TeamLasse at SemEval-2026 Task 3: A Hybrid Generative-Discriminative Framework for Dimensional Aspect-Based Sentiment Analysis

Lasse Strothe
Technical University of Munich
lasse.strothe@tum.de

Abstract

In this paper, we present our system for SemEval-2026 Task 3 Track A: Dimensional Aspect-Based Sentiment Analysis (DimABSA). The core objective is to extract structural sentiment elements—such as aspects, opinions, and categories—from text and predict their corresponding continuous Valence-Arousal (VA) scores. The primary challenge lies in simultaneously handling structural extraction and continuous numerical regression across highly imbalanced datasets encompassing multiple languages and domains. To address this complexity, we propose a decoupled, two-stage hybrid generative-discriminative framework. A generative Large Language Model first extracts structured sentiment tuples, while an encoder-based language model performs the continuous VA regression. To foster cross-lingual and cross-domain generalization, we train our models using a targeted data balancing mechanism. Our results demonstrate that this approach is highly effective for structural extraction in data-scarce scenarios, notably securing 2nd place in the underrepresented Japanese Hotel domain. However, Subtask 1 results highlight a structural limitation of our pipeline: because this subtask provides only the aspect term, our regression model—optimized for full aspect-opinion tuples—experiences a performance drop when the explicit opinion term is missing during inference.

1 Introduction

Dimensional Aspect-Based Sentiment Analysis (DimABSA) represents a significant evolution in sentiment analysis, moving beyond traditional categorical polarity toward a continuous Valence-Arousal (VA) space (Yu et al., 2026). In SemEval-2026 Task 3, participants are challenged to apply this concept across six languages and multiple domains (Lee et al., 2026) through three distinct subtasks: predicting continuous VA scores for given aspects (DimASR), extracting aspect-opinion-VA triplets (DimASTE), and predicting full aspect-category-opinion-VA quadruplets (DimASQP).

The Challenge: The core difficulty lies in the task's heterogeneity. Systems must simultaneously perform precise structural text extraction and continuous numerical regression. This complexity is compounded by severe data imbalances across domains and a drop-off in available data for low-resource languages, making cross-lingual generalization highly difficult.

Our Unified Approach: Instead of developing isolated, task-specific models, we propose a single, unified two-stage hybrid framework:

1. **Stage 1: Generative Extraction:** We utilize a Large Language Model (LLM) strictly for extracting structural elements (Aspects, Categories, Opinions).

2. **Stage 2: Discriminative Regression:** We decouple the numerical VA prediction into a dedicated encoder-based cross-encoder.

This deliberate decoupling delegates continuous scoring to a dedicated regressor, overcoming the inherent limitations of autoregressive LLMs in processing continuous mathematical variables. Our pipeline dynamically adapts to different subtasks—bypassing Stage 1 when aspects are pre-provided (Subtask 1)—though our analysis shows this introduces structural inference challenges when explicit opinion anchors are missing.

Finally, to address the skewed dataset distributions, we employ a dynamic data balancing mechanism. By strategically upsampling minority domains and low-resource languages during training, our framework fosters robust generalization across the entire multidimensional spectrum of Task 3.

2 Background and Related Work

Dimensional ABSA & Prior Approaches: Traditional ABSA focused on discrete polarities (Pon-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2155–2162

tiki et al., 2014), whereas Dimensional ABSA (DimABSA) models emotion in a continuous Valence-Arousal (VA) space (Russell, 1980; Lee et al., 2024; Yu et al., 2026). Early complex sentiment extraction relied on sequence labeling (Peng et al., 2020), which later shifted toward generative sequence-to-sequence models to better handle implicit targets (Zhang et al., 2021). More recent hybrid and retrieval-augmented approaches further demonstrate the effectiveness of combining structured knowledge, retrieval, and large language models for complex NLP tasks (Kolli et al., 2025; Üyük et al., 2024). However, while autoregressive LLMs excel at structural extraction, they are fundamentally ill-suited for continuous VA regression because they process numbers as discrete tokens rather than optimizing for continuous distance metrics like Mean Squared Error.

Situating Our Work: To address this limitation, we synthesize generative structure extraction (`Qwen2.5` (Yang et al., 2025)) with discriminative continuous regression (`XLM-RoBERTa` (Conneau et al., 2020)). Instead of training isolated models, our primary contribution is a **unified cross-lingual generalization strategy**. We employ targeted data balancing to mitigate severe domain and language disparities. Furthermore, our error analysis exposes the structural vulnerabilities of decoupled pipelines when dealing with missing inference contexts.

3 System Overview

To address the diverse requirements of Dimensional Aspect-Based Sentiment Analysis (DimABSA) across multiple languages and domains, we propose a unified, two-stage hybrid generative-discriminative framework. Instead of training separate models for each subtask, our pipeline leverages a Large Language Model (LLM) for the structural extraction of sentiment elements (Stage 1) and an encoder-based language model for the continuous regression of Valence-Arousal (VA) scores (Stage 2).

3.1 Stage 1: Generative Extraction (Qwen2.5-32B)

The first stage identifies categorical elements (Aspects, Categories, Opinions) using the instruction-tuned `Qwen2.5-32B` (Yang et al., 2025) via direct generative extraction. We employ strict prompt constraints to output a structured JSON list, requiring exact substring matches from the source text to prevent hallucinated synonyms. Depending on the subtask, the model extracts either Aspect-Opinion pairs (Subtask 2) or Aspect-Category-Opinion triplets (Subtask 3).

3.2 Stage 2: Discriminative VA Regression (XLM-RoBERTa)

While LLMs excel at text generation and structural extraction, predicting precise, continuous floating-point values is inherently challenging for purely autoregressive models, as they treat numbers as discrete text tokens rather than continuous mathematical variables. Therefore, we decouple the numerical prediction and formulate it as a sequence regression task using XLM-RoBERTa-large (Conneau et al., 2020).

For every (Aspect, Opinion) pair identified in Stage 1, we construct a cross-encoder input sequence. The original sentence serves as the primary context, while the concatenated string of the extracted aspect and opinion provides the target-specific auxiliary information. The input is structured as follows:

```
<s> Original Text </s></s> Aspect
Opinion </s>
```

The final hidden state representation of the `<s>` token, denoted as $h_{<s>} \in \mathbb{R}^d$, is passed through a linear regression head with weights $W \in \mathbb{R}^{d \times 2}$ and bias $b \in \mathbb{R}^2$. The model predicts the continuous Valence (V) and Arousal (A) scores as $\hat{y} = h_{<s>}W + b$. During training, we optimize the model using the Mean Squared Error (MSE) loss: $\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}((V_i - \hat{V}_i)^2 + (A_i - \hat{A}_i)^2)$, where N is the batch size. During inference, to ensure compliance with the task definition, the final predictions are clamped to strictly fall within the mandatory $[1.00, 9.00]$ range (Lee et al., 2026).

3.3 Subtask Adaptation and Post-Processing

A major advantage of our decoupled architecture is its zero-shot adaptability. For **Subtasks 2 and 3**, we dynamically map the generated categories using fuzzy string matching (Levenshtein distance) to ensure strict compliance with the predefined taxonomy (Yu et al., 2026). This corrects minor formatting inconsistencies. To ensure compliance with extraction metrics, a *Safe Capitalization Fix* performs a case-insensitive search of the extracted terms against the original input and reverts them to their exact original casing, preventing penalties

from LLM-induced normalization. For **Subtask 1 (DimASR)**, where aspects are already provided, we bypass Stage 1 entirely and feed the aspects directly into our XLM-R regressor, which predicts VA scores.

4 Experimental Setup

4.1 Datasets and Data Balancing

The framework is evaluated on the official SemEval 2026 DimABSA dataset (Lee et al., 2026), which spans six languages (English, Chinese, Japanese, Russian, Tatar, and Ukrainian). While the full competition evaluation includes four domains (*Restaurant*, *Laptop*, *Hotel*, and *Finance*), we strictly confined our training phase to the official Subtask 3 (DimASQP) subset. We did not incorporate any external data.

The Subtask 3 training data is inherently limited to three domains: *Restaurant* (available for all languages except Japanese), *Laptop* (English and Chinese), and *Hotel* (Japanese). To maximize the available training signals, we concatenated the official training and development sets of this subtask. Consequently, relying solely on Subtask 3 for training meant that our pipeline had to generalize zero-shot to the unseen *Finance* domain when evaluated on Subtask 1 (for Japanese).

A significant challenge is the severe imbalance across languages and domains (e.g., 6,050 Chinese *Restaurant* instances vs. 1,240 for Tatar) (Lee et al., 2026). To prevent bias toward high-resource categories, we implemented a two-step upsampling strategy. First, we abstracted the available training domains into two meta-categories: **Hospitality** (encompassing *Restaurants* and *Hotels*) and **Technology** (*Laptops*). Within each language, we balanced these meta-categories by upsampling the minority domain to match the sample count of the majority domain. Second, we applied language-level balancing by upsampling the concatenated data of lower-resource languages (e.g., Tatar and Ukrainian) to match the size of the highest-resource language. This dynamic balancing ensures robust cross-lingual and cross-domain generalization.

4.2 Training Configuration

The two stages of our pipeline were trained on an NVIDIA A100 (40GB) GPU using the following configurations:

Stage 1 (Generative Extraction): We fine-tune `Qwen2.5-32B-Instruct` (Yang et al., 2025) in 4-bit NormalFloat (NF4) quantization. To fit the 32-billion parameter model into the 40GB VRAM, we employ Unsloth's gradient checkpointing (Han and Han, 2023). We use LoRA (Hu et al., 2022) with a rank (r) of 16 and a scaling factor (α) of 16, targeting all linear layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). The model is trained for 1 epoch using the `adamw_8bit` optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2×10^{-4}. We set the maximum sequence length to 2048 tokens and use an effective batch size of 8.

Stage 2 (Discriminative Regression): The `xlm-roberta-large` model (Conneau et al., 2020) is trained as a sequence regressor with 2 output labels. We use a learning rate of 1×10^{-5} and a weight decay of 0.01. The model is trained for 4 epochs with an effective batch size of 16. Input sequences are truncated at 256 tokens. The best model is selected based on the lowest validation loss.

To represent a standard training environment, no fixed random seeds were enforced during fine-tuning, meaning results may exhibit minor variance across different initializations.

4.3 Evaluation Metrics

Following the official evaluation framework (Lee et al., 2026), we report **Root Mean Square Error (RMSE)** for Subtask 1 to measure continuous prediction error. For the joint extraction tasks (Subtasks 2 and 3), we use the **Continuous F1-score (cF1)**. This metric unifies categorical and continuous evaluation: a predicted tuple is initially counted as a categorical True Positive only if all structural elements exactly match the gold labels, and is subsequently weighted by the VA prediction error to ensure structural and sentiment precision are evaluated jointly.

5 Results and Analysis

We evaluate our unified hybrid framework across all three official DimABSA subtasks. The results demonstrate that our architecture, despite being trained solely on Subtask 3 (DimASQP) data, achieves highly competitive performance across various languages and domains, validating our data-balancing strategies and zero-shot capabilities.

5.1 Generative Extraction (Subtasks 2 and 3)

Table 1 summarizes the continuous F1-scores (cF1) and corresponding ranks achieved by our system in

the structural extraction tasks.

		Subtask 2 (cF1)		Subtask 3 (cF1)	
Lang	**Domain**	**Ours (Rank)**	**Baseline**	**Ours (Rank)**	**Baseline**
ENG	Rest	0.6391 (8th)	0.4920	0.5937 (6th)	0.3746
ENG	Lap	0.5513 (8th)	0.4424	0.3049 (9th)	0.2795
ZHO	Rest	0.5320 (6th)	0.3529	0.5026 (4th)	0.2859
ZHO	Lap	0.4807 (6th)	0.2494	0.3478 (9th)	0.1900
JPN	Hotel	**0.5694 (2nd)**	0.3464	0.3992 (4th)	0.1943
RUS	Rest	0.5253 (8th)	0.4242	0.4991 (6th)	0.2963
UKR	Rest	0.5270 (8th)	0.4220	0.4879 (5th)	0.2971
TAT	Rest	0.4496 (8th)	0.3577	0.4113 (6th)	0.2380

Table 1: Official cF1 scores and ranks for DimASTE (Subtask 2) and DimASQP (Subtask 3), compared against the official Kimi-K2 Thinking baseline.

In the *Japanese Hotel* domain, our system secured the **2nd place** overall in Subtask 2 (0.5694 cF1) and the 4th place in Subtask 3. This result indicates the effectiveness of our language-level data balancing strategy. By balancing the dataset across all six languages, the initially scarce Japanese instances were upsampled to constitute exactly one-sixth of the total training corpus. Since *Hotel* was the only available domain for Japanese in the Subtask 3 training data, the 32B-parameter LLM received concentrated exposure to this specific language-domain pair. This upsampling prevented the model from biasing towards the heavily represented English or Chinese data, enabling it to robustly internalize the structural extraction patterns for Japanese hotels.

Furthermore, our system demonstrated robust generalization across non-Latin script languages of varying resource levels. The framework secured competitive positions such as 4th place in the high-resource *Chinese Restaurant* category and 5th place in the lower-resource *Ukrainian Restaurant* category (Subtask 3).

Ablation Observation: The Necessity of Data Balancing. To validate the efficacy of our cross-lingual data balancing strategy, we observed the model's training dynamics prior to implementing the upsampling mechanism. Without balancing, the system exhibited a severe performance discrepancy across the dataset. The generative LLM heavily biased its internal representations toward the dominant English and Chinese distributions, which caused the extraction quality for low-resource languages (such as Japanese) to degrade significantly. By enforcing an equalized distribution across all languages and domains, we successfully reduced this cross-lingual variance, demonstrating that deliberate data homogenization is a critical prerequisite for a unified multilingual architecture.

Error Analysis: Category Taxonomy and the "Laptop" Drop-off. While our generative extraction performed well overall, a comparative analysis between Subtask 2 (ASTE) and Subtask 3 (ASQP) reveals a significant bottleneck in fine-grained category prediction. As shown in Table 1, performance in the English *Laptop* domain suffers a severe degradation, dropping from a cF1 of 0.5513 in Subtask 2 to 0.3049 in Subtask 3. A similar pattern is observable in the Chinese *Laptop* domain. Through error analysis, we attribute this to the high semantic overlap in the predefined aspect categories of the technical domains. While the LLM reliably extracts the literal aspect and opinion terms, it frequently generates false positives when mapping these terms to the strict Subtask 3 taxonomies (e.g., confusing `LAPTOP#OPERATION_PERFORMANCE` with `LAPTOP#USABILITY`). In contrast, the *Restaurant* domains exhibit a much lower ST2-to-ST3 degradation, suggesting their categorical taxonomy is more distinct.

Qualitative Analysis: Implicit Targets and Boundary Mismatches. A qualitative comparison between our model predictions and the gold labels in the *English Restaurant* domain highlights both the theoretical strengths and the practical vulnerabilities of the generative extraction stage. On the positive side, the LLM successfully and consistently handled implicit sentiment targets. In instances where an opinion was expressed without a clear physical noun, the model correctly generated the "NULL" aspect token and mapped it to the appropriate category. This confirms our hypothesis that generative models can bypass the strict token-dependency that bottlenecks traditional sequence labelers. However, the analysis also revealed a recurring source of error that negatively impacts the strict exact-match evaluation metric (cF1): opinion boundary mismatch. While the LLM comprehends the sentiment accurately, its generative nature causes it to frequently extract opinion spans that are slightly broader or narrower than the human gold annotations (e.g., generating "very friendly" instead of "friendly", or including adjacent punctuation). While semantically valid, these generative artifacts trigger strict-match penalties.

Lang	Domain	RMSE_VA		PCC_V	PCC_A	Rank
		Ours	Baseline	Ours	Ours	
JPN	Fin	**0.9982**	1.6396	0.7937	0.2114	15th
ZHO	Lap	**1.0931**	1.6440	0.6877	0.4656	21st
ZHO	Rest	**1.1601**	1.8959	0.7343	0.5934	21st
ENG	Rest	**1.4265**	2.1461	0.8342	0.5362	27th
RUS	Rest	**1.5991**	1.7768	0.8131	0.4923	14th
UKR	Rest	**1.6039**	1.7805	0.8059	0.4745	13th
TAT	Rest	2.0212	**1.9380**	0.5829	0.3676	12th

Table 2: Official results for VA prediction in DimASR (Subtask 1). We compare our RMSE against the official *Kimi-K2 Thinking* baseline. The stark contrast between Valence correlation (PCC_V) and Arousal correlation (PCC_A) highlights the model's struggle to predict emotional intensity without explicit opinion anchors.

5.2 Subtask 1 (DimASR): The Impact of Structural Input Mismatch

Subtask 1 evaluates the model's ability to predict continuous Valence and Arousal (VA) scores given the original text and the target aspect term. Table 2 presents our Root Mean Square Error (RMSE) alongside the Pearson Correlation Coefficients for Valence (PCC_V) and Arousal (PCC_A).

As the rankings indicate, our system's overall RMSE performance in this subtask was suboptimal compared to top-tier competitors. However, a deeper analysis of the individual PCC metrics reveals a stark divergence: the model maintained strong predictive accuracy for Valence (e.g., PCC_V of 0.8342 for English Restaurant and 0.8059 for Ukrainian Restaurant), while its performance on Arousal degraded significantly (PCC_A of 0.5362 and 0.4745, respectively). In the zero-shot Japanese Finance domain, this gap was most extreme, with a PCC_V of 0.7937 contrasted by a PCC_A of only 0.2114.

We attribute this pronounced discrepancy to a critical **structural input mismatch** inherent in our decoupled pipeline. During the training phase—which was optimized for the full extraction outputs of Subtasks 2 and 3—our XLM-RoBERTa cross-encoder learned to predict VA scores by relying on a concatenated input of the sentence, the aspect, and the explicitly extracted *opinion term*.

Because Subtask 1 only provides the aspect term, the explicit opinion anchor was absent during inference. The metric split clearly indicates that *Valence* (general polarity) can still be robustly inferred from the global sentence context. *Arousal* (emotional intensity), however, is highly localized and intrinsically tied to specific opinion modifiers (e.g., distinguishing between "good" and "absolutely fantastic"). Forcing the regressor to predict intensity without its learned explicit opinion anchor starved the model of the necessary signal, causing the Arousal predictions to collapse and severely inflating the overall RMSE.

6 Conclusion

In this paper, we presented a hybrid generative-discriminative framework for the SemEval-2026 Task 3 on Dimensional Aspect-Based Sentiment Analysis (DimABSA). Our approach centered on a strategic decoupling of tasks: utilizing the generative capabilities of `Qwen2.5-32B` for structural sentiment extraction and the discriminative precision of `XLM-RoBERTa` for continuous Valence-Arousal regression.

Our experimental results demonstrate that this two-stage pipeline, supported by a targeted cross-lingual data balancing mechanism, is highly effective for structural extraction across diverse languages and domains. This was evidenced by our 2nd place ranking in the Japanese *Hotel* domain for Subtask 2. However, our analysis also revealed a significant trade-off: while our regressor is highly precise when provided with full aspect-opinion context, its performance degrades in Subtask 1 scenarios where explicit opinion terms are omitted. This structural mismatch between training and inference highlights a critical area for future optimization in decoupled sentiment pipelines.

Future work will focus on enhancing the robustness of the regression stage by incorporating latent opinion representations when explicit anchors are missing. Additionally, we aim to refine the generative stage to minimize boundary mismatches, further closing the gap between deep semantic understanding and strict token-level extraction metrics.

7 Ethical Considerations

While our framework employs data balancing to mitigate language-level disparities, the underlying pre-trained models (`Qwen2.5`, `XLM-RoBERTa`) may still harbor inherent biases that affect Valence-Arousal predictions across different cultures. Furthermore, as a generative system, Stage 1 is susceptible to hallucinations or structural mismatches, which can lead to misinterpretation of sentiment. Finally, we acknowledge the sensitive nature of automated affect recognition and emphasize that

such systems should be used responsibly to avoid intrusive emotional profiling or the reinforcement of stereotypical biases in sentiment analysis.

Acknowledgments

This work was conducted as part of a Bachelor's Thesis at the Technical University of Munich (TUM). The authors would like to thank the organizers of SemEval-2026 Task 3 for providing the framework and datasets for this competition. We also acknowledge the Chair for Human-Centered Computing for providing the computational resources and supporting environment for this research.

References

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

Daniel Han and Michael Han. 2023. Unsloth: Accelerating large language model fine-tuning. https://github.com/unslothai/unsloth.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Shaghayegh Kolli, Richard Rosenbaum, Timo Cavelius, Lasse Strothe, Andrii Lata, and Jana Diesner. 2025. Hybrid fact-checking that integrates knowledge graphs, large language models, and search-based retrieval agents improves interpretable claim verification. In *Proceedings of the 9th Widening NLP Workshop*, pages 106–115, Suzhou, China. Association for Computational Linguistics.

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. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:8600–8607.

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.

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

Cem Üyük, Danica Rovó, Shaghayeghkolli, Rabia Varol, Georg Groh, and Daryna Dementieva. 2024. Crafting tomorrow's headlines: Neural news generation and detection in English, Turkish, Hungarian, and Persian. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 271–307, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

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.

A Appendix

A.1 Detailed Experimental Results

To provide a comprehensive overview of our system's performance and to support the error analysis

discussed in Section 5, we report the extended evaluation metrics across all subtasks.

Table 3 details the performance for the continuous Valence-Arousal regression (Subtask 1). Table 4 provides the in-depth structural extraction metrics for Subtasks 2 and 3, including continuous Precision (cPrec), continuous Recall (cRec), True Positives (TP), False Positives (FP), and False Negatives (FN). The stark increase in False Positives and False Negatives in the *Laptop* domains between Subtask 2 and Subtask 3 quantitatively highlights the LLM's struggle with the fine-grained technical category taxonomy.

Lang	Domain	RMSE	PCC_V	PCC_A	Rank
ENG	Restaurant	1.4265	0.8342	0.5362	27th
ZHO	Restaurant	1.1601	0.7343	0.5934	21st
ZHO	Laptop	1.0931	0.6877	0.4656	21st
JPN	Finance	0.9982	0.7937	0.2114	15th
RUS	Restaurant	1.5991	0.8131	0.4923	14th
UKR	Restaurant	1.6039	0.8059	0.4745	13th
TAT	Restaurant	2.0212	0.5829	0.3676	12th

Table 3: Detailed results for VA prediction in DimASR (Subtask 1).

A.2 Hyperparameter and Model Details

Table 5 details the specific configurations for both stages of our hybrid pipeline. All experiments were conducted using the Unsloth library for efficient LLM fine-tuning and the Hugging Face Transformers library for regression.

A.3 Prompt Engineering and Constraints

To ensure complete reproducibility of the generative extraction stage, we provide the exact instruction constraints embedded within the Alpaca-style prompt template. The system was instructed to act as a specialized ABSA extractor with the following logic:

System Instruction:
Analyze the provided text to perform Aspect-Based Sentiment Analysis (ABSA). Extract all aspect-opinion pairs and their corresponding categories. Return the result strictly as a JSON list of objects with the keys: 'Aspect', 'Category', 'Opinion'.

Constraint Rules:

1. **Aspect:** Extract the EXACT substring from the text representing the target.
   - Maintain original capitalization and spelling.
   - DO NOT replace with synonyms (e.g., if text says 'coffee', output 'coffee', NOT 'DRINKS').
   - If the aspect is implicit (not mentioned), use 'NULL'.
2. **Category:** Assign a semantic category strictly from: [*entities*] combined with [*attributes*] (Format: ENTITY#ATTRIBUTE).
3. **Opinion:** Extract the EXACT substring expressing the sentiment from the text.
4. **Output format:** Ensure valid JSON. Do not add explanations.

While the LLM was strictly prompted to maintain original capitalization, autoregressive generation inherently introduces minor normalizations. To strictly align with the SemEval exact-match evaluation, we applied an algorithmic post-processing step (*Safe Capitalization Fix*) that mapped the generated substrings back to their original casing in the source text using case-insensitive matching.

		Subtask 2 (DimASTE)						Subtask 3 (DimASQP)					
Lang	**Domain**	**cF1**	**cPrec**	**cRec**	**TP**	**FP**	**FN**	**cF1**	**cPrec**	**cRec**	**TP**	**FP**	**FN**
ENG	Restaurant	0.6391	0.6609	0.6186	1424	569	705	0.5937	0.6162	0.5728	1317	662	812
ENG	Laptop	0.5513	0.5950	0.5136	1095	609	879	0.3049	0.3794	0.2838	606	1096	1200
ZHO	Restaurant	0.5320	0.5638	0.5035	1530	1025	1331	0.5026	0.5366	0.4727	1437	1083	1424
ZHO	Laptop	0.4807	0.5037	0.4597	924	833	1001	0.3478	0.3650	0.3321	667	1085	1258
JPN	Hotel	0.5694	0.5777	0.5613	847	1555	596	0.3992	0.4083	0.3905	589	791	854
RUS	Restaurant	0.5253	0.5265	0.5241	784	520	526	0.4991	0.5043	0.4939	739	544	571
UKR	Restaurant	0.5270	0.5264	0.5276	788	525	522	0.4879	0.4905	0.4853	726	570	584
TAT	Restaurant	0.4496	0.4511	0.4480	678	623	632	0.4113	0.4193	0.4036	609	652	701

Table 4: Detailed extraction performance, illustrating the drop in precision and the increase of False Positives (FP) in complex taxonomies (Subtask 3).

Parameter	**Value**
Stage 1: Qwen2.5-32B-Instruct	
Quantization	4-bit NormalFloat (NF4)
LoRA Rank (r) / Alpha (α)	16 / 16
LoRA Target Modules	All linear layers
Learning Rate	2×10^{-4}
Training Epochs	1
Max Sequence Length	2048
Optimizer	adamw_8bit
Stage 2: XLM-RoBERTa-Large	
Model Architecture	Sequence Classification
Learning Rate	1×10^{-5}
Training Epochs	4
Batch Size	16
Weight Decay	0.01
Max Sequence Length	256

Table 5: Hyperparameter settings for Stage 1 (Extraction) and Stage 2 (Regression).

Lang/Domain	**Input Text**	**Target Quadruplets (Aspect, Category, Opinion, VA)**
English Laptop	Great perforemce at a great price	1. (perforemce, LAPTOP#OPERATION_PERFORMANCE, Great, 6.88#7.25) 2. (price, LAPTOP#PRICE, great, 7.12#7.50)
Japanese Hotel	部屋の清掃も行き届いていてリラックスできました。	1. (清掃, ROOMS#CLEANLINESS, 行き届いて, 6.33#6.00) 2. (NULL, ROOMS#COMFORT, リラックス, 6.50#5.83)
Tatar Restaurant	Порция бик яхшы иде!	1. (Порция, FOOD#STYLE_OPTIONS, бик яхшы, 7.17#6.33)

Table 6: Examples from the DimASQP (Subtask 3) development sets. The task requires extracting structural elements—handling noise (e.g., typos like "perforemce"), implicit targets ("NULL"), and multiple sentiments per sentence—while simultaneously predicting continuous Valence-Arousal (VA) scores across diverse scripts and domains.