hdharpure at SemEval-2026 Task 3: BERT-Based Modeling and Prediction Behavior Analysis for Multilingual Valence–Arousal Scoring

Harshal Dharpure[1], **Nicolay Rusnachenko**[2]
[1] Indian Institute of Technology, Patna
[2] Centre for Applied Creative Technologies (CFACT+), Bournemouth University, UK
harshal_2511ai30@iitp.ac.in, nrusnachenko@bournemouth.ac.uk

Abstract

The SemEval-2026 Task 3 is a *Dimensional aspect-based sentiment analysis* (DimABSA) task which extends traditional ABSA by predicting continuous regression in two dimensions: valence (V) and arousal (A). The Track A/Subtask 1 represent multilingual task in which for a given text and aspects mentioned in it, there is a need to predict V/A scores for each aspect. Our approach is based on the pretraining-finetuning concept: we first pretrain multilingual model (M') followed by its fine-tuning ($M''_{l,d}$) on the training data of specific domain (d) and language (l). We adopt XLM-RoBERTa (M) as the encoder with separate regression heads for valence and arousal prediction. Our experiments on manual split of official SemEval-2026 Task 3 dataset ($D^{20\%}_{train}$) demonstrate that fine-tuning model in two stages ($M''_{l,d}$) results in average $\approx$ 1.36 times improvement by RMSE_{VA} over approach of direct fine-tuning ($M_{l,d}$). To investigate limitations of the existing approach we visualize and discuss limitations of our system. Our code is publicly available[1].

1 Introduction

Aspect-based sentiment analysis (ABSA) is a task of predicting sentiment for specific aspects mentioned in text (Liu, 2012). *Dimensional* Aspect-Based Sentiment Analysis (DimABSA) (Yu et al., 2026; Lee et al., 2026a) is the task introduced at SemEval-2026 as an extension of traditional ABSA with the prediction of continuous scores (Russell, 1980). The scores correspond to the following parameters: *valence* and *arousal*.

The Transformer architecture (Vaswani et al., 2017) caused a significant impact on the field of Natural Language Processing (NLP), including field of Sentiment analysis. Encoder-based language models features a robust framework for establishing the connection between texts and labels (Rogers et al., 2020). BERT and its derivative variants (Devlin et al., 2019; Liu et al., 2019) has become widely distributed in various tasks that required classification output.

The DimABSA dataset (D) and its training part particularly (D_{train}) shares a large amount of annotated data that unites texts in different languages: English (`eng`), Japanese (`jpn`), Russian (`rus`), Ukrainian (`ukr`), Tatar (`tat`), Chinese (`zho`). For each language, the data from the following set of domains presented: restaurants, laptop, finance. To maximise the use of cross-lingual and cross-domain data, in this paper we employ *pretraining-finetuning* (Wang and Qu, 2024) technique to adapt encoder-based transformers for V/A scoring. We choose XLM-RoBERTa (Liu et al., 2019) for experiments as it is one of the most widely used multilingual transformers. According to our experiments on manually arranged split of training data ($D^{20\%}_{train}$), we demonstrate that computing pretraining techniques with further fine-tuning result in $\approx$ 1.36 improvement. To study prediction behaviour of the result systems by visualizing alignments in two directions: (i) discrepancies of predicted results from *etalon* (ground-truth) annotations and (ii) discrepancies of etalon annotations from data distribution in predictions. By relying on the observed results on $D^{20\%}_{train}$, we discuss limitations of our approach and suggest further directions to address them.

2 Methodology

Given $C_{l,d}$ is a collection of texts for the particular language $l \in \mathbf{L}$ and domain $d \in \mathbf{D}$.

Task description. For the particular text $t \in C_{l,d}$ with set of aspects mentioned in it $\{\alpha_1, \alpha_2, ..., \alpha_n\}$, for each aspect α_i, there is a need to predict V/A pair (v'_i, a'_i) scores. The v'_i and a'_i are in the range $[1.0, 9.0]$.

[1] https://github.com/harshalDharpure/SemEval-2026-Task-3

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2228–2232

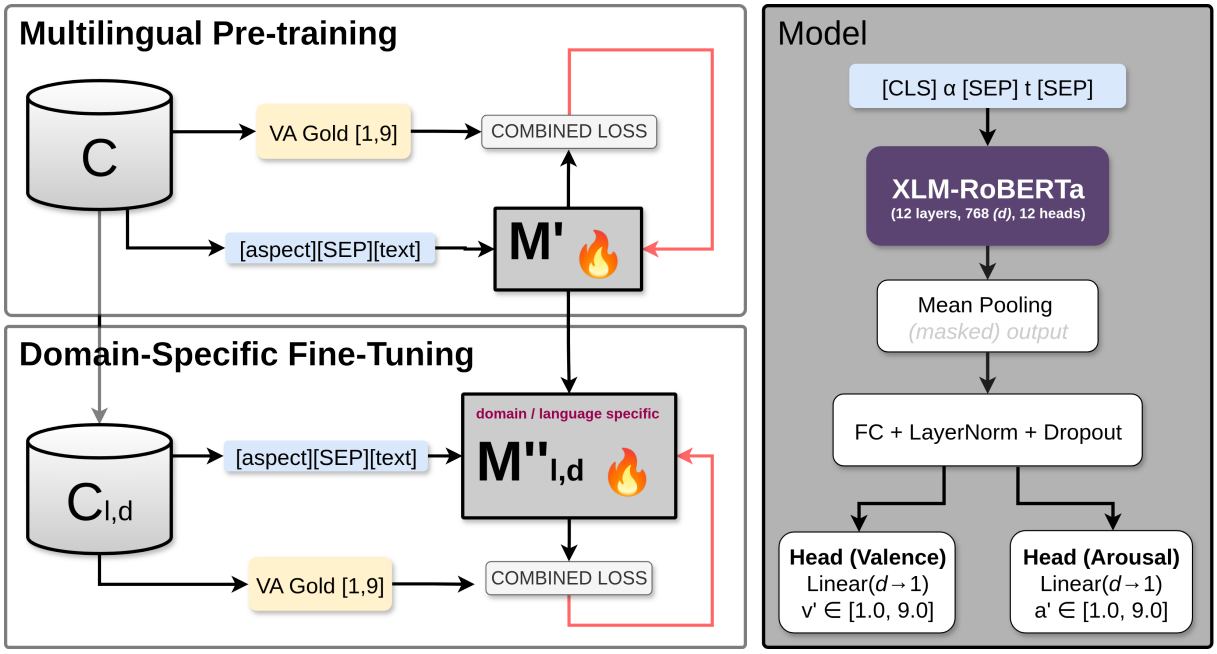

Figure 1: Pretraining-finetuning concept adaptation for model M (left), architecture of the model M (right)

Figure 1 (left) illustrates the pretraining and fine-tuning concept adaptation for model M. Given model M we compose its domain specific version (M''_{l_i,d_j}) by following steps:

1. **Pre-training**: we unite $C_{l,d}$ across all the languages $l \in \mathbf{L}$ and domains $d \in \mathbf{D}$ into single collection C to fine-tune M, with result model denoted as M'.
2. **Fine-tuning**: for each and particular language $l_i \in \mathbf{L}$ and domain $d_j \in \mathbf{D}$ we use M' to prepare domain specific version (M''_{l_i,d_j}).

In our methodology, M is a machine learning model that for a given input context outputs valence / arousal pair scores, with values cropped in accordance with task requirement. We use Mean Squared Error (MSE) loss for both valence and arousal prediction and compose COMBINED LOSS (see Figure 1) as follows:

$$\mathcal{L} = \text{MSE}(\hat{v}, v) + \text{MSE}(\hat{a}, a)$$

where v and a are ground truth valence and arousal scores, respectively.

Figure 1 (right) illustrates the architecture of the model M. We use BERT-based transformer (Liu et al., 2019), which supports input for two sentences separated by meta-token (`[SEP]`) and prefixed by token of class (`[CLS]`). For the given text t and particular aspect mentioned in it (α) we compose input context as follows:

`[CLS]` α `[SEP]` t `[SEP]`

To obtain V/A scores we use mean pooling over the input context token sequence, weighted by the attention mask to ignore padding. The pooled vector is passed through a fully connected layer (FC) with LayerNorm and dropout:

$$h_t = \text{LayerNorm}(\text{FC}(h_{pooled}))$$

and then through two heads:

$$\hat{v} = W_v \cdot h_t + b_v$$
$$\hat{a} = W_a \cdot h_t + b_a$$

where W_v and W_a are weight matrices for valence and arousal respectively; $\hat{v}$ and $\hat{a}$ are predicted valence and arousal scores, $\hat{v}, \hat{a} \in \mathbb{R}$. To meet the requirement of the task, valence and arousal scores additionally cropped to the range $[1.0, 9.0]$ as follows:

$$v = \max(1.0, \min(\hat{v}, 9.0))$$
$$a = \max(1.0, \min(\hat{a}, 9.0))$$

3 Dataset

The official dataset (DimABSA) (Lee et al., 2026b; Yu et al., 2026) and available amount of sentences in its training part (D_{train}) presented in Table 2. In this paper we construct splits of the D_{train} that correspond to:

- $D^{80\%}_{train}$ - dataset for pre-training and fine-tuning;

Model	eng			rus	ukr	jpn			tat	zho			
	Rest.	Laptop	Δ	Rest.	Rest.	Hotel	Fin.	Δ	Rest.	Rest.	Laptop	Fin.	Δ
$D^{20\%}_{train}$ (Non-official Split)													
Exp1	0.8123	0.8693	0.8408	0.8399	1.0074	0.4770	0.6078	0.5424	1.2861	0.4736	0.6084	**0.3299**	0.4706
Exp2	0.8841	0.7751	0.8175	1.0062	1.0685	0.4589	0.5887	0.5059	1.2505	0.4707	**0.6058**	0.4681	0.5190
Exp3	**0.5854**	**0.5755**	**0.5804**	**0.4699**	**0.5191**	**0.3867**	**0.3807**	**0.3837**	**0.9173**	**0.4316**	0.6166	0.3428	**0.4637**
D_{test} (Official Split)													
Exp3 (Ours)	**1.5003**	**1.5412**	**1.5208**	**1.6515**	**1.7172**	**0.8378**	**1.0292**	**0.9335**	2.0463	**0.9847**	**0.7902**	**0.5704**	**0.7818**
Kimi-K2$_{\text{thinking}}$	2.1461	2.1893	2.1677	1.7768	1.7805	1.7553	1.6396	1.6975	**1.9380**	1.8959	1.6440	1.9652	1.8350
Qwen-3 14B	2.6427	2.8089	2.7258	2.1528	2.2121	2.2906	1.8964	2.0935	2.6367	2.0073	1.7706	1.4707	1.7495

Table 1: Results on $D^{20\%}_{train}$ (Non-official Split) and D_{test} (Official Split) for the XLM-RoBERTa$_{\text{base}}$ model in various experiments set (Exp1, Exp2, Exp3); Δ denotes the average across all the domains of each language. results section for D_{test} includes comparison to Qwen-3 14B and Qwen-3 14B baseline results provided by competition organizers

Lang.	Domain	D_{train}	$D^{80\%}_{train}$	$D^{20\%}_{train}$	D_{test}
eng	Restaurant	2,284	1,827	457	1000
	Laptop	4,076	3,261	815	1000
jpn	Hotel	1,600	1,280	320	800
	Finance	1,024	819	205	800
rus	Restaurant	1,240	992	248	1072
ukr	Restaurant	1,240	992	248	1072
tat	Restaurant	1,240	992	248	1072
zho	Restaurant	6,050	4,840	1,210	1000
	Laptop	3,490	2,792	698	1000
	Finance	1,000	800	200	842
Total		**23,244**	**18,595**	**4,649**	**9,658**

Table 2: Statistics of the total amount of sentences in D_{train} and D_{test} parts of the official dataset (DimABSA), separately per each language and domain; $D^{80\%}_{train}$ and $D^{20\%}_{train}$ correspond to D_{train} split into train:test with 80/20 ratio

- $D^{20\%}_{train}$ - dataset for non-official experiments.

The splitting is performed at the sentence level with inclusion of corresponding aspects mentioned in it. We apply a random 80/20 split to construct $D^{80\%}_{train}$ and $D^{20\%}_{train}$; we keep all aspects of the same sentence together.

3.1 Experimental Setup

The full model (encoder and regression heads) is trained for 5 epochs with learning rate 1.5×10^{-5} with 16 bit float precision. For the pre-training we use 5 epochs. At the Fine-tuning trained with 3 epochs. We use XLM-RoBERTa$_{\text{base}}$ as the encoder (Liu et al., 2019) and its initial state further in this section refered as M.

For the described methodology in Section 2 we set up the following experiments that involve application of the original model M and its pre-trained version M' pre-trained on the $D^{80\%}_{train}$ (see Section 3). For the given language l_i and domain d_j we set up the following experiments:

- **Exp1**: Direct fine-tuning M on target language/domain (no pretraining)
- **Exp2**: Inference only M' (no fine-tuning)
- **Exp3**: Fine-tuning M' to obtain M''_{l_i,d_j} used to infer results.

For this task, organizers provided the following evaluation metric: RMSE$_{VA}$, which is defined as:

$$\text{RMSE}_{\text{VA}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(v'_i - v_i)^2 + (a'_i - a_i)^2\right]}$$

where n corresponds to a total number of pairs $\langle t, \alpha \rangle$, and v_i and a_i are ground truth valence and arousal scores, respectively. Lower values indicate better performance.

Hyperparameter	Value
Batch size	4
Max sequence length	192 tokens
Learning rate	1.5×10^{-5}
Dropout	0.1 (encoder), 0.3 (FC layer)
Optimizer	AdamW
Mixed precision	bfloat16
Early stopping	Patience 3 epochs (fine-tuning only)

Table 3: XLM-RoBERTa$_{\text{base}}$ training configuration

4 Result Analysis and Discussion

Table 1 provides results by RMSE$_{\text{VA}}$ for the experiments (Exp1, Exp2, Exp3), separately for $D^{20\%}_{train}$ and D_{test} and each language and domain individually.

According to the results observed on $D^{20\%}_{train}$ for the XLM-RoBERTa$_{\text{base}}$ (M), using adapting

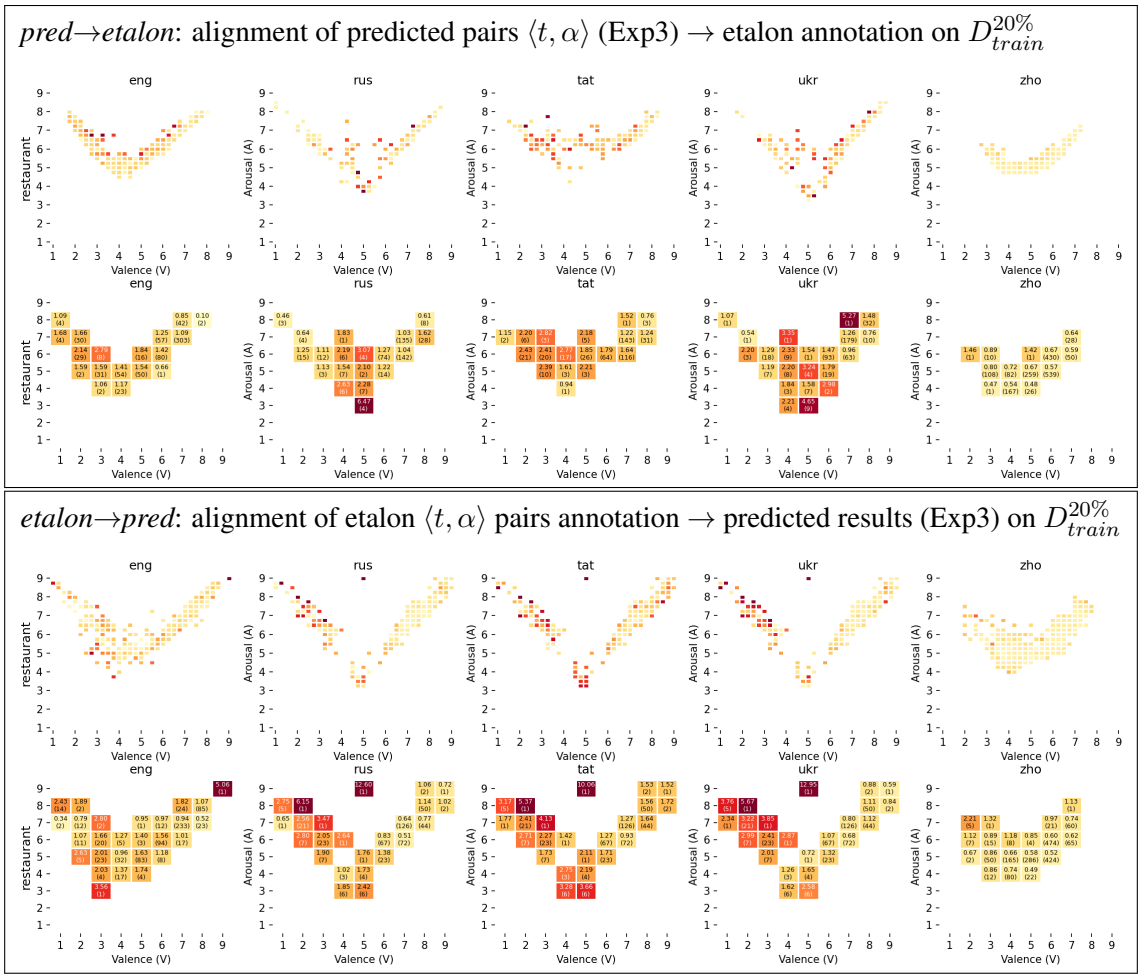

Figure 2: RMSE_{VA} results for M''_l models (Exp3) on $D^{20\%}_{train}$ for restaurant domain and coresponding languges $l \in \{\text{eng}, \text{rus}, \text{ukr}, \text{tat}, \text{zho}\}$, visulalizaed with bins of different sizes: 0.25 and 1.0; bar palette varies by RMSE_{VA} from 0 (light yellow) to 5 (dark red); *pred→etalon:* alignment of scores obtained from M''_{l_i} towards etalon scores; *etalon→pred:* alignment of etalon scores towards scores classified by M''_{l_i}.

pretraining-finetuning (Exp3) results in $\approx$ 1.36 times improvement[2] by RMSE_{VA} over results from approach Exp 1.

Our final submission for evaluation on the official split (D_{test}) corresponds to the results of the M''_{l_i,d_j} models (Exp3). For the comparison reason, we include results of baseline approaches shared by DimABSA organizers, obtained via Kimi-$\text{K2}_{\text{thinking}}$[3] and Qwen-3 14B[4]. We found that results of our final submission strongly outperform results of Qwen-3 14B and Kimi-$\text{K2}_{\text{thinking}}$ in $\approx$ 2.1 and $\approx$ 1.9 times by RMSE_{VA} respectively. The exception reprsent the subset of D_{test} for Tatar language (Restaurant domain), in which Kimi-$\text{K2}_{\text{thinking}}$ demonstate slightly superior performance to our approach ($\approx$ 1.06 times by RMSE_{VA}).

Result Analysis. For the predicted scores of the particular model (pred) and related etalon scores (etalon), we investigate the behavior by visualizing alignment of:

1. *pred→etalon* – alignment of scores obtained from M''_{l_i,d_j} towards etalon scores; this allow us to locate source of errors etalon dataset;
2. *etalon→pred* – alignment of etalon scores towards scores classified by M''_{l_i,d_j};

To make comparison of the related analysis consistent across different languages, we choose 'restaurants' domain as the most common domain (See Table 2).

Figure 2 presents the related analysis for M''_{l_i} models (Exp3) on restaurants domain from $D^{20\%}_{train}$

[2] ratio of results obtained in Exp1 to one in Exp3

[3] https://huggingface.co/moonshotai/Kimi-K2-Thinking

[4] https://huggingface.co/Qwen/Qwen3-14B

dataset. Given scores obtained from M''_{l_i} ($pred$), and ground-truth scores ($etalon$), we analyse the alignment of *pred* and *eval* in two directions: $pred \rightarrow etalon$, and $etalon \rightarrow pred$. According to the related analysis, we identified the following problems that impact the performance:

1. **cropped etalon space**: the etalon space has no information for arousal (a) scores in range $a \in [0, 2]$ (pred→etalon block in Figure 2);
2. **presence of outliers**: examples valence/arousal scores of $\langle 5.0, 9.0 \rangle$ across various languages (rus, tat, ukr) (etalon→pred block in Figure 2);
3. **formation of error clusters**; example: areas $v \in \langle 0, 3 \rangle, a \in \langle 6, 9 \rangle$ has relatively high RMSE (etalon→pred block in Figure 2);

To address these problems, we believe that the research around proper exploit of `[CLS]` token output of BERT-based is important. As a one approach, the Natural Language Inference (NLI) technique (Sun et al., 2019) to (i) supply prompt hint after separator token (`[SEP]`), and (ii) exploit output from `[CLS]` token in area classification and fact-checking. To provide numerical definitions with their lexical definitions, one approach we provisionally see is to re-center V/A space by defining larger areas to be referred as high / middle / low subareas of related V/A scores.

5 Conclusion

This paper presents a study on adapting encoder-based transformers in a novel task, dubbed as dimensional aspect-based sentiment analysis (DimABSA) presented at SemEval-2026 as Task 3. This task extends the traditional ABSA with the prediction of continuous scores for valence and arousal. DimABSA has a multilingual and multidomain setup, and to leverage the use of cross-lingual and cross-domain data, we this paper contribute by investigating effect of pre-training on the model performance. According to our experiments that involve application of XLM-RoBERTa$_{\text{base}}$ model, we demonstrate that pretraining-finetuning concept results in ≈ 1.36 improvement by RMSE$_{\text{VA}}$ over direct domain and language specific fine-tuning.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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. 2026a. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis.

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

Bing Liu. 2012. Sentiment analysis and opinion mining.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

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

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yu Wang and Wen Qu. 2024. A tutorial on the pretrain-finetune paradigm for natural language processing.

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.