YNU-ABSA at SemEval-2026 Task 3: A Unified Pipeline for Continuous and Structured Dimensional ABSA

Qimao He,Xiaobing Zhou
School of Information Science and Engineering
Yunnan University
Kunming, China
12025115087@stu.ynu.edu.cn, zhouxb@ynu.edu.cn

Abstract

Dimensional Aspect-Based Sentiment Analysis (DimABSA) aims to jointly model continuous Valence–Arousal (VA) regression and structured sentiment extraction at the aspect level in multilingual settings, requiring both fine-grained emotion modeling and structural consistency. Prior approaches often separate regression and extraction or rely on stage-wise pipelines, which may limit numerical stability and structural alignment. To address this challenge, we propose a unified pipeline for all three subtasks of DimABSA Track A. Although Task 1 and Task 2/3 use different backbone architectures, they are integrated through consistent preprocessing, a shared dimensional sentiment perspective, and unified post-processing principles. For Task 1, we enhance aspect–context interaction via aspect-conditioned cross-attention and attention pooling, together with bounded output mapping and lightweight calibration for stable VA prediction. For Task 2/3, we formulate triplet and quadruplet prediction as constrained conditional generation with LoRA fine-tuning and structural validation. Experiments show consistent improvements across languages, including lower RMSE, higher correlation, and better cF1. Error analysis further shows that Arousal remains more difficult than Valence.

1 Introduction

Dimensional emotion modeling represents affective states in a continuous Valence–Arousal (VA) space (Lee et al., 2022), enabling fine-grained intensity distinctions beyond discrete polarity labels. DimABSA(Yu et al., 2026) Track A extends this paradigm to multilingual and multi-domain aspect-level sentiment analysis (Lee et al., 2026). The track comprises three subtasks (Lee et al., 2026): Task 1 predicts continuous VA values for a given aspect, while Task 2 and Task 3 extract structured units from raw text—including aspect, opinion, and (for Task 3) category—together with their associated VA values (Zhang et al., 2021).

This setting presents several challenges. Continuous regression is more sensitive to representation learning and calibration than discrete classification (Yu et al., 2016), and structured extraction is tightly coupled with numerical prediction, where errors in span detection, pairing, or value estimation directly affect overall evaluation. Furthermore, cross-lingual and cross-domain variation increases modeling difficulty and amplifies error propagation.

To address these issues, we propose a unified pipeline covering all three subtasks. For Task 1, we use aspect-conditioned cross-attention and attention pooling for stable VA regression. For Task 2/3, we adopt LoRA-based constrained generation with JSON validation, parsing repair, and VA calibration. The unification is not based on a single shared backbone, but on a coherent workflow for dimensional sentiment prediction, structured output normalization, and robust post-processing across Track A.

2 Background

Early sentiment analysis mainly focused on sentence- or document-level polarity classification, which lacks aspect awareness and fine-grained distinctions in multi-aspect settings. Aspect-Based Sentiment Analysis (ABSA) addresses this limitation by modeling sentiment at the aspect level (Lee et al., 2026). Existing approaches include sequence labeling and span extraction, multi-stage pairing pipelines (Peng et al., 2020), and unified generative frameworks (Zhang et al., 2021). However, extending ABSA from discrete polarity labels to continuous Valence–Arousal (VA) prediction substantially increases complexity, as models must jointly handle structured extraction and stable regression (Mohammad et al., 2018).

For Task 1, typical baselines derive sentence

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1100–1105

representations from pretrained encoders and feed them into a regression head, which may dilute aspect-specific signals in multi-aspect contexts (Lee et al., 2026). For Task 2/3, extractive pipelines identify spans and perform pairing, with Task 3 additionally predicting categories; such multi-stage designs are prone to error accumulation (Zhang et al., 2021), and multilingual settings further complicate span detection due to syntactic and morphological variation (Muhammad et al., 2025). Unlike traditional ABSA settings (Buechel and Hahn, 2017), Track A requires continuous emotion prediction together with structured extraction, motivating models that improve both numerical stability and structural consistency.

3 System Description

Task 1 The baseline for Task 1 encodes the input as a sentence or sentence–aspect pair(Mohammad and Bravo-Marquez, 2017), applies mean pooling over token representations, and feeds the resulting embedding into an MLP regression head to predict VA values. While simple, this design fails to model explicit aspect–context interaction.

We introduce three improvements. At the data level, we locate the first occurrence of the aspect span during preprocessing and explicitly indicate when alignment is unavailable. In addition, we apply lightweight augmentation to extremely short sentences by appending semantically compatible context templates that preserve the original sentiment orientation while providing slightly richer lexical context. The goal is not to change the polarity or intensity label itself, but to reduce the representation sparsity of very short inputs. This is particularly useful for dimensional sentiment regression, where subtle valence–arousal cues are often expressed through a small number of modifiers, intensifiers, or local contextual patterns. By exposing the model to short inputs with minimally enriched but semantically consistent context, we improve its sensitivity to fine-grained intensity cues without altering the underlying meaning.

On top of backbone token representations, we construct masks for the aspect span and full sentence. A masked mean over the aspect tokens yields an aspect representation, which serves as a query for multi-head cross-attention over sentence tokens(Feng et al., 2022). We further apply learnable attention pooling to obtain a focused global emotional representation. The cross-attention output and pooled sentence representation are concatenated and passed to an MLP regression head to produce VA logits.

To ensure numerical validity, we constrain predictions as:

$$\hat{\mathbf{y}} = 1 + 8 \cdot \sigma(\mathbf{z}), \qquad \hat{\mathbf{y}} = (\hat{v}, \hat{a}), \tag{1}$$

guaranteeing $\hat{v}, \hat{a} \in [1, 9]$.

We optimize with Huber loss augmented by correlation-aware objectives. We choose Huber loss because it is robust to occasional large deviations while preserving stable optimization for small errors, which is suitable for noisy VA regression labels and cross-domain variation.

$$\mathcal{L} = \mathcal{L}_{Huber} - \lambda \cdot PCC(\hat{v}, v, \hat{a}, a) + \mu \cdot \mathcal{L}_{rank}(\hat{a}, a), \tag{2}$$

where λ and μ balance regression accuracy, correlation consistency, and ranking preservation for Arousal.

Task 2 and Task 3 We unify Task 2 and Task 3 as a constrained conditional generation problem(Feng et al., 2022). Given input text x, the model generates an output sequence y composed of multiple structured objects, each corresponding to a triplet or quadruplet formatted as a JSON array(Peng et al., 2020).

We adopt parameter-efficient fine-tuning via LoRA on a pretrained large language model(Citron et al., 2020). For a concatenated prompt–answer sequence, the language modeling loss is computed only over assistant output tokens, focusing optimization on structured answer generation rather than instruction reproduction. Let $y = (y_1, \ldots, y_T)$ denote the target output sequence. The training objective maximizes:

$$\max_{\theta} \sum_{t=1}^{T} \log P_{\theta}(y_t \mid y_{<t}, x)\,, \tag{3}$$

where prompt tokens are masked to avoid gradient interference.

To mitigate structurally invalid outputs (e.g., malformed JSON or missing fields), we introduce inference-time parsing and repair. Generated text is first strictly parsed as JSON; if parsing fails, a tolerant fallback procedure recovers usable structures. Each object is validated and repaired: missing fields are set to NULL, and textual fields are constrained to exact substrings of the source text or NULL, reducing hallucination-induced precision loss.

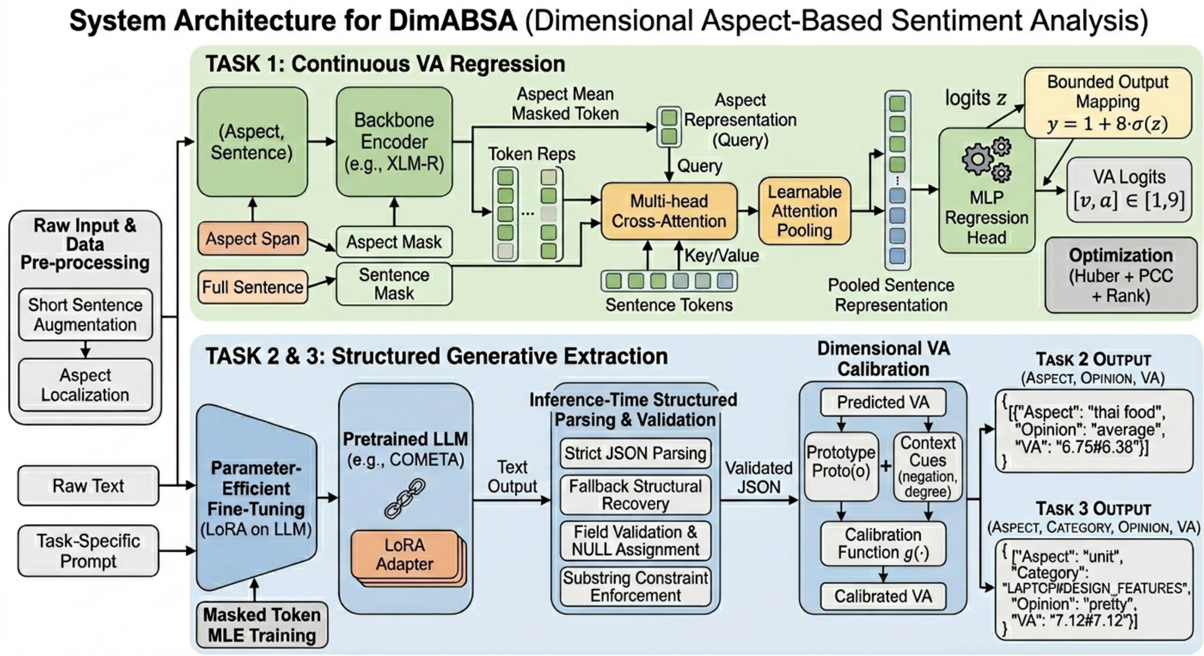

Figure 1: Overall system architecture for DimABSA Track A. Task 1 adopts an aspect-conditioned regression framework with cross-attention and attention pooling. Task 2/3 employ instruction-based generative extraction with LoRA fine-tuning, complemented by structured parsing and VA calibration.

To stabilize continuous VA predictions, we further apply lightweight calibration based solely on training statistics. For each opinion term o, we aggregate its observed VA values in training and define a prototype:

$$\mathrm{Proto}(o) = \mathrm{median}(V(o)), \tag{4}$$

where $V(o)$ denotes the associated VA set. The prediction $\widehat{\mathrm{VA}}$ is calibrated as:

$$\widehat{\mathrm{VA}}' = g\Big(\widehat{\mathrm{VA}},\, \mathrm{Proto}(o),\, x\Big), \tag{5}$$

where $g(\cdot)$ applies numerical constraints and contextual intensity cues (e.g., negation, degree adverbs). This strategy reduces numerical drift and distributional collapse without external data, empirically improving cTP and cF1 while preserving structural validity.

4 Experimental Setup

Dataset For Task 1, we train separate regression models for each language–domain pair. The official training data is split into 90% training and 10% development for checkpoint selection and model comparison. We use a learning rate of 3×10^{-5}, batch size 32, maximum length 512, and domain-specific epoch settings. To reduce randomness, we train with three seeds (42/43/44) and average predictions for ensemble inference. Implementation is based on PyTorch and HuggingFace Transformers.

For Task 2/3, we fine-tune `Qwen2.5-3B-Instruct` using LoRA under an instruction-based generation formulation. We use a learning rate of 1×10^{-4}, train for 1.5 epochs with gradient accumulation (4 steps), adopt a cosine scheduler, and enable FP16 training. During inference, strict JSON parsing, repair, and deduplication are applied to improve structural validity and effective prediction rates.

Evaluation Metric Task 1 is evaluated by $\mathrm{RMSE}_{\mathrm{VA}}$ (lower is better), with $\mathrm{PCC}_{\mathrm{V}}$ and $\mathrm{PCC}_{\mathrm{A}}$ reported as auxiliary correlation metrics (higher is better). Task 2/3 is evaluated by cF1, alongside cPrecision, cRecall, and continuous true positives (cTP).

Evaluation strictly requires exact structural matching, while numerical deviations in VA values incur continuous penalties. Therefore, both structural correctness and stable value prediction are essential for strong performance.

5 Results

Task 1 Results As shown in Figure 2, Japanese in the hotel domain achieves the lowest RMSE (0.7554), suggesting that aspect-conditioned mod-

Setting	RMSE_VA ↓	PCC_V ↑	PCC_A ↑
Mean Pooling	0.8054	0.8564	0.6033
Attention Pooling	0.7748	0.8710	0.6599
Attention Pooling + Cross-Attn	0.7554	0.9010	0.6807

Table 1: Ablation study of pooling and interaction mechanisms on Task 1. Lower RMSE_{VA} is better, while higher PCC_{V} and PCC_{A} indicate stronger correlation.

Language	Domain	Backbone	RMSE_VA ↓	PCC_V ↑	PCC_A ↑
ENG	laptop	bert-base-multilingual-cased	1.4603	0.7954	0.4917
ENG	laptop	xlm-roberta-base	1.4198	0.8296	0.5155
JPN	hotel	xlm-roberta-base	1.0064	0.7411	0.4413
JPN	hotel	bert-base-japanese-wwm	0.7554	0.9010	0.6807

Table 2: Comparison of different pretrained backbones across language–domain settings. Lower RMSE_{VA} indicates better regression accuracy, while higher PCC_{V} and PCC_{A} reflect stronger correlation.

eling benefits from well-aligned language-specific representations.

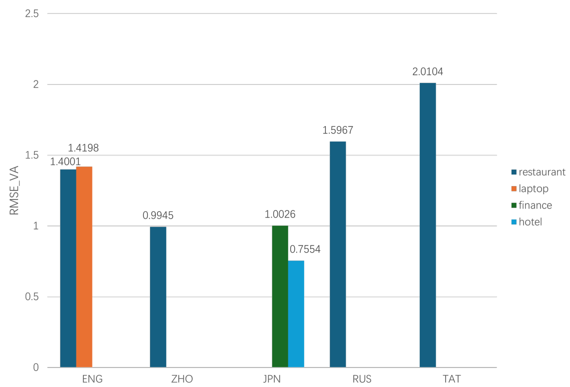

Figure 2: RMSE_{VA} comparison across languages and domains. Different colors denote different domain settings, as indicated by the legend. Lower values indicate better regression performance.

Task 1 Ablation We compare three architectures sequentially: Mean Pooling, Attention Pooling, and Attention Pooling with Aspect-Conditioned Cross-Attention. Results in Table 1 show that replacing Mean Pooling with Attention Pooling reduces RMSE_VA, indicating that learnable weighted aggregation better captures emotion-related tokens and intensity cues.

Adding Aspect-Conditioned Cross-Attention further improves performance, decreasing RMSE_VA and increasing both PCC_V and PCC_A. This demonstrates that explicit aspect–context interaction enhances aspect-specific emotional alignment. The improvements follow a consistent layer-wise trend, with each additional module yielding monotonic gains across metrics.

To assess backbone effects, we compare pretrained models on English (*laptop*) and Japanese (*hotel*), as shown in Table 2. On English–laptop, XLM-RoBERTa-base outperforms `bert-base-multilingual-cased`, likely due to stronger multilingual pretraining and more stable semantic representations. On Japanese–hotel, performance differences are more substantial: the language-specific pretrained model achieves markedly lower regression error and higher correlation scores, suggesting that in-language pretraining better captures lexical and structural patterns for aspect-conditioned emotion modeling.

Task 2 and Task 3 Results On the English extraction tasks, our system slightly outperforms the baseline in both Triplet and Quadruplet extraction. The Quadruplet task is inherently more challenging due to the inclusion of the *category* field and stricter structural matching requirements. Nevertheless, by incorporating instruction-level constraints and an inference-time repair mechanism, we are able to maintain a relatively high proportion of valid predictions.

Task 2 and Task 3 Ablation To evaluate the generative framework for triplet extraction, we compare an extractive baseline with our LoRA-based generative approach on the ENG–restaurant subset (Table 4). Both precision and recall improve simultaneously, indicating that the gains arise from increased true positives and reduced false positives rather than one-sided optimization.

These results demonstrate the structural advantage of generative modeling. Unlike the extractive baseline, which relies on local label predictions,

Task 2 (DIMASTE, ENG)							
Domain	**cF1 ↑**	**cPrec ↑**	**cRec ↑**	**cTP**	**TP**	**FP**	**FN**
laptop	0.4952	0.5372	0.4594	906.8002	989	699	985
restaurant	0.5240	0.5531	0.4978	1059.7220	1163	753	966
average	0.5096	0.5451	0.4786	983.2611	1076.0	726.0	975.5
Task 3 (DIMASQP, ENG)							
Domain	**cF1 ↑**	**cPrec ↑**	**cRec ↑**	**cTP**	**TP**	**FP**	**FN**
restaurant	0.5183	0.5489	0.4909	1045.0646	1141	763	988
average	0.5183	0.5489	0.4909	1045.0646	1141.0	763.0	988.0

Table 3: Results on English (ENG) for Task 2 (DIMASTE) and Task 3 (DIMASQP). We report cF1, cPrecision (cPrec), cRecall (cRec), continuous true positives (cTP), and the discrete TP/FP/FN counts.

Task	**Method**	**cF1 ↑**	**cPrec ↑**	**cRec ↑**	**cTP ↑**	**TP**	**FP**	**FN**
Task 2	Extractive Baseline	0.5877	0.5621	0.6158	251.2425	287	160	121
Task 2	Ours (Generative + LoRA)	0.7253	0.7235	0.7271	296.6406	323	87	85

Table 4: Comparison between the extractive baseline and our generative approach with LoRA on Task 2. Higher cF1, cPrecision (cPrec), cRecall (cRec), and cTP indicate better performance. Data are from dev experiments.

the generative model produces *Aspect*, *Opinion*, and VA values within a unified generation space, enabling global coordination under full contextual awareness. LoRA fine-tuning further adapts the model to domain data while preserving the pre-trained semantic capacity, leading to improved structural matching quality.

Error Analysis In Task 1, PCC_A is consistently lower than PCC_V across languages and backbones, indicating an inherent modeling difficulty. Valence aligns with explicit polarity cues, whereas Arousal relies on subtle and dispersed contextual signals (e.g., intensity modifiers, negation, punctuation) and exhibits higher subjectivity and annotation variance. Consequently, even with aspect-conditioned cross-attention, gains on Arousal remain limited, suggesting structurally different predictability across emotion dimensions.

For Task 2/3, we observe cases where Aspect and Opinion are correctly identified but incorrectly paired. Unlike extractive models with explicit pairing modules, generative models depend on implicit semantic associations, which can yield locally plausible yet globally inconsistent combinations in complex sentences. Such mismatches directly reduce cPrecision and cF1, as structural evaluation requires joint alignment of aspect, opinion, and VA values. Although the generative framework unifies structure, it remains vulnerable to implicit pairing instability under complex syntactic conditions.

6 Conclusion

We address all three sub-tasks of DimABSA Track A within a unified pipeline. For Task 1, we adopt a regression model with aspect-conditioned cross-attention and learnable attention pooling. For Task 2/3, we employ instruction-based generative extraction with LoRA fine-tuning, enhanced by parsing repair, deduplication, and VA calibration for structural and numerical stability.

Results show stable performance across multilingual VA regression and English structural extraction. Ablation studies indicate that interaction modeling and learnable aggregation drive improvements in Task 1, while inference-time stabilization is critical for maintaining valid outputs and improving cF1 in generative extraction.

Limitations Our approach has several limitations. In low-resource languages, regression errors remain relatively high, suggesting that single-language training does not fully address data scarcity and distribution shift. In generative extraction, performance still partially depends on inference-time repair and calibration, which may require refinement under stronger domain shifts.

Future work will explore more effective cross-lingual transfer strategies and investigate end-to-end joint learning frameworks that optimize structural extraction and VA prediction under a unified objective, reducing reliance on post-processing and improving robustness.

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.

Francesca M. M. Citron, Mollie Lee, and Nora Michaelis. 2020. Affective and psycholinguistic norms for german conceptual metaphors (cometa). *Behavior Research Methods*, 52(3):1056–1072.

Ao Feng, Xuelei Zhang, and Xinyu Song. 2022. Unrestricted attention may not be all you need–masked attention mechanism focuses better on relevant parts in aspect-based sentiment analysis. *IEEE Access*, 10:8518–8528.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(4):65:1–65:18.

Lung-Hao Lee, Liang-Chih Yu, Natalia Loukachevitch, Ilyas 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. *arXiv preprint arXiv:2601.23022*.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Lima Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikir Aji, Fermin Dario Mario Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 others. 2025. Semeval-2025 task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2558–2569, Vienna, Austria. Association for Computational Linguistics.

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. AAAI Press.

Liang-Chih Yu, Jonas Becker, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Lung-Hao Lee, Ying-Lung Lin, Jin Wang, Jan Philip Wahle, Terry Ruas, Natalia Loukachevitch, Alexander Panchenko, Ilseyar Alimova, Lilian Wanzare, Nelson Odhiambo, Bela Gipp, Kai-Wei Chang, and Saif M. Mohammad. 2026. Semeval-2026 task 3: Dimensional aspect-based sentiment analysis (dimabsa). *arXiv preprint arXiv:2604.07066*.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K Robert Lai, and Xuejie Zhang. 2016. Building chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 540–545.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.