CLRG at SemEval-2026 Task 3: One Size Does Not Fit All: A Resource Adaptive Framework for Dimensional Sentiment Regression

Wardat Shams Iqbal[1], Ruwad Naswan[2], Swakkhar Shatabda[1]

[1]BRAC University, Dhaka, Bangladesh
[2]Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh
wardat.shams.iqbal@g.bracu.ac.bd, ruwad45678@gmail.com, swakkhar.shatabda@bracu.ac.bd

Abstract

Predicting continuous Valence and Arousal scores across diverse languages poses significant challenges due to typological differences and the difficulty of modeling affective intensity. We introduce AdaptStance, a parameter-efficient framework designed for the SemEval-2026 Task 3 benchmark. To address cross-lingual disparities, AdaptStance routes inputs through resource-specific pipelines: direct regression with a hybrid concordance loss for high-resource languages, and an auxiliary multi-task mechanism to stabilize regression in low-resource and non-Western contexts. Architectural analysis reveals that decoupling task heads benefits morphologically related languages, whereas joint representations act as crucial regularizers for distant language families. Ultimately, this lightweight approach achieves competitive performance over generative baselines, demonstrating the efficacy of targeted architectural alignment while identifying Valence as the primary bottleneck in continuous affect prediction. Our code is available on GitHub.[1]

1 Introduction

While the proliferation of Large Language Models (LLMs) has enabled the extraction of public opinion at scale, traditional sentiment analysis remains constrained by categorical classification. Labeling a stance strictly as "favor" or "against" oversimplifies semantic reality and fails to capture the continuous spectrum of human emotion.

Modeling dimensional affect introduces further challenges, particularly in cross-lingual settings. While standard models easily detect sentiment polarity (Valence), they consistently struggle to quantify its intensity (Arousal). Furthermore, the reliance on massive, English-centric LLMs exacerbates this disparity for languages like Nigerian Pidgin and Swahili, where data scarcity renders standard pre-training ineffective.

SemEval-2026 Task 3: Dimensional Stance Analysis (DimStance) addresses these limitations by shifting the objective to the regression-based prediction of continuous Valence and Arousal scores (Becker et al., 2026; Yu et al., 2026).

In response, we propose **AdaptStance**, a parameter-efficient BERT based, dual-pipeline framework that dynamically adapts to linguistic resource availability. Recognizing that "one size does not fit all," we orchestrate two distinct methodologies. For high-resource languages (English, German), we employ direct regression using Attention Pooling and a custom hybrid log-loss function. Conversely, for non-Western languages (Nigerian Pidgin, Swahili, Chinese), we introduce an **Auxiliary Multi-Task Head**. This forces the simultaneous learning of coarse-grained classification and fine-grained regression, acting as a critical regularizer that stabilizes Arousal prediction where standard encoders fail to converge. Our system achieves highly competitive performance, securing **Rank 5** in Chinese (ZHO) and **Rank 7** in Swahili (SWA), outperforming massive generative baselines (e.g., Mistral-314B) with a fraction of the parameters.

2 Related Work

2.1 From Categorical Stance to Dimensional Analysis

Stance detection has traditionally been formulated as a categorical classification task, where the objective is to determine a user's position (e.g., Favor, Against, or None) toward a specific target (Mohammad et al., 2016; Schiller et al., 2020). While effective for broad categorization, the discrete approach fails to capture the intensity of the sentiment or the degree of emotional activation involved in the stance.

To address this limitation, recent research has integrated affective science theories, specifically the

[1]https://github.com/pineconedad/AdaptStance

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3225–3231
July 3-4, 2026 ©2026 Association for Computational Linguistics

Circumplex Model of Affect (Russell, 1980), which maps emotion onto continuous dimensions of Valence (negativity-positivity) and Arousal (calm-excited) . This shift has given rise to Dimensional Aspect-Based Sentiment Analysis (DimABSA), where stance targets are treated analogously to aspects (Lee et al., 2026). The recently introduced Stance-as-DimABSA framework (Becker et al., 2026) formalizes this by requiring models to predict real-valued scores for both valence and arousal. However, existing work often focuses on high-resource languages or treats valence and arousal as equally learnable tasks, often obscuring the specific challenges inherent in modeling arousal from text alone.

2.2 Pre-trained Encoders in Low-Resource Settings

The dominant paradigm for multilingual NLP relies heavily on massive pre-trained language models (PLMs) like XLM-R (Conneau et al., 2020) or generative LLMs. While effective, these models are computationally expensive and often struggle with precise regression tasks without extensive instruction tuning. In contrast, our work revisits lightweight, encoder-only architectures (Devlin et al., 2019). We build on prior findings that domain-adaptive pre-training significantly boosts downstream performance (Gururangan et al., 2020), demonstrating that a well-aligned, smaller encoder generalizes more effectively in low-resource dimensional regression than generic, massive baselines.

2.3 The Valence-Arousal Discrepancy

Standard evaluations for dimensional sentiment, including the DimABSA shared task (Lee et al., 2026), typically report performance by averaging the Root Mean Square Error (RMSE) of Valence and Arousal into a single metric. This aggregation masks the distinct difficulty of modeling affective intensity. Unlike prior studies, our framework explicitly investigates this gap, treating the severe performance discrepancy between Valence and Arousal as a core distributional bottleneck rather than a simple capacity issue.

3 System Overview

Our system, illustrated in Figure 1, employs a Resource-Adaptive Framework designed to address the linguistic disparity between high-resource and low-resource languages. We hypothesize that while high-resource languages (English, German) benefit from direct regression optimization on large-scale pre-trained models, low-resource and non-Western languages (Pidgin, Swahili, Chinese) require auxiliary structural guidance to stabilize training. We note that Chinese, while not low-resource in the strict sense, is grouped with the non-Western track due to its distinct linguistic typology and the architectural needs of its dedicated encoders; its strong performance under Pipeline B is examined further in Section 5.

Consequently, our framework routes inputs through two distinct architectural pipelines: Pipeline A (Regularized Direct Regression) for Western/High-Resource languages, and Pipeline B (Decoupled Multi-Task Learning) for Non-Western/Low-Resource languages.

3.1 Problem Formulation

Formally, given an input text sequence $T = \{t_1, t_2, \ldots, t_n\}$ and a target aspect term a, the goal is to predict a continuous bi-dimensional sentiment score $Y = (v, \alpha)$, where v represents Valence and α represents Arousal. Both target variables are continuous values normalized to the range $[0, 1]$ (or $[1, 9]$ depending on the specific dataset scale). We model this as a conditional probability regression problem $P(Y \mid T, a; \theta)$, where θ represents the model parameters.

3.2 Feature Extraction and Pooling

Across both pipelines, we utilize a Transformer-based encoder to generate contextual embeddings. Unlike standard approaches that rely solely on the `[CLS]` token, we employ Attention Pooling to aggregate the sequence representation dynamically. Let $\mathbf{H} \in \mathbb{R}^{L \times d}$ be the output hidden states of the last transformer layer, where L is the sequence length and d is the hidden dimension. We compute a learnable weight vector $\mathbf{w}_{\text{att}}$ to derive the final sentence embedding $\mathbf{e}$:

$$\beta_i = \text{softmax}(\mathbf{H}_i \cdot \mathbf{w}_{\text{att}}^{\top}) \tag{1}$$

$$\mathbf{e} = \sum_{i=1}^{L} \beta_i \mathbf{H}_i \tag{2}$$

3.3 Pipeline A: Regularized Direct Regression (ENG, DEU)

For high-resource languages, we employ a streamlined architecture focused on optimizing the correlation metric directly.

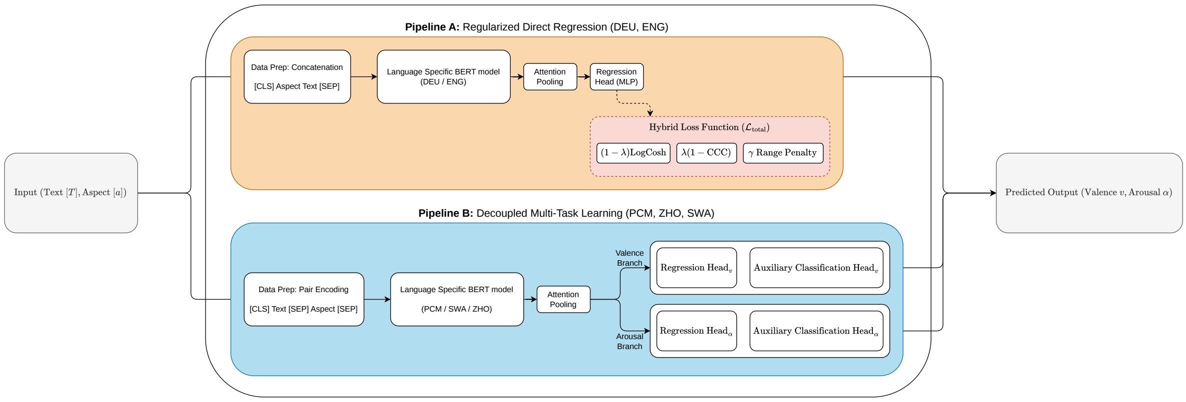

Figure 1: Adaptive System Architecture

Input Representation We treat the aspect and text as a unified semantic unit by concatenating them without special separator tokens, allowing the self-attention mechanism to flow naturally across the boundary: `[CLS] Aspect Text [SEP]`.

Hybrid Loss Function To mitigate the sensitivity of Mean Squared Error (MSE) to outliers and to explicitly encourage agreement-based correlation through the Concordance Correlation Coefficient (CCC) as a training objective, we propose a compound loss function $\mathcal{L}_{\text{total}}$:

$$\mathcal{L}_{\text{total}} = (1-\lambda)\mathcal{L}_{\text{LogCosh}} + \lambda\mathcal{L}_{\text{CCC}} + \gamma\mathcal{L}_{\text{Range}} \tag{3}$$

Here, $\mathcal{L}_{\text{LogCosh}}$ approximates the Huber loss to reduce outlier penalties, $\mathcal{L}_{\text{CCC}} = 1 - \rho_c$ maximizes the concordance correlation, and $\mathcal{L}_{\text{Range}}$ is a penalty term that regularizes predictions to stay within valid bounds. This pipeline utilizes BERT-Large and BERT-Base-German (see Table 1) to leverage their robust pre-trained representations.

3.4 Pipeline B: Decoupled Multi-Task Learning (PCM, ZHO, SWA)

For low-resource languages, we observed that direct regression frequently fails to converge or suffers from high variance, particularly in the Arousal dimension. To address this, we introduce Auxiliary Classification Heads to guide the encoder.

Input Representation We utilize standard pair encoding to explicitly distinguish the aspect from the context: `[CLS] Text [SEP] Aspect [SEP]`. The two pipelines were developed independently and adopted different input formats; we retained both because each performed well on its respective pipeline during validation. Pipeline A's concatenation allows self-attention to flow uninterrupted across the aspect-text boundary, complementing the attention-pooling aggregation, while Pipeline B's pair encoding produces explicit segment embeddings that help the auxiliary classification heads localize aspect-specific signal.

Decoupled Auxiliary Heads Early experiments revealed a "bottleneck" where a joint head for Valence and Arousal caused negative transfer, particularly degrading Arousal performance ($\text{RMSE}_{\mathcal{A}}$). In our final system, we decouple the tasks completely. The shared sentence embedding $\mathbf{e}$ is fed into two separate projection layers, each branching into a regression head (primary) and a classification head (auxiliary). The auxiliary heads predict coarse-grained labels (e.g., Low, Neutral, High for Arousal; Against, Neutral, Favor for Valence) using Cross-Entropy loss ($\mathcal{L}_{\text{CE}}$). The final loss is a weighted multi-task objective:

$$\mathcal{L}_B = \alpha(\mathcal{L}^v_{\text{CE}} + \mathcal{L}^\alpha_{\text{CE}}) + \beta(w_v\mathcal{L}^v_{\text{MSE}} + w_\alpha\mathcal{L}^\alpha_{\text{MSE}}) \tag{4}$$

This formulation forces the model to learn discriminative features for Valence and Arousal independently before fusing them for the final continuous prediction.

3.5 Model Specifications

To implement our Resource-Adaptive Framework, we selected pre-trained checkpoints tailored to the linguistic profiles of each target language. Table 1 details the mapping of languages to their respective pipelines and base encoders. For the high-resource track (Pipeline A), we leveraged high-capacity monolingual models to maximize representation quality. Conversely, for the non-Western track (Pipeline B), we prioritized language-specific or dialect-aware encoders. Chinese benefits from a mature pre-training ecosystem, while Pidgin and Swahili rely on more recent dialect-adapted encoders.

Lang.	Pipeline	Base Model Checkpoint
Western / High-Resource		
ENG	A	`bert-large-uncased`
DEU	A	`bert-base-german-cased`
Non-Western / Low-Resource		
PCM	B	`Davlan/bert-base-...-naija`
SWA	B	`Davlan/bert-base-...-swahili`
ZHO	B	`bert-base-chinese`

Table 1: Mapping of target languages to their respective pipelines and base models.

4 Experimental Setup

We evaluate our system on the DimABSA 2026 shared task dataset for Track B, Subtask 1, which spans five languages across two topical domains: politics (PCM, DEU, SWA) and environmental protection (ENG, ZHO). The dataset provides continuous Valence and Arousal annotations on a 1–9 scale for each aspect term within a given text. We report results using the official evaluation metric, $RMSE_{VA} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(v_i^p - v_i^g)^2 + (\alpha_i^p - \alpha_i^g)^2]}$, along with its per-dimension decomposition into $RMSE_V$ and $RMSE_A$.

All models are fine-tuned using the AdamW optimizer with a learning rate of 3×10^{-5} and a weight decay of 0.01 for a maximum of 25 epochs, with early stopping based on validation $RMSE_{VA}$. Input sequences are tokenized with a maximum length of 128 tokens. For Pipeline A (ENG, DEU), we use a batch size of 32, while Pipeline B (PCM, SWA, ZHO) uses a batch size of 64 to compensate for smaller training sets. Dropout rates of 0.3 and 0.2 are applied to the shared representation layer and the joint VA regression head, respectively, and gradient norms are clipped at 1.0 to stabilize training.

For Pipeline B, the multi-task loss (Eq. 4) is parameterized to balance the regression and classification terms at a 0.7 to 0.3 ratio. The higher weight on regression reflects that it is the primary objective of the task, while the auxiliary classification heads serve a regularization role. Within the regression component, we apply an elevated weight to Valence ($w_v = 2.0$) compared to Arousal ($w_\alpha = 1.0$). This reflects our preliminary observation that Valence constitutes the dominant error bottleneck ($RMSE_V > RMSE_A$). All language-specific models are trained independently on their respective partitions using a single NVIDIA P100 GPU.

4.1 Pipeline Selection

Rather than assigning languages to pipelines based purely on *a priori* linguistic assumptions, we adopted a validation-driven selection approach. Each language was independently evaluated under both Pipeline A and Pipeline B on the official development split, and assigned to the configuration with the lower $RMSE_{VA}$. Table 2 reports the per-language results.

Lang.	Pipeline A	Pipeline B	Selected
ENG	**2.065**	2.114	A
DEU	**2.092**	2.730	A
PCM	2.066	**1.911**	B
SWA	2.400	**2.132**	B
ZHO	0.774	**0.617**	B

Table 2: Per-language $RMSE_{VA}$ under both pipelines. Bold indicates the selected configuration.

Pipeline B was decisively superior for ZHO, SWA, and PCM, while Pipeline A was substantially better for DEU. For ENG the gap was marginal (0.05 $RMSE_{VA}$). Notably, Pipeline A on SWA produced a near-zero arousal correlation ($\mathrm{PCC}_A = -0.047$), indicating that Pipeline B's auxiliary classification heads regularize the encoder where direct regression fails to learn the arousal dimension. The linguistic framing in Section 3 is therefore best understood as a post-hoc interpretation of these empirical findings rather than an a priori hypothesis.

5 Results and Discussion

5.1 Official Results

Table 3 presents the official test set results for Subtask 1. Our system achieved competitive standings, particularly in non-Western languages, ranking 5th in Chinese (ZHO) and 7th in Swahili (SWA). Notably, our specialized Pipeline B architecture exhibited a stark performance inversion relative to massive large language model baselines. In ZHO and SWA, our system significantly outperformed both the mBERT and Mistral-314B baselines. Conversely, in Western or morphologically related languages—English (ENG), German (DEU), and

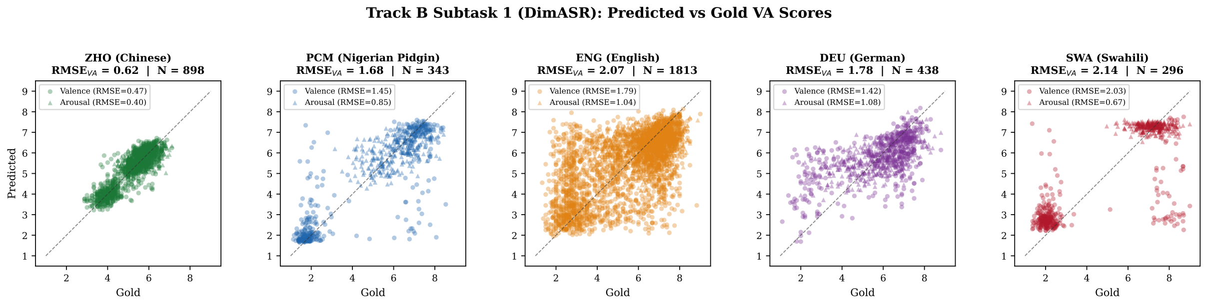

Figure 2: Predicted versus Gold scores for Valence and Arousal across the five target languages. Arousal predictions (triangles) demonstrate tighter clustering, while Valence predictions (circles) reveal pronounced variance and heteroscedasticity at scale extremes.

Language	Track	Rank	Our System	Mistral-314B	mBERT
ZHO	B	5	**0.617**	0.740	1.275
SWA	B	7	**2.132**	2.299	2.783
PCM	B	10	1.911	**1.739**	3.215
DEU	A	10	2.092	**1.591**	2.329
ENG	A	11	2.065	**1.643**	2.698

Table 3: Official System Performance ($RMSE_{VA}$). Bold values highlight the superior score between our official submission and the Mistral-314B baseline.

Language	Joint (Official)	Decoupled (Refined)	Δ
DEU	2.092	**1.780**	$+0.312$
PCM	1.911	**1.680**	$+0.231$
SWA	**2.132**	2.140	-0.008
ZHO	**0.617**	0.620	-0.003

Table 4: Effect of Decoupled Auxiliary Heads on $RMSE_{VA}$. Positive Δ indicates an improvement in error reduction.

Nigerian Pidgin (PCM)—our official submission underperformed Mistral-314B. This discrepancy highlights distinct architectural dependencies based on linguistic topology, which we investigate below. We base the ensuing analysis on the official $RMSE_{VA}$ metric, decomposing it into per-dimension $RMSE_V$ and $RMSE_A$ to expose the relative difficulty of each affective axis.

5.2 Post-Evaluation Refinement

Our official submission utilized a joint auxiliary head for predicting Valence and Arousal in Pipeline B. Post-evaluation analysis revealed a persistent gap between $RMSE_V$ and $RMSE_A$, indicating that a shared representation space forced a detrimental trade-off between the two dimensions. Consequently, we conducted an ablation study by decoupling the Valence and Arousal auxiliary classification heads.

As shown in Table 4, decoupling the heads yielded immediate improvements for Western and linguistically similar languages. German $RMSE_{VA}$ improved by 0.312, and Pidgin (PCM) improved by 0.231. Crucially, this architectural refinement brought our PCM score to **1.680**, allowing our lightweight dialect-specific model to surpass the Mistral-314B baseline (1.739). However, applying the decoupled architecture to ZHO and SWA resulted in performance degradation, establishing a clear divergence in optimal topologies. Notably, the post-evaluation decoupled Pipeline B score for DEU (1.780) is also lower than the officially submitted Pipeline A score (2.092), suggesting that decoupled auxiliary heads on a language-specific encoder may be the stronger configuration for DEU as well — an observation only available after the evaluation phase. English was excluded from this analysis as decoupling produced marginal degradation, consistent with the narrow Pipeline A advantage observed during selection (Table 2).

5.3 Error Analysis and Bottlenecks

Figure 2 illustrates the error distribution of our refined system across all five languages. Three critical bottlenecks emerge from this visualization:

First, Valence constitutes the primary regression bottleneck. With the notable exception of Chinese (which achieved a highly clustered $RMSE_{VA}$ of 0.62), $RMSE_V$ consistently exceeds $RMSE_A$ across all languages. Arousal predictions cluster in a narrow, confident band along the diagonal, whereas Valence predictions exhibit high scatter, indicating that Arousal is more readily extracted from textual semantics than nuanced Valence.

Second, the models demonstrate severe variance compression and regression to the mean. The Ordinary Least Squares (OLS) slope of predicted versus gold values falls substantially below 1.0. For in-stance, the Swahili Arousal slope is effectively flat (-0.02), indicating the model collapses predictions toward the center of the scale rather than tracking ground-truth variance.

Third, heteroscedasticity is prominent in low-resource environments. In PCM and SWA, the Mean Absolute Error (MAE) at high-valence extremes is approximately twice the MAE at low-valence inputs. The models systematically fail to capture intense positive sentiment, likely due to a lack of high-valence representation in the training distributions.

5.4 Architectural Implications for Cross-Lingual Transfer

The divergence in architectural preference between language families aligns with recent literature demonstrating the limitations of cross-lingual transfer in multilingual LLMs for non-Western languages (Tanwar et al., 2025).

Our findings suggest that decoupled task heads successfully stabilize regression for languages with robust, nuanced contextual embeddings (DEU, PCM). However, for distant linguistic families (ZHO, SWA), where base models lack deep cultural or domain-specific semantic mapping, the system relies heavily on the joint Valence-Arousal representation. In these cases, forcing the model to share parameters between dimensions acts as a necessary regularizer, compensating for weaker underlying contextual representations.

6 Conclusion

In this paper, we presented AdaptStance, a resource-adaptive framework for SemEval-2026 Task 3 that dynamically routes languages through distinct architectural pipelines. Our results demonstrate that no single architecture universally optimizes dimensional affect: while decoupling Valence and Arousal heads improves regression for morphologically related languages, joint representation remains critical for regularizing distant language families. Furthermore, our analysis highlights Valence as the primary regression bottleneck across all languages. Ultimately, our competitive performance against massive generative baselines underscores that for fine-grained affective regression, precise architectural topology and pre-training alignment are more effective—and vastly more computationally accessible—than sheer model scale.

Limitations

This work has several limitations. First, pipeline assignment was determined through single-run validation comparisons without repeated trials or seed-variance analysis; reported differences — particularly the marginal ENG gap — should not be interpreted as statistically robust. Second, we did not conduct controlled component-level ablations on the hybrid loss terms ($\mathcal{L}_{\text{LogCosh}}$, $\mathcal{L}_{\text{CCC}}$, $\mathcal{L}_{\text{Range}}$), the attention pooling mechanism, or the auxiliary classification heads in isolation; the full system was evaluated as a unit under shared task time constraints. Third, the routing between pipelines was determined manually from validation results rather than through a learned gating mechanism. Future work could explore data-driven routing criteria, multi-seed evaluation, and isolated component contribution analysis.

References

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.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

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.

Saif 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.

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

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2020. Stance detection benchmark: How robust is your stance detection? *Preprint*, arXiv:2001.01565.

Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang Wang, and Tanmoy Chakraborty. 2025. Do you know about my nation? investigating multilingual language models' cultural literacy through factual knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14956–14979, Suzhou, China. Association for Computational Linguistics.

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.