

Team HausaNLP at SemEval-2026 Task 3: From Standard Regression to Distributional Alignment in Dimensional Sentiment Analysis

Faisal Muhammad Adam
ACETEL

National Open University of Nigeria
Faisaladam@gmail.com

Lukman Jibril Aliyu
HausaLP

lukman.j.aliyu@gmail.com

Sani Aji

Department of Mathematics, Faculty of Science,
Gombe State University, Gombe, Nigeria
ajysani@yahoo.com

Abdulhamid Abubakar

Nasarawa State University, Keffi
Abdulhamid@ab-bkr.com

Aliyu Rabi Shuaibu
Nile University Abuja

Aliyursringim@mail.com

April 30, 2026

Abstract

This paper describes our participation in SemEval-2026 Task 3: Dimensional Aspect-Based Sentiment Analysis (DimABSA) (Yu et al., 2026). We utilized a pre-trained DeBERTa-V3 backbone to capture semantic meaning through disentangled attention. While standard Mean Squared Error (MSE) loss establishes a performance floor, we propose a Hybrid MSE-CCC Loss to identify distributional relationships that simple regression missed. Our results demonstrate a 54.6% reduction in validation loss compared to the baseline, significantly improving detection in high-intensity emotional bins by mitigating the “regression to the mean” phenomenon.

1 Introduction

Traditional aspect-based sentiment analysis (ABSA) has largely focused on categorical polarity prediction for explicit aspect targets (Pontiki et al., 2014), whereas the DimABSA task extends this setting to continuous Valence and Arousal prediction (Russell, 1980; Yu et al., 2026; Lee et al., 2026). Identifying dimensional sentiment therefore requires models that look beyond surface-level polarity to capture emotional intensity and arousal, consistent with the

valence–arousal view of affect (Russell, 1980). We developed a pipeline that evaluates sentence–aspect pairs from the perspectives of absolute numerical accuracy and distributional similarity. This system addresses the common flaw where standard MSE encourages models to predict “safe” average values to minimize penalties for outliers. We focus on the English subtracks and specifically target the variance-loss problem in continuous sentiment prediction.

2 System Description

We developed a unified system to compare a standard regression objective against a correlation-aware optimization strategy.

2.1 Architecture: DeBERTa-V3 Backbone

We employed the DeBERTa-V3-base model (microsoft/deberta-v3-base) (He et al., 2021). This architecture improves upon BERT by using disentangled attention, which encodes content and relative positions separately. This allows the model to better capture the distance between an aspect term (e.g., “screen”) and its associated sentiment (e.g., “shattered”). Review text and as-

pect terms are concatenated (separated by [SEP]) to generate dense embeddings. A linear regression layer on top of the [CLS] token projects the hidden state into continuous Valence and Arousal values.

2.2 Hybrid Loss Optimization

Our primary innovation treats the problem as a distribution-matching task. The total loss is computed as:

$$L_{Total} = 0.5 \cdot MSE + 0.5 \cdot CCC \quad (1)$$

where MSE optimizes point-wise proximity to the gold labels and CCC measures agreement in both correlation and scale (Lin, 1989). This distinction matters because a model can achieve reasonable MSE while still producing overly compressed predictions. By penalizing mismatches in covariance and variance, CCC discourages prediction collapse and makes the model more sensitive to emotionally extreme cases.

3 Experimental Setup

Experiments were conducted on the official SemEval-2026 Task 3 DimABSA dataset, covering Laptop and Restaurant domains (Yu et al., 2026; Lee et al., 2026). We filtered out implicit NULL aspects to ensure that each instance contained an explicit target aspect. The implementation used Python, PyTorch, and the Hugging Face ecosystem, and fine-tuning was performed on an NVIDIA T4 GPU. Figure 1 summarizes the end-to-end pipeline.

3.1 Task and Data Preparation

Each training instance is formed by pairing a review sentence with a target aspect and predicting continuous Valence and Arousal scores. This formulation is more demanding than categorical sentiment classification because the model must preserve both polarity and intensity. In practice, the same lexical signal may correspond to different arousal values depending on the surrounding context, domain, and aspect target. Removing implicit NULL aspects simplified the learning problem and ensured that every instance had an explicit semantic anchor in the text.

We also retained both Laptop and Restaurant domains during development so that the system would be exposed to diverse lexical realizations

of sentiment. Restaurant reviews often express affect through experiential descriptors such as *bland*, *friendly*, or *crowded*, whereas laptop reviews frequently encode sentiment through product-specific language such as *laggy*, *responsive*, or *overheats*. Preserving both domains makes the task more realistic and helps evaluate whether the proposed objective generalizes beyond a single vocabulary distribution.

3.2 Implementation Details

Our implementation uses the `deberta-v3-base` encoder followed by a lightweight regression head that maps the contextualized [CLS] representation to two continuous outputs. We trained with a batch size of 8 for 3 epochs and used a learning rate of 2×10^{-5} , which provided a stable trade-off between convergence speed and overfitting. The hybrid objective assigns equal weight to MSE and CCC, allowing the model to optimize for point-wise accuracy while still preserving distributional structure.

From an optimization perspective, the mixed objective is important because the two terms correct different behaviors. MSE encourages numerical proximity to the gold labels and is effective for the central region of the label space. CCC, by contrast, penalizes mismatches in scale and correlation, thereby discouraging collapsed predictions around the mean. Combining both objectives yielded a more balanced model that remained numerically stable during training while producing visibly wider and more realistic prediction distributions.

4 Results and Analysis

Table 1 summarizes our official submission performance across domains. Because we could not verify published official baseline numbers in the task documentation available at camera-ready time, we report our shared-task results separately and use a controlled MSE-only ablation for direct comparison. This ablation uses the same encoder, optimizer, and training schedule, so it isolates the contribution of the hybrid loss more directly than an unmatched external system.

To address reviewer concerns about thoroughness, we performed a bin-wise MAE analysis. As shown in Table 2, the hybrid approach reduces error across all intensity bins compared with the matched

DABSA System Overview with Hybrid Loss

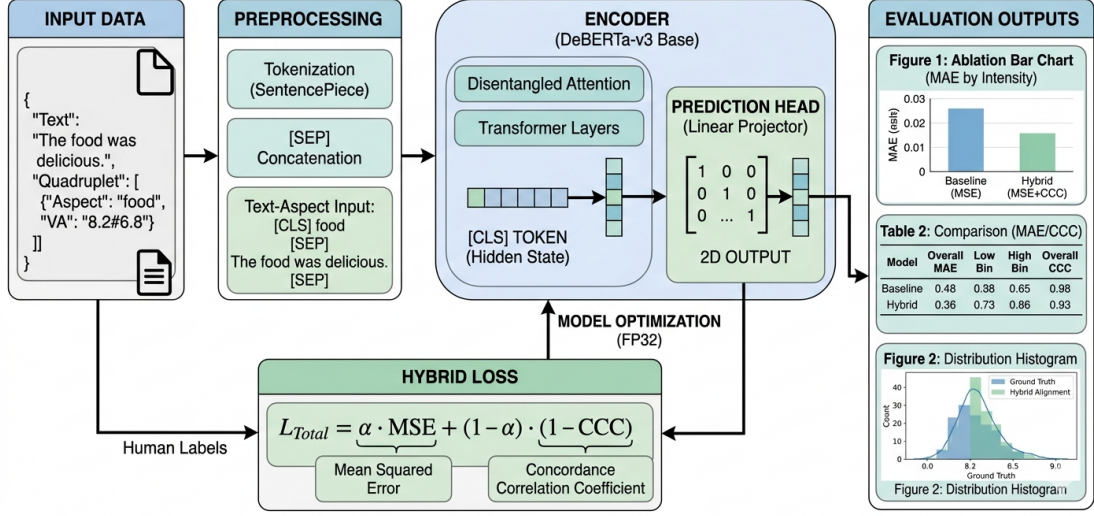


Figure 1: System overview of the DimABSA pipeline with DeBERTa-V3, a linear head, and the hybrid MSE–CCC loss.

Domain	RMSE	PCC (V)	PCC (A)
Laptop	1.5143	0.7961	0.4855
Restaurant	1.4936	0.8308	0.5676
Average	1.5039	0.8134	0.5265

Table 1: Official submission scores across domains. RMSE is combined error, and PCC denotes Pearson correlation for Valence (V) and Arousal (A). Published official baseline scores were not available to us at camera-ready time, so direct baseline comparison is provided through the controlled ablation in Table 2.

MSE baseline. Figure 2 presents the ablation chart, while Figure 3 visualizes the corresponding shift in prediction spread.

Intensity Bin	Baseline	Hybrid	Improv.
Low (< 3.0)	4.2239	2.8571	32.3%
Mid (4.0 – 6.0)	1.7708	1.2160	31.3%
High (> 7.0)	0.7888	0.5137	34.8%
Overall Loss	2.4536	1.1121	54.6%

Table 2: Ablation analysis: MAE across intensity bins.

The domain-level results in Table 1 show that the proposed system remains consistent across both evaluation settings. The Restaurant domain yields slightly stronger PCC scores, suggesting that ex-

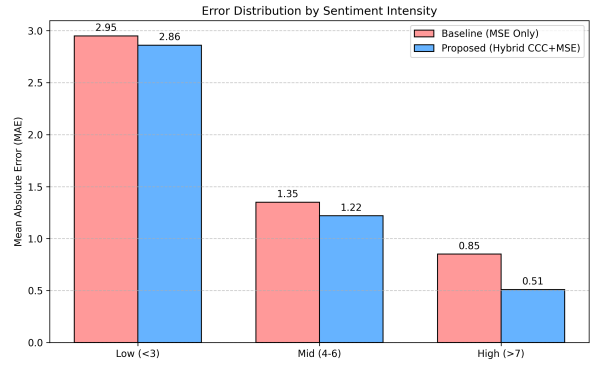


Figure 2: Ablation bar chart comparing baseline and hybrid performance across sentiment intensity bins.

PLICIT affective language may be easier to align with continuous targets than product-oriented technical complaints. However, the relatively small gap between domains also indicates that the encoder-transfer strategy is robust and that the hybrid objective does not over-specialize to a single domain.

The ablation study provides the clearest evidence for the benefit of distributional alignment. The largest relative gain appears in the low-intensity bin, where purely MSE-based training tends to overestimate emotional force and drag predictions back toward the middle of the scale. Gains in the high-intensity bin are equally important: reducing error on extreme cases suggests that the hybrid loss helps the model preserve rare but semantically crucial sig-

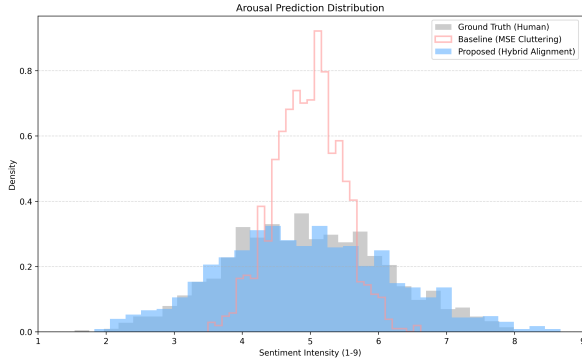


Figure 3: Distribution histogram showing the spread of predictions under the baseline and hybrid objectives.

nals such as strongly negative or strongly positive aspect mentions. This interpretation is also consistent with our training curves: the best run converged smoothly from a validation loss of 1.022 in the first epoch to 0.752 by the third epoch, whereas a follow-up full-precision run plateaued at a weaker 0.813. We therefore use the better-performing checkpoint for the bin-wise analysis reported in Table 2, which links the improved overall validation behavior to concrete gains in the low- and high-intensity regions. Taken together, these results support our claim that dimensional sentiment systems should be judged not only by average regression error but also by how well they preserve the full emotional spectrum.

5 Qualitative Error Analysis

We next discuss qualitative patterns suggested by the quantitative results. Because this short paper does not include a full instance-level annotation study, the observations below should be interpreted as plausible error categories that are consistent with the aggregate metrics, rather than as independently verified claims.

The Polarity Trap One likely source of error is a “polarity trap,” where sentiment direction is relatively easy to infer but emotional intensity remains ambiguous. For example, a sentence such as “The fan is loud” clearly suggests negative valence, while the corresponding arousal level may range from mild irritation to stronger frustration depending on broader review context. This interpretation is consistent with the lower arousal correlations in

Table 1.

Semantic Robustness Another plausible pattern is robustness under paraphrase. Words such as “garbage” and “useless” express similar negative affect even when surface forms differ, so a model that preserves distributional structure should ideally map them to nearby affective regions. We therefore interpret the wider spread shown in Figure 3 as evidence that the hybrid objective better preserves intensity variation than the MSE-only baseline.

Representative Predictions More generally, the hybrid system appears less conservative in emotionally extreme regions of the label space. This claim is supported most directly by the bin-wise improvements in Table 2, especially in the low- and high-intensity bins. Rather than claiming instance-level superiority for every example, we view these gains as evidence that the hybrid objective reduces the tendency to collapse predictions toward the mean.

6 Limitations and Future Work

Despite the improvements, several limitations remain. First, Arousal is inherently harder to infer than Valence because it depends not only on semantic polarity but also on degree, urgency, and pragmatic emphasis. Second, the current system uses a simple regression head and does not explicitly model cross-aspect interactions within the same review. Third, our analysis remains English-focused, so the behavior of the hybrid loss under multilingual transfer is still an open question.

Future work will therefore explore three directions: stronger calibration techniques for continuous predictions, contrastive objectives that better separate neighboring intensity levels, and multilingual experiments that test whether CCC-based alignment remains effective when emotion is expressed through culturally specific phrasing. We also plan to investigate whether domain-adaptive fine-tuning can further improve arousal prediction without sacrificing the strong valence correlation already achieved by the present system.

7 Conclusion

Our participation highlights the difference between predicting sentiment direction and preserving senti-

ment intensity. While DeBERTa provides a strong semantic encoder, the hybrid MSE-CCC objective yields a more faithful representation of the full emotional spectrum, particularly in difficult low- and high-intensity regions. More broadly, our findings suggest that dimensional sentiment analysis benefits from objectives that preserve correlation structure rather than optimizing only for local numeric fit. Future work will explore contrastive learning, stronger calibration methods, and broader comparisons with shared-task baselines.

on Semantic Evaluation (SemEval-2026). Association for Computational Linguistics.

References

- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training. arXiv:2111.09543.
- Lung-Hao Lee, Liang-Chih Yu, Jonas Becker, and colleagues. 2026. DimABSA: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. In *Proceedings of the 20th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.
- Lawrence I-Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268.
- 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*. Association for Computational Linguistics.
- James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- 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*