ALPS-Lab at SemEval-2026 Task 3: A Multilingual Generative LLM Approach for Dimensional Aspect Sentiment Analysis

Songqian Dai **Wei Lin**
Institude of Computing and Mathematics
Fujian University of Technology
2251316002@smail.fjut.edu.cn, wlin@fjut.edu.cn

Abstract

We present a system for Dimensional Aspect-Based Sentiment Analysis (DimABSA) in SemEval-2026 Task 3, focusing on predicting aspect-level sentiment quadruplets with continuous valence–arousal scores across multiple languages and domains. Our approach leverages generative large language models (LLMs), specifically Gemma-3 27B, fine-tuned using QLoRA for efficient adaptation under limited GPU memory. By merging multilingual datasets within each domain, we enable cross-lingual transfer and improve performance in low-resource settings. Post-processing scripts are used to address duplicate predictions and ensure accurate evaluation. Experimental results demonstrate that our system achieves competitive cF1 scores, outperforming official baselines in most domains. We discuss the impact of multilingual training, hyper-parameter choices, and limitations, highlighting directions for future work in data augmentation and optimization.

1 Introduction

Aspect-based sentiment analysis (ABSA) (Pontiki et al., 2016) aims to identify sentiment polarity toward specific aspects or entities within a sentence. Over time, ABSA has evolved to encompass a range of subtasks, including aspect sentiment triplet extraction (ASTE) and aspect sentiment quadruplet prediction (ASQP) (Zhang et al., 2021). ASQP is the most comprehensive, capturing the full relationship between aspects, opinions, and categories. Earlier approaches relied on rule-based methods (Hu and Liu, 2004), while recent work has adopted BERT and encoder–decoder architectures (Cai et al., 2021; Gou et al., 2023). Unlike these tasks, which treat affective states as discrete classes (positive, neutral, negative), the dimensional approach provides more fine-grained emotional information.

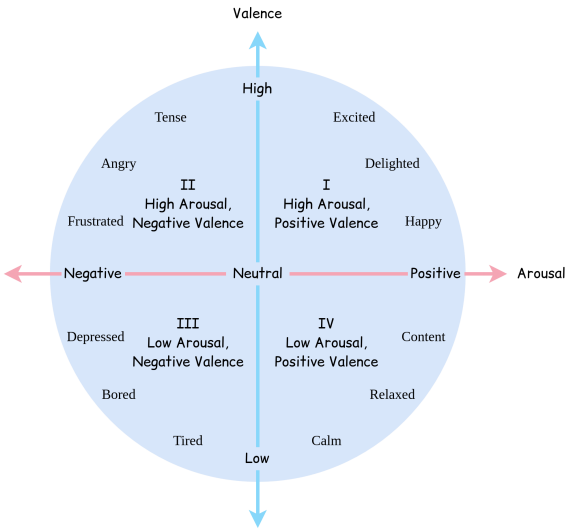

Figure 1: This figure (Yu et al., 2016) represents emotions in dimensional Valence-Arousal space.

The Dimensional Aspect-Based Sentiment Analysis (DimABSA) task involves identifying specific aspects within a text and predicting their sentiment intensities along continuous dimensions. As illustrated in Figure 1, unlike traditional ABSA, which uses categorical labels, DimABSA employs numeric representations—specifically Valence (degree of pleasure) and Arousal (degree of activation). This approach is crucial as it provides a more granular and comprehensive framework for sentiment analysis. capturing the intensity of emotions, it offers richer information for downstream applications.

DimABSA extends classical aspect-based sentiment analysis by requiring systems to predict continuous sentiment scores—valence and arousal (Russell, 2003)—and, where relevant, discrete aspect labels such as categories or opinion spans (Buechel and Hahn, 2016). Early resources for dimensional sentiment include the EmoBank corpus (Buechel and Hahn, 2017). The shared task pro-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1652–1658

vides multilingual datasets to support cross-lingual transfer research.

This year's task (Lee et al., 2026) features a multilingual dataset, facilitating research into cross-lingual transfer and under-resourced languages.

Following the shared task, performance for DimASQP is evaluated using the continuous F1 (cF1) metric, which treats a prediction as correct only if all categorical elements A, C, O exactly match the gold annotation and then weights the true positive by the normalized Euclidean distance between predicted and gold valence–arousal scores.

We propose a supervised fine-tuning approach on generative large language models, adapted to support Gemma-3 (Team, 2025) 27B. To deal with limited GPU memory, we use QLoRA (Dettmers et al., 2023) to fine-tune low-rank adapters on top of quantized base models instead of updating all parameters. As the Gemma-3 models support multilingual inference, we merge the datasets from different languages within the same domain and train a single model per domain. This allows the model to leverage shared structure across languages while still specializing to each domain. Our system achieved approximately **6th** place in the official leaderboard ranking.

A key challenge was some limited dataset size—the training set was less than twice the size of the test set. Without additional strategies such as data augmentation or in-context learning, it was difficult to further improve performance. The code will be released on https://github.com/connectionRst/DimABSA2026.

2 Background

Dimensional Aspect Sentiment Quadruplet Prediction (DimASQP), subtask 3 of the DimABSA task (Yu et al., 2026), is defined as a joint extraction–classification–regression task on the DimABSA datasets. The shared task defines several language–domain combinations and provides training, development, and test sets for each. Systems are evaluated on how accurately they recover the gold continuous scores and aspect-level targets. Given an input sentence , the goal is to recover a set of sentiment quadruplets:

$$(A, C, O, \mathrm{VA})$$

where each tuple contains:

- an aspect term A, i.e., a word or phrase denoting the opinion target;
- an aspect category C from a predefined Entity-Attribute inventory associated with ;
- an opinion term O, i.e., a sentiment-bearing expression (including possible modifiers) describing the attitude toward ; and
- a pair of continuous valence–arousal scores:

$$(V, A) \in \mathbb{R}^{[1,9]\times[1,9]}$$

where $[1, 9]$ denote extreme negative/positive valence or low/high arousal and 5 denotes neutrality.

The shared task adopts a continuous evaluation framework to assess predictions of aspect-level sentiment quadruplets. Unlike traditional categorical metrics, the continuous F1 (cF1) metric accounts for both exact categorical matches and the proximity of predicted valence–arousal (VA) scores to gold annotations.

Specifically, for each gold quadruplet, the evaluator identifies matching predictions based on the aspect term A, opinion term O, and category C. When a match is found, the continuous true positive score is computed as:

$$\mathrm{cTP}_t = 1 - \frac{d_{\mathrm{VA}}}{D_{\max}}$$

where d_{VA} denotes the Euclidean distance between the predicted and gold VA values, and $D_{\max} = \sqrt{128}$ represents the maximum possible distance in the $[1, 9] \times [1, 9]$ space. If the predicted VA values are outside this space, the cTP_t would be 0. Specially, when the system produce more than one quadruplet with same (A, C, O), the evaluator would indicate it as false negative match as a penalty.

The final performance metrics are then defined as:

$$\mathrm{cPrecision} = \sum \frac{\mathrm{cTP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{cRecall} = \sum \frac{\mathrm{cTP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{cF1} = \frac{2 \cdot \mathrm{cPrecision} \cdot \mathrm{cRecall}}{\mathrm{cPrecision} + \mathrm{cRecall}}$$

where TP, FP, and FN denote the counts of true positives, false positives, and false negatives, respectively.

With the development of large language models (LLMs), LLM-based methods have gained significant popularity, particularly with the advent of techniques like Chain-of-Thought reasoning (Kim et al., 2024). Furthermore, parameter-efficient fine-tuning (PEFT) approaches, such as LoRA and QLoRA (Xu et al., 2024), have been introduced to address the challenges of limited GPU memory

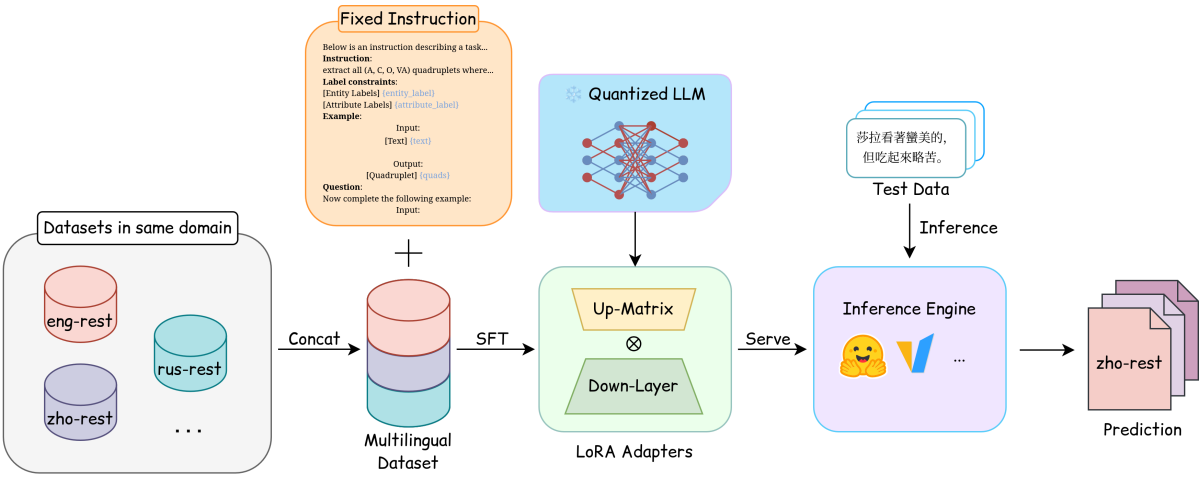

Figure 2: The architecture of our system. Given an input sentence, a generative LLM (Qwen-3 4B or Gemma-3 27B) with QLoRA adapters produces (A, C, O, VA) quadruplets. During SFT we embed the allowed category list and a single example into a fixed prompt and train the model to emit the structured assistant response. At inference the example is replaced by the test sentence and the decoded output is parsed (e.g., with a regex) to recover (A, C, O, VA) tuples.

and small labeled datasets. Data augmentation (Xu et al., 2025) and in-context learning methods (Zhu et al., 2024) have also been developed to enhance performance in low-resource scenarios.

3 System Overview

Our system follows a generative large language model (LLM) paradigm for dimensional aspect-based sentiment analysis. As is showned in Figure 2, Given an input sentence, the model generates structured output containing aspect quadruplets (A, C, O, VA) in a predefined schema. In the SFT stage we build a one-turn dialog for each domain by embedding the allowed category list and an example into a fixed prompt template; the model is trained to emit the assistant response containing the quadruplets. At inference time the test sentence is plugged into the same template (replacing the training example), and the decoded output—formatted according to a simple schema—can be matched with a regular expression to recover the predicted (A, C, O, VA) tuples. After the competition we observed that LLMs sometimes produced duplicated outputs; we mitigated this by implementing a post-processing script. This part is also mentioned in Implementation Details.

We initially explored Qwen-3 4B following the official example baseline, but preliminary development set evaluation suggested limited capacity for fine-grained quadruplet extraction. We subsequently adopted Gemma-3 (Team, 2025) 27B, which demonstrated stronger performance in later experiments.

We fine-tuned low-rank adapters with QLoRA on quantized base models, allowing a 27B-parameter model to fit within a 24 GB NVIDIA GPU with Ampere architecture. The final base model used is `unsloth/gemma-3-27b-it-unsloth-bnb-4bit`.

We revised the dialog generation pipeline to produce the final prompt text in JSON format, rather than concatenating hardcoded special symbols (chat templates). This enables compatibility with a broader range of models, rather than restricting the system to a single architecture.

Notably, the required JSON schema for multimodal LLMs (e.g., Gemma-3) differs from that used by text-only models (e.g., Qwen-3). To accommodate this, we adapt the prompt construction process: by parsing the Unsloth-style model name, we infer the model type and generate a schema tailored to each model's requirements. This approach ensures that prompts are correctly formatted for both text-only and multimodal LLMs, supporting seamless fine-tuning across diverse model architectures.

4 Experimental Setup

4.1 Datasets

For most experiments we use the original train/dev/test splits as provided by the organizers. For the final submission, we merge the train and dev sets to partially overcome the data shortage and train a model on the combined data, which is then evaluated on the official test set. There are no

more foreign or augmented datasets being used in training.

4.2 Implementation Details

Preprocessing. we mostly follow the original data format and tokenize with each model's native tokenizer. We standardize output formats so that the model always predicts aspects and dimensional scores in a consistent schema.

Parameter tuning. we fine-tuned models with batchsize = 8 as a balance between GPU memory usage and training stability. Smaller batches train more slowly, while larger batches made results less stable. For the learning rate, we observe that 10^{-4} is better than 10^{-5} for the `hotel_jpn` subtask, while 10^{-5} is more generally robust across other subtasks. We decode with temperature = 0.1, which is more suitable for the information extraction. Nevertheless, exact reproducibility may be challenging: when running inference on different servers or GPUs, results can vary slightly. Interestingly, a higher decoding temperature is more suitable for the ukr-rest and tat-rest domains, which are constructed via machine translation. Other SFT parameters—including LoRA, training seed, weight decay, and learning rate scheduler—are kept at their default values.

Post-Processing Scripts After Submission. Generative models sometimes produce repeated quadruplets (A, C, O, VA) in their outputs, typically due to decoding artifacts. Such duplicates can artificially lower the true positive count under the official continuous F1 metric. To address this, we implement a simple post-processing step that removes exact duplicates—quadruplets are considered identical if all elements match. In rare cases, the same (A, C, O) may be associated with different VA scores; This scenario was observed in `tat-rest`, which may be due to Gemma-3's limited support for Cyrillic script languages. while our current approach retains these, further refinement to handle such cases is left for future work. In our experiments, this scenario did not occur. The script is included in the repo.[1]

4.3 Baseline

Official baselines are provided by the shared task organizers (Lee et al., 2026). These include:

- **Kimi Baseline:** Following the provided startup example, this baseline utilizes the Kimi-K2 Thinking model in a one-shot in-context learning (ICL) setting. A single demonstration from the training set is included in the prompt to guide the model in generating structured quadruplets and continuous valence–arousal scores.
- **SFT Baseline:** This baseline represents performance using large-scale open-source models (Llama-3.3 70B and GPT-OSS 120B) fine-tuned with 4-bit QLoRA.

5 Results

Note: Unless otherwise specified, all reported scores below are post-hoc results—predictions were processed with our post-processing script after the competition, and evaluated using the official test set and metrics provided by the organizers.

5.1 Main Result

Table 1 shows the performance of our proposed method on the subtask 3. As the task includes multiple language–domain combinations, we report average cF1 scores within each domain and the official ranking.

Our reported cF1 is slightly higher than the official submission because that submission lacked post-processing to handle LLM issues—duplicate entries for the same (A, C, O) and missing VA scores. After cleaning the outputs, we re-evaluated with the official metrics script against the gold data and report the cleaned score.

Overall, our system achieves the highest cF1 in seven of eight domains; the sole exception is `jpn-hot`, where the SFT baseline outperforms us. This may caused by two union point: the `hotel` We outperform the Kimi-K2 Thinking baseline on every domain, and we exceed the SFT baseline on eng-rest, eng-lap, rus-rest, tat-rest, ukr-rest, zho-rest, and zho-lap.

The lower cF1 score for `jpn-hot` can be attributed to two main factors:

- For the hotel domain, only Japanese data is available, which limits the effectiveness of our multilingual training strategy.
- The dataset size for `jpn-hot` is relatively small (1600 examples), compared to domains like `eng-lap` (2284 examples), making it more challenging for the model to generalize.

5.2 Analysis

To better understand the contribution of multilingual training within a domain, we compares *inde-*

[1] `VA.ipynb`.

Domain	Ours	Kimi Baseline	SFT Baseline
eng-rest	0.6212	0.3746	0.5048
eng-lap	0.3441	0.2795	0.2483
jpn-hot	0.3679	0.1943	0.4151
rus-rest	0.5169	0.2963	0.4118
tat-rest	0.4404	0.2380	0.3702
ukr-rest	0.5163	0.2971	0.4070
zho-rest	0.4864	0.2859	0.4391
zho-lap	0.3969	0.1900	0.3551

Table 1: Main experimental results (cF1) comparing our system ('Ours'), the Kimi-K2 Thinking baseline, and the SFT baseline (Lee et al., 2026). Domains jpn-hot and zho-lap used GPT-OSS 120B runs; other domains used Llama-3.3 70B.

Domain	Multilingual	Independent
eng-rest	0.759	0.7431
eng-lap	0.5977	0.5859
rus-rest	0.5488	0.5399
tat-rest	0.5222	0.5512
ukr-rest	0.4281	0.5136
zho-rest	0.5851	0.5921
zho-lap	0.4302	0.4136

Table 2: Development-set cF1 scores comparing models trained independently on each language–domain pair versus single multilingual model that merges all languages for each domain.

pendent models (one per language–domain pair) with a *multilingual* model that sees all languages for the same domain during fine-tuning. This is motivated by the scarcity of data: most language–domain datasets are too small (less than 10000 examples) and the aspect/VA distributions are not balanced. Table 2 reports the development-set cF1 scores for each setting.

We observe that merging languages generally improves performance in low-resource conditions. The largest gains occur for **ukr-rest** and **rus-rest**, where the independent models were weakest; combining data allows the model to learn shared patterns of aspect expressions and sentiment, partially offsetting the limited size. Some domains, such as **eng-rest** and **zho-rest**, also see moderate improvements despite having more data, suggesting that cross-lingual transfer is broadly beneficial.

However, the multilingual strategy is not uniformly better. For the two laptop domains (**eng-lap** and **zho-lap**) the multilingual model slightly underperforms the separate models. We attribute this to mismatched valence–arousal distributions across languages—when the joined dataset exhibits different centers or variances in VA space, the model struggles to fit both simultaneously. This limitation points to a direction for future work: designing more sophisticated multilingual fine-tuning methods that account for distributional shifts.

6 Conclusion

In this work, we participated in the DimABSA shared task on dimensional aspect-based sentiment analysis. We focused on leveraging multilingual generative LLMs (Gemma-3 27B) together with parameter-efficient fine-tuning (QLoRA) to build systems that can predict aspect-level sentiment scores across several languages and domains under limited-resource conditions.

Our experiments indicate that this combination of modern LLMs and lightweight adaptation yields competitive results, placing our systems around 6th in the official ranking despite the relatively small amount of labeled data. Multilingual training within domains and careful hyper-parameter tuning (e.g. batch size and learning rate) are particularly important to stabilize training and achieve good performance.

Limitation

Several limitations remain in our current approach. Firstly, data scarcity is a significant challenge. We plan to address this by augmenting the dataset using generative LLMs and statistical methods to incorporate additional resources. Secondly, to improve the concrete performance

specific for this task, fine-tuning other untouched hyper-parameters like LoRA params is needed. Finally, while reinforcement learning for semantic analysis is a promising direction, it was not explored in this work and may be considered in future research.

LoRA Hyper-Parameters: The LoRA hyper-parameters were chosen from preliminary experiments and not exhaustively tuned; it remains unclear whether these settings are optimal for low-resource conditions or whether alternative ranks, learning rates, or regularization schemes would yield improvements. A deeper discussion of model size and architectural differences is also missing.

Data Scarcity: Our strategy for handling limited data is straightforward. More advanced techniques —data augmentation, retrieval-augmented in-context learning (Zhu et al., 2024), and LLM-based data synthesis (Xu et al., 2025)—could better leverage unlabeled examples and reduce overfitting. Careful attention to valence–arousal distribution is necessary to avoid introducing synthetic-data bias.

Pipeline: We evaluated only generative LLMs; hybrid solutions that combine encoder-based models (Xu et al., 2024) with structured prediction or discrete VA-level modeling may offer complementary strengths and merit exploration.

Objective-aligned Learning: We did not attempt to directly optimize the continuous F1 objective (e.g., through reinforcement learning, PPO, or other policy-gradient methods). Exploring such objective-aligned training remains a promising direction for future work.

Inference Speed & Continuous Training: QLoRA was applied via the Unsloth framework to simplify development. Inference throughput under this setup is limited. Moreover, re-training from scratch cost extra time; Optimizing the serving pipeline (for example by adopting VLLM or other high-throughput backends) is left to future work.

Acknowledgements

This paper was supported by the Scientific Research and Development Foundation of Fujian University of Technology (Grantno.GY-Z220209).

References

Sven Buechel and Udo Hahn. 2016. Emotion Analysis as a Regression Problem - Dimensional Models and Their Implications on Emotion Representation and Metrical Evaluation. In *ECAI 2016 - 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, pages 1114–1122.

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.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-Category-Opinion-Sentiment Quadruple Extraction with Implicit Aspects and Opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLORA: efficient finetuning of quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Jieyong Kim, Ryang Heo, Yongsik Seo, SeongKu Kang, Jinyoung Yeo, and Dongha Lee. 2024. Self-Consistent Reasoning-based Aspect-Sentiment Quad Prediction with Extract-Then-Assign Strategy. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7295–7303, Bangkok, Thailand.

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. arXiv: 2601.23022 [cs.CL].

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.

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110 1: 145-72 .

Gemma Team. 2025. Gemma 3.

Hongling Xu, Delong Zhang, Yice Zhang, and Ruifeng Xu. 2024. HITSZ-HLT at SIGHAN-2024 dimABSA Task: Integrating BERT and LLM for Chinese Dimensional Aspect-Based Sentiment Analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 175–185, Bangkok, Thailand.

Hongling Xu, Yice Zhang, Qianlong Wang, and Ruifeng Xu. 2025. DS^2-ABSA: Dual-Stream Data Synthesis with Label Refinement for Few-Shot Aspect-Based Sentiment Analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15460–15478, Vienna, Austria.

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)*.

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, San Diego, California.

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.

Senbin Zhu, Hanjie Zhao, Xingren Wang, Shanhong Liu, Yuxiang Jia, and Hongying Zan. 2024. ZZU-NLP at SIGHAN-2024 dimABSA Task: Aspect-Based Sentiment Analysis with Coarse-to-Fine In-context Learning. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 112–120, Bangkok, Thailand.