looploop at SemEval-2026 Task 3: A Dimensional Aspect-Based Sentiment System with DeBERTa Regression and Qwen3 Instruction Fine-Tuning

Liu Yang, Gang Hu and Jing Li
School of Information Science and Engineering,
Yunnan University, Kunming 650091, China
{12025215185, 12025215204, lijing_fyjo}@stu.ynu.edu.cn

Abstract

Aspect-Based Sentiment Analysis (ABSA) has evolved to capture continuous affective states, posing challenges for traditional classification models. We adopt a hybrid approach tailored to the varying complexities of the subtasks. For Task 1 (Valence-Arousal Regression), we employ a discriminative architecture using pre-trained DeBERTa encoder with a MeanPooling mechanism to directly regress continuous sentiment scores. For Tasks 2 and 3, which require complex structural extraction of opinion triplets and quadruplets, we utilize a generative approach by fine-tuning the Qwen3-4B-Instruct large language model via 4-bit QLoRA. Our system effectively handles both precise numerical regression and complex structural text generation, achieving competitive results across the English laptop and restaurant domains.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) has long been a pivotal research area in Natural Language Processing, traditionally focusing on predicting discrete sentiment polarities (e.g., positive, negative, neutral) for specific aspects within a text (Pontiki et al., 2014, 2016). However, human emotions are inherently nuanced, continuous, and highly complex, which rigid categorical labels often fail to fully capture.

To address this limitation, the Circumplex Model of Affect is widely adopted in psychology and affective computing (Russell, 1980; Buechel and Hahn, 2017). As illustrated in Figure 1, this model maps emotions onto a continuous two-dimensional space defined by two orthogonal axes: Valence and Arousal (VA). Valence represents the degree of pleasantness (ranging from highly negative to highly positive), while Arousal represents the intensity of the physiological activation (ranging from calm to highly excited). For instance,"thrilled" corresponds to high valence and high arousal, whereas "depressed" is characterized by low valence and low arousal. This continuous VA space provides a much richer and more fine-grained representation of user opinions than traditional polarity labels.

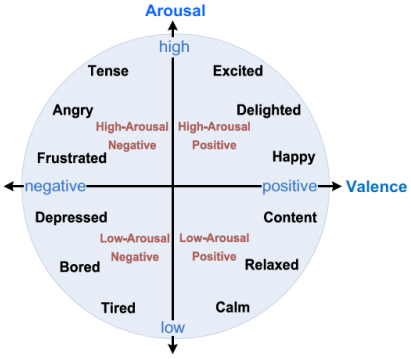


Figure 1: The Valence-Arousal (VA) Coordinate Space for Emotion Representation

Bridging these two paradigms, the SemEval-2026 Task 3 (Yu et al., 2026) introduces Dimensional Aspect-Based Sentiment Analysis (DimABSA), providing datasets annotated with precise continuous VA scores at the aspect level across multiple domains (Lee et al., 2026). The task is divided into three progressive subtasks: predicting Valence-Arousal intensity for a given aspect (Task 1), extracting aspect-opinion-VA triplets (Task 2), and extracting aspect-category-opinion-VA quadruplets (Task 3).

To tackle the distinct complexities of these subtasks, we propose a tailored hybrid framework. Rather than forcing a single model architecture to handle both continuous regression and complex structural generation, our system elegantly decouples the problem. For Task 1, we treat VA prediction as a continuous regression problem, employing a discriminative architecture using pre-trained language models (DeBERTa) (He et al., 2021) enhanced with a MeanPooling mechanism (Reimers

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1511–1517

and Gurevych, 2019) and Huber Loss (Huber, 1964). For Tasks 2 and 3, which require aligning multiple overlapping entities, we frame the extraction as an autoregressive generation task (Zhang et al., 2021). We utilize the Qwen3-4B-Instruct large language model (Bai et al., 2023), heavily fine-tuned via 4-bit Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023), combined with a strict post-processing span recovery algorithm.

Our extensive experiments demonstrate the efficacy of this hybrid approach, effectively balancing the demands of precise numerical regression and complex structured text generation.

2 Related Work

2.1 Aspect-Based Sentiment Analysis

Aspect-Based Sentiment Analysis (ABSA) traditionally focused on predicting discrete sentiment polarities for predefined aspects using pipeline-based extraction methods (Pontiki et al., 2014). The introduction of pre-trained language models (PLMs) such as BERT, RoBERTa, and DeBERTa has revolutionized these systems by providing deep contextualized representations. Recently, the field has transitioned toward end-to-end generative paradigms, framing ABSA as a sequence generation problem to extract multiple elements simultaneously (Zhang et al., 2021). Building upon this evolution, our system strategically leverages the robust contextualized representations of DeBERTa for precise numerical regression in Task 1, and the advanced generative capabilities of modern LLMs for multi-element tuple extraction in Tasks 2 and 3.

2.2 Dimensional Sentiment Analysis

While categorical sentiment is intuitive, it frequently fails to capture the nuanced intensity and complexity of human emotions. The Circumplex Model of Affect addresses this by mapping emotions onto a continuous two-dimensional space defined by Valence and Arousal (VA) (Russell, 1980; Buechel and Hahn, 2017). Recent datasets, including DimStance and DimABSA (Becker et al., 2026; Lee et al., 2026), extend this continuous paradigm to aspect-level evaluations. This shift demands precise numerical regression capabilities from existing models. To meet these strict requirements, we deliberately shift away from traditional classification heads and formulate Task 1 as a direct regression problem, optimizing the continuous VA predictions using Huber Loss (Huber, 1964) to maintain stability against outliers.

2.3 Large Language Models in Extraction

The emergence of Large Language Models (LLMs) has redefined information extraction. Despite their strong zero-shot reasoning capabilities, LLMs often struggle with the strict structural formatting and exact-match constraints required for complex tuple extraction (Zhang et al., 2023). Instruction tuning combined with Parameter-Efficient Fine-Tuning (PEFT) methods, particularly QLoRA (Dettmers et al., 2023), offers a highly effective solution. QLoRA enables quantized LLMs to generate complex structured outputs efficiently without catastrophic forgetting. Inspired by these findings, we specifically select the highly capable Qwen3-4B-Instruct model, paired with 4-bit QLoRA, to successfully tackle the strict formatting and complex structural multi-element generation required in Tasks 2 and 3.

3 System Overview

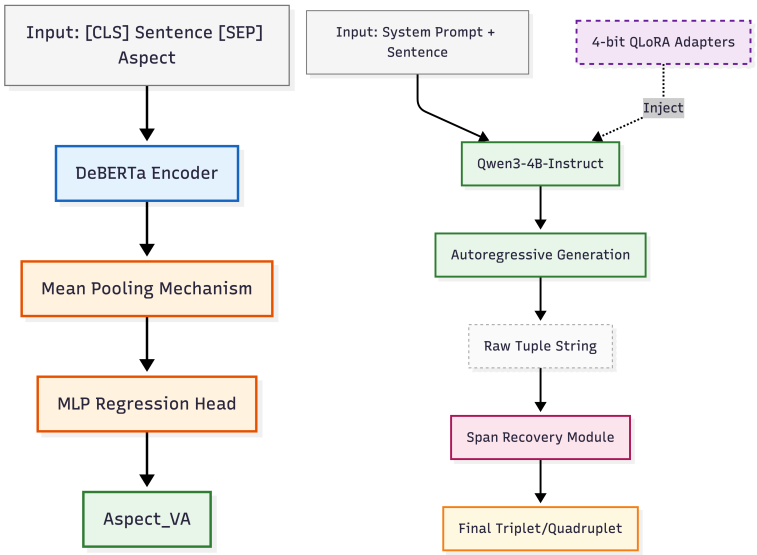
To address the distinct complexities of the three subtasks, we design specialized architectures tailored to either continuous regression or structured generation. Figure 2 presents the overall architecture for our system, including (a) Task 1 and (b) Tasks 2–3.

3.1 Task 1: Encoder-Based Regression

As depicted in Figure 2(a), Task 1 requires predicting continuous Valence and Arousal scores. We formulate this as a regression problem. The input is constructed as a text pair: [CLS] Sentence [SEP] Aspect: {aspect} [SEP]. Instead of conventionally relying solely on the [CLS] token, we introduce a MeanPooling layer (Reimers and Gurevych, 2019) to aggregate the sequence representations. Given the last hidden states $H \in \mathbb{R}^{L \times d}$ and the attention mask $M \in \{0, 1\}^L$, the pooled representation $v \in \mathbb{R}^d$ is calculated as:

$$v = \frac{\sum_{i=1}^{L} h_i \cdot m_i}{\max(\sum_{i=1}^{L} m_i, \epsilon)}$$

where $\epsilon = 10^{-6}$ prevents division by zero. This mechanism ensures that the contextual information of the entire sequence, rather than just the aggregated [CLS] token, contributes evenly to the final prediction. The pooled vector v is then passed through an MLP regression head consisting of a



(a) Framework for Task 1. (b) Framework for Task 2 & 3.

Figure 2: Overall framework of our proposed method. (a) illustrates the pipeline for predicting Aspect-VA intensity, while (b) shows the generation and recovery process for extracting tuples.

Dropout layer ($p = 0.20$), a Linear reduction layer mapping from hidden size d to $d/2$, a GELU activation function, another Dropout layer, and a final Linear layer outputting 2 logits. A Sigmoid function squashes the logits to a $(0, 1)$ range, which are then linearly mapped back to the target $(1, 9)$ VA scale using $y = 1.0 + 8.0 \times x$.

3.2 Tasks 2 & 3: QLoRA-Based Generative Extraction

Tasks 2 and 3 require extracting tuples (Peng et al., 2020): (Aspect, Opinion, VA) and (Aspect, Category, Opinion, VA) (Cai et al., 2021), respectively. As shown in Figure 2(b), traditional token classification struggles with the arbitrary ordering and overlapping spans of these elements. Therefore, we utilize Qwen3-4B-Instruct and reframe the extraction as an autoregressive text generation task (Vaswani et al., 2017; Raffel et al., 2020).

We construct a strict instructional prompt specifying the dimensional bounds (1.00 to 9.00) and the precise output format (e.g., [Quadruplet] (Aspect, CATEGORY#ATTRIBUTE, Opinion, VA)). To enable efficient training, the base LLM is loaded in 4-bit NormalFloat (NF4) precision with double quantization. We inject LoRA (Hu et al., 2022) trainable rank decomposition matrices into the attention and feed-forward layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj).

3.3 Post-Processing and Span Recovery

LLMs are prone to minor hallucinatory deviations, such as altering the capitalization of extracted aspect terms or omitting whitespaces. To align the generated tokens exactly with the original text for exact-match evaluation, we implemented a robust post-processing module. First, regular expressions dynamically parse the generated strings into structured dictionaries. Then, a Span Recovery algorithm searches the lowercased original text for the lowercased generated aspect/opinion. If a match is found, the exact original span (preserving correct casing and punctuation) replaces the generated string. Furthermore, the VA scores are strictly normalized and clipped: values are truncated to [1.00, 9.00] and formatted to exactly two decimal places to ensure formatting compliance.

4 Experimental Setup

4.1 Datasets and Preprocessing

We utilized the official DimABSA datasets for English laptop and restaurant domains (Lee et al., 2026). During preprocessing for Tasks 2 and 3, we constructed a Supervised Fine-Tuning (SFT) dataset mapping system prompts, user inputs, and target outputs into conversational structures. Sentences lacking valid target tuples were deliberately filtered out during SFT to prevent the model from learning a bias towards generating empty sets [Triplet]/[Quadruplet]. For all datasets, 10% of

Task	System	Laptop	Restaurant	Average
Task 1 ($\text{RMSE}_{\text{VA}} \downarrow$)	Baseline (Kimi-K2 Thinking)	2.1893	2.1461	2.1677
	Baseline (Qwen-3 14B)	2.8089	2.6427	2.7258
	looploop (Ours)	**1.3021**	**1.2048**	**1.2534**
Task 2 (cF1 ↑)	Baseline (Kimi-K2 Thinking)	0.4424	0.4920	0.4672
	Baseline (Qwen-3 14B)	0.3827	0.4483	0.4155
	looploop (Ours)	**0.4799**	**0.5799**	**0.5299**
Task 3 (cF1 ↑)	Baseline (Kimi-K2 Thinking)	0.2795	0.3746	0.3271
	Baseline (Qwen-3 14B)	0.1529	0.2673	0.2101
	looploop (Ours)	**0.2781**	**0.5562**	**0.4171**

Table 1: English test-set results and comparison with the official baselines (Laptop/Restaurant) across all subtasks.

the training data was held out as a validation set using GroupShuffleSplit based on sentence IDs to prevent data leakage.

4.2 Training Details

Task 1. We employ DeBERTa-v3-large as the backbone (He et al., 2021). The maximum sequence length is dynamically padded to multiples of 8 to leverage Tensor Cores. The model is trained for 20 epochs with batch size 16 using AdamW (weight decay = 0.5). We apply a freeze–unfreeze strategy: the encoder is frozen for the first epoch (head learning rate 1×10^{-4}), then unfrozen with encoder learning rate 4.5×10^{-6}. We optimize with Huber loss (Huber, 1964), Automatic Mixed Precision (AMP), and gradient clipping at 1.0. Early stopping is triggered after 3 epochs without improvement in validation RMSE_{VA}.

Tasks 2 & 3. We fine-tune Qwen3-4B-Instruct using the HuggingFace Trainer. LoRA hyperparameters are set to rank $r = 16$, $\alpha = 32$, and dropout 0.05. For Task 2, we train for 3 epochs with per-device batch size 2, gradient accumulation 4, and learning rate 8×10^{-5}. For Task 3, we use batch size 4, gradient accumulation 1, and learning rate 6×10^{-5}. Both tasks use a linear scheduler with warmup ratio between 0.03 and 0.06. During inference, we adopt greedy decoding (temperature $= 0.0$) for determinism.

Reproducibility. Detailed environment specifications (library versions and hardware) as well as training/inference scripts are provided in our public repository [1]. We do not use any external training data beyond the official DimABSA datasets.

[1] https://github.com/kiss-the-rain/DimABSA2026

5 Results

Table 1 reports our official English test-set performance and compares it with the official leaderboard baselines across all subtasks and both domains. For Task 1 (VA regression), we evaluate RMSE_{VA} (lower is better). Our system achieves 1.3021 on Laptop and 1.2048 on Restaurant, with an average of 1.2534, substantially reducing error compared to the baselines (2.1677 and 2.7258).

For Tasks 2–3, performance is measured by cF1 (higher is better). On Task 2, our system obtains 0.4799 (Laptop) and 0.5799 (Restaurant), yielding an average of 0.5299 and outperforming both baselines on average (0.4672 and 0.4155). On Task 3, our system achieves 0.2781 (Laptop) and 0.5562 (Restaurant), with an average of 0.4171; the Restaurant domain shows a large gain over the strongest baseline (0.5562 vs. 0.3746).

According to the official leaderboard, our submission ranks 6th/5th on Task 1 (Laptop/Restaurant), 16th/14th on Task 2, and 11th/10th on Task 3 within the English track. These ranks are based on the leaderboard snapshot at the time of submission.

Overall, Table 1 reveals a consistent trend: extraction-based subtasks (Tasks 2–3) are more difficult than regression (Task 1), and the difficulty increases from triplet to quadruple extraction. The Laptop domain is particularly challenging for Task 3, motivating future improvements such as stronger constraint decoding, better category grounding, and domain-adaptive prompt designs.

Architecture / Module	Task 1 ($\text{RMSE}_{\text{VA}} \downarrow$)	Task 2 (cF1 $\uparrow$)	Task 3 (cF1 $\uparrow$)
Our Full System	1.2534	0.5299	0.4171
w/o MeanPooling (Use [CLS])	1.2725	–	–
w/o Huber Loss (Use MSE)	1.2679	–	–
w/o Span Recovery Algorithm	–	0.4305	0.3682

Table 2: Ablation Study Results on Validation Set.

6 Analysis

6.1 Ablation Studies

To assess the contribution of key components in our system, we conduct ablation experiments on the validation set. Table 2 summarizes the performance changes when individual modules are removed.

MeanPooling vs. [CLS] (Task 1). Replacing MeanPooling with the standard [CLS] representation increases the validation RMSE_{VA} from 1.2534 to 1.2725. This indicates that averaging token representations (masked by the attention mask) yields a more robust sentence–aspect representation for continuous VA regression than relying on a single classification token.

Loss function (Task 1). Training with MSE instead of Huber loss leads to worse RMSE_{VA} ($1.2534 \rightarrow 1.2679$), suggesting that Huber loss provides more stable optimization under outliers and extreme VA values.

Span recovery (Tasks 2–3). Removing the span recovery module degrades extraction performance on both Tasks 2 and 3. On Task 2, cF1 drops from 0.5299 to 0.4305, confirming that projecting generated mentions back to the exact surface forms in the original text is crucial under strict exact-match evaluation. On Task 3, performance also decreases ($0.4171 \rightarrow 0.3682$), suggesting that exact-span alignment becomes even more important when additional structural constraints (aspect, category, and opinion) must be satisfied simultaneously. Overall, these results demonstrate that span recovery is a key component for robust structured generation under exact-match evaluation.

6.2 Error Analysis

An in-depth manual analysis of our predictions revealed several persistent challenges. For **Task 1**, the regression model tends to under-predict extreme boundary values. Sentences with very high gold intensity (e.g., VA = 9.00) are often predicted closer to the mean (e.g., 8.50), suggesting a conservative mean-reversion bias induced by the training distribution and robust loss optimization.

For **Tasks 2–3**, the dominant errors come from (i) *structural and formatting deviations* in generation (e.g., minor bracket/label inconsistencies or missing separators), and (ii) *span-level mismatches* under strict exact-match evaluation. Although our post-processing and span recovery module mitigates many surface-form variations (e.g., casing/whitespace), span mismatches can still occur when multiple similar candidates appear in the sentence, especially for quadruple extraction. The ablation results (Table 2) indicate that span recovery provides a clear net benefit under the strict exact-match evaluation protocol.

In **Task 3**, a major source of false positives stems from *category confusion*. The model frequently struggles to differentiate semantically adjacent predefined categories, such as misclassifying FOOD#QUALITY as RESTAURANT#GENERAL when the sentence expresses broad praise without specific attributes. Additionally, *implicit aspects* (unstated but implied) can lead the model to generate plausible yet non-textual aspect terms, which are penalized by exact-match constraints.

7 Conclusion

This paper describes our system for SemEval-2026 Task 3 (DimABSA). We show that continuous VA prediction benefits from a discriminative encoder-based regressor with MeanPooling and Huber loss, while complex tuple extraction can be effectively addressed by instruction fine-tuning a quantized LLM with QLoRA combined with strict post-processing. Our analysis highlights that exact-span matching and category grounding remain key bottlenecks for generative extraction, motivating future work on more reliable constraint-aware decoding and span alignment strategies.

Limitations

Autoregressive LLM inference is computationally expensive and incurs higher latency than traditional

sequence-labeling approaches, which may limit deployment in real-time settings. Our extraction pipeline also relies on post-processing (regex parsing, span recovery, and numeric normalization) to satisfy strict evaluation requirements; errors in any stage can cascade, especially for quadruple extraction. Finally, the regression model is sensitive to the distribution of VA scores and may exhibit conservative predictions on rare extreme intensities.

Acknowledgments

We would like to thank the anonymous reviewers and the SemEval-2026 Task 3 organizers for their hard work in providing the DimABSA datasets and maintaining the evaluation platform.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, and Luo Ji. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, Jan Philip Wahle, Terry Ruas, Lung-Hao Lee, and Saif M. Mohammad. 2026. Dimstance: Multilingual datasets for dimensional stance analysis. *Preprint*, arXiv:2601.21483.

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.

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 (ACL)*, pages 340–350.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations (ICLR)*.

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

Peter J. Huber. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

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.

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 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8600–8613.

Maria Pontiki, Dimitrios 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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.