ssurface3 at SemEval-2026 Task 3: Efficient Methods for Multilingual Dimensional Aspect-Based Sentiment Analysis

Anatolii Frolov[1] **Elisei Rykov**[1]
[1]Skoltech
Anatolii.Frolov@skoltech.ru

Abstract

This paper describes our submission to the Dimensional Aspect-Based Sentiment Analysis (dimABSA) Shared Task (Subtask 1), which requires predicting continuous Valence and Arousal scores for target aspects in multilingual reviews. We evaluate three approaches: prompting-based baselines, a multilingual encoder model, and a decoder-only LLM with supervised fine-tuning. Our main focus is efficient adaptation under multilingual data scarcity. We show that compact encoder and decoder models, when properly fine-tuned, achieve strong performance across languages and domains. To improve training stability and enforce valid predictions, we use a bounded regression formulation that maps outputs to the target score range. We also explore parameter-efficient fine-tuning and intermediate training on external affective data. Results show that prompting-based baselines are substantially weaker than supervised models. The multilingual encoder provides a strong and efficient baseline, while the best performance is achieved by a compact decoder model with parameter-efficient fine-tuning. Overall, our findings highlight the importance of careful fine-tuning and training design for multilingual dimensional sentiment analysis.The code is available in the GitHub repository[1]

1 Introduction

Traditionally, Aspect-Based Sentiment Analysis (ABSA) (Lee et al., 2026) has been formulated as a discrete classification task, where opinions are categorized as positive, negative, or neutral. However, this coarse-grained formulation fails to capture the complexity and intensity of human emotions. For example, a review stating **The soup was okay** and one stating **The soup was absolutely life-changing** may both be classified as "Positive," despite conveying substantially different emotional intensities.

To address this limitation, recent work has shifted toward Dimensional Aspect-Based Sentiment Analysis (dimABSA). Instead of assigning a single discrete label, dimABSA maps sentiment to a continuous affective space defined by Valence (degree of pleasure) and Arousal (degree of excitement or intensity). As a result, the task becomes a fine-grained regression problem, requiring models to predict precise real-valued scores on a scale from 1.00 to 9.00.

In this work, we address the `dimABSA` challenge (Yu et al., 2026) with a robust framework that combines multilingual encoders and LLMs. First, we stabilize training with a bounded regression architecture. Second, we improve the model's affective reasoning via intermediate pre-training on an external affective dataset EmoBank (Buechel and Hahn, 2017). The combined pipeline helps mitigate the "curse of multilinguality" and yields strong performance across English, Chinese, and Cyrillic-script languages.

2 Related Work

The dimABSA task bridges the gap between discrete linguistic labels and the continuous nature of human affect. By utilizing the Valence-Arousal space, we can model nuanced emotional states that traditional categorical systems miss. This approach is particularly vital for cross-lingual applications, where the same sentiment label may carry different cultural or linguistic intensities. The development of this problem began with SemEval-2014 Task 4 (Pontiki et al., 2014), which highlighted the ambiguity of aspect terms and annotator disagreement. These tasks originally focused on classification (Positive, Negative, Conflict, Neutral). The continuous nature of ABSA was later emphasized based on the circumplex model of affect (Russell, 1980).

[1]https://github.com/ssurface3/dimabsa

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2910–2919
July 3-4, 2026 ©2026 Association for Computational Linguistics

Recently, SIGHAN-2024 shared tasks (Xu et al., 2024) reached high scores on Chinese datasets; however, Chinese is not a low-resource language. The dimABSA 2025 shared task expands this to Japanese, Russian, Tatar, and Ukrainian. Our work aims to address the performance gap in continuous ABSA for these diverse language families.

3 System Overview

Our approach to the dimABSA task consists of three complementary methods: a prompting-based baseline, an encoder-based model, and a decoder-based model. Together, these methods allow us to compare lightweight inference-time prompting with supervised encoder and decoder architectures in a multilingual setting.

3.1 Prompting Baseline

The prompting-based baseline serves as a simple and transparent reference point with little to no task-specific training. It allows us to assess how far instruction-following behavior alone can go on the dimABSA task.

As shown in Figure 3, the model receives an input sentence and a target aspect (see Figure 3), from which it is prompted to predict continuous Valence and Arousal (VA) scores on the 1.0–9.0 scale.

In the zero-shot setting, the model receives only a task description, the input sentence, and the target aspect, without any labeled examples. This setting tests the model's internal alignment with the VA scale based solely on its instruction-following capabilities.

In the few-shot (K-shot) setting, we additionally provide k high-quality demonstrations, i.e., [Sentence, Aspect, Gold VA score] triples, selected from the training set. These demonstrations help calibrate the model's numerical output range and adapt it to the target domain (e.g., restaurant or laptop reviews) before predicting the VA scores for the target aspect.

Zero-Shot Prompt is shown in the Figure 5. It is a basic prompt for generation, containing all the needed information. We provided four diverse examples covering different sentiment polarities and intensities to calibrate the model's numerical range Figure 6.

3.2 Encoder-Based Baselines

Our primary encoder-based model is built on the mmBERT architecture (Marone et al., 2025). We choose this checkpoint because it combines a strong pre-training recipe with broad multilingual coverage, including support for low-resource languages.

Specifically, we use the `mmBERT-base` configuration, which consists of 12 Transformer layers, a hidden size of 768, and 12 attention heads. The encoder is fine-tuned for dimABSA with a task-specific regression head.

To adapt the pre-trained encoder to the dimensional ABSA task, we attach a linear regression head to the final hidden state of the special classification token (`[CLS]`). Let $h_{\texttt{[CLS]}} \in \mathbb{R}^{768}$ denote the encoder representation of the `[CLS]` token. The prediction is computed as

$$\hat{y} = W h_{\texttt{[CLS]}} + b,$$

where $\hat{y} \in \mathbb{R}^2$ and $\hat{y} = [\hat{V}, \hat{A}]$ corresponds to the predicted Valence and Arousal scores.
To ensure a comprehensive and meaningful comparison against industry-standard bidirectional encoders, we also benchmark our approach against several heavily vetted multilingual models. These include:

- **XLM-RoBERTa (Base and Large):** We include both configurations of XLM-R (Conneau et al., 2020) to evaluate the impact of model scale on dimensional ABSA and to provide a benchmark against the current industry workhorse for cross-lingual transfer.
- **mDeBERTa-v3-base:** We evaluate mDeBERTa-v3 (He et al., 2021), which utilizes disentangled attention and ELECTRA-style pre-training. This represents the current state-of-the-art for transformer encoders in multilingual NLU tasks.

For a fair comparison, all encoder baselines utilize the same regression head architecture and input template described above. This allows us to isolate the architectural advantages of the decoder-based LoRA approach from the supervised encoding paradigm.

The input is formatted as a single sequence containing the review text and the target aspect term. To preserve aspect-centricity, we use a standard BERT-style input template: `[CLS] Text [SEP] Aspect [SEP]` (and <s> for XLM-R).

3.3 Decoder-based Method

Beyond the encoder-based baseline, we also employ a decoder-only Large Language Model (LLM)

to leverage its cross-lingual reasoning capabilities and broad pre-trained knowledge of affective expressions. We use Qwen-2.5-1.5B as the backbone model (Qwen Team, 2024) due to its strong performance-to-size ratio, which enables efficient training on consumer-grade hardware.

To align the causal decoder with the regression objective, we use a descriptive prompt template that frames the task as an instruction-following problem. The input sequence is constructed as follows: "`Predict the valence and arousal scores (range 1-9) for the aspect {aspect}' in the following sentence: '{text}' Valence and Arousal:`'.

By embedding the target `{aspect}` and the context `{text}` into this instructional scaffolding, we provide the model with explicit task semantics and the expected numerical range. This configuration allows the self-attention mechanism to perform cross-interaction between the aspect term and the sentimental cues present in the sentence. We extract the hidden state $\mathbf{h}_{last}$ from the final colon token in the sequence, using it as the contextual representation for the regression head.

Linear Head Fine-Tuning In this configuration, the 1.5B-parameter backbone is kept frozen, and only the final linear regression head is trained. Specifically, we update the weights of a linear layer that maps the 1536-dimensional hidden representation to a 2-dimensional output.

Low-Rank Adaptation (LoRA) We employ LoRA (Hu et al., 2022) by injecting trainable low-rank matrices into the attention and MLP layers. This enables the model to adapt to multi-domain sentiment nuances while avoiding the memory overhead of full-parameter updates.

Full Fine-Tuning In the most computationally intensive configuration, we unfreeze all 1.5 billion parameters. This allows full adaptation to the dimABSA data distribution, but it also increases the risk of catastrophic forgetting, where the model's general-purpose multilingual knowledge may be overwritten by patterns specific to the training subsets.

4 Experimental Setup

4.1 Data

The official dimABSA-2025 dataset covers six languages; however, several subsets suffer from severe data scarcity, especially the Cyrillic-script languages (Russian, Tatar, and Ukrainian). Detailed split statistics are provided in Table 4 in the Appendix.

4.2 Hyperparameter Ablation

We conduct a hyperparameter search over mmBERT fine-tuning settings to identify an effective configuration. The search space is summarized in Table 7, and selected results are reported in Table 9.

To reduce the risk of catastrophic forgetting (i.e., degradation of the model's pre-trained multilingual capabilities), we perform checkpoint-wise evaluation throughout training. Specifically, we run inference after each epoch and select the checkpoint that provides the best fit on the evaluation split before performance begins to deteriorate, particularly on low-resource languages. This strategy improves robustness across both higher-resource domains and the more challenging Cyrillic-script and Tatar subsets.

For the decoder-based setup, we additionally study LoRA hyperparameters. The parameters used are shown in Table 10.

To ensure that the predictions remain within the valid annotation range (1.0–9.0), we use a bounded regression formulation instead of hard clipping. Let z_V and z_A denote the unconstrained outputs for Valence and Arousal. We transform them as follows:

$$\hat{V} = 1.00 + (9.00 - 1.00)\,\sigma(z_V),$$

$$\hat{A} = 1.00 + (9.00 - 1.00)\,\sigma(z_A),$$

where $\sigma(\cdot)$ is the sigmoid function. This preserves the ordering and relative differences of predictions while enforcing the target range. For both methods, we primarily use Mean Squared Error (MSE) as the training objective.

5 Results

Overall results are reported in Table 1. We observe a clear performance hierarchy across the evaluated paradigms: prompting-based baselines perform worst, the encoder-based model provides a strong supervised baseline, and decoder-based fine-tuning achieves the best overall results.

Among prompting baselines, few-shot prompting substantially improves over zero-shot prompting for Qwen2.5-1.5B (average $RMSE_{VA}$ drops from 2.31 to 1.86), confirming the importance of in-context calibration for the 1.0–9.0 VA scale. However, even the strongest prompting baselines

Method	Setup	English		Japanese		Russian	Tatar	Ukrainian	Chinese			Avg.
		Laptop	Rest.	Finance	Hotel	Rest.	Rest.	Rest.	Finance	Laptop	Rest.	
Baseline	Zero-Shot (Qwen2.5-1.5B)	2.14	1.81	2.83	1.93	2.20	2.59	2.37	2.63	2.39	2.18	2.31
	Few-Shot (Qwen2.5-1.5B)	1.40	1.39	2.14	1.74	1.75	2.43	1.88	2.12	1.91	1.82	1.86
	Few-Shot (Kimi-k2)	2.19	2.15	1.64	1.76	1.78	1.94	1.78	1.97	1.64	1.90	1.87
	QLoRA (Qwen3-14B)	2.81	2.64	1.90	2.29	2.15	2.64	2.21	1.47	1.77	2.01	2.19
Encoder	Full FT (mmBERT-base)	1.31	1.34	1.12	0.75	1.60	1.89	1.63	0.67	0.89	1.05	1.22
	Full FT (XLM-R-base)	1.49	1.38	1.17	0.90	1.78	2.22	1.83	0.67	1.01	1.14	1.36
	Full FT (XLM-R-large)	1.36	1.28	1.13	0.75	1.78	2.14	1.81	0.62	0.86	1.00	1.27
	Full FT (mDeBERTa-v3-base)	2.13	2.08	1.42	1.33	2.17	2.22	2.17	0.87	1.28	1.36	1.70
Decoder	Linear Head (Qwen2.5-1.5B)	**1.04**	**0.91**	0.96	0.65	**1.08**	1.63	1.29	0.75	0.81	0.91	1.00
	Full FT (Qwen2.5-1.5B)	**1.04**	0.94	1.27	0.66	1.34	1.70	1.50	1.02	0.88	0.85	1.12
	LoRA (Qwen2.5-1.5B)	1.06	0.99	1.31	0.65	1.30	1.73	1.51	1.09	0.88	**0.84**	1.14
Leaderboard	Top-1 unofficial results	1.24	1.10	**0.65**	**0.55**	1.21	**1.53**	**1.18**	**0.48**	**0.61**	0.92	0.95
	Official Best (Team Name)	LogSigma	LogSigma	TeleAI	TeleAI	PAI	PAI	PAI	HUS@NLP-VNU	TeleAI	ICT-NLP	

Table 1: Comprehensive comparison across all experimental methods, reported in $RMSE_{VA}$ (lower is better).single-run estimates selected on the development set. Best result per subset in **bold**.

(Qwen2.5-1.5B few-shot: 1.86; Kimi-k2 few-shot: 1.87) remain far behind supervised training. Notably, the larger Qwen3-14B QLoRA baseline still underperforms the smaller few-shot prompting baseline (2.19 vs. 1.86), highlighting that model scale alone is insufficient without task-specific adaptation. Table 8 summarizes the performance of the proposed prompting strategies.

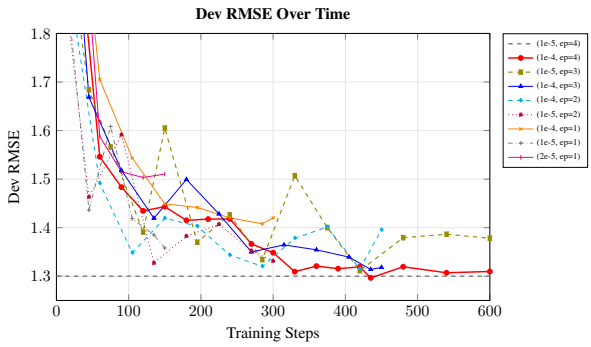
The encoder-based models provide a robust supervised baseline. As shown in Table 11, XLM-RoBERTa-Large achieves the strongest performance among bidirectional encoders with an average RMSE of 1.27. Interestingly, the mmBERT-base model (1.22) remains highly competitive despite its smaller parameter count, validating its specialized pre-training for multilingual tasks. In contrast, mDeBERTa-v3 exhibited higher error rates (1.70), which we attribute to the inherent instability of its disentangled attention mechanism when applied to continuous regression in high-variance multilingual settings. Crucially, all supervised encoders substantially outperform prompting-based baselines, confirming that instruction-following alone is insufficient for precise dimensional sentiment mapping.

Compared with the *Top-1 System* (included as a leaderboard reference), our best model (LoRA on Qwen2.5-1.5B) achieves a comparable overall average (1.14 vs. 0.95), remaining competitive while using only 1.18% of the backbone parameters. At the per-subset level, our LoRA model outperforms the reference on four subsets (both English domains, Russian, and Ukrainian), while the reference remains stronger on Japanese and Chinese subsets.

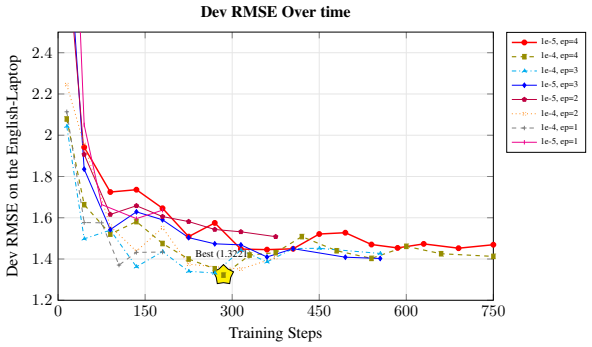
5.1 Distributional Mismatch and Augmentation Failure

As noted in our initial experiments, intermediate training on EmoBank and Facebook datasets improved development performance but led to a degradation on the hidden test set. To investigate this, we visualize the density distributions of the DimABSA training data compared to the EmoBank augmentation data The KDE plots (graphs 7) reveal a severe Neutrality Bias in the augmentation data. EmoBank's distribution is characterized by a sharp, narrow spike at the mean (5.0 on a 1–9 scale). In contrast, the DimABSA target domain is significantly more polarized and bimodal , particularly in the Valence dimension. For Arousal, the mismatch is even more pronounced: while EmoBank treats most sentences as neutral intensity, DimABSA labels are widely distributed with a significant shift toward higher intensity values (6.0–8.0). By injecting EmoBank data, the model's prior was inadvertently "pulled" toward the neutral center, causing it to lose the sensitivity required to predict the extreme aspect-level sentiment values found in reviews.This explains the drop in test performance and justifies our final decision to use a LoRA-tuned model trained exclusively on in-domain data. The mismatch likely stems from domain differences: EmoBank contains sentence-level annotations of general text (news, blogs), whereas DimABSA targets aspect-level sentiment in product and hotel reviews — a substantially more polarized and opinion-dense register. Future work could address this via domain-adaptive filtering or re-weighting of augmentation samples to better match the target distribution.

The distribution of Arousal in Figure 2 further explains the performance gap compared to Valence.



(a) Encoder-based approach trained on the original dimABSA dataset only.



(b) Encoder-based approach trained on the augmented dataset.

Figure 1: Convergence curves across hyperparameter configurations for the encoder-based approach under (left) original and (right) augmented training data.

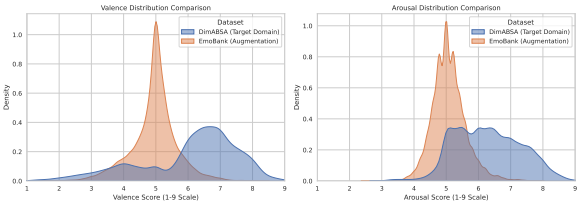


Figure 2: Kernel Density Estimation (KDE) plots of Valence (left) and Arousal (right) distributions. The EmoBank augmentation data shows a extreme "neutrality bias" (spike at 5.0), whereas the DimABSA target domain is significantly more polarized and bimodal.

While Valence shows clear bimodal peaks (showing distinct positive and negative clusters), Arousal in the DimABSA data forms a wide, flat plateau. This lack of clear density peaks suggests that Arousal is more linguistically diverse and lacks the high-frequency 'anchor' words that Valence possesses. The model struggles to find a correlation because the intensity of an aspect is often context-dependent rather than explicitly lexicalized

5.2 Encoder Baseline and Data Augmentation Ablation

As the number of epochs increases, most configurations show a clear trend toward convergence, suggesting that mmBERT needs a larger number of optimization steps to reliably align multilingual representations with the continuous Valence–Arousal (VA) space. Although the higher learning rate of 1×10^{-4} achieves lower RMSE on the development set in some runs, this improvement does not transfer to the held-out English-Laptop dev split used for final model selection. We therefore select 1×10^{-5} as the preferred learning rate due to its more stable generalization behavior.

Despite different early-stage trajectories, most configurations eventually converge to a narrow error band (approximately 1.30–1.40 RMSE) on the English-Laptop test set. In practice, this regime is typically reached within 400–600 optimization steps when using smaller learning rates.

To improve the encoder's sentiment modeling in English and Chinese, we augment training with additional data from EmoBank and Chinese-EmoBank. Fine-tuning on the augmented corpus yields mixed results:

1. Performance on the development set improves, but test-set performance slightly degrades, suggesting a distribution mismatch.

2. Training time increases substantially due to the larger corpus size.

3. The development set does not always appear to be representative of the corresponding test-set distribution.

As shown in the convergence curves for the augmented-data variant in Figure 1b, training exhibits substantially higher optimization noise than the baseline runs. The larger and more heterogeneous training corpus creates a more complex optimization landscape, which increases sensitivity to the learning rate. This behavior is especially pronounced for the 1×10^{-4} configurations (olive curve), which show frequent RMSE spikes. We interpret this as evidence that the model struggles to find a stable minimum that generalizes across the English and Chinese data distributions. In addition, as shown in Table 5, augmentation with large English/Chinese resources further reduces the relative share of low-resource languages in the training

mixture, which may weaken adaptation to Cyrillic-script and Tatar subsets.

5.3 Decoder-Based Method

Since the prompting baseline improves when provided with examples, we next evaluate supervised adaptation of a decoder-only backbone to better capture task-specific Valence–Arousal semantics.

We first consider a regression-head-only setting to probe the backbone capacity of Qwen2.5-1.5B. In this configuration, all backbone layers are frozen, and only the final linear layer is fine-tuned on the full training set for 5 epochs with a learning rate of 1×10^{-5}. Even this minimal adaptation yields strong results, indicating that the Qwen2.5-1.5B backbone already contains useful multilingual affective representations.

We then apply LoRA to specialize the model for VA regression while preserving its pre-trained multilingual knowledge. In our setup, only 1.18% of the parameters are updated. This parameter-efficient adaptation provides a favorable trade-off between specialization and stability, and helps reduce the risk of catastrophic forgetting compared to full fine-tuning.

Full fine-tuning also produces strong results, but it is less stable and tends to lose some of the model's generality. Figure 4 shows the convergence dynamics of Qwen2.5-1.5B under full parameter updates. The average $RMSE_{VA}$ decreases consistently during training and eventually stabilizes around 1.12. At the same time, convergence speed differs across domains: higher-resource subsets such as English-Laptop and Japanese-Hotel reach a plateau earlier, whereas the low-resource Tatar subset improves more slowly and remains a persistent bottleneck.

Overall, decoder-based fine-tuning improves performance across languages relative to the encoder baseline. The largest gains are observed on Chinese-Finance, and the decoder-based models also outperform mmBERT on Tatar, suggesting better adaptation to underrepresented languages when the decoder backbone is fine-tuned appropriately.

5.4 Statistical Significance of LoRA solution

To address potential variance resulting from random initialization, we conducted an empirical stability test for our best-performing method (LoRA-tuned Qwen2.5-1.5B). As shown in Table 12, we report the performance across three independent random seeds ($s \in \{7, 42, 1337\}$). The results demonstrate remarkable stability, with an average overall RMSE of 1.14 ± 0.01. The standard deviation across all subsets remains extremely low (ranging from 0.01 to 0.06), confirming that the performance gains are an inherent property of the model architecture and the parameter-efficient fine-tuning strategy, rather than a result of favorable random initialization. This statistical consistency provides strong evidence for the reliability of our proposed method in multilingual settings.

6 Conclusion

We evaluated encoder- and decoder-based approaches to multilingual dimABSA across ten domains. LoRA-tuned Qwen2.5-1.5B achieves the best overall performance (mean $RMSE_{VA} = 1.14$, $\sigma \leq 0.06$ across three seeds), outperforming full fine-tuning and head-only tuning. The central finding is that adapting only 1.18% of parameters via LoRA avoids catastrophic forgetting while delivering strong cross-lingual transfer, particularly on low-resource Cyrillic-script subsets. Parameter-efficient fine-tuning of compact LLMs is a practical and robust strategy for multilingual dimensional sentiment analysis.

7 Limitations

Despite the competitive performance achieved by our `LoRA-adapted` framework, we identify several key limitations that warrant further investigation.

Hardware and Scaling Constraints The primary bottleneck of this study was the computational environment. Our experiments were restricted to NVIDIA RTX 2080 Ti GPUs with 11GB of VRAM, which dictated the selection of the 1.5B parameter variant of `Qwen2.5`. While this model provides an excellent performance-to-size ratio, existing literature suggests that larger variants (e.g., 7B or 14B) possess significantly higher native reasoning capabilities for low-resource and distant languages

Linguistic Bottlenecks in Low-Resource Domains A persistent resource disparity remains evident in our results. Even with optimal fine-tuning, the model's performance on **Tatar (TAT)** and **Japanese Finance** remained significantly lower than on English and Chinese benchmarks. This suggests that the underlying pre-training distribution of the backbone model still heavily favors Indo-European and East Asian linguistic structures, leav-

ing minority languages with a lower "performance floor"

Logical Semantics and Sentiment Flipping Qualitative analysis revealed that our system still struggles with complex linguistic structures such as contrastive conjunctions and long-range negation. The "Sentiment Flip" error observed in Table 2 highlights that the model occasionally relies on local keywords (e.g., “heaven”) rather than global sentence logic. This indicates a need for more sophisticated prompt engineering or specific "logic-aware" pre-training objectives to ensure models can handle contrastive affective shifts.

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*, pages 578–585.

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. *Preprint*, arXiv:1911.02116.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

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.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Maria Pontiki, Dimitris Galani, 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.

Qwen Team. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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

Jun Xu, Jiayi Zhang, Jiaming Chen, Xiao Zhang, and Ruifeng Xu. 2024. HITSZ-HLT at SIGHAN-2024 dimABSA task: Ensemble multi-task learning for dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing*, pages 184–189.

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.

A Qualitative Error Analysis

To understand the model’s failure modes, we performed a qualitative analysis of high-error predictions. A primary issue identified is the **Sentiment Flip** in contrastive sentences. As we can see in the table 2

Category	Details
Sentence	*“Unfortunately, the waiter did sell the shrimp and grits with the additional crab as heaven in his mouth, but that it was not.”*
Aspect	*shrimp and grits*
Gold (V, A)	(1.88, 5.70)
Pred (V, A)	(7.88, 5.93)
Observation	**Sentiment Flip**: The model was distracted by the local positive keyword “heaven” and failed to process the global negation provided by “Unfortunately” and “but that it was not.”

Table 2: Case Study: Qualitative analysis of a Valence prediction error.

The performance ceiling of the encoder-based architectures was happening exclusively due to this aspect of the sentences. There were some ideas to include the opinion or sentiment(positive/negative) but our LLM inference attempts couldn’t produce useful results.

Finally, there was a problem of catastrophic forgetting in the full-fine-tuning process: Table 3

Language	Domain	PCC_V ↑	PCC_A ↑	RMSE_VA ↓
English (ENG)	Laptop	0.7954	0.4602	1.0376
English (ENG)	Restaurant	0.8461	0.5775	0.9365
Japanese (JPN)	Finance	0.4173	-0.0074	1.2744
Japanese (JPN)	Hotel	0.8614	0.5826	0.6558
Russian (RUS)	Restaurant	0.7000	0.2288	1.3422
Tatar (TAT)	Restaurant	0.1266	0.0628	1.6953
Ukrainian (UKR)	Restaurant	0.5760	0.1441	1.5004
Chinese (ZHO)	Finance	0.4355	0.1041	1.0171
Chinese (ZHO)	Laptop	0.6143	0.2558	0.8799
Chinese (ZHO)	Restaurant	0.7055	0.4618	0.8504
Average	**All Tasks**	**0.6078**	**0.2870**	**1.1190**

Table 3: Performance results for **Full Parameter Fine-tuning** (Qwen-2.5-1.5B) using the optimal checkpoint at Epoch 6.

While effective for English, this method exhibits a significant performance collapse in low-resource and specialized domains. Even scoring lower than mmBERT. This indicates that the optimal amount of hyperparameters has to be chosen in order to fight this problem.

Dataset Example (JSONL)

```
{
  "ID": "lap26_aspect_va_test_10",
  "Text": "For my purposes, the
      addition of a standard HDMI
      vice last years mini-HDMI is
      an improvement requiring no
      adapter",
  "Aspect_VA": [{
    "Aspect": "adapter",
    "VA": "6.50\#6.67"
  },
  {
    "Aspect": "standard HDMI",
    "VA": "7.00\#7.25"
  }]
}
```

Figure 3: A formatted example of a single entry

B Dataset data

Language	Domain	Train	Dev	Test
English	Laptop/Rest.	6,360	400	2,000
Chinese	Lap/Rest/Fin.	10,540	800	2,842
Russian	Restaurant	1,240	56	1,072
Tatar	Restaurant	1,240	56	1,072
Ukrainian	Restaurant	1,240	56	1,072
Japanese	Finance	1,024	200	800
Japanese	Hotel	1,600	200	800

Table 4: Statistics of dimABSA shared task datasets.

Task	Zero-Shot			Few-Shot (k=4)		
	$PCC_V \uparrow$	$PCC_A \uparrow$	$RMSE \downarrow$	$PCC_V \uparrow$	$PCC_A \uparrow$	$RMSE \downarrow$
Eng. Laptop	0.4673	-0.0860	2.1429	0.7368	0.3631	1.4008
Eng. Rest.	0.5586	-0.0300	1.8112	0.7811	0.3481	1.3944
Jpn. Finance	0.3477	0.0962	2.8303	0.5429	0.1158	2.1442
Jpn. Hotel	0.6214	0.0166	1.9282	0.8510	0.3565	1.7401
Rus. Rest.	0.5418	0.1114	2.2030	0.7754	0.2226	1.7451
Tat. Rest.	0.1044	0.0628	2.5857	0.1602	0.1204	2.4319
Ukr. Rest.	0.4275	0.0799	2.3704	0.6603	0.2421	1.8754
Zho. Finance	0.4579	0.0300	2.6308	0.6104	0.2156	2.1156
Zho. Laptop	0.3941	0.0572	2.3916	0.6885	0.3337	1.9058
Zho. Rest.	0.4942	-0.0697	2.1780	0.7053	0.4481	1.8196
Average	**0.4415**	**0.0268**	**2.3072**	**0.6512**	**0.2766**	**1.8573**

Table 8: Comparison of Zero-Shot and Few-Shot (k=4) performance across all tasks. Metrics include Pearson Correlation (PCC_V, PCC_A) and $RMSE_{VA}$.

Language Group	Total Samples	Global %
English (ENG)	8,760	29.15%
Chinese (ZHO)	14,182	47.20%
Cyrillic (RUS, TAT, UKR)*	7,104	23.65%
Total Dataset	**30,046**	**100.00%**

*Each Cyrillic language represents exactly 7.88% of the global data.

Table 5: Language distribution across the global dataset (Total $N = 30,046$).

Dataset	Language	Original Scale	Size
EmoBank	English	1.0 - 5.0	10,062
Facebook Posts	English	1.0 - 9.0	2,895
SIGHAN TrainSet1	Chinese	1.0 - 9.0	6,000
Total Added	**Mixed**	**Scaled to 1-9**	**18,957**

Table 6: External affective datasets used for intermediate pre-training.

C Ablated Hyperparameters

Parameter	Value
Model	`jhu-clsp/mmBERT-base`
Batch Size	64
Grad. Accumulation	1
Learning Rates	1e-5, 2e-5, 1e-4
Epochs	1, 2, 3, 4

Table 7: Hyperparameter search space for mmBERT fine-tuning.

D Inference results for models

Parameter	Value
Rank (r)	16
Alpha (α)	32
Dropout	0.05
Target Modules	All Linear Layers

Table 10: LoRA configuration parameters for the Qwen-2.5-1.5B backbone.

LR	Epochs	Dev RMSE	Test ENG RMSE
1e-5	1	1.1223	1.5034
1e-5	2	1.0414	1.4013
1e-5	3	0.9861	1.3140
1e-5	4	0.9621	1.3068
1e-4	4	0.9409	1.3113

Table 9: Hyperparameter tuning results and RMSE performance for mmBERT models.

E Full Fine Tune

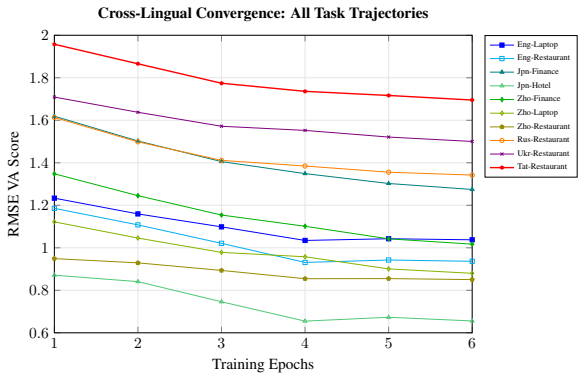


Figure 4: Trajectory analysis of all 10 task domains over 6 epochs. The visualization highlights that high-resource domains (English, Japanese-Hotel) achieve early stability, while low-resource languages (Tatar, Ukrainian) maintain a steeper learning gradient throughout the training process.

F Baseline Prompts

Zero-Shot Template

```
You are a sentiment analysis system.
Predict two scores (1.00-9.00):
Valence (V): positivity.
Arousal (A): intensity.
Respond ONLY in format V#A.
Text: {text} | Aspect: {aspect} | V#A:
```

Figure 5: Zero-shot prompt for Qwen2.5-1.5B.

Few-Shot Template

```
[Instructions as above]
Examples:
Text: The touchscreen works very well |
 Aspect: touchscreen | V#A: 7.80#7.60
Text: The hardware... is really cheap |
 Aspect: hardware... | V#A: 3.25#7.50
...
Text: {text} | Aspect: {aspect} | V#A:
```

Figure 6: Few-shot prompt template for Qwen2.5-1.5B.

https://box.skoltech.ru/index.php/s/raUdSKzhPGDmVuE?path=

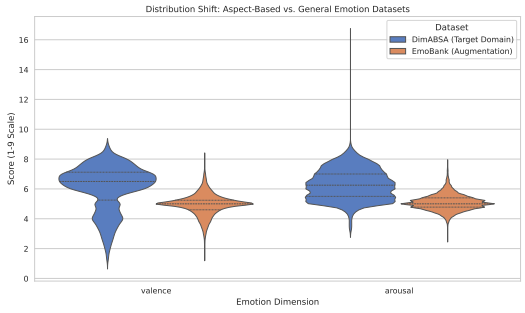


Figure 7: Violin plots of emotion dimensions. The width of the blue DimABSA violins at the extremes (low and high scores) illustrates why models trained on the neutral-heavy EmoBank (orange) fail to generalize to the target test set.

Method	Setup	English		Japanese		Russian	Tatar	Ukrainian	Chinese			Avg.
		Laptop	Rest.	Finance	Hotel	Rest.	Rest.	Rest.	Finance	Laptop	Rest.	
Encoder	Full FT (mmBERT-base)	1.31	1.34	1.12	0.75	1.60	1.89	1.63	0.67	0.89	1.05	1.22
	Full FT (XLM-R-base)	1.49	1.38	1.17	0.90	1.78	2.22	1.83	0.67	1.01	1.14	1.36
	Full FT (XLM-R-large)	1.36	1.28	1.13	0.75	1.78	2.14	1.81	0.62	0.86	1.00	1.27
	Full FT (mDeBERTa-v3-base)	2.13	2.08	1.42	1.33	2.17	2.22	2.17	0.87	1.28	1.36	1.70

Table 11: Comparison of RMSE_VA across different languages and domains for bidirectional encoder baselines. Lower is better.

Method	Setup	English		Japanese		Russian	Tatar	Ukrainian	Chinese			Avg.
		Laptop	Rest.	Finance	Hotel	Rest.	Rest.	Rest.	Finance	Laptop	Rest.	
LoRA	Seed 7	1.07	0.99	1.31	0.64	1.24	1.74	1.49	1.15	0.90	0.85	1.14
	Seed 42	1.06	1.00	1.28	0.65	1.30	1.72	1.52	1.03	0.86	0.83	1.13
	Seed 1337	1.06	0.97	1.35	0.66	1.36	1.74	1.53	1.09	0.88	0.84	1.15
	Mean ± SD	**1.06±.01**	**0.99±.01**	**1.31±.03**	**0.65±.01**	**1.30±.06**	**1.73±.01**	**1.51±.02**	**1.09±.06**	**0.88±.02**	**0.84±.01**	**1.14±.01**

Table 12: Statistical validation of the LoRA-tuned Qwen2.5-1.5B model across three independent random seeds. Results reported in $RMSE_{VA}$.