Scmhl5 at SemEval-2026 Task 3: Uncertainty-Aware Adversarial Learning for Embedding Enhancement in Dimensional Aspect-Based Sentiment Analysis

Haohuan Chen and **Han Liu**
College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
chenhaohuan@szdx.wecom.work, han.liu@szu.edu.cn

Abstract

This paper presents an uncertainty-aware adversarial learning framework developed for SemEval-2026 Task 3, a shared task focusing on Dimensional Aspect-Based Sentiment Analysis (ABSA). Our framework involves three key components: Uncertainty modeling, Heterogeneous Mixture-of-Experts (HMoE) architecture, and embedding-level adversarial training. Experimental results demonstrate that our framework effectively reduces the Root Mean Square Error (RMSE), thereby validating the synergistic advantages of uncertainty modeling and heterogeneous fusion strategies in fine-grained sentiment regression tasks.

1 Introduction

Dimensional ABSA is a fine-grained affective computing task that provides a more nuanced understanding of human emotions than traditional classification. It requires predicting continuous Valence and Arousal scores rather than discrete class labels. Valence represents the nature of an emotion, while Arousal represents its intensity. For instance, although "frustrated" and "depressed" both share a negative Valence, they differ significantly in their Arousal levels.

In this shared task, we address the challenge of undertaking the regression task across a diverse set of languages, ranging from widely spoken languages like English and Chinese to underrepresented African languages such as Swahili and Nigerian Pidgin.

The main challenges of this task lie in the representational difference between high-resource and low-resource languages, and the inherent noise in the human sentiment scoring process. Standard multilingual models may not sufficiently capture the unique features of African languages while performing well on languages like English or German. Furthermore, as human labels may be subjective and inconsistent, traditional methods that perform point estimation can lead to overfitting instead of learning the true underlying sentiment trends.

To address these challenges, we propose an uncertainty-aware adversarial framework that strategically combines heterogeneous encoders with probabilistic modeling. The main contributions of our work are as follows.

Language-Specific HMoE Framework We propose a HMoE architecture that is trained independently for each target language. By creating training instances for separate expert models, we enable the gating mechanism to specialize in the unique linguistic characteristics of each domain. This allows the system to adaptively fuse representations from a general expert and a language-specific expert, maximizing the representational power on a per-language basis.

Uncertainty-Aware Adversarial Learning We implement a robust regression strategy that synergizes Gaussian distribution modeling with adversarial training based on Fast Gradient Method (FGM). While the Gaussian Negative Log-Likelihood (NLL) loss enables the model to attenuate the impact of subjective annotation noise by quantifying aleatoric uncertainty, FGM enforces a smoother decision boundary in the embedding manifold. The synergy between uncertainty modeling and adversarial regularization effectively prevents overfitting to inconsistent human labels within each specific language track[1].

2 Background

2.1 Dataset and Preparation

We participated in both Track A and Track B of the shared task, and our experiments were conducted only on the official datasets provided by the organizers (Yu et al., 2026). These resources consist of

[1]Our code is publicly available at https://github.com/haohuanchen/UUA

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1748–1754
July 3-4, 2026

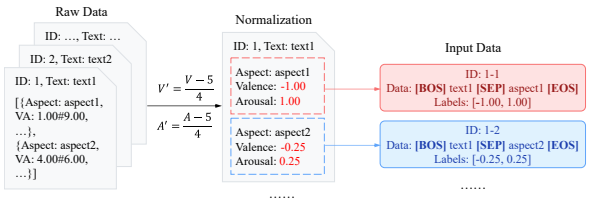

Figure 1: Data preprocessing pipeline and aspect-level sample construction

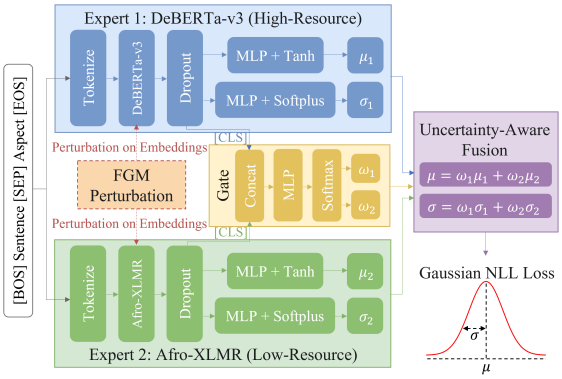

Figure 2: Overview of the uncertainty-aware adversarial learning framework.

the Track A dataset from Lee et al. (2026) and the Track B dataset from Becker et al. (2026). Both tracks employ a consistent core schema comprising ID, text, aspects, and continuous Valence and Arousal scores. Each ID corresponds to a unique text and multiple aspects, with each aspect mapped to a single pair of VA scores. Consequently, we treat the "Aspect" as the primary input unit to ensure that the model produces a unique VA prediction for each instance.

As illustrated in Figure 1, all VA scores are first normalized from their original scale to a range of [-1, 1] using a linear transformation to facilitate stable training. During the inference stage, the predicted values are rescaled to the original interval [1, 9] via a corresponding inverse mapping. Subsequently, we decompose text entries containing multiple aspects into distinct samples by concatenating the text with each individual aspect. This process ensures that each target aspect is treated as an independent input unit paired with its specific VA label. Our framework is primarily developed and iteratively refined on the Track B dataset, and was subsequently extended to generate the final predictions for Track A.

2.2 Related Work

Traditional approaches to ABSA have predominantly focused on categorical classification, assigning discrete polarity labels to specific targets (Pontiki et al., 2016). However, such coarse-grained schemas often fail to encapsulate the nuance and intensity of human affect. Consequently, the field has increasingly adopted the dimensional emotion model rooted in the circumplex theory of affect (Russell and A., 1980), which quantifies sentiment via continuous Valence and Arousal variables. Although earlier architectures based on LSTMs established foundational baselines for this regression task (Wang et al., 2016), they struggle to capture long-range semantic dependencies in complex multilingual contexts. Recent advancements leveraging Transformer-based architectures have demonstrated superior capability in feature extraction (Mendes and Martins, 2023), yet precise regression across typologically diverse languages remains challenging due to the inherent subjectivity of human annotation.

The paradigm of cross-lingual understanding has been revolutionized by massive multilingual pre-trained language models such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). These models project diverse languages into a shared semantic space, enabling zero-shot transfer. Nevertheless, they frequently encounter the "curse of multilinguality," where limited model capacity leads to performance degradation in low-resource languages when the vocabulary is dominated by high-resource data (Lauscher et al., 2020). To mitigate this issue, language-specific adaptive pre-training has emerged as a vital strategy. Alabi et al. (2022) introduced Afro-XLMR, which was designed by adapting the XLMR architecture specifically for African languages by performing masked language modeling in a curated corpus of 17 languages. Conversely, for high-resource scenarios, DeBERTa (He et al., 2020) significantly improves natural language understanding through its disentangled attention mechanism. Our framework strategically fuses these heterogeneous experts to maximize representational power across differing linguistic domains.

3 Framework Overview

Our proposed framework is an uncertainty-aware one built upon a HMoE architecture. As illustrated in Figure 2, the framework addresses the linguistic disparity of the shared task by combining high-

resource and low-resource encoders, while simultaneously modeling affective ambiguity through a probabilistic distribution.

3.1 Heterogeneous Expert Modeling

To optimize text representation across diverse languages, we treat the backbones as specialized experts. The input sequence is formatted as `[BOS] Sentence [SEP] Aspect [EOS]`, which is processed by two parallel streams: DeBERTa-v3-large serves as a high-resource expert to capture general semantic features, while Afro-XLMR-large acts as a language-specific expert to bridge the vocabulary gap in low-resource African languages. Each expert stream consists of a dedicated tokenizer, a transformer backbone and a dropout layer, ensuring that the model extracts complementary features from distinct linguistic semantic spaces.

3.2 Uncertainty Modeling

Consistent with the Gaussian distribution-based uncertainty modeling strategy, our framework does not treat sentiment as a deterministic point. Instead, each expert branch bifurcates into two MLP-based regression heads to estimate the parameters of a normal distribution $\mathcal{N}(\mu, \sigma^2)$.

The mean head μ involves a Tanh activation function to bound the predicted sentiment intensity, and the uncertainty head σ is provided with a Softplus activation to produce a positive standard deviation. This probabilistic formulation allows the model to explicitly represent the aleatoric uncertainty inherent in subjective human annotations.

3.3 Learnable Gated Fusion

The fusion of expert outputs is managed by a learnable gating network. This module takes the [CLS] tokens from both encoders as inputs, concatenates the inputs and passes the resulting vector through an MLP followed by a Softmax layer to generate expert weights w_1 and w_2. These weights dynamically determine the contributions of various experts based on the input context. The final parameters are calculated as: $\mu = w_1\mu_1 + w_2\mu_2$ and $\sigma = w_1\sigma_1 + w_2\sigma_2$. This uncertainty-aware fusion ensures that the predicted distribution is an adaptive mixture of the two experts' confidence levels. For final inference, we adopt this fused mean μ as the deterministic VA score.

3.4 Embedding-Level Adversarial Training

To enhance the robustness of HMoE, we incorporate embedding-level adversarial training using the FGM. As indicated by the dashed arrows in Figure 2, FGM calculates a norm-bounded perturbation based on the gradient of the Gaussian NLL Loss. This perturbation is injected directly into the word embedding layers of both encoders during training. By minimizing the loss on both clean and adversarial examples, the model is trained to maintain stable sentiment predictions even under minor linguistic perturbations, which is crucial for generalizing to noisy and low-resource datasets.

4 Experimental setup

4.1 Training Implementation

Our framework is implemented using PyTorch[2] and the Hugging Face Transformers library[3]. We utilize **microsoft/deberta-v3-large** and **Davlan/afro-xlmr-large** as our core backbones. We adopt a per-language training strategy, where a dedicated model is developed independently for each dataset. This approach enables the gating mechanism to specifically optimize for the fine-grained linguistic nuances inherent in each language.

4.2 Hyperparameter Settings

The key hyperparameters used across all tracks are summarized as follows.

Learning Rate: We employ the AdamW optimizer with a constant learning rate of 5×10^{-6}.

Batch Size: The batch size is set to 16. For the more complex HMoE architectures, the batch size is adjusted based on memory constraints[4].

Regularization: Dropout of 0.2 and early stopping with a patience of 5 epochs based on validation RMSE.

Adversarial Settings: FGM is applied to the word_embeddings layer with a step size $\epsilon = 1.0$.

Reproducibility: The random seed is fixed to 42.

[2] https://pytorch.org/

[3] https://github.com/huggingface/transformers

[4] All experiments were conducted on an NVIDIA GeForce RTX 4090 (24GB VRAM). Due to memory limitations, the batch size was set to 8 for HMoE and HMoE_{U}, and 4 for HMoE_{A} and HMoE_{UA}.

Model	deu↓	eng↓	pcm↓	swa↓	zho↓
DeB	1.5696	1.7498	1.5104	2.3440	0.6873
DeB_{U}	1.6782	1.8063	1.5903	2.4166	0.6765
DeB_{A}	1.5407	1.7411	1.6500	2.4932	0.6844
DeB_{UA}	1.5487	1.7181	1.4072	2.5741	0.6940
Afro	1.8272	2.0272	1.6717	2.0170	0.9318
Afro_{U}	1.7066	1.9861	1.5921	2.6237	0.7146
Afro_{A}	1.6242	2.0090	1.6053	1.9883	0.8730
Afro_{UA}	1.5860	1.9558	1.6650	**1.9391**	0.7753
HMoE	1.5566	1.7877	1.5083	2.0069	0.6636
HMoE_{U}	1.5166	1.7857	1.5051	2.2192	0.6757
HMoE_{A}	1.4688	1.6872	1.5024	2.3040	**0.6079**
HMoE_{UA}	**1.4375**	**1.6612**	**1.3697**	1.9426	0.6737

Table 1: Main results on Track B. Subscripts U and A denote uncertainty modeling and adversarial training, respectively, and UA represents the combination of both U and A components.

4.3 Evaluation Metrics

The primary evaluation metric for this shared task is the RMSE, calculated as:

$$\text{RMSE}_{VA} = \sqrt{\sum_{i=1}^{N} \frac{\Delta V_i^2 + \Delta A_i^2}{N}} \quad (1)$$

where N is the total number of instances, $\Delta V_i = V_p^{(i)} - V_g^{(i)}$ and $\Delta A_i = A_p^{(i)} - A_g^{(i)}$, $V_p^{(i)}$ and $A_p^{(i)}$ denote the predicted valence and arousal values for the i-th instance, and $V_g^{(i)}$ and $A_g^{(i)}$ denote the corresponding gold values.

5 Results

In this section, we present a comprehensive analysis of our experimental results. We first discuss the performance on Track B, which serves as our primary testbed for architectural optimization and ablation studies. We then evaluate the generalization of our framework on Track A.

5.1 Performance on Track B

The results for Track B across five diverse languages are summarized in Table 1. Our iterative development focused on the interplay between the HMoE backbone, Gaussian NLL (U), and adversarial training (A).

Efficacy of HMoE The results show that HMoE generally performs better than using a single backbone. In Track B, HMoE_{UA} achieves the lowest RMSE in most languages, including German (deu), English (eng), and Nigerian Pidgin (pcm). This phenomenon indicates that the gating mechanism can successfully combine the strengths of both encoders to improve overall accuracy.

However, HMoE is not always the best choice. In Swahili (swa), the single model Afro_{UA} performs slightly better than HMoE_{UA}. This suggests that for certain specific languages, the specialized encoder alone is already quite effective, and adding a second expert may not provide additional benefits or might even introduce slight interference.

Ablation of U and A Components The impacts of uncertainty modeling (U) and adversarial training (A) vary across languages and backbones.

Adversarial Training (A): FGM consistently provides a regularization benefit across most settings. In particular, HMoE_{A} achieves the overall best RMSE in Chinese (zho). This phenomenon suggests that adversarial perturbations effectively help the model learn more robust features in the embedding space for this language.

Uncertainty Modeling (U): While U is designed to handle potential label noise, the effect of U can be double-edged. In Chinese (zho), adding U to HMoE_{A} actually increases the RMSE. A possible reason is that the Chinese dataset possesses highly consistent and high-quality annotations. In such cases, the probabilistic softening of labels by the Gaussian loss may be unnecessary and could lead to slight underfitting. Conversely, in Nigerian Pidgin (pcm), where labels may be more subjective, the synergy of U and A remains highly effective.

Language-Specific Discussion Although Afro-XLMR is primarily designed for African languages, it contributes significantly to non-African tracks when fused. For instance, in German (deu) and Nigerian Pidgin (pcm), the HMoE variants consistently outperform the standalone DeB models. This validates the effectiveness of our choice of experts, the inclusion of Afro-XLMR provides representation diversity that captures linguistic features overlooked by the general semantic encoder.

In Nigerian Pidgin (pcm), while HMoE_{U} and HMoE_{A} lead to marginal improvements over the base model, their combination HMoE_{UA} yields a substantial performance leap. This suggests that U and A target different aspects of the problem, where U mitigates label noise and A enhances embedding stability. Their interaction is thus important for languages with subjective labels.

Model	eng-lap↓	eng-res↓	jpn-fin↓	jpn-hot↓	rus-res↓	tat-res↓	ukr-res↓	zho-fin↓	zho-lap↓	zho-res↓
DeB_{UA}	**1.3696**	1.3168	**0.9344**	**0.6811**	1.4609	2.3184	1.5928	0.5557	0.7494	**0.9838**
Afro_{UA}	1.4965	1.4353	0.9495	0.8454	1.5170	2.0689	1.5815	**0.5158**	0.7165	1.3707
HMoE_{UA}	1.4275	**1.2941**	0.9538	0.7324	**1.4172**	**2.0043**	**1.5196**	0.5458	**0.6821**	0.9849

Table 2: Main results on Track A.

Subtask	Source	Model	eng-lap↑	eng-res↑	jpn-hot↑	rus-res↑	tat-res↑	ukr-res↑	zho-lap↑	zho-res↑
Subtask 2	GPT-4.1	DeB_{UA}	0.5132	0.6127	0.3357	0.3960	0.3609	0.4206	0.3090	0.3955
		Afro_{UA}	0.5133	0.6036	0.3315	0.3951	0.3661	0.4192	0.3111	0.3858
		HMoE_{UA}	0.5136	0.6150	0.3363	0.3986	0.3649	0.4236	0.3096	0.3948
		Mean	**0.5134**	**0.6104**	**0.3345**	**0.3966**	**0.3640**	**0.4211**	**0.3099**	**0.3920**
	Gold	DeB_{UA}	0.9051	0.9101	0.9407	0.8937	0.8473	0.8841	0.9434	0.9303
		Afro_{UA}	0.9055	0.8983	0.9307	0.8888	0.8584	0.8798	0.9474	0.9098
		HMoE_{UA}	0.9044	0.9150	0.9405	0.8957	0.8544	0.8957	0.9478	0.9301
		Mean	**0.9050**	**0.9078**	**0.9373**	**0.8927**	**0.8534**	**0.8865**	**0.9462**	**0.9234**
Subtask 3	GPT-4.1	DeB_{UA}	0.2750	0.5119	0.2195	0.3138	0.2594	0.3338	0.1986	0.3309
		Afro_{UA}	0.2754	0.5040	0.2168	0.3127	0.2628	0.3326	0.1996	0.3226
		HMoE_{UA}	0.2752	0.5134	0.2199	0.3159	0.2629	0.3358	0.1990	0.3302
		Mean	**0.2752**	**0.5098**	**0.2187**	**0.3141**	**0.2617**	**0.3341**	**0.1991**	**0.3279**
	Gold	DeB_{UA}	0.9051	0.9101	0.9498	0.8937	0.8473	0.8841	0.9434	0.9303
		Afro_{UA}	0.9056	0.8983	0.9396	0.8888	0.8584	0.8798	0.9474	0.9098
		HMoE_{UA}	0.9044	0.9150	0.9496	0.8957	0.8544	0.8957	0.9478	0.9301
		Mean	**0.9050**	**0.9078**	**0.9463**	**0.8927**	**0.8534**	**0.8865**	**0.9462**	**0.9234**

Table 3: Results on Track A Subtasks 2 and 3. Bold values indicate the mean of each group.

5.2 Generalization on Track A

As UA models outperform others on Track B, we further evaluate their performance on Track A, with results shown in Table 2.

Strengths of HMoE HMoE_{UA} demonstrates the most balanced performance, achieving the lowest RMSE in 5 out of 10 subtasks. The model performs particularly well in Russian (rus), Tatar (tat), and Ukrainian (ukr). This performance suggests that for languages with complex morphology or limited resources, the gated fusion mechanism effectively combines the general semantic knowledge of DeBERTa with the multilingual strengths of Afro, leading to more stable predictions.

Expert Advantages in Specific Domains We also observe that the single-backbone models maintain strong advantages in certain areas. DeB_{UA} remains highly competitive in high-resource languages such as English and Japanese, achieving the best scores in eng-lap and both Japanese subtasks. On the other hand, Afro_{UA} shows unique potential in specific contexts, such as the Chinese financial domain (zho-fin). These variations indicate that while our HMoE framework is robust on average, the choice of the best expert is still influenced by the specific language and domain of the text.

5.3 Evaluation on Subtasks 2 and 3

Unlike Subtask 1, which focuses solely on VA regression, Subtask 2 and 3 require the extraction of additional sentiment elements. In Subtask 2, it is necessary to identify the "Aspect" and "Opinion" terms first, followed by the prediction of VA scores. Subtask 3 further adds the requirement for "Category" identification. Since our primary framework is optimized for VA regression, we implemented a two-stage pipeline to participate in these subtasks.

Pipeline with GPT-4.1 In the first stage, we design prompts to guide the GPT-4.1 model to extract the required entities. In the second stage, the models previously trained on Subtask 1 are used to predict the VA scores. The results for this pipeline are summarized in the "GPT-4.1" section of Table 3. The nearly identical scores across models suggest that the performance bottleneck lies in LLM extraction rather than VA regression precision.

Verification via Gold Extraction We conduct an additional experiment for verification by using

the gold labels of Aspects, Opinions, and Categories to perform VA prediction. As shown in the "Gold" rows of Table 3, the scores get better drastically, reaching nearly identical high-performance levels regardless of the underlying model. This phenomenon confirms that the accuracy of element extraction is far more critical than the precision of VA scores. Although the performance of GPT-4.1 under the utilized few-shot prompts was relatively limited, we provide these prompts in Appendix B for future reference and comparison.

6 Conclusion

This paper presents an uncertainty-aware adversarial learning framework developed by our team for the SemEval-2026 Task 3 Dimensional ABSA task. By integrating a HMoE architecture with Gaussian distribution-based uncertainty modeling and FGM-based adversarial training, our framework effectively addresses the challenges of linguistic disparity and label subjectivity. Experimental results demonstrate that the proposed framework consistently reduces the RMSE, validating the synergistic advantages of heterogeneous fusion and uncertainty modeling in fine-grained affective computing tasks.

Acknowledgement

This paper is supported in part by Natural Science Foundation of Shandong Province (Grant ZR2025QC1591) and Shenzhen Science and Technology Program (Grant ZDSYS20220527171400002).

References

Jesujoba Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th international conference on computational linguistics*, pages 4336–4349.

Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, Jan Philip Wahle, Terry Ruas, Idris Abdulmumin, Lung-Hao Lee, Wen-Ni Liu, Tzu-Mi Lin, Zhe-Yu Xu, Ying-Lung Lin, Jin Wang, Maryam Ibrahim Mukhtar, Bela Gipp, and Saif M. Mohammed. 2026. Dimstance: Multilingual datasets for dimensional stance analysis.

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.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

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. Mohammed. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis.

Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *European Conference on Information Retrieval*, pages 84–100. Springer.

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.

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

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.

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

	eng-lap	eng-res	jpn-fin	jpn-hot	rus-res	tat-res	ukr-res	zho-fin	zho-lap	zho-res
Leaderboard Score↓	N/A	1.3168	0.9292	0.6811	1.4609	2.0142	1.4732	N/A	0.7165	0.9838
Leaderboard Rank	N/A	20th	10th	10th	9th	12th	8th	N/A	11th	14th
Final Score↓	1.3696	1.2941	0.9344	0.6811	1.4172	2.0043	1.5196	0.5158	0.6821	0.9838
ScoreΔ	—	0.0227	0.0052	—	0.0437	0.0099	0.0464	—	0.0344	—

Table 4: Leaderboard results for Track A Subtask 1. Due to technical anomalies during the submission phase, our predictions for **eng-lap** and **zho-fin** were not successfully recorded on the leaderboard.

		eng-lap	eng-res	jpn-hot	rus-res	tat-res	ukr-res	zho-lap	zho-res
Subtask 2	Leaderboard Score↑	0.5136	0.6127	0.3357	0.3960	0.3649	0.4267	0.3111	0.3955
	Leaderboard Rank	14th	13th	13th	15th	12th	13th	12th	12th
	Final Score↑	0.5134	0.6104	0.3345	0.3966	0.3640	0.4211	0.3099	0.3920
	ScoreΔ	0.0002	0.0023	0.0012	0.0006	0.0009	0.0056	0.0012	0.0035
Subtask 3	Leaderboard Score↑	0.2752	0.5119	0.2195	0.3138	0.2629	0.3384	0.1996	0.3309
	Leaderboard Rank	12th	16th	9th	10th	9th	9th	11th	11th
	Final Score↑	0.2752	0.5098	0.2187	0.3141	0.2617	0.3341	0.1991	0.3279
	ScoreΔ	—	0.0021	0.0008	0.0003	0.0012	0.0043	0.0005	0.0030

Table 5: Leaderboard results for Track A Subtasks 2 and 3.

	deu	eng	pcm	swa	zho
Leaderboard Score↓	1.4375	1.6612	1.4072	1.9391	0.6765
Leaderboard Rank	4th	8th	5th	3rd	8th
Final Score↓	1.4375	1.6612	1.3697	1.9391	0.6079
ScoreΔ	—	—	0.0375	—	0.0686

Table 6: Leaderboard results for Track B.

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 Leaderboard Results

To keep the standard and repeatability of the code, we cleaned up the code after the task. Because of setting updates and parameter changes, some results changed slightly. All results shown in this paper are from the final cleaned code. In this section, we provide a comprehensive comparison of our leaderboard scores, rankings, and final results for Track A Subtask 1 in Table 4, Track A Subtasks 2 and 3 in Table 5, and Track B in Table 6, respectively.

B Prompt template

For Subtasks 2 and 3, we design a few-shot prompting strategy for GPT-4.1 to automatically extract aspects, opinions and categories. The prompt template is as follows.

```
You are a strict aspect-opinion extractor.
Task: Extract all aspect-opinion pairs from the
    sentence and assign an entity and attribute.

Instructions:
- Extract all aspects and their corresponding opinions
    from the sentence.
- Output format must strictly be: aspect, opinion,
    entity, attribute
- Multiple pairs separated with ' | '
- If an aspect has no opinion, output 'NULL' as the
    opinion
- If no aspect exists, output 'NULL, NULL, entity,
    attribute '.
- Aspect and opinion must be exact words from the
    sentence.
- Do NOT add explanations or extra text.

Entity must be one of: { entity_list }
Attribute must be one of: { attribute_list }
- Entity and Attribute must be chosen strictly from
    the lists above.
- Do NOT invent new entity or attribute names.

Examples:
Sentence: {sentence sample1}
Output: {output sample1}

Sentence: {sentence sample2}
Output: {output sample2}

Sentence: {sentence sample3}
Output: {output sample3}

Now extract from the following sentence:
Sentence: {sentence}
Output:
```