved for early stopping.

Table 3 reports the resulting test scores alongside the organizer-provided baselines. Among the baselines, Kimi-K2 Thinking achieves lower RMSE_{VA} on four of five datasets, while Qwen-3 14B leads only on Chinese Finance (1.4707 vs. 1.9652). XLM-RoBERTa outperforms both baselines on all five datasets, with improvements over the per-dataset stronger baseline ranging from 31% to 63% across the five language–domain pairs. The largest gains occur on Chinese datasets, where relative improvements reach 50–63%, notably higher than the 31–34% gains on English datasets, where all three Chinese datasets achieve RMSE_{VA} below 1.0 while both baselines remain above 1.4.

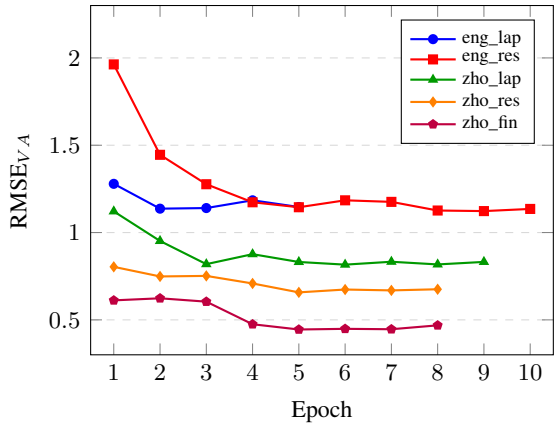

Figure 2: Validation RMSE$_{VA}$ across training epochs.

5.2 Training Convergence Analysis

Figure 2 shows the validation RMSE$_{VA}$ across training epochs for all five datasets when trained on the merged train+dev data. Several observations can be made: (1) Chinese datasets converge to lower error levels than English ones, consistent with the larger training sizes; (2) early stopping effectively prevents overfitting, with most models stopping between epochs 5–9; (3) English Restaurant requires the most epochs (10) to converge, likely due to its smaller training set combined with higher annotation variability.

5.3 Comparison with LLM-based Methods

To compare against prompting-based approaches, several LLMs are evaluated under a few-shot setting on the same 20% held-out evaluation set:

- **GPT-5.2**: A proprietary OpenAI model, accessed via API.
- **LLaMA-3-70B-Instruct**: A 70B-parameter open-source model by Meta.
- **LLaMA-4-Maverick-Instruct**: A 400B-parameter mixture-of-experts model by Meta.

For the LLM methods, a structured prompt is used with system instructions defining valence and arousal, followed by 6 few-shot examples sampled from the 80% training portion. A low temperature of 0.1 is used to encourage deterministic outputs. The prompt asks the model to output VA scores in the format V#A. Predicted values are clipped to the [1, 9] range. The same prompt format and few-shot examples are applied uniformly across all LLMs and datasets. Table 4 presents the results.

Fine-tuning vs. LLMs. Under the few-shot prompting setup, the fine-tuned XLM-RoBERTa models outperform the evaluated LLM-based methods across all datasets. Figure 3 makes this gap visually explicit: the XLM-RoBERTa bars are consistently the lowest in every language–domain pair. The average improvement over the best LLM per dataset is 0.78 RMSE$_{VA}$ points (a relative reduction of approximately 46%). The performance gap is particularly striking for Chinese datasets, where the best LLM achieves RMSE$_{VA}$ of 1.5825 on Chinese Laptop while the fine-tuned model achieves 0.8034, and similarly 1.7279 vs. 0.6727 on Chinese Restaurant.

Method	eng_lap	eng_res	zho_lap	zho_res	zho_fin
GPT-5.2	1.9663	1.9203	1.5825	1.7279	1.7311
LLaMA-3-70B	1.6243	1.6748	1.9249	1.9589	2.2312
LLaMA-4-Mav.	1.6987	1.7955	1.6480	1.7718	1.7858
XLM-RoBERTa	**1.1084**	**1.3402**	**0.8034**	**0.6727**	**0.5118**

Table 4: RMSE$_{VA}$ comparison on the 20% held-out evaluation set (lower is better).

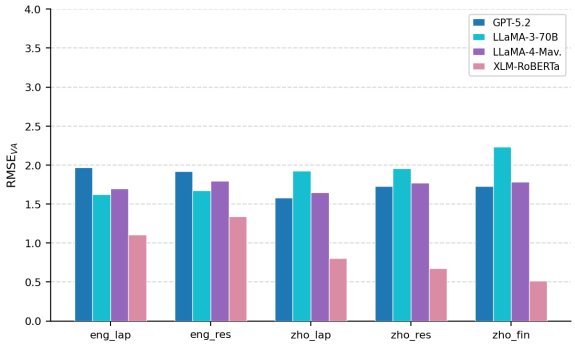

Figure 3: Visual comparison of RMSE$_{VA}$ across methods on the 20% held-out evaluation set.

LLM Comparison. Among the LLMs, no single model dominates across all datasets. Within the prompting-based group, LLaMA-3-70B achieves the lowest RMSE$_{VA}$ on both English datasets, whereas GPT-5.2 performs best on all three Chinese datasets, suggesting language-specific calibration differences. The gap between the best- and worst-performing LLM can be substantial (e.g., 1.7311 to 2.2312 on Chinese Finance), suggesting that VA regression performance varies considerably across models and languages.

Analysis of LLM Limitations. The substantial gap between LLM-based approaches and fine-tuning can be attributed to several factors: (1) VA regression requires predicting precise numerical values on a continuous scale, which is inherently challenging for LLMs that generate text token by token; (2) the few-shot setting provides very limited

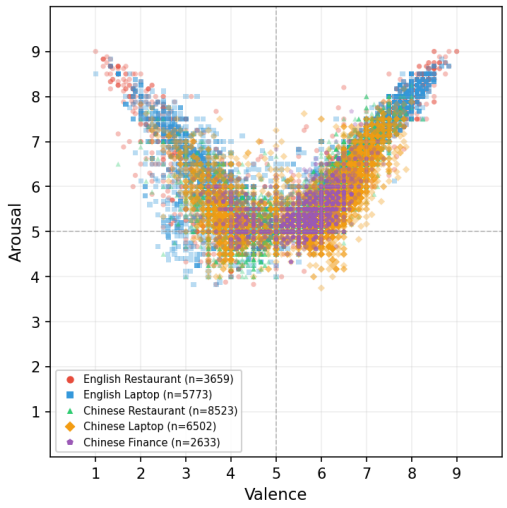

Figure 4: Training data distribution in VA space. English datasets cluster in the high-valence, high-arousal region, while Chinese datasets are more centered.

calibration signal (only 6 examples) for the model to learn the dataset-specific distribution; (3) fine-tuning directly optimizes MSE loss, providing a clear gradient signal for regression, while LLMs must infer numerical patterns from text.

5.4 Error Analysis

Cross-lingual Observations. The fine-tuned model performs better on Chinese datasets than on English ones (average test RMSE_{VA} of 0.7485 vs. 1.4712 for Chinese and English, respectively). This may be attributed to the larger training sets available for Chinese domains (e.g., 6,050 sentences for Chinese Restaurant); annotation consistency differences between languages are plausible but have not been directly verified in this study. Chinese Finance achieves the lowest error (0.5391). The zho_fin annotations are more uniformly distributed across the VA space than the other four datasets, producing the smallest per-dimension errors on both valence (0.44) and arousal (0.31) in Table 5.

Distributional Bias. Figure 4 shows the training data distribution in the VA space across all five datasets. English datasets are heavily skewed toward positive valence and high arousal (mean $V \approx 6.1$, $A \approx 6.8$), while Chinese datasets are more centered (mean $V \approx 5.7$, $A \approx 5.5$). This distributional imbalance directly explains several test prediction patterns, as the model tends to predict values closer to the positive training mean, failing to capture extreme negativity. For example, for the sentence "If the line is longer I would highly recommend skipping this place and heading down the street to Fido – insanely good coffee and great breakfast food," the model predicts nearly identical valence for all three aspects (*place*: $V = 7.38$; *coffee*: $V = 7.43$; *breakfast food*: $V = 7.71$), while gold labels differ substantially ($V = 2.88$, 8.38, and 7.88, respectively). The overall positive tone of the sentence ("highly recommend", "insanely good") dominates the encoding, causing the model to miss the negative sentiment toward *place*. As this example illustrates, although each aspect is individually concatenated to the input, the prediction relies on the [CLS] representation, which is a global mixture of all input tokens via self-attention. Without a dedicated mechanism to focus on tokens relevant to the target aspect, the positive tone of surrounding context can dominate this global encoding regardless of which aspect is being evaluated, causing the model to assign near-identical predictions to aspects with substantially different gold values. This failure is a direct consequence of the [CLS]-based architecture: replacing global pooling with aspect-span attention or explicit aspect markers in the input would be a natural architectural response to such over-smoothing errors.

Figure 5 provides a spatially resolved view of these patterns. For English datasets, the highest errors concentrate in the low-valence region ($V < 3$), with RMSE_{VA} of 2.43 for restaurant and 2.20 for laptop, compared to only 1.09 and 1.08 in the high-valence region $[7, 9]$, confirming the distributional bias. Chinese datasets exhibit more uniform error surfaces, with considerably less extreme variation across the entire VA space, consistent with their more balanced training distributions. Notably, the heatmap cells with the fewest training samples (low-valence, high-arousal) consistently show the largest errors, reinforcing that data scarcity in underrepresented VA regions is the primary driver of prediction errors.

Furthermore, valence is generally harder to predict than arousal across language–domain pairs. Table 5 shows the per-dimension RMSE: valence RMSE exceeds arousal RMSE in four out of five datasets, with the largest gap on English Laptop (1.25 vs. 0.75); the sole exception is Chinese Restaurant, where arousal error is higher (0.74 vs. 0.61). This suggests that arousal, being more closely tied to surface-level cues such as exclamation marks and intensifiers, is easier for the model to capture, while valence requires deeper semantic understanding of sentiment polarity.

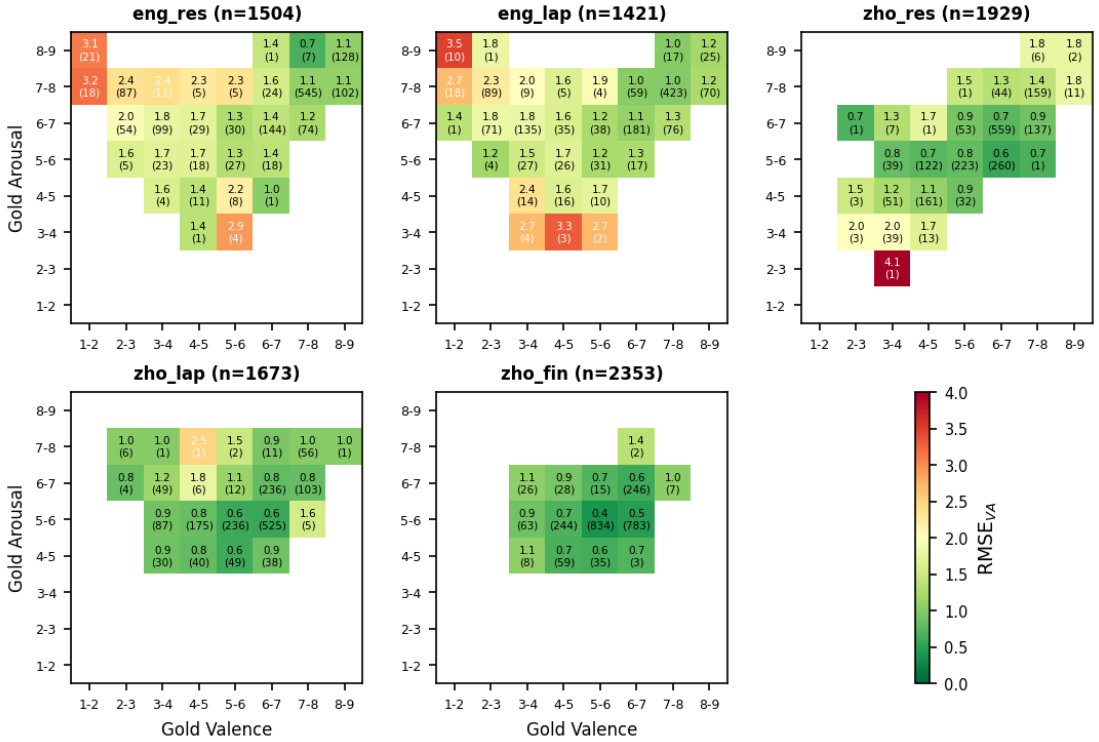

Figure 5: Test-set $RMSE_{VA}$ heatmap across the VA space. Each cell shows the RMSE and sample count (n) for gold samples in that valence–arousal bin. Red indicates higher error; green indicates lower error.

Language	Domain	$RMSE_V$	$RMSE_A$
eng	Laptop	1.25	0.75
eng	Restaurant	1.11	0.98
zho	Laptop	0.57	0.48
zho	Restaurant	0.61	0.74
zho	Finance	0.44	0.31

Table 5: Test-set RMSE for valence ($RMSE_V$) and arousal ($RMSE_A$).

6 Conclusion

This paper presented a fine-tuning approach based on XLM-RoBERTa-base with dual regression heads for dimensional aspect sentiment regression. Experiments on five language–domain pairs demonstrate that task-specific fine-tuning outperforms few-shot prompting with several LLMs under the evaluated conditions. Valence is generally harder to predict than arousal. The error analysis identifies key error patterns: [CLS]-based oversmoothing for co-occurring aspects, and distributional bias toward positive training values. Incorporating these insights into an improved system was precluded by shared-task time constraints and the need to preserve comparability with submitted results; aspect-aware pooling and data augmentation in data-sparse VA regions remain natural directions for future work. Future work includes using larger pretrained models, multi-task learning across subtasks, and incorporating affective lexicons for low-resource domains.

Limitations

The model uses only the [CLS] representation without aspect-level attention. The LLM comparison uses a single prompting setup with 6 few-shot examples, which may not fully represent LLM capabilities. The evaluation covers only two languages, leaving generalizability untested. Each language–domain model is also trained independently without parameter sharing. Joint or multi-task training across domains and languages may improve data efficiency, particularly for smaller datasets such as Chinese Finance. No ablation studies were conducted comparing, e.g., a single regression head with a 2-dimensional output against two independent scalar heads, or aspect-aware pooling against plain [CLS] encoding; such comparisons are left for future work.

Acknowledgments

Thanks to the organizers for providing the dataset and evaluation infrastructure.

References

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.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

Alexis Conneau, Karttikeya Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Marius 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.

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. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(4).

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.

Lung-Hao Lee, Liang-Chih Yu, Suge Wang, and Jian 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 (SIGHAN-10)*, pages 165–174, Bangkok, Thailand. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

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.

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

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.

Chen Zhang, Qiuchi Li, Dawei Song, and Linqi Song. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE*

Transactions on Knowledge and Data Engineering, 35(11):11019–11038.

A Prompt

The following prompt is used for all LLM-based comparison methods. It consists of a system prompt and a set of 6 few-shot examples.

A.1 System Prompt

System Prompt

You are an expert in sentiment analysis. Your task is to predict Valence and Arousal scores for aspects in sentences.

Definitions:

- **Valence**: emotional positivity/negativity (1.0 = very negative, 5.0 = neutral, 9.0 = very positive)
- **Arousal**: emotional intensity/excitement (1.0 = very calm/sluggish, 5.0 = moderate, 9.0 = very excited)

Output format: valence#arousal (e.g., 7.50#6.80)

A.2 Few-shot Examples

Few-shot Examples

1. **Text:** "the food was absolutely amazing!!"
   Aspect: "food"
   Answer: 8.50#8.25
2. **Text:** "but the staff was so horrible to us."
   Aspect: "staff"
   Answer: 1.33#8.67
3. **Text:** "food was just average... if they lowered the prices just a bit, it would be a bigger draw."
   Aspect: "food"
   Answer: 5.00#5.00
4. **Text:** "i love this macbook."
   Aspect: "macbook"
   Answer: 7.10#6.90
5. **Text:** "horrible product."
   Aspect: "product"
   Answer: 2.60#5.70
6. **Text:** "it has and does everything it should."
   Aspect: "NULL"
   Answer: 5.67#5.50

B Prediction Examples

Table 6 shows selected test predictions. Near-exact cases match gold values closely; failure cases involve sarcasm or implicit negativity where the model predicts positive values.

Text (Aspect)		Pred	Gold
Near-exact predictions			
"I enjoy real flavor, real fruits" (*flavor*)	V	7.75	7.75
	A	7.75	7.75
"The keyboard is full size and the spacing . . . comfortable" (*keyboard*)	V	6.99	7.00
	A	7.02	7.00
"We shared the big easy breakfast and it was okay" (*big easy breakfast*)	V	6.49	6.50
	A	6.31	6.33
"The battery still lasts about 3.5 hours . . . " (*battery*)	V	5.99	6.00
	A	5.99	6.00
Failure cases			
"I ordered a burger medium and got a charred, tasteless hockey puck" (*burger*)	V	4.03	1.38
	A	5.22	8.50
"My wife also had the pleasure of adding spoiled creamer" (*creamer*)	V	6.80	1.67
	A	6.75	8.17
"I'm being generous by giving this restaurant 2 stars" (*restaurant*)	V	7.08	2.17
	A	7.08	7.67
"Let me note that this waitress is giving the customers at the 3 tables surrounding us WAY better customer service" (*waitress*)	V	5.84	1.83
	A	5.82	8.00

Table 6: Selected test predictions.

C Error Distribution

Table 7 shows the distribution of per-instance RMSE_{VA} errors on the test set, where each error is the Euclidean distance between predicted and gold VA points on the $[1, 9]$ scale. Over 94% of Chinese Finance predictions fall below 1.0 (accurate), while English datasets have 16–17% with error above 2.0 (large errors); the median error remains below 1.0 across all three Chinese datasets.

Lang	Domain	Median	%<1.0	%>2.0
eng	Laptop	0.87	55.6	16.9
eng	Restaurant	0.78	60.1	16.2
zho	Laptop	0.44	86.6	2.2
zho	Restaurant	0.66	72.2	3.4
zho	Finance	0.38	94.8	0.1

Table 7: Per-instance RMSE_{VA} error distribution on the test set.