

# NTNU-SMIL at SemEval-2026 Task 3: Logistic-Loss Regression with Same-Language Transfer for Valence–Arousal Stance Prediction in Dimensional Stance Analysis (DimStance)

Siang-Ting Lin, Tien-Hong Lo, Yun-Ting Sun, Jhih-Rong Guo, Tung-Yen Hao, Fong-Chun Tsai, and Berlin Chen

The Speech and Machine Intelligence Laboratory (SMIL)  
Department of Computer Science and Information Engineering  
National Taiwan Normal University, Taiwan

{61347114S, teinhonglo, 61347015s, jhihrong, 41247050, fongchun.tsai, berlin}@ntnu.edu.tw

## Abstract

Dimensional Stance Analysis (DimStance) requires aspect-conditioned regression of valence and arousal (VA) across languages and domains, where distribution shift often causes scale mismatch and variance shrinkage, especially for arousal. We describe NTNU-SMIL’s system for SemEval-2026 Task 3 Track B (DimStance) Subtask 1 (DimASR) (Yu et al., 2026; Becker et al., 2026). Our model uses sentence-pair encoding ([CLS] Text [SEP] Aspect [SEP]), dual VA heads, and a logistic-loss regression objective implemented by min–max normalizing labels to  $[0, 1]$  and mapping logits back to  $[1, 9]$  via a sigmoid. For English and Chinese, we further apply same-language transfer by pretraining on Track A and fine-tuning on Track B (Lee et al., 2026). We add lightweight out-of-fold linear calibration and multi-seed ensembling to reduce scale mismatch under shift (Guo et al., 2017; Kuleshov et al., 2018; Lakshminarayanan et al., 2017). Post-hoc analysis on the released test gold indicates that same-language transfer and logistic-loss regression account for most of the gains in English and Chinese, while calibration provides secondary adjustments; arousal variance collapse remains a key failure mode in lower-resource settings such as Swahili.

## 1 Introduction

Traditional stance detection is typically formulated as a categorical classification problem (favor/against/neutral), as in SemEval-2016 Task 6 (Mohammad et al., 2016). Subsequent work expanded stance modeling to target-conditioned and cross-lingual settings (Augenstein et al., 2016; Vamvas and Sennrich, 2020). However, most prior benchmarks remain label-discrete and do not capture fine-grained intensity variation.

Dimensional affect modeling instead represents affect using continuous variables such as valence and arousal (Mehrabian and Russell, 1974; Russell, 1980; Bradley and Lang, 1994). SemEval-2026 Task 3 introduces DimABSA, a multilingual, multi-domain benchmark for aspect-level VA prediction (Yu et al., 2026; Lee et al., 2026). DimStance further extends the dimensional paradigm from sentiment to stance, requiring aspect-conditioned VA regression across languages and domains (Becker et al., 2026). In Track B, Subtask 1 (DimASR) corresponds to target-conditioned valence–arousal regression for stance, where the stance target is treated as an aspect.

As illustrated in Figure 1, DimStance predicts aspect-specific VA scores and is highly sensitive to cross-lingual distribution shift. Even strong multilingual encoders can yield miscalibrated scales and under-dispersed predictions under cross-lingual transfer, which degrades both RMSE and correlation (Guo et al., 2017; Kuleshov et al., 2018; Ovidia et al., 2019). This issue is often most severe for arousal, especially in low-resource languages.

7.00#6.00 4.83#3.83  
 $i_1$   $i_2$   
S: Nuclear energy plays a role in mitigating climate change.  
 $a_1$   $a_2$

Subtask	Input	Output	Task Type
VA Prediction	$S + a_1$ $S + a_2$	$\{i_1\}$ $\{i_2\}$	Regression

Figure 1: Illustration of DimStance Subtask 1 (DimASR) as aspect-conditioned valence–arousal regression. Given a text  $S$  and a target/aspect  $a_j$ , the system predicts a VA pair  $i_j = (V_j, A_j)$ .

In this paper, we focus on English and Chinese, the two languages for which Track A provides same-language cross-domain supervision

(Lee et al., 2026). Our system combines sentence-pair encoding, a final-layer [CLS] representation, dual regression heads, logistic-loss regression, same-language transfer, and lightweight calibration-aware ensembling; post-hoc analysis suggests that same-language transfer and logistic-loss regression account for most of the gains.

## 2 Related Work

### 2.1 Stance Detection and Aspect/Target Conditioning

Early stance detection benchmarks mainly cast the task as categorical classification, with SemEval-2016 Task 6 as a representative shared task (Mohammad et al., 2016). Target-conditioned encoding has been shown to be effective for stance prediction when target/aspect information is explicitly provided (Augenstein et al., 2016). Recent benchmarks further emphasize multilingual and cross-target stance modeling (Vamvas and Senrich, 2020).

### 2.2 Dimensional Affect Modeling and Dimensional ABSA/Stance

Dimensional affect modeling represents affect using continuous variables such as valence and arousal (Mehrabian and Russell, 1974; Russell, 1980; Bradley and Lang, 1994). DimABSA provides multilingual, multi-domain aspect-level VA annotations for dimensional sentiment learning (Yu et al., 2026; Lee et al., 2026). DimStance extends the dimensional paradigm from sentiment to stance by requiring aspect-conditioned VA regression across languages and domains (Becker et al., 2026).

### 2.3 Calibration and Ensembling for Robust Regression

Calibration has been widely studied for modern neural networks (Guo et al., 2017). For regression, calibrated approaches are known to improve reliability under distribution shift (Kuleshov et al., 2018). Ensembling is a strong and simple strategy for improving generalization and predictive stability (Lakshminarayanan et al., 2017), and uncertainty under shift has been empirically analyzed for deep models (Ovadia et al., 2019). Motivated by these observations, our system combines aspect-conditioned encoding, dimension-specific heads, same-language cross-domain pretraining for En-

glish and Chinese, and lightweight post-hoc calibration for robust VA regression.

## 3 Methods

### 3.1 Overall Architecture

Our approach follows a two-stage training strategy: (1) pretraining on Track A for aspect-aware dimensional regression (English and Chinese only), and (2) fine-tuning on Track B with dual-head regression and a logistic-loss regression formulation.

Figure 2 illustrates the overall architecture of our system.

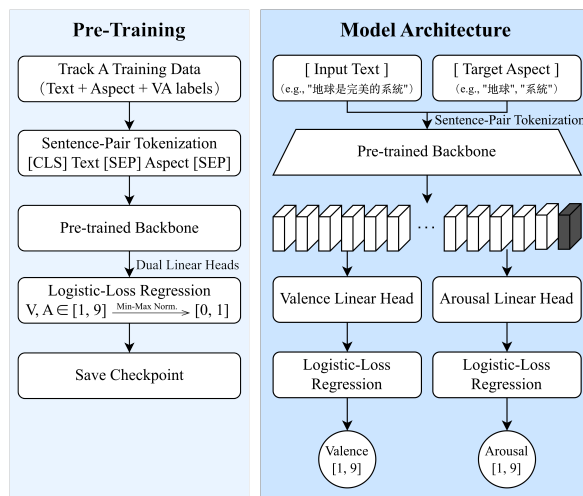


Figure 2: Overall architecture of our system. Left: same-language pretraining on Track A for English and Chinese only. Right: fine-tuning on Track B with sentence-pair encoding and two regression heads for valence and arousal.

**Sentence-Pair Encoding.** For both training stages, we construct sentence-pair inputs in the form [CLS] Text [SEP] Aspect [SEP]. This formulation allows the backbone encoder to model text–aspect interactions implicitly through self-attention over the concatenated sequence.

**Backbone Encoder.** We employ a pre-trained transformer backbone, with the exact model chosen per language (Appendix A.4). The encoder produces contextualized token representations for the joint text–aspect sequence. We use the final-layer [CLS] token embedding as the aggregated aspect-aware representation for downstream valence and arousal prediction.

**Dual-Head Regression Design.** Valence and arousal are modeled using two independent linear

heads. Let  $\mathbf{h}$  denote the final hidden representation. The predictions are computed as:

$$z_V = \mathbf{w}_v^\top \mathbf{h} + b_v, \quad z_A = \mathbf{w}_a^\top \mathbf{h} + b_a.$$

In our implementation, these head outputs are treated as logits and are mapped back to the original score range via a sigmoid transformation.

**Logistic-Loss Regression Formulation.** Instead of directly optimizing mean squared error, we apply min-max normalization to the gold scores:

$$y' = \frac{y - 1}{8}, \quad y \in [1, 9], \quad y' \in [0, 1].$$

We then optimize the regression objective using binary cross-entropy on continuous soft targets:

$$\mathcal{L}_{\text{BCE}} = - (y' \log \sigma(z) + (1 - y') \log(1 - \sigma(z))),$$

where  $\sigma(\cdot)$  denotes the sigmoid function. Under this formulation, the model learns logits whose sigmoid outputs correspond to normalized continuous scores rather than discrete class probabilities.

**Two-Stage Training Strategy (English and Chinese).** In Stage 1, the model is pretrained on Track A data (English and Chinese only) to learn aspect-aware dimensional representations (Lee et al., 2026). In Stage 2, the model is fine-tuned on Track B data with the same regression formulation.

**OOF Calibration and Ensembling.** To reduce scale mismatch under distribution shift, we fit a lightweight linear calibration model on out-of-fold (OOF) predictions from 5-fold training splits, and apply it to the corresponding test predictions (Kuleshov et al., 2018). Final predictions are obtained by averaging calibrated outputs across folds and random seeds (Lakshminarayanan et al., 2017).

## 4 Experiments

### 4.1 Experimental Settings

**Datasets** Table 1 summarizes the dataset statistics used in our experiments.

**Data Usage and Splitting Strategy.** Although we officially participate only in Track B (DimStance), we leverage Track A (DimABSA) Subtask 1 data for same-language supervised pretraining for English and Chinese (Yu et al., 2026; Lee et al., 2026). For Track B, we perform 5-fold cross-validation on training data with sentence-level grouping to avoid leakage across aspects.

Track	Language	Train Inst.	Dev Inst.
A	ENG	9,432	615
A	ZHO	17,658	1,806
B	ENG	2,059	339
B	DEU	1,335	75
B	ZHO	1,091	149
B	PCM	1,118	122
B	SWA	1,622	145

Table 1: Dataset statistics used in our experiments. Track A training data are used only for same-language pretraining (English and Chinese).

Out-of-fold (OOF) predictions from 5-fold training are used for linear calibration and ensembling. Calibration is performed exclusively on Track B training data, and no test labels are accessed during training, calibration, or model selection. The full protocol is summarized in Table 7 in Appendix A.3.

**Evaluation Metrics.** For Track B Subtask 1 (DimASR), we follow the official evaluation metric defined in the DimStance dataset paper (Becker et al., 2026). System performance is measured using Root Mean Square Error (RMSE) in the valence-arousal (VA) space:

$$\text{RMSE}_{VA} = \sqrt{\frac{\sum_{i=1}^N \left( V_p^{(i)} - V_g^{(i)} \right)^2 + \left( A_p^{(i)} - A_g^{(i)} \right)^2}{N}}. \quad (1)$$

where  $N$  is the total number of instances;  $V_p^{(i)}$  and  $A_p^{(i)}$  denote the predicted valence and arousal values for instance  $i$ , respectively; and  $V_g^{(i)}$  and  $A_g^{(i)}$  denote the corresponding gold valence and arousal values.

We also report Pearson correlation coefficients (PCC) for valence and arousal.

**Training Objective.** We model valence and arousal with two independent logits  $z_V$  and  $z_A$ . Gold scores  $y \in [1, 9]$  are min-max normalized as  $\tilde{y} = (y - 1)/8 \in [0, 1]$ . Predictions are mapped back to the original scale via  $\hat{y} = 8\sigma(z) + 1$ . Training uses binary cross-entropy on the normalized soft targets, as described in Section 3.1.

**Hyperparameters.** All hyperparameter settings are listed in Table 9 in Appendix A.5.

## 5 Results

Table 3 presents the results of our official final submission on the Track B test set, evaluated using

Category	System	ENG	DEU	ZHO	PCM	SWA	AVG
PLM Regressors	RemBERT	1.993	2.262	0.795	2.149	2.322	1.904
	LaBSE	2.288	1.918	0.715	1.720	2.103	1.749
	XLM-R	2.005	1.667	0.747	1.732	2.054	1.641
	mBERT <sup>†</sup>	2.699	2.329	1.276	3.215	2.784	2.461
Closed LLMs (Prompting, 16-shot)	GPT-5 mini	1.819	1.754	1.492	1.389	2.321	1.755
	Gemini-2.5 Flash	1.771	1.701	1.284	1.353	2.259	1.674
	Kimi K2	1.643	1.671	1.044	1.284	2.242	1.577
LLM Regressors (Fine-tuning)	Mistral-3 14B <sup>†</sup>	1.643	1.591	0.740	1.739	2.299	1.602
	Phi-4 14B	1.575	1.535	0.691	1.409	2.055	1.453
	Qwen-2.5 72B	1.520	1.493	0.681	1.228	2.012	1.387
	Llama-3.3 70B	<b>1.468</b>	1.415	0.679	<b>1.157</b>	<b>1.859</b>	<b>1.316</b>
PLM Regressors	<b>Ours (NTNU-SMIL)</b>	1.521	<b>1.347</b>	<b>0.556</b>	1.567	1.960	1.390

Table 2: Comparison with baseline and reference systems on the official test set ( $RMSE_{VA}$ ; lower is better). <sup>†</sup>mBERT and Mistral-3 14B are official baselines. Bold indicates the best score for each language and the overall average within this table.

the official evaluation script.  $RMSE_{VA}$  is the ranking metric, and  $PCC_V$  and  $PCC_A$  measure linear association.

Our system performs best on the Chinese test set (ZHO;  $RMSE_{VA}=0.5561$ ), whereas the Swahili test set (SWA) shows the largest degradation, particularly for arousal ( $PCC_A=-0.0088$ ).

Lang	Dom	$RMSE_{VA} \downarrow$	$PCC_V \uparrow$	$PCC_A \uparrow$
ENG	Env	1.5207	0.8016	0.4778
DEU	Pol	1.3467	0.8736	0.6339
ZHO	Env	<b>0.5561</b>	<b>0.9127</b>	<b>0.6592</b>
PCM	Pol	1.5674	0.8745	0.6175
SWA	Pol	1.9602	0.6763	-0.0088

Table 3: Official test-set results for Track B Subtask 1 (DimASR), evaluated using the official  $RMSE_{VA}$  metric.  $RMSE_{VA}$  is the ranking metric (lower is better).

Table 2 compares our approach with official baselines and other reference results reported in the shared task materials (Becker et al., 2026). Although large LLM regressors achieve lower average  $RMSE_{VA}$ , our PLM-based system remains competitive, especially on DEU and ZHO, highlighting the effectiveness of task-specific regression design and calibration-aware ensembling.

## 6 Analysis

### 6.1 Same-Language Cross-Domain Transfer (English and Chinese)

To examine whether same-language cross-domain knowledge helps DimStance, we compare two official submission settings: (i) **Direct FT**, trained only on Track B, and (ii) **A→B FT**, pretrained on

Track A and then fine-tuned on Track B. Table 4 reports the corresponding official test-set results and shows consistent gains from Track A pretraining for both English and Chinese (Lee et al., 2026).

Lang	Setting	$RMSE_{VA} \downarrow$	$PCC_V \uparrow$	$PCC_A \uparrow$
ENG	Direct FT	1.6163	0.7683	0.4447
	A→B FT	<b>1.5207</b>	<b>0.8016</b>	<b>0.4778</b>
ZHO	Direct FT	0.6392	0.8732	0.5360
	A→B FT	<b>0.5561</b>	<b>0.9127</b>	<b>0.6592</b>

Table 4: Same-language cross-domain transfer (Track A → Track B) for English and Chinese. **Direct FT** denotes training on Track B only. **A→B FT** denotes Track A pretraining followed by Track B fine-tuning. Both settings correspond to official submissions evaluated on the official test set.

### 6.2 Loss and Calibration Ablations (Post-hoc)

Using released test gold, we run controlled post-hoc experiments while varying (i) the regression loss (MSE vs. logistic-loss regression) and (ii) whether OOF linear calibration is applied (Kuleshov et al., 2018; Guo et al., 2017). Table 5 shows that regression loss choice substantially affects stability and cross-lingual robustness.

### 6.3 Distribution Shift Analysis

To better understand the performance gap across languages, we analyze prediction behavior across languages. Figure 3 visualizes gold versus predicted VA scores for ZHO and SWA on the released test gold.

As shown in Figure 3, predictions for ZHO closely follow the diagonal reference line ( $y = x$ ),

Setting (scored on released test gold)	ENG	DEU	ZHO	PCM	SWA
MSE + 5-fold CV	1.7095	<b>1.3555</b>	0.6923	1.4845	1.8940
MSE + 5-fold CV + OOF calib	1.7095	1.3569	0.6923	1.4845	<b>1.8928</b>
BCE + 5-fold CV + OOF calibration	<b>1.6163</b>	1.3762	<b>0.6392</b>	1.5406	2.0627

Table 5: Post-hoc ablations on the released test gold ( $RMSE_{VA}$ ; lower is better), scored with the official evaluation script. OOF calibration is fitted on Track B training predictions only and then applied to test predictions.

indicating good calibration and dispersion. In contrast, SWA predictions collapse into a narrow range around the mean, especially for arousal. The fitted regression line becomes nearly flat and the correlation is close to zero, suggesting severe variance shrinkage under cross-lingual distribution shift.

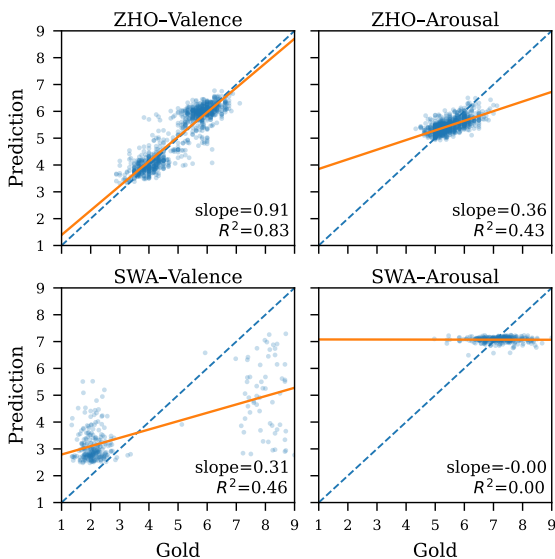


Figure 3: Gold vs. predicted scores for ZHO and SWA on the released test gold. The dashed line denotes the identity line ( $y = x$ ), and the solid line shows a least-squares linear fit. Predictions for ZHO are well aligned with the gold scores, whereas SWA exhibits pronounced variance shrinkage, particularly for arousal.

Our model achieves the strongest performance on ZHO among the five languages. One possible explanation is that the ZHO subset exhibits a relatively compact VA distribution, with most instances concentrated in a narrower middle region of the score space. Such a distribution may make the regression mapping easier to learn. Moreover, ZHO benefits from same-language Track A  $\rightarrow$  Track B transfer, which provides additional in-language supervision.

By contrast, SWA lacks such auxiliary supervision and shows substantially weaker generalization. This observation is consistent with the dataset-level analysis in the DimStance paper. Their Figure 2

shows that although all languages broadly follow a U-shaped valence–arousal relationship, the spread and shape of the VA space vary considerably across languages and domains (Becker et al., 2026). Taken together, these results suggest that under lower-resource conditions and stronger distribution shift, the model tends to produce under-dispersed predictions and loses sensitivity to arousal variation.

## 7 Conclusion

We presented NTNU-SMIL’s system for SemEval-2026 Task 3 Track B (DimStance) Subtask 1 (DimASR) (Yu et al., 2026; Becker et al., 2026). Our approach integrates sentence-pair aspect conditioning, a final-layer [CLS] representation, dual VA heads, logistic-loss regression, same-language Track A  $\rightarrow$  Track B transfer, and lightweight calibration-aware ensembling.

Results and post-hoc analyses show that same-language cross-domain transfer and logistic-loss regression are the primary drivers of robust VA prediction when additional supervision is available.

Loss formulation has a stronger impact on stability than post-hoc calibration, while distribution-shift analysis reveals arousal variance collapse as a key failure mode in low-resource settings (SWA).

Future work will explore variance-preserving objectives and distribution-robust calibration strategies to mitigate cross-lingual degradation.

## Limitations

Our system relies on Track A DimABSA data for same-language pretraining, which is available only for English and Chinese (Lee et al., 2026). Arousal prediction also remains vulnerable to variance shrinkage in lower-resource languages.

## Acknowledgments

This work was conducted at the Speech and Machine Intelligence Laboratory (SMIL), National Taiwan Normal University. We thank the SemEval-2026 Task 3 organizers and SMIL members for their support and discussions.

## References

- Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. [Stance detection with bidirectional conditional encoding](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 876–885. Association for Computational Linguistics.
- Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, Jan Philip Wahle, Terry Ruas, Idris Abdulmumin, Lung-Hao Lee, Nelson Odhiambo, Lilian Wanzare, Wen-Ni Liu, Tzu-Mi Lin, Zhe-Yu Xu, Ying-Lung Lin, Jin Wang, Maryam Ibrahim Mukhtar, Bela Gipp, and Saif M. Mohammad. 2026. [Dimstance: Multilingual datasets for dimensional stance analysis](#). *Preprint*, arXiv:2601.21483.
- Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330. PMLR.
- Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. [Accurate uncertainties for deep learning using calibrated regression](#). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2796–2804. PMLR.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. [Simple and scalable predictive uncertainty estimation using deep ensembles](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Lung-Hao Lee, Liang-Chih Yu, Natalia Loukashchev, 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.
- Albert Mehrabian and James A. Russell. 1974. *An Approach to Environmental Psychology*. MIT Press, Cambridge, MA.
- Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. [Semeval-2016 task 6: Detecting stance in tweets](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. [Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- Jannis Vamvas and Rico Sennrich. 2020. [X-stance: A multilingual multi-target dataset for stance detection](#). *Preprint*, arXiv:2003.08385.
- 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.

## A Appendix

### A.1 Abbreviations

Table 6 summarizes abbreviations used in the paper.

Notation	Abbrev.	Meaning
Language ID ( <b>Lang</b> )	ENG / DEU / ZHO / PCM / SWA	English / German / Chinese / Nigerian Pidgin / Swahili
Domain ( <b>Dom</b> )	Env. / Pol.	Environmental Protection / Politics

Table 6: Abbreviations used throughout the paper. **Lang** denotes language identifiers, and **Dom** denotes domains in Track B Subtask 1.

### A.2 Pearson Correlation Coefficient

In addition to  $\text{RMSE}_{VA}$ , we report Pearson correlation coefficients (PCC) for valence and arousal:

$$\text{PCC}_V = \text{corr}\left(\{V_p^{(i)}\}_{i=1}^N, \{V_g^{(i)}\}_{i=1}^N\right), \quad (2)$$

$$\text{PCC}_A = \text{corr}\left(\{A_p^{(i)}\}_{i=1}^N, \{A_g^{(i)}\}_{i=1}^N\right). \quad (3)$$

### A.3 Training and Evaluation Protocol

Stage	Data Usage
Stage 1	Track A train (ENG/ZHO only)
Stage 2 Train	Track B train (5-fold group-based CV)
Calibration	OOF predictions on Track B train
Model Selection	Track B dev RMSE
Final Evaluation	Official hidden test set

Table 7: Training and evaluation protocol.

### A.4 Backbone Models

Lang	Dom	Backbone (Hugging Face)
ENG	Env.	microsoft/deberta-v3-large
DEU	Pol.	deepset/gbert-large
ZHO	Env.	hfl/chinese-roberta-wwm-ext-large
PCM	Pol.	Davlan/afro-xlmr-large
SWA	Pol.	Davlan/afro-xlmr-large

Table 8: Language-domain specific backbone models used in our system.

### A.5 Hyperparameter Settings

Hyperparameter	Value
Random seeds	{42, 52, 62, 72, 82}
# CV folds ( $K$ )	5
Early stopping patience	3
Batch size	8
Gradient accumulation steps	4
Learning rate	$1.5 \times 10^{-5}$
Weight decay	0.01
Dropout	0.0
Max epochs	20
Fixed folds	Yes

Table 9: Hyperparameter settings used in our experiments.