hllwan at SemEval-2026 Task 3: Dimensional Aspect-Based Sentiment Analysis via LLM Feature Fusion and Test-Time Adaptation

Jinglong Li and Yang Yang
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, Jiangsu, China
{jinglong_555, yyang}@njust.edu.cn

Abstract

This paper describes the system developed by the hllwan team for SemEval-2026 Task 3: Dimensional Aspect-Based Sentiment Analysis (DimABSA). Unlike traditional categorical sentiment analysis, predicting continuous Valence and Arousal (VA) scores across multiple languages and domains poses significant theoretical and engineering challenges. To systematically address data scarcity and cross-domain distribution shifts, we propose a highly robust framework. First, we implement a translation-based data augmentation strategy with precise HTML-tag alignment to mitigate low-resource constraints. Second, we introduce an unsupervised opinion extraction module based on syntactic dependency parsing to explicitly capture sentiment-bearing words. Third, we design a Tripartite Feature Fusion architecture built upon both encoder-only (DeBERTa-v3) and causal LLM (Qwen2.5) models to dynamically aggregate global and localized aspect-opinion embeddings. Finally, we apply an unsupervised Test-Time Adaptation (TTA) mechanism to calibrate normalization layers on the fly. Our system demonstrates highly competitive performance while offering critical insights into the limitations of LLMs in cross-lingual sentiment transfer.

1 Introduction

Sentiment analysis has evolved significantly over the past decade, shifting from coarse-grained document-level polarity classification to fine-grained Aspect-Based Sentiment Analysis (ABSA). Recently, Dimensional ABSA (DimABSA) has emerged to represent sentiments as continuous coordinates in a Valence-Arousal (VA) space, providing a much more nuanced and psychologically grounded understanding of human emotions. SemEval-2026 Task 3 pushes this boundary further by introducing multilingual datasets spanning highly specialized domains such as political discourse and environmental protection campaigns.

The core challenges of this multidimensional and cross-lingual task are multifaceted. Firstly, the **data scarcity** in low-resource languages (e.g., Swahili, Nigerian Pidgin) strictly limits the generalization capabilities of deep neural networks. Without sufficient annotated data, modern over-parameterized models easily overfit to spurious lexical correlations.

Secondly, predicting precise real-valued VA intensities requires models to capture the **subtle semantic interactions** between specific aspect terms and their surrounding opinion expressions. Since opinion spans are not explicitly annotated in the provided dataset, standard '[CLS]' token pooling often fails to capture these highly localized interactions adequately. When a sentence contains multiple aspects with conflicting sentiments, global pooling mechanisms suffer from severe information bottleneck and semantic blending.

Lastly, the **domain shift** between training sets and unseen test sets inevitably leads to out-of-distribution performance degradation. A model trained on generic environmental discussions will face drastically different syntactic structures when evaluated on political debates.

Our primary contributions are summarized as follows:

- A highly reliable data augmentation pipeline leveraging machine translation with targeted HTML-tag wrappers to elegantly solve the notorious aspect-misalignment problem.
- A rule-based syntactic dependency parsing algorithm using `spaCy` to automatically extract implicit opinion expressions.
- A Tripartite Feature Fusion module compatible with both DeBERTa (He et al., 2023) and Qwen2.5 (Team et al., 2025) architectures, explicitly enforcing token-level span attention.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1147–1152
July 3-4, 2026

- An unsupervised Test-Time Adaptation (TTA) strategy to dynamically optimize consistency loss on the test set, significantly improving out-of-distribution robustness without labeled target data.

2 Related Work

2.1 Dimensional Sentiment Analysis

Traditional sentiment analysis formulates the problem as a discrete classification task (e.g., classifying text into positive, negative, or neutral categories) (Pontiki et al., 2014). However, psychological research indicates that human emotions are inherently continuous. The Valence-Arousal (VA) circumplex model (Russell, 1980) represents emotions in a two-dimensional continuous space, where Valence denotes the degree of pleasure and Arousal denotes the level of physiological activation. Previous works in dimensional sentiment analysis have primarily focused on word-level (Warriner et al., 2013) or sentence-level VA prediction (Wang et al., 2016). Our work extends this to the aspect level (DimABSA), which requires a deeper understanding of the syntactic and semantic relationships between the target aspect and its specific surrounding context.

2.2 Large Language Models in ABSA

The introduction of Pre-trained Language Models (PLMs) like BERT (Devlin et al., 2019) shifted the paradigm for ABSA. Standard approaches often utilize the final hidden state of the '[CLS]' token for downstream classification. More recently, Large Language Models (LLMs) based on causal decoder architectures, such as LLaMA (Touvron et al., 2023) and Qwen (Team et al., 2025), have demonstrated remarkable zero-shot reasoning capabilities.

However, predicting continuous VA dimensions is inherently a regression problem. Directly prompting autoregressive LLMs for continuous numerical outputs often yields high variance. Therefore, instead of generation-based prompting, we adopt Parameter-Efficient Fine-Tuning (PEFT) techniques—specifically Low-Rank Adaptation (LoRA) (Hu et al., 2022)—to adapt decoder-only LLMs as powerful feature extractors for downstream regression heads.

2.3 Test-Time Adaptation

Test-Time Adaptation (TTA) (Wang et al., 2021) has gained significant traction in computer vision to combat distribution shifts by updating model parameters on unlabeled test data during inference. We extend TTA to continuous sentiment regression in Natural Language Processing by employing a self-supervised consistency loss constrained solely to the normalization layers of our PLMs, ensuring safe adaptation without representational collapse.

3 Task Description and Evaluation

The DimABSA task requires participating systems to predict the continuous Valence (V) and Arousal (A) scores for a specifically designated aspect within a text. Formally, given an input sentence consisting of a sequence of words $X = \{w_1, w_2, \ldots, w_n\}$ and an extracted aspect term $A_t \in X$, the goal is to learn a mapping function $\mathcal{F}_\theta$ parameterized by neural network weights θ that outputs a two-dimensional continuous vector:

$$\mathcal{F}_\theta(X, A_t) \rightarrow (V, A) \tag{1}$$

where $V \in [1, 9]$ represents the degree of pleasure (from negative to positive), and $A \in [1, 9]$ represents the degree of physiological activation (from calm to excited).

The primary evaluation metrics for this shared task are the Root Mean Square Error (RMSE) and the Pearson Correlation Coefficient (PCC).

RMSE measures the absolute prediction error:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\{V,A\}}(y_{i,j} - \hat{y}_{i,j})^2} \tag{2}$$

PCC measures the linear correlation between the predicted and actual scores, evaluating whether the relative ranking of sentiments is preserved:

$$\text{PCC} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{3}$$

where N is the total number of evaluation samples, y_i is the ground truth label, and $\hat{y}_i$ is the predicted score. Our system aims to minimize RMSE while simultaneously maximizing PCC.

4 System Overview

Our proposed framework consists of four main interconnected components: data augmentation via

Algorithm 1 Syntactic Opinion Extraction

Require: Input Text X, Aspect Span A_t, Dependency Parser $\mathcal{M}$
Ensure: Opinion String O

$Tree \leftarrow \mathcal{M}(X)$
$Opinions \leftarrow$ Empty List
for each $token$ **in** A_t **do**
  // Rule 1: Extract direct modifiers
  for each $child$ **of** $token$ **do**
    if $child$ is ADJ/ADV **and** dependency is `amod`, `acomp`, or `advmod` **then**
      $Opinions$.append($child$)
    end if
  end for
  // Rule 2: Extract predicates and complements
  if $token$ is a nominal subject (`nsubj`) **then**
    $verb \leftarrow token$.head
    if $verb$ is a meaningful VERB (e.g., not *be, have, do*) **then**
      $Opinions$.append($verb$)
    end if
    $Opinions$.append(`acomp` children of $verb$)
  end if
end for
$O \leftarrow$ Join(RemoveDuplicates($Opinions$))
return O

machine translation, syntactic opinion extraction, tripartite feature fusion architecture, and unsupervised test-time adaptation.

4.1 Data Augmentation via Translation

For low-resource languages in the dataset, such as Nigerian Pidgin (PCM) and Swahili (SWA), we translate training samples into English. A major bottleneck in translating ABSA data is that the exact aspect and opinion spans often shift, merge, or disappear entirely in the translated text.

To solve this problem without relying on complex word-alignment models, we inject HTML tags into the source text before passing it to the translation API. By wrapping aspects in `<span id="i">...</span>`, we force the translation engine to retain these boundary markers. Post-translation, we use regular expressions to extract the translated aspect and subsequently clean the text. This guarantees perfect aspect-level alignment.

4.2 Opinion Term Extraction

While target aspects are explicitly provided in the DimABSA dataset, the subjective opinion words modifying these aspects are often implicit and unannotated. Accurately locating these opinion words is crucial for precise dimensional sentiment regression. To automatically extract aspect-specific opinion expressions, we leverage the `spaCy` library (Honnibal et al., 2020) for robust syntactic dependency parsing.

Given an input text X and an aspect span A_t, we apply a set of heuristic linguistic rules to traverse the syntactic tree, which is detailed in Algorithm 1. We extract direct adjectival and adverbial modifiers (`amod`, `acomp`, `advmod`), as well as predicates and their complements if the aspect acts as a nominal subject (`nsubj`). All extracted opinion words are concatenated into a single string O, serving as an auxiliary input for our feature fusion module.

4.3 Tripartite Feature Fusion Architecture

We employ two distinct backbone architectures to ensure model diversity: `deberta-v3-large` (an Encoder-only model) and `Qwen2.5-1.5B` (a Decoder-only causal language model).

Low-Rank Adaptation (LoRA): To ensure extreme parameter efficiency, we freeze the backbone weights $W_0 \in \mathbb{R}^{d\times k}$ and update only a low-rank decomposition matrix $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ with a rank factor $r \ll \min(d, k)$.

Token Offset Masking: During tokenization, we utilize the 'offset_mapping' feature to locate the exact character-to-token boundaries of the Aspect (A_t) and Opinion (O) spans. This allows us to construct precise binary masks $M_{asp} \in \{0, 1\}^L$ and $M_{opi} \in \{0, 1\}^L$.

Let $H \in \mathbb{R}^{L\times d}$ be the last hidden states output by the PLM. We compute three distinct contextual vectors:

1. Global Context (h_{global}): For DeBERTa, this is the representation of the '[CLS]' token. For Qwen2.5, it is the representation of the final valid token in the sequence:

$$h_{global} = H_{cls} \quad \text{or} \quad H_{last} \tag{4}$$

2. Aspect Focus (h_{aspect}): We apply mean-pooling over the hidden states corresponding exclusively to the aspect tokens:

$$h_{aspect} = \frac{\sum_{i=1}^{L} H_i \cdot M_{asp}^{(i)}}{\max(\sum_{i=1}^{L} M_{asp}^{(i)}, \epsilon)} \tag{5}$$

Algorithm 2 Test-Time Adaptation (TTA) per Batch

Require: Test Batch B, Model $\mathcal{F}_\theta$, Steps K, LR α

Freeze all parameters in θ except Norm Layers θ_{norm} (e.g., LayerNorm/RMSNorm)
Set $\mathcal{F}_\theta$ to `train()` mode (enable Dropout)
for $k = 1$ to K **do**
  $\hat{Y}_1 \leftarrow \mathcal{F}_\theta(B)$ // First stochastic pass
  $\hat{Y}_2 \leftarrow \mathcal{F}_\theta(B)$ // Second stochastic pass
  $\mathcal{L}_{TTA} \leftarrow \text{MSE}(\hat{Y}_1, \hat{Y}_2)$
  $\theta_{norm} \leftarrow \theta_{norm} - \alpha \nabla_{\theta_{norm}} \mathcal{L}_{TTA}$
end for
Set $\mathcal{F}_\theta$ to `eval()` mode (disable Dropout)
Return Final Prediction $\hat{Y} \leftarrow \mathcal{F}_\theta(B)$

3. Opinion Focus ($h_{opinion}$): Similarly, we calculate the opinion-focused representation:

$$h_{opinion} = \frac{\sum_{i=1}^{L} H_i \cdot M_{opi}^{(i)}}{\max(\sum_{i=1}^{L} M_{opi}^{(i)}, \epsilon)} \quad (6)$$

The final aggregated feature h_{fuse} is the arithmetic mean of these three distinct semantic components:

$$h_{fuse} = \frac{1}{3}(h_{global} + h_{aspect} + h_{opinion}) \quad (7)$$

h_{fuse} is then passed through a Dropout layer ($p = 0.1$) followed by a linear projection head with a Sigmoid activation, rescaled to the $[1, 9]$ range.

4.4 Test-Time Adaptation (TTA)

To counteract domain shifts between the training and testing phases, we introduce an unsupervised adaptation step prior to the final prediction, detailed in Algorithm 2.

During inference, we freeze all model parameters except for the normalization layers. For each incoming test batch, we temporarily enable Dropout. We perform two stochastic forward passes to obtain $\hat{Y}_1$ and $\hat{Y}_2$, computing a consistency loss $\mathcal{L}_{TTA} = \|\hat{Y}_1 - \hat{Y}_2\|^2$. By backpropagating this loss to update the normalization parameters for K steps, the model learns to produce consistent representations for the out-of-distribution inputs.

5 Experimental Setup

5.1 Datasets

We evaluate our system on the official DimABSA 2026 dataset (Lee et al., 2026; Becker et al., 2026). The dataset distributions are summarized in Table 1. We split 10% of the training data for validation.

Lang.	Domain	Train Size	Val Size
ENG	Env. Protection	1122	112
DEU	Politics	717	71
ZHO	Env. Protection	800	80
PCM	Politics	1168	116
SWA	Politics	1498	149

Table 1: Statistics of the DimABSA dataset splits.

5.2 Implementation Details

All models were implemented using the PyTorch framework (Paszke et al., 2019) and the Hugging Face Transformers library (Wolf et al., 2020). We summarize our core hyperparameters in Table 2. We employed Early Stopping with a patience of 4 epochs based on the validation RMSE metric and utilized the AdamW optimizer (Loshchilov and Hutter, 2019).

Hyperparameter	Value
Max Sequence Length	256
Optimizer	AdamW
Learning Rate	1e-4
Batch Size (DeBERTa / Qwen)	8 / 4
LoRA Rank (r)	32
LoRA Alpha (α)	64
Dropout Rate	0.1
TTA Learning Rate	1e-4
TTA Steps	3
Epochs	15

Table 2: Hyperparameter settings utilized for both the main training phase and the Test-Time Adaptation.

6 Results and Analysis

6.1 Main Results

Table 3 presents the quantitative performance. The Qwen2.5-7B model slightly outperformed DeBERTa across most high-resource languages (ENG, DEU, ZHO). This highlights the superior contextual understanding of modern large language models. However, an ensemble averaging the predictions of both architectures yielded the best overall robustness, particularly stabilizing the predictions for low-resource languages (PCM, SWA).

6.2 Language-Specific Insights and Error Analysis

While our framework excels in high-resource scenarios, a critical analysis reveals several negative in-

Model	RMSE on Different Languages (↓)					Average
	ENG	DEU	ZHO	PCM	SWA	
Baseline (Mistral-3 14B)	1.6430	1.5910	0.7400	1.7390	1.7020	1.4830
Baseline (mBERT)	2.6985	1.3420	2.3254	1.2756	3.2152	2.1713
XLM-RoBERTa-large (Conneau et al., 2020)	2.1816	1.5611	0.9668	1.5064	2.0266	1.6485
DeBERTa-v3-large (He et al., 2023)	1.5898	1.5740	0.6720	**1.2232**	1.9522	1.4022
Qwen2.5-7B (Team et al., 2025)	**1.5122**	**1.4937**	**0.6154**	1.2974	**1.8452**	**1.3528**

Table 3: Main quantitative results on the DimABSA 2026 dataset evaluated using Root Mean Square Error (RMSE). Lower scores indicate better performance.

sights regarding low-resource and domain-specific settings.

First, our HTML-based translation augmentation struggled heavily with morphologically rich languages like Swahili. The translation API frequently broke the tags or aggressively altered the grammatical structure, causing the mapped aspects to lose their contextual opinion modifiers.

Furthermore, in the German Politics subset, the sentiment toward a political entity is frequently conveyed through sarcastic rhetorical questions or cultural-specific metaphors. Because these sentences lack explicit adjectival modifiers that our spaCy parser can cleanly catch, the $h_{opinion}$ representation became sparse. This forced the models to fall back onto the global context, which ultimately misinterpreted the sarcasm as neutral factual reporting.

7 Conclusion

We described the comprehensive framework tailored for SemEval-2026 Task 3. By combining HTML-based cross-lingual translation augmentation, syntactic dependency-based opinion extraction, Tripartite Feature Fusion, and dynamic Test-Time Adaptation (TTA), our system effectively models dimensional sentiment intensities with high precision. Our analysis highlights that while LLM feature extractors are powerful, handling implicit political sarcasm and preserving morphology during cross-lingual transfer remain significant open challenges. Future work will explore Mixture-of-Experts (MoE) architectures to further isolate domain-specific knowledge during the TTA phase.

References

Jonas Becker, Liang-Chih Yu, Shamsuddeen Hassan Muhammad, Jan Philip Wahle, and 1 others. 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.

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*, pages 4171–4186.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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

Lung-Hao Lee, Liang-Chih Yu, Natalia Loukashevich, Ilseyar Alimova, and 1 others. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *Preprint*, arXiv:2601.23022.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In

Advances in neural information processing systems, volume 32.

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.

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

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and 1 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*.

Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional cnn-lstm model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 225–230.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Thomas Wolf, Lysandre Debut, Victor Sanh, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.