Habib University at SemEval-2026 Task 3: A Pipeline Approach for Dimensional Aspect-Based Sentiment Analysis

Muhammad Affan, Muhammad Hassan Shahzad, Mikaal Imam
Moiz Zulfiqar, Sandesh Kumar, Abdul Samad
Dhanani School of Science & Engineering, Habib University, Pakistan
{ma08910, ms09070, mi08753, mz08229}@st.habib.edu.pk
{sandesh.kumar, abdul.samad}@sse.habib.edu.pk

Abstract

Aspect-based sentiment analysis has evolved from categorical polarity classification to fine-grained modeling of continuous affective dimensions. Dimensional Aspect-Based Sentiment Analysis (DimABSA) extends this paradigm by requiring both structured sentiment extraction and continuous valence–arousal (VA) regression in multilingual settings. In this paper, we present our system for SemEval-2026 Task 3, which evaluates this challenge across six languages and four domains, requiring systems to extract aspect–category–opinion quadruplets and predict VA scores on a 1–9 scale. We propose a modular four-stage multilingual transformer pipeline for element extraction, aspect–opinion pairing, category prediction, and VA regression. We conduct experiments over multiple models and training configurations, including VA rescaling to [-1,1], Gaussian label noise injection, Concordance Correlation Coefficient (CCC) loss, and Savitzky–Golay smoothing. Among all languages, our system achieves the lowest RMSE of 0.5333 on Subtask 1 and the highest cF1 of 0.5492 on Subtask 2. We further investigate data augmentation to improve low-resource performance and address label imbalance. Ultimately, our modular architecture demonstrated highly competitive cross-lingual transfer, achieving top-tier placements in low-resource settings, including 2nd place for Tatar and 6th place for Russian in dimensional regression.

1 Introduction

Sentiment analysis enables an automated understanding of opinions in text, from product reviews to financial reports. Traditional Aspect-Based Sentiment Analysis (ABSA) (Verma et al., 2024) assigns discrete polarity labels positive, negative, or neutral that are practical but coarse.

SemEval-2026 Task 3 (Yu et al., 2026) addresses this limitation through *Dimensional ABSA* (DimABSA), replacing categorical labels with continuous *valence-arousal* (VA) scores on a 1–9 scale (Russell, 1980). A visual illustration of the VA space is provided in (Figure 3). This richer representation enables systems to distinguish not just polarity but also depth of emotion.

The task covers four domains—restaurants, laptops, hotels, and finance—across six diverse languages: English, Chinese, Japanese, Russian, Tatar, and Ukrainian. Three subtasks form a progression of difficulty:

- **ST1 (DimASR):** Given a sentence and an aspect, predict its continuous (v, a) scores.
- **ST2 (DimASTE):** From a sentence alone, extract all (aspect, opinion, VA) triplets.
- **ST3 (DimASQP):** Extract complete (aspect, category, opinion, VA) quadruplets.

Our pipeline approach utilizes mDeBERTa-v3 as the primary encoder, fine-tuned for structured extraction and dimensional regression. To address low-resource language scarcity, category imbalance, and VA score skewness, we incorporate external in-domain data and apply controlled data augmentation strategies. For regression stability, we rescale VA scores to [-1,1] and inject Gaussian label noise during training. This design improves robustness across languages while maintaining distributional consistency in continuous sentiment prediction.

2 Dataset

Each instance in the dataset consists of an ID, a sentence or phrase, and a set of (aspect, opinion, category, VA) quadruplets. The data spans six languages (English, Chinese, Russian, Japanese, Tatar, and Ukrainian) and four domains: Restaurant, Laptop, Hotel, and Finance. Table 4 summarizes the domain distribution across languages and the number

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3449–3459

of instances per task. Overall, the dataset contains 23,244 instances and 196 unique entity–attribute categories.

The data exhibits significant imbalance in both VA score distribution and category frequency. Valence scores are skewed toward the mid-to-high range (6–8), while arousal values cluster around 5–7, limiting representation of extreme emotional states. Category imbalance is also pronounced: a single category (FOOD#QUALITY) accounts for 28.2% of all instances, whereas 31% of categories individually contribute less than 2% of the data (Lee et al., 2026).

3 Related Work

Aspect-Based Sentiment Analysis(ABSA) has seen rapid architectural evolution over the last decade. It originated as the task of Aspect-Level Sentiment Classification, designed to identify the discrete sentiment label (e.g., positive, negative, neutral) of an aspect-entity pair within a sentence (Pontiki et al., 2016). Earlier solutions relied heavily on recurrent architectures, specifically CNN-LSTMs (Wang et al., 2016), to capture local information within sentences. However, they suffered from the fundamental rigidity of recurrent processing.

With the advent of the transformer architecture and finetuning of these models, the focus of ABSA shifted towards this domain. Pre-trained models like BERT bypassed the recurrent bottleneck through self-attention mechanisms, demonstrating superior performance over legacy LSTM techniques (Sun et al., 2019). This leap in computational power led to researchers introducing complex fine-grained extraction tasks, such as triplet and quadruplet extraction(Cai et al., 2021). Subsequently, further refined techniques were used for contextual representations using Graph Neural Networks (Tian et al., 2021) and contrastive learning (Liang et al., 2021).

Despite these structural advancements, sentiment was historically modelled as a rigid discrete classification problem. Recent literature has started to shift towards dimensional representation, with (Wang et al., 2016) first introducing the Valence-Arousal(VA) space within ABSA. This continuous quantification of sentiment was also formalized in SemEval-2025 task 11, which evaluated sentiment on a continuous scale (Muhammad et al., 2025), and then in the SIGHAN-2024 shared task, which introduced Dimensional Aspect Based Sentiment Analysis (DimABSA) for Chinese texts (Lee et al., 2024). Our task, SemEval-2026 task 3, is also a DimABSA task like this shared task in SIGHAN, but this one is multilingual extending onto review texts from 6 languages and 5 domains (Yu et al., 2026). The top team in this shared task employed a BERT-based pipeline and an LLM approach, ensembling both. The approach of the BERT pipeline achieved significantly better results (Xu et al., 2024).

4 Methodology

In subtask 1, given the sentence $S = [w_1, w_2, ..., w_n]$ and a term $aspect$, predict the valence $(valence)$ and $(arousal)$ values which are continuous on the scale of $[1, 9]$. In subtask 2, only the sentence S is given and the objective is to extract $(aspect, opinion, valence - arousal)$ triplets. Whereas in subtask 3, the objective is to extract $(aspect, opinion, valence - arousal, category)$ quadruplets from S. The $aspect$ and $opinion$ terms can be implicit or a substring within S, in case of them being implicit, they are represented by 'NULL'. On the other hand, the category has two parts and is divided like this $entity\#attribute$ (Yu et al., 2026).

Our system is a four-stage pipeline: (1) Extractor, (2) Pairer, (3) Category Classifier, (4) Regressor. Each stage is trained independently. The three subtasks use different subsets of this pipeline, as described below.

4.1 Subtask 1: Dimensional Sentiment Regression (DimASR)

ST1 is a multilingual valence-arousal (VA) regression problem. Given an input sentence S and aspect A, the objective is to predict continuous valence and arousal scores (v, a) mapped to the two-dimensional affective space by Russell (1980).

Following prior work on dimensional sentiment analysis (Xu et al., 2024), encoder based multilingual transformers are selected due to their numerical stability and strong crosslingual generalization. They show that BERT-style encoders provide stable continuous score prediction compared to generative LLMs, while (E. G. P. et al., 2025) demonstrate that XLM-RoBERTa effectively handles multilingual sentiment tasks due to large-scale multilingual pretraining.

Based on these findings, four backbone mod-

els are benchmarked under a unified framework: mBERT, mDeBERTa, XLM-RoBERTa-base and XLM-RoBERTa-large.

To analyze model capacity effects, both base and large variants are evaluated. For each model, the input sequence is constructed as:

$$Input = [CLS] \oplus A \oplus : \oplus S \oplus [SEP] \quad (1)$$

A linear regression head predicts valence and arousal jointly:

$$(\hat{v}, \hat{a}) = W_{reg} h_{cls} + b_{reg} \quad (2)$$

Initial benchmarking shows that mDeBERTa consistently outperforms mBERT and XLM-RoBERTa variants across languages and domains. Although XLM-RoBERTa-large improves over the base version, the performance gain is smaller than switching to mDeBERTa. Therefore, mDeBERTa is selected as the primary backbone for further refinement.

To stabilize regression training, VA labels originally in the range $[1, 9]$ are linearly rescaled to $[-1, 1]$:

$$y' = \frac{y - 5}{4} \quad (3)$$

Rescaling is motivated by (Atmaja and Akagi, 2021), who demonstrate that correlation-based affect recognition benefits from normalized target ranges. Mapping to $[-1, 1]$ improves interpretability, stabilizes optimization, and aligns prediction magnitudes with bounded regression outputs.

To improve robustness and reduce overfitting to exact annotation values, Gaussian noise is injected into the rescaled targets during training:

$$y_{noise} = y' + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad (4)$$

Noise injection is inspired by (Lu et al., 2021), who show that controlled perturbation enhances generalization in fine-grained sentiment tasks. While their approach injects noise at the embedding level, the adaptation here regularizes regression targets directly.

Two loss formulations are investigated. The first is Mean Squared Error (MSE):

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} \|y_i - \hat{y}_i\|^2 \quad (5)$$

The second is Concordance Correlation Coefficient (CCC) loss:

$$\mathcal{L}_{CCC} = 1 - CCC(y, \hat{y}) \quad (6)$$

where

$$CCC = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2} \quad (7)$$

(Atmaja and Akagi, 2021) argue that CCC is more suitable than pure error-based metrics for dimensional affect recognition, as it penalizes both correlation mismatch and mean bias. Experiments show that while CCC improves alignment in certain domains, the best overall performance is achieved using rescaled targets with Gaussian noise under MSE optimization.

Finally, Savitzky–Golay filtering (Gallagher, 2020) is applied to smooth VA labels prior to training.

This filtering technique reduces high-frequency noise while preserving overall signal shape. Although smoothing stabilizes some domain predictions, improvements remain moderate compared to rescaling and noise regularization.

The final system therefore consists of mDeBERTa with label rescaling to $[-1, 1]$ and Gaussian noise injection, trained using a regression objective. This configuration achieved the best balance between correlation performance and prediction stability across languages.

4.2 Subtask 2: Dimensional Triplet Extraction (DimASTE)

ST2 requires extracting all (aspect, opinion, VA) triplets from a sentence with no aspect provided as input. This necessitates two pipeline stages prior to regression: extraction and pairing.

4.2.1 Stage 1 – Aspect and Opinion Extraction

We frame extraction as BIO sequence labeling over the label space {B-ASP, I-ASP, B-OPI, I-OPI, O}. Standard BIO cannot represent *implicit* elements (aspects or opinions not literally present in text). We address this by prepending two special tokens:

$$S' = [\texttt{NULL}_{asp},\ \texttt{NULL}_{opi},\ w_1, \ldots, w_n]$$

When the ground-truth aspect is implicit, $\texttt{NULL}_{asp}$ receives the B-ASP label; when the opinion is implicit, $\texttt{NULL}_{opi}$ receives B-OPI. This **Double [NULL] Token** strategy handles implicitness without any architectural modification. A BERT based model encodes S' and a token classification head is trained with cross-entropy loss. While the overall

ST2 pipeline is evaluated on the official cF1 metric, we isolated and evaluated this extractor component using standard Token-Level F1. Based on superior isolated F1 performance, we selected mDeBERTa as the backbone for the extractor model, while XLM-RoBERTa was chosen for the subsequent pairing model.

4.2.2 Stage 2 – Aspect-Opinion Pairing

The Extractor yields candidate aspect spans $\mathcal{A}$ and opinion spans $\mathcal{O}$. We form the full Cartesian product $\mathcal{P} = \mathcal{A} \times \mathcal{O}$ and filter it with a binary classifier. Each candidate pair is encoded as:

$$[\texttt{CLS}]\ S'\ [\texttt{SEP}]\ \textit{aspect}\ [\texttt{SEP}]\ \textit{opinion}\ [\texttt{SEP}]$$

Hard negatives (within-sentence mis-pairings) and **soft negatives** (cross-sentence random pairings) are constructed during training so the model learns both local and global pairing cues, as the negative pair-construction technique defined by (Xu et al., 2024).

4.2.3 Stage 3 – VA Regression

Valid pairs proceed to the Regressor. The input is extended to include the opinion:

$$Input = [CLS] \oplus A \oplus [SEP] O \oplus : \oplus S \oplus [SEP] \tag{8}$$

The same regression head and training procedure from ST1 is applied but the opinion is also added here.

4.3 Subtask 3: Dimensional Quadruplet Extraction (DimASQP)

ST3 adds category prediction to ST2. After valid aspect-opinion pairs are identified, each pair must be assigned a category label in `Entity#Attribute` form from a predefined set of 196 labels before regression.

Rather than flat 196-class classification, we decompose prediction into two sequential steps to reduce confusion between entity-sharing attributes:

1. **Entity Model.** Input: [CLS] S' [SEP] *aspect* [SEP]
   Predicts the entity component $\hat{e}$ (e.g. FOOD, SERVICE).

2. **Attribute Model.** Input: [CLS] S' [SEP] *aspect* [SEP] $\hat{e}$ [SEP]
   Conditions on $\hat{e}$ to predict the attribute $\hat{a}$ (e.g. QUALITY, PRICE).

The final category is assembled as $C = \hat{e}\#\hat{a}$. Conditioning the attribute prediction on the entity reduces cross-entity attribute confusion—for example distinguishing `FOOD#STYLE` from `DRINKS#STYLE`—and allows each attribute classifier to be trained on a more balanced entity-specific subset of the label space.

After category assignment, the validated quadruplet proceeds to the same Regressor used in ST2, with the same input format and training procedure.

4.4 Data Augmentation

Details about ablation experiments conducted using these augmentation techniques can be found in the B and B.1 sections.

SIGHAN-2024 Integration. We convert the SIGHAN-2024 Chinese ABSA dataset (Xu et al., 2024) from its parallel-list format into DimABSA quadruplets. This provides high-quality in-domain Chinese training data that significantly boosts performance on Chinese language-domain pairs.

LLM-Based Synthetic Generation. Using Gemini 2.5 Flash with batch prompting, we generate synthetic instances that preserve original aspect terms and category labels while diversifying sentence text, opinion expressions, and VA scores—targeting underrepresented VA regions.

Mention Replacement Following the mention replacement strategy proposed by (Dai and Adel, 2020), we constructed independent banks of aspect and opinion spans from the training data. For augmentation, an aspect A in a sentence is randomly replaced with another aspect A' sampled uniformly from the aspect bank. This augmentation was done for training the extractor model as it is essentially a NER task and the strategy proposed in the paper targets NER specifically.

Contextualized Word Replacement via Masked Language Modelling To introduce syntactic diversity while maintaining strict label integrity, we employed this Contextualized augmentation strategy as suggested by (Wu et al., 2019) and (Kumar et al., 2020). Given a sentence S, we first align character level boundaries for the aspect and opinion labels using offset mapping, these mapped tokens are strictly frozen. Meanwhile, 25% of the remaining tokens are masked. Finally, the pretrained XLM-RoBERTa model evaluates bidirectional context of the sentence and predicts logically consistent

replacements for the masked positions. This results in newly synthesized sentences with consistent aspect-opinion pairs.

5 Experiment Setup

5.1 ST1: Regression Setup

All experiments are implemented using the HuggingFace Transformers library. Each backbone model is fine-tuned under identical conditions to ensure a fair comparison.

Training is performed using the AdamW optimizer with a learning rate of 2×10^{-5}. The maximum input length is set to 128 tokens. Early stopping is applied based on the RMSE. All models are trained using the same batch size of 16, number of epochs 10, and identical optimization settings to maintain consistency across comparisons. The internal dropout is also disabled, as it improves regression performance as suggested in this paper (Xu et al., 2024). These hyperparameter values follow standard fine-tuning practice for encoder-based transformers on sequence labeling and regression tasks (Devlin et al., 2019), and were validated against the development split using early stopping. No extensive grid search was conducted due to computational constraints; the learning rate of 2×10^{-5} was chosen as it consistently avoids instability in multilingual transformer fine-tuning across prior work.

The dataset includes multiple languages: English, Japanese, Russian, Tatar, Ukrainian, and Chinese, and multiple domains: Laptop, Restaurant, Finance, and Hotel (Yu et al., 2026). Training is conducted in a multilingual setting by combining all available training data.

Evaluation follows the official Task 1 metrics:

- Pearson Correlation Coefficient (PCC) for Valence
- PCC for Arousal
- Root Mean Squared Error (RMSE)

Since the official ranking emphasizes RMSE, model selection is primarily based on average RMSE across languages and domains.

The final system, incorporating label rescaling and Gaussian noise regularization, achieves the highest results.

5.2 ST2: Extraction and Pairing Setup

Consistent with the hyperparameter selection strategy established for ST1, each backbone of the extractor and pairing model is trained using AdamW optimizer with a learning rate of 2×10^{-5}, the extractor is trained for 15 and pairer for 8 epochs, the batch size is kept 16 for both. Negative-pair construction is applied, mixing hard negatives and soft negatives. The Regressor uses the ST1 final configuration. Early stopping is applied based on the validation F1 for both models. The maximum input length is set to 128 tokens.

The models with the best overall F1 scores on the test split are selected and inference is made by pipelining all the models together to generate triplets. The metric used for triplets is cF1, whereas the metric used for the seperate training of extraction and pairing models is F1 score.

5.3 ST3: Full Pipeline Setup

Subtask 3 (ST3) of DimABSA 2026 is addressed using a two-stage classification pipeline consisting of an **Entity Classifier** and an **Attribute Classifier**. Both classifiers are initialized with `XLM-RoBERTa-base` models and fine-tuned separately for 10 epochs using cross-entropy loss. The Attribute Classifier is trained per-entity to leverage the balanced entity-conditional label distribution.

Consistent with the overarching hyperparameter strategy detailed in Section 5.1, both models are trained using the AdamW optimizer with a learning rate of 2×10^{-5} and a batch size of 16. Training is conducted with early stopping applied based on validation performance. The entity and attribute models use DataLoader objects with dynamic padding and batch size consistent across experiments.

Evaluation metrics include Continuous F1-score (cF1), Continuous Precision (cPrec), and Continuous Recall (cRec). The pipeline is evaluated end-to-end by first predicting the entity and then using it as context for attribute prediction.

6 Results

Table 1 presents the official evaluation results ST1 for our system across all languages and domains. Chinese domains achieve the lowest RMSE overall, benefiting from additional training data, which boosts system performance. However, this focus slightly impacts English, which attains higher RMSE. Japanese performs reliably, demonstrating stable predictions even with limited data. Low re-

Language	Domain	RMSE	Rank
English	Laptop	1.3654	13
English	Restaurant	1.3059	18
Japanese	Finance	0.8907	8
Japanese	Hotel	0.6680	9
Russian	Restaurant	1.4344	6
Tatar	Restaurant	1.6041	2
Ukrainian	Restaurant	1.4661	7
Chinese	Restaurant	0.9898	17
Chinese	Finance	0.5333	8
Chinese	Laptop	0.7311	13

Table 1: SubTask 1

Language	Domain	cF1	Rank
English	Laptop	0.4770	15
English	Restaurant	0.5202	17
Japanese	Hotel	0.3311	13
Russian	Restaurant	0.5492	6
Tatar	Restaurant	0.4839	5
Ukrainian	Restaurant	0.5324	6
Chinese	Restaurant	0.4622	11
Chinese	Laptop	0.4159	11

Table 2: SubTask 2

Language	Domain	cF1	Rank
English	Laptop	0.0000	18
English	Restaurant	0.0000	16
Japanese	Hotel	0.1853	9
Russian	Restaurant	0.2963	11
Tatar	Restaurant	0.2500	11
Ukrainian	Restaurant	0.2938	12
Chinese	Restaurant	0.3139	10
Chinese	Laptop	0.4199	10

Table 3: SubTask 3

source languages such as Tatar (1.6041, rank 2), Russian (1.4344, rank 6), and Ukrainian (1.4661, rank 7) remain challenging, yet they achieve competitive ranks among all languages. Overall, RMSE and rank trends highlight the trade-off between leveraging rich high-resource data and maintaining consistent performance across diverse languages and domains.

Table 2 and 3 present ST2 and ST3 results respectively, where Russian Restaurant achieves the highest cF1, while English Restaurant and Ukrainian restaurant have comparable performance. For English, both Laptop and Restaurant domains in ST3 yielded a cF1 of 0.0000. Upon reviewing our system outputs, predictions were successfully generated by the pipeline; however, we hypothesize this catastrophic drop stems from a silent serialization or formatting mismatch specific to the English submission file, which caused the evaluation script to reject the predicted quadruplets. This is supported by the fact that ST2 English performance (cF1 $\approx$ 0.50) was highly competitive, isolating the failure to the final file generation stage rather than an architectural collapse of the classification models. Apart from this, Chinese performance is lower on average (cF1 $\approx$ 0.4390) due to complex BIO segmentation, whereas low resource languages such as Tatar (0.4839, rank 5), Russian (0.5492, rank 6), and Ukrainian (0.5324, rank 6) remain competitive through cross-lingual transfer. Domains with more consistent and explicit aspects, such as Chinese Laptop (0.4199, rank 10) and Restaurant (0.3139, rank 10), are less affected, highlighting that ST2 predictions directly mediate ST3 outcomes in the pipeline, and cascading errors propagate less severely for stable domains.

6.1 Discussion

As detailed in Appendix B and B.1, our ablation studies reveal critical insights into the pipeline's behavior. First, we observed a "cross-lingual paradox": integrating external Chinese data (SIGHAN-2024) slightly degraded native Chinese performance but acted as a powerful cross-lingual regularizer, significantly boosting zero-shot alignment in other languages. Second, experiments demonstrated that Contextual Masked Language Modeling (MLM) actively harms sequence labeling performance, likely because it perturbs fragile local syntactic cues necessary for precise BIO boundary detection. Consequently, semantic substitution via Mention Replacement proved to be a far more robust augmentation strategy.

Limitations

While our modular pipeline achieves strong cross-lingual performance, the sequential four-stage architecture introduces computational overhead and the risk of cascading errors. Misclassifications in the Stage 1 sequence tagger irrecoverably bound the performance of the subsequent pairing and regression stages. Furthermore, despite targeted data augmentation and mention replacement strategies, the system continues to struggle with extreme minority categories and severe label imbalance, particularly for data points occupying the extreme edges of the valence-arousal affective space.

Acknowledgments

We thank the organizers of SemEval-2026 Task 3 for curating the DimABSA dataset and providing the comprehensive evaluation framework (Yu et al., 2026) that made this shared task possible. We also gratefully acknowledge the support for computational resources from Google Colab and Kaggle which enabled our fine-tuning processes.

References

Bagus Tris Atmaja and Masato Akagi. 2021. Evaluation of error and correlation-based loss functions for multitask learning dimensional speech emotion recognition. *Journal of Physics: Conference Series*, 1896(1):012004.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Aathira E. G. P., Adarsh Firoz, Meera Murali, Suraj Rajamanickam, and Balasubramanian Palani. 2025. Hermes@DravidianLangTech 2025: Sentiment analysis of dravidian languages using XLM-RoBERTa. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 330–334, Albuquerque, New Mexico. Association for Computational Linguistics.

Neal B. Gallagher. 2020. Savitzky-Golay smoothing and differentiation filter. Technical report, Eigenvector Research, Inc.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. 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. 2026. Dimabsa: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. *Preprint*, arXiv:2601.23022.

Lung-Hao Lee, Liang-Chih Yu, Suge Wang, and Jian Liao. 2024. Overview of the SIGHAN 2024 shared task for Chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 165–174, Bangkok, Thailand. Association for Computational Linguistics.

Bin Liang, Wangda Luo, Xiang Li, Lin Gui, Min Yang, Xiaoqi Yu, and Ruifeng Xu. 2021. Enhancing aspect-based sentiment analysis with supervised contrastive learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3242–3247, New York, NY, USA. Association for Computing Machinery.

Hao Lu, Jian Yang, Changzhi Hu, and Wei Fang. 2021. One for "All": A unified model for fine-grained sentiment analysis under three tasks. *PeerJ Computer Science*, 7:e816.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Lima Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Dario Mario Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 others. 2025. SemEval-2025 task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2558–2569, Vienna, Austria. Association for Computational Linguistics.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, and 1 others. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuanhe Tian, Guimin Chen, and Yan Song. 2021. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2910–2922, Online. Association for Computational Linguistics.

J. Verma, A. Gupta, S. Garg, and P. Kumar. 2024. Aspect-based sentiment analysis: An extensive study of techniques, challenges, and applications. In *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 576–581.

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, Berlin, Germany. Association for Computational Linguistics.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science – ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV*, page 84–95, Berlin, Heidelberg. Springer-Verlag.

H. Xu, D. Zhang, Y. Zhang, and R. Xu. 2024. HITSZ-HLT at SIGHAN-2024 dimABSA task: Integrating BERT and LLM for chinese dimensional aspect-based sentiment analysis. In *Proceedings of the 10th SIGHAN Workshop on Chinese 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.

A Dataset Distribution

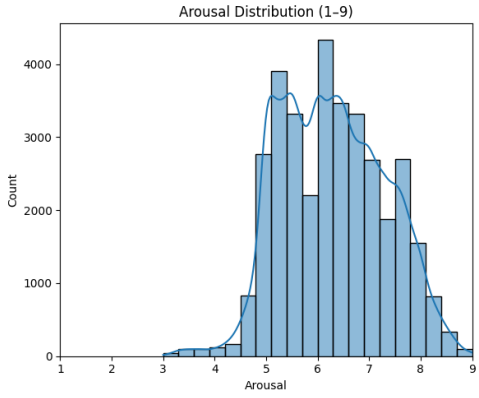

Figure 1: Arousal distribution.

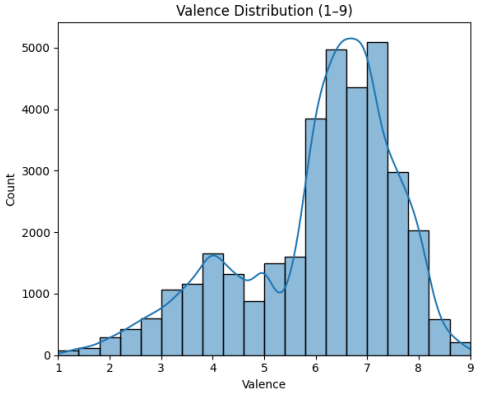

Figure 2: Valence distribution.

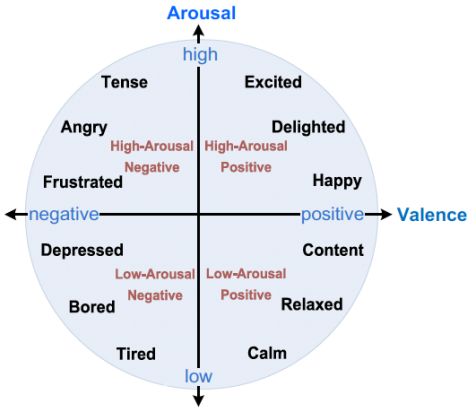

Figure 3: Valence-Arousal spatial representation.

Table 4: Domains and data point counts per language across tasks.

Language	**Task 1**	**Task 2**	**Task 3**
English	Restaurant: 2284 Laptop: 4076	Restaurant: 2284 Laptop: 4076	Restaurant: 2284 Laptop: 4076
Japanese	Hotel: 1600 Finance: 1024	Hotel: 1600	Hotel: 1600
Russian	Restaurant: 1240	Restaurant: 1240	Restaurant: 1240
Tatar	Restaurant: 1240	Restaurant: 1240	Restaurant: 1240
Ukrainian	Restaurant: 1240	Restaurant: 1240	Restaurant: 1240
Chinese	Restaurant: 6050 Laptop: 3490 Finance: 1000	Restaurant: 6050 Laptop: 3490	Restaurant: 6050 Laptop: 3490

B Subtask 1: Extended Ablation and Model Selection

To empirically justify the architectural decisions detailed in the methodology, we present an exhaustive domain-level ablation study in Table 5. It is critical to note that all benchmarking and hyperparameter tuning presented in this section were conducted strictly on the development data split provided by the organizers.

The evaluation tracks the primary ranking metric, Root Mean Squared Error (RMSE), across five critical phases of our pipeline development:

1. **Baseline (mDeBERTa):** Established as the optimal backbone after initial testing against mBERT and XLM-RoBERTa variants.

2. **+ SIGHAN-2024[1] (The Cross-Lingual Paradox):** The integration of external Chinese DimABSA data yielded a counter-intuitive phenomenon known as negative interference. While the influx of Chinese data slightly degraded native Chinese performance (e.g., Chinese Restaurant RMSE increased from 0.7619 to 0.7942), it acted as a powerful cross-lingual regularizer. The enhanced underlying sentiment geometry massively boosted zero-shot alignment in other languages, reducing English Laptop RMSE from 0.9661 to 0.9150 and Russian Restaurant from 1.2751 to 1.1958.

3. **+ CCC Loss:** Attempting to optimize for the Concordance Correlation Coefficient failed to yield universal improvements, degrading performance in highly skewed domains.

4. **+ 2D VA Binning:** This augmentation strategy caused a catastrophic failure in English continuous tracking, more than doubling the error rate.

5. **+ Target Rescale & Gaussian Noise (Final):** The mathematical regularization of scaling labels to $[-1, 1]$ coupled with noise injection proved to be the most robust global configuration, correcting the Chinese degradation caused by the SIGHAN integration while preserving the cross-lingual gains.

Table 5: Domain-level RMSE ablation study evaluated on the development split. Lower is better. Bold text indicates the optimal configuration.

Lang	Domain	Base	+SIGHAN	+CCC	+2D Bin	Final
Eng	Laptop	0.9661	0.9150	0.9821	2.2242	**0.9542**
Eng	Rest.	0.9352	0.9215	0.8721	1.6634	**0.9111**
Jpn	Finance	0.8353	0.8287	0.8607	0.9099	**0.7520**
Jpn	Hotel	0.9452	0.9560	1.0009	0.9648	**0.9484**
Rus	Rest.	1.2751	1.1958	1.2268	1.2702	**1.2554**
Tat	Rest.	1.3481	1.3308	1.3469	1.3523	**1.2840**
Ukr	Rest.	1.3132	1.2706	1.2820	1.3019	**1.3166**
Zho	Finance	0.4944	0.5186	0.5345	0.5105	**0.4792**
Zho	Laptop	0.6979	0.7309	0.7289	0.7317	**0.7114**
Zho	Rest.	0.7619	0.7942	0.7338	0.7888	**0.7536**
Avg	**All**	0.9572	0.9462	0.9569	1.1718	**0.9366**

[1]The external Chinese DimABSA dataset utilized in our experiments is accessible at: https://github.com/NYCU-NLP/SIGHAN2024-dimABSA/blob/main/DataSets/dimABSA2024/Simplified

B.1 Subtask 2: Extractor Ablation and Data Strategies

The aspect and opinion extraction module functions as the foundational sequence tagger for both ST2 and ST3. To optimize its boundary detection capabilities, we conducted a rigorous series of local ablation studies, presented in Table 6. All metrics reflect the macro F1-score evaluated on a 10% of data seperated from training data as a test set.

The empirical data reveals three critical insights regarding sequence labeling for DimABSA:

1. **Backbone Superiority:** Consistent with Subtask 1, mDeBERTa-base (0.8034) outperformed both XLM-RoBERTa-base (0.7908) and mBERT-base (0.7709) on the original dataset, proving its superior contextual representation for token-level classification.

2. **The Contextual Perturbation Failure:** While we hypothesized that Entity-Preserving Contextualized Masked Language Modeling (MLM) would improve syntactic robustness, experiments on a trilingual subset (English, Chinese, Russian) proved otherwise. Even with strict entity freezing, the synthetic contextual noise degraded the baseline F1 from 0.7846 to 0.7693. We hypothesize that perturbing the immediate grammatical neighborhood around an aspect destroys the local syntactic cues the model relies on for BIO boundary detection. Consequently, this strategy was abandoned.

3. **Mention Replacement and SIGHAN Synergy:** Semantic substitution via Mention Replacement (MR) successfully forced the model to learn structural patterns rather than memorizing entity lexicons, raising the F1 score to 0.8110. However, experiments revealed that the highest validation F1 score (0.8531) was achieved by integrating the high-quality, human-annotated SIGHAN-2024 dataset without synthetic MR augmentation. Combining MR with SIGHAN introduced slight noise, marginally dropping the score to 0.8488, proving that pure human-annotated data remains the most robust cross-lingual anchor.

Table 6: Subtask 2 Extractor ablation study. Metrics reflect the local validation standard macro-F1 score. **Note: As this ablation isolates the discrete BIO sequence tagger prior to the regression stage, standard F1 is reported rather than the pipeline-level (cF1)**. Results demonstrate the negative impact of Contextual MLM and that integrating human-annotated SIGHAN data without synthetic Mention Replacement yields the optimal configuration. Bold text indicates the optimal configuration.

Model Architecture	Data Configuration	Augmentation Strategy	Val F1 ↑
Phase 1: Backbone Selection			
mBERT-base	Original	None	0.7709
XLM-RoBERTa-base	Original	None	0.7908
mDeBERTa-base	Original	None	0.8034
Phase 2: Trilingual Subset Analysis (Eng, Zho, Rus)			
mDeBERTa-base	Subset	None	0.7846
mDeBERTa-base	Subset	Contextual MLM (XLM-R)	0.7693
mDeBERTa-base	Subset + SIGHAN	None	0.8127
mDeBERTa-base	Subset + SIGHAN	Contextual MLM (XLM-R)	0.8091
Phase 3: Final Augmentation and Integration			
mDeBERTa-base	Original	Mention Replacement (MR)	0.8110
XLM-RoBERTa-base	Original + SIGHAN	Human Annotated	0.8444
mDeBERTa-base	Original + SIGHAN	MR + SIGHAN	0.8488
mDeBERTa-base	**Original + SIGHAN**	**Human Annotated (Final)**	**0.8531**