DimABSA: Building Multilingual and Multidomain Datasets for Dimensional Aspect-Based Sentiment Analysis

Lung-Hao Lee[1,*], **Liang-Chih Yu**[2,*], **Natalia Loukachevitch**[3], **Ilseyar Alimova**[4]
Alexander Panchenko[4,5], **Tzu-Mi Lin**[1], **Zhe-Yu Xu**[1], **Jian-Yu Zhou**[1], **Guangmin Zheng**[6]
Jin Wang[6], **Sharanya Awasthi**[7], **Jonas Becker**[8], **Jan Philip Wahle**[8], **Terry Ruas**[8]
Shamsuddeen Hassan Muhammad[9], **Saif M. Mohammad**[10]

[1]National Yang Ming Chiao Tung University, [2]Yuan Ze University
[3]Moscow State University, [4]Skoltech, [5]AIRI, [6]Yunnan University, [7]University of Cincinnati
[8]University of Göttingen, [9]Imperial College London, [10]National Research Council Canada

***Contact:** lhlee@nycu.edu.tw, lcyu@saturn.yzu.edu.tw*

Abstract

Aspect-Based Sentiment Analysis (ABSA) focuses on extracting sentiment at a fine-grained aspect level and has been widely applied across real-world domains. However, existing ABSA research relies on coarse-grained categorical labels (e.g., positive, negative), which limits its ability to capture nuanced affective states. To address this limitation, we adopt a dimensional approach that represents sentiment with continuous valence–arousal (VA) scores, enabling fine-grained analysis at both the aspect and sentiment levels. To this end, we introduce DimABSA, the first multilingual, dimensional ABSA resource annotated with both traditional ABSA elements (aspect terms, aspect categories, and opinion terms) and newly introduced VA scores. This resource contains 76,958 aspect instances across 42,590 sentences, spanning six languages and four domains. We further introduce three subtasks that combine VA scores with different ABSA elements, providing a bridge from traditional ABSA to dimensional ABSA. Given that these subtasks involve both categorical and continuous outputs, we propose a new unified metric, continuous F1 (cF1), which incorporates VA prediction error into standard F1. We provide a comprehensive benchmark using both prompted and fine-tuned large language models across all subtasks. Our results show that DimABSA is a challenging benchmark and provides a foundation for advancing multilingual dimensional ABSA. We publicly released the DimABSA dataset, which was used for Track A of SemEval-2026 Task 3, attracting over 300 participants.[1]

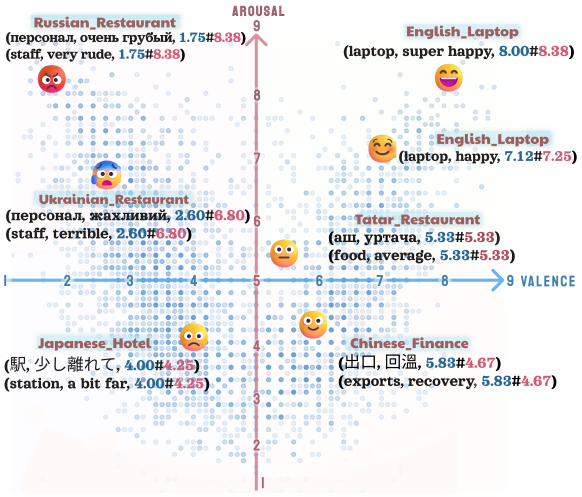


Figure 1: **Valence–Arousal (VA) space**. Illustrated examples of aspect instances across multiple languages and domains. Selected instances are visualized as triplets (aspect, opinion, paired VA score). Blue dots show a U-shaped distribution of VA scores.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) aims to analyze opinion structures at the fine-grained aspect level, rather than at the sentence or document level (Pontiki et al., 2014, 2015, 2016). It is formulated as the individual or joint extraction of sentiment elements, including aspect terms, aspect categories, opinion terms, and sentiment polarity. For example, given the sentence *I am happy with the laptop.*, an ABSA system is expected to extract the explicit aspect term *laptop*, the opinion term *happy*, assign the aspect category `LAPTOP#GENERAL` from a predefined set, and predict `Positive` sentiment polarity. These elements can form different

[1]https://github.com/DimABSA/DimABSA2026

structural combinations (e.g., aspect pairs, triplets, and quadruplets), enabling fine-grained analysis that provides deeper insights into user opinions across various applications. (D'Aniello et al., 2022; Zhang et al., 2023; Hua et al., 2024).

Despite substantial progress in ABSA, existing work remains limited by its reliance on coarse-grained, categorical sentiment labels (e.g., positive, negative, neutral). This categorical approach fails to capture subtle affective differences, such as those conveyed by variations in lexical intensity (e.g., *good*, *excellent*) or by sentiment-modifying adverbs (e.g., *slightly*, *very*).

To address these limitations, we draw inspiration from affective science, which provides nuanced sentiment representation frameworks (Russell, 1980, 2003) that model sentiment along real-valued dimensions of valence (negative–positive) and arousal (sluggish–excited). As illustrated by the aspect instances in Figure 1, this dimensional representation captures diverse affective states across languages and domains, rather than being limited to categorical labels. Moreover, even affective states that share the same polarity (e.g., *super happy* vs. *recovery*, and *very rude* vs *a bit far*) can be distinguished through their fine-grained valence–arousal (VA) scores. This expressive capability further extends ABSA to enable fine-grained analysis at both the aspect and sentiment levels. The VA framework has been used to support various applications, such as stance detection (Upadhyaya et al., 2023), misinformation identification (Liu et al., 2024), identifying biosocial markers for mental health (Teodorescu et al., 2023), dialogue generation (Wen et al., 2024), hate speech detection (Sariyanto et al., 2025), and emotion dynamics analysis (Hipson and Mohammad, 2021; Teodorescu et al., 2023; Wang et al., 2025).

To move beyond the categorical sentiment representation, we introduce the DimABSA, the first multilingual dimensional ABSA resource manually annotated with both traditional ABSA annotations (aspect terms, aspect categories, opinion terms) and newly introduced VA scores. DimABSA comprises 76,958 aspect instances across 42,590 sentences, covering six languages (Chinese, English, Japanese, Russian, Tatar, and Ukrainian) and four domains, including customer reviews (Hotel, Laptop, Restaurant) and financial reporting (Finance).

To explore the utility of dimensional sentiment in ABSA, we introduce three novel subtasks by integrating continuous VA scores with traditional ABSA components, each corresponding to distinct application scenarios, as shown in Table 1.

Subtask (Input → Output)	Prediction Type	Metric
DimASR (text + aspects → V#A)	Regression	RMSE
DimASTE (text → A, O, V#A)	Extraction, Regression	cF1
DimASQP (text → A, C, O, V#A)	Extraction, Classification, Regression	cF1

Table 1: **DimABSA subtasks** with input–output structure, task type, and metrics.

- **Subtask 1 - Dimensional Aspect Sentiment Regression (DimASR):** A regression task that predicts VA scores for each aspect in a sentence.
- **Subtask 2 - Dimensional Aspect Sentiment Triplet Extraction (DimASTE):** A hybrid task that jointly extracts aspect (A) and opinion (O) terms and predicts their associated VA scores.
- **Subtask 3 - Dimensional Aspect Sentiment Quadruplet Prediction (DimASQP):** An extension of DimASTE that additionally includes aspect category (C) classification, thus combining extraction, classification, and regression.

As DimASTE and DimASQP are hybrid tasks requiring both categorical prediction and VA regression, the standard F1 score, widely used in ABSA, is insufficient to jointly assess these components. Thus, we propose a unified metric: the *continuous F1 (cF1)* score, which incorporates VA prediction error into the F1 formulation. Experiments on the DimABSA subtasks with various large language models (LLMs) show that closed-source LLMs provide data-efficient baselines, while fine-tuned large-scale LLMs (70B and 120B) further improve performance. Both model families show performance variability across languages and domains, and still face challenges, particularly for low-resource languages.

Our contributions are summarized as follows:

- We introduce the DimABSA datasets with continuous VA scores, enabling fine-grained analysis across languages and domains.
- We design three subtasks to facilitate the transition from categorical to dimensional ABSA.
- We propose the cF1 score, a unified metric that integrates categorical and continuous evaluation into a single measure.

The DimABSA datasets were used for Track A of

SemEval-2026 Task 3, attracting over 300 participants (Yu et al., 2026).

2 The DimABSA Construction

We collect data from six languages and four domains. Each instance is annotated with ABSA components and VA ratings, following the specifications of the three target subtasks: DimASR, DimASTE, and DimASQP.

2.1 Data Collection

As shown in Table 2, we used four widely studied ABSA domains: restaurant (`rest`), laptop (`lap`), hotel (`hot`), and finance (`fin`). For these domains, we consider four high-resource languages: English (`eng`), Japanese (`jpn`), Russian (`rus`), and Chinese (`zho`). To extend coverage to under-resourced languages, we translated the Russian data into Tatar (`tar`) and Ukrainian (`ukr`), leveraging their linguistic and cultural relatedness. While this approach has inherent limitations, it enables the inclusion of languages that would otherwise be unavailable. All translations were reviewed by native speakers and manually corrected to ensure the quality of ABSA-relevant structural annotations.

We collect data from multiple sources, including existing labeled ABSA datasets and newly curated unlabeled data. The existing labeled datasets are used solely for training, while the newly curated data are annotated and split into training, development, and test sets. The data sources for each language are described below.

English. We use the training split of the ACOS dataset (Cai et al., 2021), manually annotating the restaurant and laptop quadruplets in this split with VA scores to replace the original sentiment polarity labels. For the development and test sets, we collect restaurant reviews from Yelp Open Dataset[2] and laptop reviews from Amazon Reviews 2023 (Hou et al., 2024).

Japanese. For the finance domain, the training set is sampled from the chABSA dataset.[3] We manually annotate VA scores for each aspect in these samples, replacing the original sentiment polarity labels. The development and test sets are collected from the same EDINET[4] sources as chABSA, with samples involving the same companies removed to avoid overlap. For the hotel domain, we crawl reviews from Rakuten Travel.[5]

Russian. The SemEval-2016 restaurant review dataset (Pontiki et al., 2016) serves as the data source. The labeled portion contains annotated aspects, their categories, and sentiment polarity. We annotate opinion terms and VA values manually. The unlabeled portion of reviews is used for the development and test sets.

Tatar. We automatically translate the Russian dataset into Tatar using Yandex Translate. The resulting translations are reviewed by a native speaker. Manual corrections were applied to 45.5% of the translated instances to assure data quality.

Ukrainian. Similar to Tatar, we translate the Russian dataset into Ukrainian and adopt the same review procedure, with 35.6% of the instances corrected by native speakers.

Chinese. For the restaurant domain, we use the SIGHAN-2024 dataset (Lee et al., 2024) for training, and construct the development and test sets from Google Reviews[6] and the PTT platform[7]. For the laptop domain, we crawl reviews from Mobile01[8]. For the finance domain, we collect annual reports of Taiwanese companies from MOPS[9].

We preprocess all collected data by removing duplicate entries, invisible characters, and corrupted encodings. The cleaned texts are then segmented into sentences and prepared for annotation.

2.2 Annotation Process

The DimABSA datasets comprise four core components: three categorical ABSA elements and a paired valence–arousal (VA) score. Each component is defined as follows:

Aspect Term (A): a word or phrase indicating an opinion target, such as *service*, *screen*, *profit*.

Aspect Category (C): a predefined Entity#Attribute label associated with an aspect term (e.g., `FOOD#QUALITY`, `SERVICE#GENERAL`) (Pontiki et al., 2015, 2016).

Opinion Term (O): a sentiment-bearing word or phrase associated with a specific aspect term. The annotation further includes sentiment modifiers to support fine-grained sentiment representation (e.g., *very good*, *extremely bad*, *a little slow*).

[2] https://business.yelp.com/data/resources/open-dataset
[3] https://github.com/chakki-works/chABSA-dataset
[4] https://disclosure2.edinet-fsa.go.jp
[5] https://travel.rakuten.co.jp
[6] https://customerreviews.google.com
[7] https://www.pttweb.cc
[8] https://www.mobile01.com/category.php?id=2
[9] https://emops.twse.com.tw/server-java/t58query

Dataset	Source(s)	Subtask	Train sent./tuple	Dev sent./tuple	Test sent./tuple	Total sent./tuple	F1 (A) / (A,C) / (A,C,O)	RMSE (V#A)
eng-rest	ACOS	ST1	2284 / 3659	200 / 340	1000 / 1504	3484 / 5503	0.892 / 0.778 / 0.703	1.542#1.883
	Yelp Open Dataset	ST2–3		200 / 408	1000 / 2129	3484 / 6196		
eng-lap	ACOS	ST1	4076 / 5773	200 / 275	1000 / 1421	5276 / 7469	0.857 / 0.767 / 0.635	1.547#2.291
	Amazon Reviews 2023	ST2–3		200 / 317	1000 / 1975	5276 / 8065		
jpn-hot	Rakuten Travel	ST1	1600 / 2846	200 / 284	800 / 1092	2600 / 4222	0.824 / 0.754 / 0.643	0.919#1.018
		ST2–3		200 / 364	800 / 1443	2600 / 4653		
jpn-fin	chABSA; EDINET	ST1	1024 / 1672	200 / 319	800 / 1302	2024 / 3293	0.761 / – / –	0.814#0.755
rus-rest	SemEval'16	ST1	1240 / 2487	56 / 81	1072 / 1637	2368 / 4205	0.865 / 0.748 / 0.726	1.030#2.041
		ST2–3		48 / 102	630 / 1310	1918 / 3899		
tat-rest	SemEval'16 (MT)	ST1	1240 / 2487	56 / 81	1072 / 1637	2368 / 4205	0.865 / 0.748 / 0.726	1.030#2.041
		ST2–3		48 / 102	630 / 1310	1918 / 3899		
ukr-rest	SemEval'16 (MT)	ST1	1240 / 2487	56 / 81	1072 / 1637	2368 / 4205	0.865 / 0.748 / 0.726	1.030#2.041
		ST2–3		48 / 102	630 / 1310	1918 / 3899		
zho-rest	SIGHAN-2024	ST1	6050 / 8523	225 / 416	1000 / 1929	7275 / 10868	0.826 / 0.713 / 0.650	0.591#0.758
	Google Reviews; PTT	ST2–3		300 / 761	1000 / 2861	7350 / 12145		
zho-lap	Mobile01	ST1	3490 / 6502	261 / 431	1000 / 2662	4751 / 9595	0.803 / 0.723 / 0.602	0.827#1.074
		ST2–3		300 / 551	1000 / 2798	4790 / 9851		
zho-fin	MOPS	ST1	1000 / 2633	200 / 563	842 / 2354	2042 / 5550	0.604 / – / –	0.910#0.810

Table 2: **DimABSA dataset overview.** For each dataset (language–domain), we report the corresponding source(s), subtask type (ST1 vs. ST2–3), and the numbers of sentences and tuples in the train/dev/test splits. Tuple agreement is evaluated using F1, and valence–arousal agreement is evaluated using RMSE. Dataset examples are provided in Appendix A.

Valence-Arousal (VA): Both the valence and arousal are rated on a 1–9 scale, where 1 denotes extreme negative valence or low arousal, 9 denotes extreme positive valence or high arousal, and 5 denotes neutral valence or medium arousal.

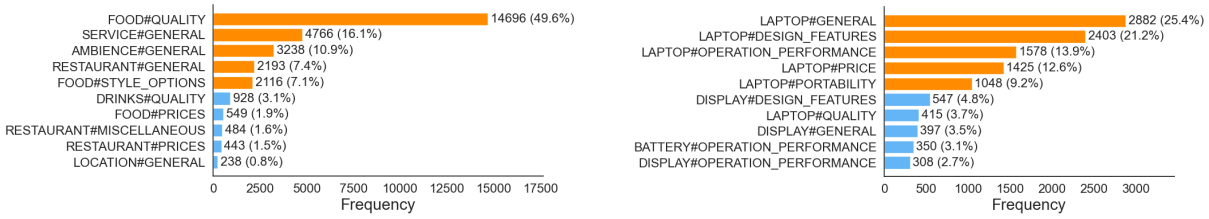
The annotated elements vary across datasets depending on the subtask setting. For finance-domain datasets focused on DimASR, we annotate `(A, VA)` pairs. For all other datasets, we annotate full `(A, C, O, VA)` quadruplets. For existing datasets, we supplement the annotations with VA scores and any missing components of the quadruplet. We construct a shared training set for all subtasks. However, we do not use a single shared development/test set across subtasks, since DimASR assumes the aspect term is given as input. Instead, we create a dedicated development/test set for DimASR, and a shared set for DimASTE and DimASQP.

The annotation process is conducted in two phases. We first extract the categorical triplet `(A, C, O)` for each quadruplet from sentences, followed by the assignment of VA scores for tuples. For tuple extraction, annotators follow guidelines with examples (see Appendix B), including extracting all valid tuples, incorporating sentiment modifiers, and ensuring exact span matching. Each sentence is annotated independently by two annotators. If both annotators agree, the tuple is accepted. Otherwise, a third annotator adjudicates the disagreement. Instances without consensus among the three annotators are discarded.

The accepted tuples are then annotated with VA ratings. Annotators are introduced to the VA dimensions with illustrative examples (see Appendix B). The annotation interface is implemented using the Self-Assessment Manikin (SAM) scale (Bradley and Lang, 1994), accompanied by VA emojis (Kutsuzawa et al., 2022) for reference. This pictorial protocol helps annotators assign VA ratings more accurately. Each tuple is annotated by five annotators and the final VA rating is obtained by discarding outlier ratings beyond the mean $\pm$ 1.5 standard deviations and averaging the remaining scores.

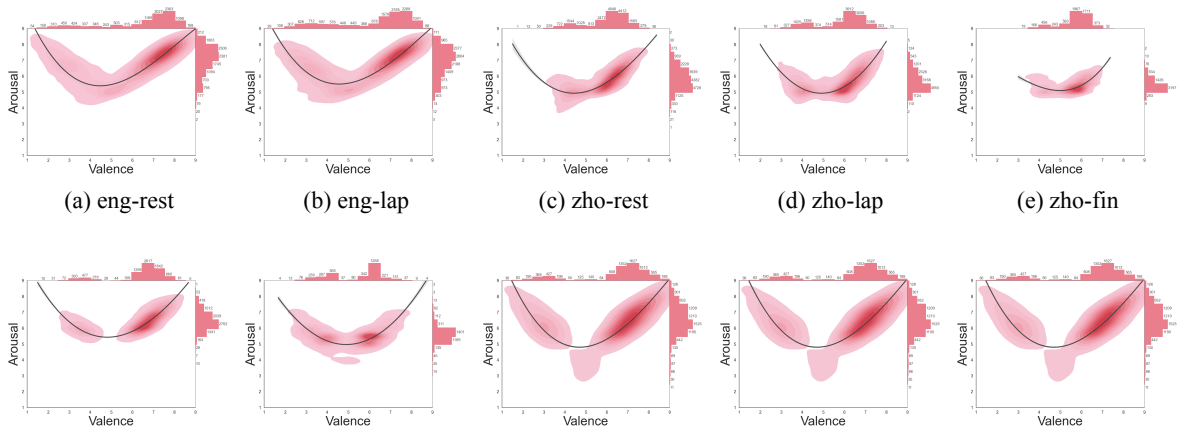
2.3 Evaluating Annotation Quality

Table 2 report tuple-level annotation agreement using the F1 score, following prior work (Chebolu et al., 2024; Wu et al., 2025). The F1 score is computed between two annotators by treating one as the predicted output and the other as the gold stan-



(a) Restaurant (b) Laptop

Figure 2: **Aspect-Category distributions** for the restaurant and laptop domains aggregated across languages.



(a) eng-rest (b) eng-lap (c) zho-rest (d) zho-lap (e) zho-fin

(f) jpn-hot (g) jpn-fin (h) rus-rest (i) tat-rest (j) ukr-rest

Figure 3: **Joint Valence-Arousal (VA) distributions** across various languages in the DimABSA dataset.

dard. For VA agreement, we report Root Mean Square Error (RMSE) separately for valence and arousal. RMSE is calculated by comparing each annotator's rating against the mean of the five annotators, and the final agreement score is the average RMSE across all annotators.

Tuple-level F1 scores show that annotation agreement tends to decrease with increasing tuple complexity. Agreement is highest for single aspect terms, while lower for finance-domain datasets than for customer reviews, likely reflecting domain-specific ambiguity. Agreement on `(A, O)` pairs remains relatively consistent across datasets. In contrast, agreement on `(A, C, O)` triplets varies more widely, likely due to the larger and more complex aspect category sets.

For VA annotation, the Japanese and Chinese datasets yield lower RMSE, especially for arousal, likely due to the more formal nature of the financial texts and potential cultural differences in emotional expression. Additionally, annotators consistently exhibit lower agreement for arousal than for valence, confirming prior findings that arousal is harder to annotate reliably (Buechel and Hahn, 2017; Mohammad, 2018; Lee et al., 2022).

3 Dataset Analysis

We analyze the distribution of aspect categories and VA scores across the datasets.

Long-tailed Aspect Category Distribution. As shown in Figure 2, the distributions across domains are highly imbalanced and exhibit a long-tailed pattern, consistent with findings from previous ABSA tasks (Pontiki et al., 2015, 2016). For the restaurant domain, the 18 categories are highly concentrated: the top five categories account for 97.89% of all aspect instances, while the remaining 13 categories contribute only 2.11%. In contrast, the laptop domain includes 148 categories, with the top five and ten categories covering 37.20% and 54.18% of instances, respectively. Similarly, in the hotel domain, the top five categories account for 51% of the 47 categories (see Appendix C)

U-shaped VA Distribution. In Figure 3, the red shaded regions illustrate the joint VA distribution of aspect instances, with darker colors indicating

a higher instance density. The black curve represents the relationship between valence and arousal. Overall, all datasets exhibit a broadly U-like VA pattern, indicating that arousal tends to be lowest around neutral valence and increases toward both negative and positive extremes.

Beyond this shared pattern,we observe variations across languages and domains. At the language level, Chinese and Japanese exhibit more compact VA distributions with lower dispersion, whereas other languages show greater coverage of both VA extremes. At the domain level, the finance domain shows a more constrained arousal distribution, with fewer instances at the arousal extremes. This pattern may be attributed to the formal nature of the source data, which consists primarily of financial reports.

4 Evaluation

4.1 Setups

Tasks. We evaluate the datasets on three subtasks: DimASR, DimASTE, and DimASQP. The data split sizes are detailed in Table 2.

Models. We evaluate various LLMs under the following two settings.

Zero-/few-shot learning. We employ two closed-source LLMs, GPT-5 mini (OpenAI, 2025a) and Kimi K2 Thinking (MoonshotAI, 2025). In few-shot settings, in-context examples are selected from the first k samples in the training set. We access both models via API to leverage their built-in reasoning capabilities. The prompts are provided in Appendix D.

Supervised fine-tuning. We fine-tune four publicly available LLMs: Qwen3 14B[10](Alibaba, 2025), Ministral-3 14B[11](MistralAI, 2025), Llama-3.3 70B[12](Meta, 2024), and GPT-OSS 120B[13](OpenAI, 2025b). All models are fine-tuned using 4-bit QLoRA (Dettmers et al., 2023) for parameter-efficient adaptation. The prompts are provided in Appendix E.

Implementation Details. All experiments are implemented using the PyTorch-based implementations of LLMs provided by the Hugging Face Transformers library. We adopt the AdamW optimizer with a linear learning rate scheduler. The learning rate is fixed at 2e-5. The batch size is set to 4 for LLMs. All models are fine-tuned for 5 training epochs. Experiments are conducted on NVIDIA H200 GPUs.

4.2 Metrics

Subtask 1. DimASR is evaluated by measuring the prediction error in the VA space using RMSE, defined as

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

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

Subtasks 2 & 3. DimASTE and DimASQP are evaluated using the *continuous F1 (cF1)* score, which unifies categorical and continuous evaluation. Following the standard F1, a predicted tuple is counted as a true positive (TP) only if all its categorical elements exactly match the gold annotation. This categorical TP is then extended as a *continuous true positive (cTP)* by incorporating the VA prediction error. Formally, let P denote the set of predicted triplets `(A, O, VA)` or quadruplets `(A, C, O, VA)`. For any prediction $t \in P$, its cTP is defined as

$$\text{cTP}^{(t)} = \begin{cases} 1 - \text{dist}(\text{VA}_p^{(t)}, \text{VA}_g^{(t)}), & t \in P_{\text{cat}} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $P_{cat} \subseteq P$ denotes the set of predictions in which all categorical elements, `(A, O)` for a triplet or `(A, C, O)` for a quadruplet, exactly match the gold annotation for the same sentence. The distance function is defined as

$$\text{dist}(\text{VA}_p, \text{VA}_g) = \frac{\sqrt{(V_p - V_g)^2 + (A_p - A_g)^2}}{\text{D}_{\text{max}}} \tag{3}$$

where $\text{dist}(\cdot)$ denotes the normalized Euclidean distance between the predicted $\text{VA}_p = (V_p, A_p)$ and gold $\text{VA}_g = (V_g, A_g)$ in the VA space, and $\text{D}_{\text{max}} = \sqrt{8^2 + 8^2} = \sqrt{128}$ is the maximum possible Euclidean distance in the VA space on the [1, 9] scale, ensuring that $\text{dist} \in [0, 1]$.

Building on per-prediction $\text{cTP}^{(t)}$, *cRecall* and *cPrecision* are defined as the total cTP divided by the number of gold and predicted triplets/quadruplets, respectively. The cF1 is computed as their

[10] https://huggingface.co/Qwen/Qwen3-14B

[11] https://huggingface.co/mistralai/Ministral-3-14B-Reasoning-2512

[12] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[13] https://huggingface.co/openai/gpt-oss-120b

harmonic mean. An illustrative example is given in Appendix F

4.3 Results

Table 3 summarizes the results of various LLMs across all subtasks. Closed-source models serve as data-efficient baselines, with Kimi-K2-Thinking achieving higher average performance than GPT-5 mini under both zero- and one-shot settings. Fine-tuned large-scale LLMs further improve performance on average over their closed-source counterparts. Overall, the results suggest that all DimABSA subtasks remain challenging.

DimASR. This subtask presents baseline results for sentiment regression. Closed-source models provide strong prompting baselines. The fine-tuned 14B and 70B models do not consistently outperform these baselines. In contrast, fine-tuning the 120B model yields a substantial and consistent improvement across languages, indicating that supervised adaptation can enhance the model's ability to capture nuanced affective information in multilingual settings.

DimASTE. This subtask presents baseline results for sentiment regression coupled with structured `(A, O)` extraction, revealing language-specific differences in LLM performance. Both closed-source models perform best on English, followed by Russian and Ukrainian, with a noticeable gap to other languages, even high-resource ones such as Chinese and Japanese. This suggests that the structural patterns required for extraction in English and Slavic languages are more effectively captured during LLM pretraining.

After fine-tuning, only the 70B and 120B models show consistent improvements, with particularly substantial gains in previously underperforming Chinese and Japanese. This indicates that supervised adaptation at sufficient scale helps compensate for underrepresented structural patterns during pretraining. Notably, although Tatar also benefits from fine-tuning, it remains the lowest-performing language, suggesting that LLMs still struggle with low-resource languages whose structural properties are difficult to capture.

DimASQP. This subtask highlights the impact of introducing aspect categories on extraction performance, revealing degradation across all models, with the fine-tuned 70B and 120B models showing smaller drops. At the domain level, the laptop domain shows a significant decrease, likely due to its larger and more diverse category set, whereas the restaurant domain drops less sharply, reflecting its smaller, more concentrated category set. Within the restaurant domain, English achieves the best performance across languages, while Tatar remains the lowest-performing language, a trend consistent with the DimASTE results.

4.4 Analysis

Effect of few-shot prompting on performance. As shown in Figure 4, DimASR behaves differently from DimASTE and DimASQP in the early few-shot setting. The one-shot performance improvement for DimASR indicates that even a single example provides a crucial reference for the continuous VA scale, allowing the model to calibrate its numerical output. In contrast, DimASTE and DimASQP show unstable early performance, with limited examples yielding marginal gains due to the complexity of structural induction. Across subtasks, performance generally reaches a plateau around 32 shots. Results up to 256 shots show that few-shot prompting underperforms fine-tuned benchmarks (GPT-OSS 120B for DimASR and Llama-3.3 70B for DimASTE and DimASQP) on all datasets except English. This suggests that fine-tuning remains a strong alternative to prompting.

Effect of few-shot prompting on VA distributions. As shown in Figure 5, zero shot prompting produces a random grid-like VA distribution across languages, likely due to its token-based generation without reference calibration. In contrast, the one-shot setting immediately begins to align with the gold distribution, indicating initial calibration of the VA scale. As more in-context examples are provided (up to 64 and 256 shots), the predicted distribution increasingly resemble the gold standard. However, although higher shot counts yield closer visual alignment, performance does not necessarily continue to improve once it reaches saturation.

Categorical version of the DimABSA dataset.[14] To align with traditional ABSA, we partition the DimABSA datasets into positive (V>5.5), neutral (4.5<=V<=5.5), and negative (V<4.5) subsets, and conduct polarity-based evaluation using the same LLMs (see Appendix J). The results show that the fine-grained DimABSA datasets are more challenging than their categorical counterparts, while this conversion also provides a new multilingual ABSA dataset in the

[14]The categorical version of the dataset is available at: https://github.com/DimABSA/DimABSA2026/tree/main/task-dataset/track_a/categorical.

Subtask	Dataset	Zero-Shot Learning		One-Shot Learning		Supervised Fine-Tuning			
		GPT-5 mini	Kimi K2 Thinking	GPT-5 mini	Kimi K2 Thinking	Qwen3 (14B)	Ministral-3 (14B)	Llama-3.3 (70B)	GPT-OSS (120B)
DimASR	eng-rest	2.9490	**2.3432**	2.3926	**2.1461**	2.6427	2.6316	2.5244	**1.4605**
	eng-lap	3.2115	**2.6546**	2.5637	**2.1893**	2.8089	2.6258	2.7354	**1.5269**
	jpn-hot	3.1406	**2.3294**	2.1607	**1.7553**	2.2906	2.2999	2.6255	**0.7188**
	jpn-fin	2.6760	**2.3379**	1.9243	**1.6396**	1.8964	2.0700	2.4191	**1.0188**
	rus-rest	2.5447	**2.0630**	2.0390	**1.7768**	2.1528	2.3617	2.5089	**1.4775**
	tat-rest	2.6645	**2.3636**	2.2308	**1.9380**	2.6367	3.0463	2.9165	**1.7153**
	ukr-rest	2.5628	**2.0782**	2.0438	**1.7805**	2.2121	2.4592	2.5709	**1.5166**
	zho-rest	2.7125	**2.2623**	2.2467	**1.8959**	2.0073	1.9373	2.4463	**1.0349**
	zho-lap	2.4790	**2.0426**	1.9380	**1.6440**	1.7706	1.8267	2.3633	**0.8032**
	zho-fin	**2.6547**	2.9662	2.0094	**1.9652**	1.4707	1.8900	2.5632	**0.6511**
	AVG	2.7595	**2.3441**	2.1549	**1.8731**	2.1889	2.3149	2.5674	**1.1924**
DimASTE	eng-rest	0.4993	**0.5101**	**0.5034**	0.4920	0.4483	0.2930	0.5418	**0.5442**
	eng-lap	0.4491	**0.4519**	**0.4874**	0.4424	0.3827	0.2736	**0.4664**	0.4515
	jpn-hot	0.1727	**0.3148**	0.2487	**0.3464**	0.1622	0.1458	0.4694	**0.5397**
	rus-rest	0.4016	**0.4031**	0.3730	**0.4242**	0.3341	0.1774	**0.4590**	0.4262
	tat-rest	0.3419	**0.3503**	0.3159	**0.3577**	0.2020	0.1154	**0.4101**	0.3578
	ukr-rest	0.4014	**0.4082**	0.3326	**0.4220**	0.3099	0.1595	**0.4517**	0.4250
	zho-rest	0.3190	**0.3728**	0.2425	**0.3529**	0.2509	0.1446	**0.4789**	0.4759
	zho-lap	**0.2372**	0.2227	**0.2815**	0.2494	0.2099	0.1182	0.4344	**0.4366**
	AVG	0.3528	**0.3792**	0.3481	**0.3859**	0.2875	0.1784	**0.4640**	0.4571
DimASQP	eng-rest	**0.4036**	0.3740	**0.3816**	0.3746	0.2673	0.2058	**0.5048**	0.5013
	eng-lap	0.2304	**0.2805**	**0.2842**	0.2795	0.1529	0.1293	**0.2483**	0.2411
	jpn-hot	0.0907	**0.1309**	0.1542	**0.1943**	0.0400	0.0311	0.3577	**0.4151**
	rus-rest	0.2508	**0.2656**	0.2505	**0.2963**	0.1682	0.1000	**0.4118**	0.3683
	tat-rest	0.1974	**0.2310**	0.2025	**0.2380**	0.0954	0.0611	**0.3702**	0.3094
	ukr-rest	0.2465	**0.2974**	0.2180	**0.2971**	0.1641	0.0975	**0.4070**	0.3663
	zho-rest	0.2481	**0.2975**	0.1891	**0.2859**	0.1605	0.0934	**0.4391**	0.4249
	zho-lap	0.1356	**0.1569**	**0.1921**	0.1900	0.1124	0.0728	0.3506	**0.3551**
	AVG	0.2254	**0.2542**	0.2340	**0.2695**	0.1451	0.0989	**0.3862**	0.3727

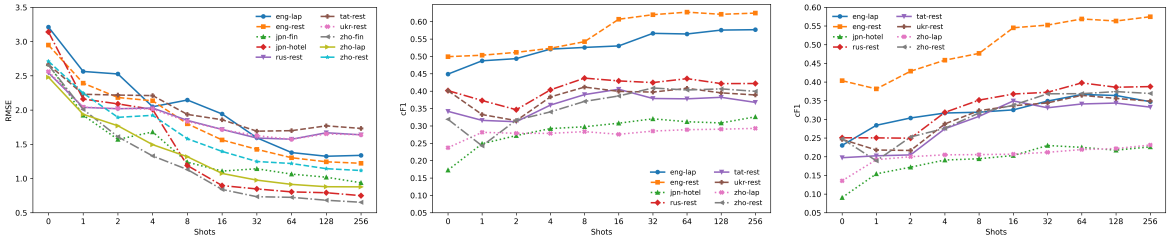
Table 3: **Comparison of LLM-based approaches**. RMSE is used for the DimASR subtask, while cF1 is used for the DimASTE and DimASQP subtasks. Bold indicates the best result within each setting for each dataset. cPrecision and cRecall results for the DimASTE and DimASQP subtasks are shown in Appendix G.

standard categorical setting.

5 Related Work

Dimensional sentiment resources. Previous studies on dimensional sentiment analysis have developed resources with single or combined dimensions across lexical, phrasal, and sentential granularities. Sentiment lexicons assign affective scores to individual words, such as SentiWordNet (Baccianella et al., 2010), SO-CAL (Taboada et al., 2011), SentiStrength (Thelwall et al., 2012), and NRC-VAD (Mohammad, 2018, 2025). Phrase-level datasets formulate sentiment composition through modifiers, including SemEval-2015 Task 10 (Rosenthal et al., 2015) and SemEval-2016 Task 7 (Kiritchenko et al., 2016). At the sentence level, affective scores are provided for texts of varying lengths (Preoţiuc-Pietro et al., 2016; Buechel and Hahn, 2017; Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018; Muhammad et al., 2025). The Stanford Sentiment Treebank (Socher et al., 2013) and Chinese EmoBank (Lee et al., 2022) provide cross-granularity resources, bridging phrase- and sentence-level representations and covering all three granularities.

Traditional ABSA datasets. Most existing ABSA datasets are English-centric and use categorical polarity labels. SemEval-2014 (Pontiki et al., 2014) initiated ABSA for English restaurant and laptop reviews, followed by extensions to more subtasks and languages (Pontiki et al., 2015, 2016). Subsequent datasets progressively enriched the annotation schema, introducing triplets of aspect, opinion, and polarity (Xu et al., 2020; Peng et al., 2020), quadruples by adding an aspect category (Zhang et al., 2021; Cai et al., 2021), and broader domain coverage (Chebolu et al., 2024). M-ABSA (Wu et al., 2025) further introduced a multilingual parallel benchmark via automatic



(a) DimASR (b) DimASTE (c) DimASQP

Figure 4: **Few-shot Performance of GPT-5 mini.** Results are summarized across all subtasks and datasets. See Appendix H for corresponding numerical scores.

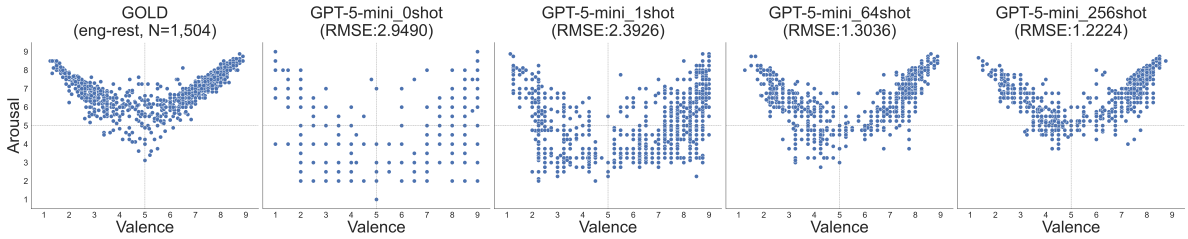


Figure 5: **Gold and predicted VA distributions for DimASR.** Results represent the English–Restaurant test set; see Appendix I for other datasets.

translation.

Dimensional ABSA datasets. Existing work is limited to a single-language, single-domain Chinese restaurant dataset from SIGHAN-2024 (Lee et al., 2024). Our DimABSA datasets are the first multilingual and multi-domain resource for dimensional ABSA.

6 Conclusions

We present DimABSA, the first multilingual dimensional ABSA resource with VA annotations. It comprises 10 datasets spanning six high- and low-resource languages across four domains. To move beyond categorical ABSA, we introduce three subtasks ranging from pure VA regression to joint extraction-regression tasks, and formulate a unified metric for categorical and continuous evaluation. Experiments with various LLMs reveal two key limitations: the constraints of token-based VA prediction in regression and persistent difficulties with hybrid extraction–regression tasks, especially for low-resource languages. These findings underscore critical challenges and opportunities for advancing multilingual dimensional ABSA.

Limitations

Although DimABSA is multilingual, interpretations of valence and arousal can vary across cultures, thereby affecting cross-lingual comparability. We mitigate this by using five native-speaker annotators per language and sample, consistent 1–9 VA scales, and shared guidelines; nonetheless, results should be interpreted as comparisons across language-community-domain settings. Expanding language coverage and testing measurement invariance are important directions for future work.

Cross-cultural comparison is a limitation of all data. For example, people from one culture may be prone to give annotations close to the centre of the scale whereas people in another culture may be more likely to spread their responses. Also, people from different cultures may use different amounts of emotional language to convey the same intensity of emotion. Therefore, cross-cultural perception is an essential issue for further research.

Ethical Considerations

People expresses attitudes, opinions, and sentiment towards entities and their aspects in complex and nuanced ways. Further, there can be considerable person-to-person variation. It should be noted that human annotated labels capture **perceived**

sentiment and attitudes, and that in several cases this may be different from the speaker's true attitudes. Nonetheless, since language is key mechanism to communicate, at an aggregate-level perceived opinions tend to correlate with actual opinions. Thus, even perceived opinions are useful at an aggregate level. However, caution must be employed when using individual inferred opinions to make decisions about individuals, especially high-stakes decisions.

Like various enabling technologies, ABSA can be abused and misused. Notably, for example, it can be used to identify various likes and dislikes of peoples and use it to manipulate people into behaviour that may not be in their best interest (e.g., purchasing products or availing services that they cannot afford or that are not particularly useful to them). This is especially concerning for vulnerable populations such as children and elderly. We expressly forbid any commercial use of our data.

For a detailed discussion of a large number of ethical considerations associated with automatic sentiment and emotion detection, we refer the reader to (Mohammad, 2022, 2023).

Use of AI Assistants

We used ChatGPT and Grammarly to correct grammatical errors, enhancing the coherence of the final manuscript. While these tools have augmented our capabilities and contributed to our findings, it's pertinent to note that they have inherent limitations. We have made every effort to use AI in a transparent and responsible manner. Any conclusions drawn are a result of combined human and machine insights.

Acknowledgments

Lung-Hao Lee and Liang-Chih Yu acknowledge the support of National Science and Technology Council, Taiwan, under the grants NSTC 113-2221-E-155-046-MY3 and NSTC 114-2221-E-A49-059-MY3.

Jonas Becker acknowledges the support of the Landeskriminalamt NRW. Jan Philip Wahle and Terry Ruas acknowledge the support of the Lower Saxony Ministry of Science and Culture, and the VW Foundation.

Shamsuddeen Hassan Muhammad acknowledges the support of Google DeepMind, whose funding made this work possible.

The work of Alexander Panchenko and Ilseyar Alimova was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the R.F. in accordance with the agreement 000000C313925P4F0002 and the agreement with Skoltech №139-10-2025-033.

Ilseyar Alimova gratefully acknowledges Dina Abdullina for the Tatar data annotation.

References

Alibaba. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 2200–2204.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 578–585.

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*, pages 340–350.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2024. OATS: A challenge dataset for opinion aspect target sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 12336–12347.

Giuseppe D'Aniello, Matteo Gaeta, and Ilaria La Rocca. 2022. KnowMIS-ABSA: An overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. *Artificial Intelligence Review*, 55(7):5543–5574.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 10088 –10115.

Will E. Hipson and Saif M. Mohammad. 2021. Emotion dynamics in movie dialogues. *PLOS ONE*, 16(9):1–19.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Yan Cathy Hua, Paul Denny, Jörg Wicker, and Katerina Taskova. 2024. A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review*, 57:296.

Lars Kaesberg, Terry Ruas, Jan Philip Wahle, and Bela Gipp. 2024. CiteAssist: A system for automated preprint citation and BibTeX generation. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 105–119, Bangkok, Thailand. Association for Computational Linguistics.

Svetlana Kiritchenko, Saif Mohammad, and Mohammad Salameh. 2016. SemEval-2016 Task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the10th International WorkshoponSemantic Evaluation*, pages 42–51.

Gaku Kutsuzawa, Hiroyuki Umemura, Koichiro Eto, and Yoshiyuki Kobayashi. 2022. Classification of 74 facial emojiś emotional states on the valence-arousal axes. *Scientific Reports*, 12:398.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(4):65.

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*, pages 165–174.

Zhiwei Liu, Tianlin Zhang, Kailai Yang, Paul Thompson, Zeping Yu, and Sophia Ananiadou. 2024. Emotion detection for misinformation: A review. *Information Fusion*, 107(102300).

Meta. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

MistralAI. 2025. Introducing Mistral 3. *mistral.ai*, pages Accessed: 2025–12–31.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pages 174–184.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pages 65–77.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 Task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Saif M. Mohammad. 2025. NRC VAD Lexicon v2: Norms for Valence, Arousal, and Dominance for over 55k English Terms. *arXiv preprint arXiv:2503.23547*, abs/2503.23547. ArXiv:2503.23547.

MoonshotAI. 2025. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 others. 2025. BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2025a. GPT-5 system card. *https://cdn.openai.com/gpt-5-system-card.pdf*.

OpenAI. 2025b. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 34(5):8600–8607.

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*, pages 19–30.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 486–495.

Maria Pontiki, Dimitris 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*, pages 27–35.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International WorkshoponSemantic Evaluation*, pages 451–463.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Happy Khairunnisa Sariyanto, Diclehan Ulucan, Oguzhan Ulucan, and Marc Ebner. 2025. Towards explainable hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 12883–12893.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Saif Mohammad. 2023. Language and mental health: Measures of emotion dynamics from text as linguistic biosocial markers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3117–3133, Singapore. Association for Computational Linguistics.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology*, 63(1):163–173.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. Toxicity, morality, and speech act guided stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4464–4478.

Xiaoyi Wang, Jiwei Zhang, Guangtao Zhang, and Honglei Guo. 2025. Feel the difference? a comparative analysis of emotional arcs in real and LLM-generated CBT sessions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19999–20017.

Zhiyuan Wen, Jiannong Cao, Jiaxing Shen, Ruosong Yang, Shuaiqi Liu, and Maosong Sun. 2024. Personality-affected emotion generation in dialog systems. *ACM Transactions on Information Systems*, 42(5):134.

ChengYan Wu, Bolei Ma, Yihong Liu, Zheyu Zhang, Ningyuan Deng, Yanshu Li, Baolan Chen, Yi Zhang, Yun Xue, and Barbara Plank. 2025. M-ABSA: A multilingual dataset for aspect-based sentiment analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2530–2557.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2339–2349.

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*. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11019–11038):101436.

Appendix

A Dataset Examples

Dataset	ST	Input	Output
eng-rest	ST 1	The food was excellent	8.00#8.12
	ST 2	Service at the bar was a little slow	(Service, a little slow, 4.10#4.30)
	ST 3	Their sodas are usually expired and flat	(sodas, DRINKS#QUALITY, usually expired, 1.90#7.20) (sodas, DRINKS#QUALITY, flat, 2.40#6.80)
eng-lap	ST 1	Very poor build quality	2.25#8.12
	ST 2	Display is bright and great	(Display, bright, 7.00#7.50) (Display, great, 7.75#8.25)
	ST 3	This HP laptop is as good as it sounds on paper	(HP laptop, LAPTOP#GENERAL, good, 5.75#5.50)
jpn-hot	ST 1	強風のせいか建物の 軋む音 がうるさすぎて眠れなかった (Because of the strong wind, the creaking sounds of the building were too loud, and I couldn’t sleep.)	2.38#7.5
	ST 2	作りは古く特別な部分は無いが、立地などなど、総合的には及第點 (The building is quite old and has nothing particularly special about it, but its location and other factors make it acceptable overall.)	(綜合的, 及第點, 6.00#5.38) (作り, 古く, 4.17#4.83)
	ST 3	駅近なのでとても便利です (Because it is close to the station, it is very convenient.)	(駅近, LOCATION#GENERAL, とても便利, 7.30#7.20)
jpn-fin	ST 1	主な減少は、投資有価証券 が 11 億 54 百万円であります。 (The main decrease amounted to 1.154 billion yen in investment securities.)	3.75#5.00
rus-res	ST 1	Замечательный ресторан!!! (A wonderful restaurant!!!)	9.00#9.00
	ST 2	Как и 3 года назад, в меню ничего нового, одни пиццы и пасты. (Just like three years ago, there is nothing new on the menu, only pizzas and pastas.)	(меню, ничего нового, 3.5#5.75) (пиццы, одни, 3.5#5.75) (пасты, одни, 3.5#5.75)
	ST 3	Еда всегда вкусная. (The food is always delicious.)	(Еда, всегда вкусная, FOOD#QUALITY, 8#7.25)
tat-res	ST 1	Тыныч музыка уйный. (Quiet music is playing.)	5.17#3.33
	ST 2	Порция бик яхшы иде! (The portion was very good!)	(Порция, бик яхшы, 7.17#6.33)
	ST 3	Зур булмаган уңайлы кафе. (A small, comfortable cafe.)	(кафе, уңайлы, AMBIENCE#GENERAL, 6.5#5.67)

Dataset	ST	Input	Output
ukr-res	ST 1	Ми вибрали страви, але їх не виявилося в наявності. (We chose dishes, but they were not available.)	3.00#6.00
	ST 2	Їжа завжди смачна. (The food is always tasty.)	(Їжа, завжди смачна, 8#7.25)
	ST 3	Їжа сподобалася, особливо десерт - наполеон, дуже смачний). (We liked the food, especially the dessert Napoleon. It's very tasty).)	(Їжа, сподобалася, FOOD#QUALITY, 6.67#5.67) (десерт, дуже смачний, FOOD#QUALITY, 7.5#6.33) (наполеон, дуже смачний, FOOD#QUALITY, 7.5#6.33)
zho-res	ST 1	但是味道 真的非常普通。 (But the taste is really very ordinary.)	4.00#6.00
	ST 2	服務那裡好? 根本爛透了。 (How was the service there? It was absolutely terrible.)	(服務, 根本爛透, 2.75#7.12)
	ST 3	鮮奶凍很濃郁、小湯圓也很好吃，環境乾淨明亮。 (The fresh milk pudding is very rich, the small glutinous rice balls are also delicious, and the environment is clean and bright.)	(鮮奶凍, FOOD#QUALITY, 很濃郁, 6.67#6.17) (小湯圓, FOOD#QUALITY, 很好吃, 7.00#6.25) (環境, AMBIENCE#GENERAL, 乾淨明亮, 7.00#5.50)
zho-lap	ST 1	確實是一台功能強大的筆電！ (It is indeed a powerful laptop!)	7.00#6.38
	ST 2	這個保固說實在不是很完美。 (Honestly, this warranty is not very perfect.)	(保固, 不是很完美, 4.38#5.25)
	ST 3	M16 的外型真的是很好看，但是售價真的挺高，感謝分享 (The M16 really looks great, but the price is quite high. Thanks for sharing.)	(M16 的外型, LAPTOP#DESIGN]_FEATURES, 真的是很好看, 6.70#6.70) (售價, LAPTOP#PRICE, 真的挺高, 3.75#6.12)
zho-fin	ST 1	利息保障倍數 由 53.16 降至 12.59。 (The interest coverage ratio decreased from 53.16 to 12.59.)	4.00#4.83

B Annotation Guidelines

Aspect-Based Sentiment Analysis (ABSA) is a fine-grained sentiment analysis task that aims to identify what people are expressing sentiments about in a text (the aspect) and how they feel about it (the opinion). Unlike general sentiment classification, which assigns a single sentiment label to an entire text, ABSA decomposes sentiment into structured representations at the aspect level for more precise interpretation. This annotation task involves extracting sentiment information as structured quadruplets, and further annotating Valence-Arousal Scores to replace the sentiment labels (Positive, Negative or Neutral). Annotators extract quadruplets of the form: **(Aspect Term, Aspect Category, Opinion Term, Valence#Arousal)**. In this phase, you will identify four key components for each sentiment expressed:

1. **Aspect Term (A)**: The specific word or phrase in the sentence that refers to an **attribute or component** of the **subject** being discussed.
2. **Aspect Category (C)**: A predefined, **broad classification** that the aspect term falls under.
3. **Opinion Term (O)**: The word or phrase that expresses sentiment (positive, negative, or neutral) towards the aspect term.

4. **Valence (V)** and **Arousal (A)** are **continuous emotion dimensions** used to describe the emotional tone of an opinion expression beyond discrete labels like “positive” or “negative.”

   - **Valence**: Valence refers to the degree of positivity or negativity associated with an emotion or sentiment and can range from 1 (very negative) to 9 (very positive). Examples: ”delightful” → **high valence** (e.g., 8.5); ”horrible” → **low valence** (e.g., 2.0)
   - **Arousal**: Arousal captures the intensity or activation level of the emotional state and can range from 1 (very calm/low energy) to 9 (very excited/high energy). Examples: ”calm” → **low arousal** (e.g., 2.5); ”furious” → **high arousal** (e.g., 8.5)

B.1 Annotation Rules for Triplets

Rule 1	There can be multiple aspect-opinion combinations in a sentence. Extract all satisfactory aspect-opinion combinations separately.
Example	Beautiful decor, great atmosphere, amazing food
Triplets	(decor, RESTAURANT#STYLE_OPTIONS, Beautiful) (atmosphere, AMBIENCE#GENERAL’, great) (food, FOOD#QUALITY, amazing)
Rule 2	If adverbs or negators are being used with an opinion, it should be included as well in the triplets because such modifiers may affect valence-arousal scores.
Example	I think the service was tremendously horrible but the food they served was absolutely delicious
Triplets	(service, SERVICE#GENERAL, tremendously horrible) (food, FOOD#QUALITY, absolutely delicious)
Rule 3	Don’t annotate words that aren’t even required. Keep it precise and only annotate what is required.
Example	I also ordered room service last night from this restaurant, and it was amazing.
Triplets	(room service, SERVICE#GENERAL, amazing)
Rule 4	Use the words in the sentences. Do not add extra words to the annotations.
Example	The staff is always friendly and helpful.
Triplets	(staff, SERVICE#GENERAL, always friendly) (staff, SERVICE#GENERAL, helpful) Notes: The opinion is “helpful”, not “always helpful”
Rule 5	Do not create new aspect categories. Only use the ones which are mentioned. Only use the following combinations, don’t use any other word as an aspect category.

B.2 Aspect Category List

- Laptop

Entity Labels
LAPTOP, DISPLAY, KEYBOARD, MOUSE, MOTHERBOARD, CPU, FANS_COOLING, PORTS, MEMORY, POWER_SUPPLY, OPTICAL_DRIVES, BATTERY, GRAPHICS, HARD_DISK, MULTIMEDIA_DEVICES, HARDWARE, SOFTWARE, OS, WARRANTY, SHIPPING, SUPPORT, COMPANY
Attribute Labels
GENERAL, PRICE, QUALITY, DESIGN_FEATURES, OPERATION_PERFORMANCE, USABILITY, PORTABILITY, CONNECTIVITY, MISCELLANEOUS

- Restaurant

Entity Labels
RESTAURANT, FOOD, DRINKS, AMBIENCE, SERVICE, LOCATION
Attribute Labels
GENERAL, PRICES, QUALITY, STYLE_OPTIONS, MISCELLANEOUS

- Hotel

Entity Labels
HOTEL, ROOMS, FACILITIES, ROOM_AMENITIES, SERVICE, LOCATION, FOOD_DRINKS
Attribute Labels
GENERAL, PRICE, COMFORT, CLEANLINESS, QUALITY, DESIGN_FEATURES, STYLE_OPTIONS, MISCELLANEOUS

B.3 Combined Interpretation for Valence-Arousal Score

The **Valence–Arousal** pair provides a more nuanced understanding of sentiment:

Expression	**V**	**A**	**Interpretation**
”very courteous”	7.5	5.7	(7.5, 5.7) Moderately positive and moderately active
”terrible”	2.8	6.8	(2.8, 6.8) Strongly negative and emotionally intense
”okay”	5.0	3.0	(5.0, 3.0) Neutral valence and low emotional intensity

The score distribution of valence and arousal often follows a **smile-shaped or U-shaped curve**.

- The higher the valence, the higher the arousal.
- The lower the valence, the higher the arousal as well.
- When valence is neutral, arousal tends to be lower.

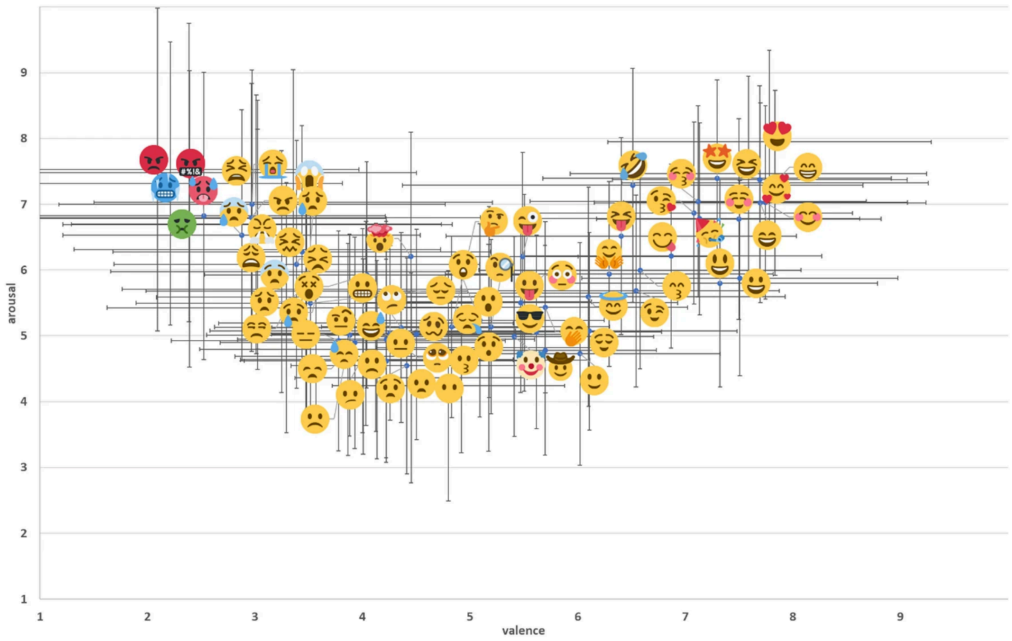
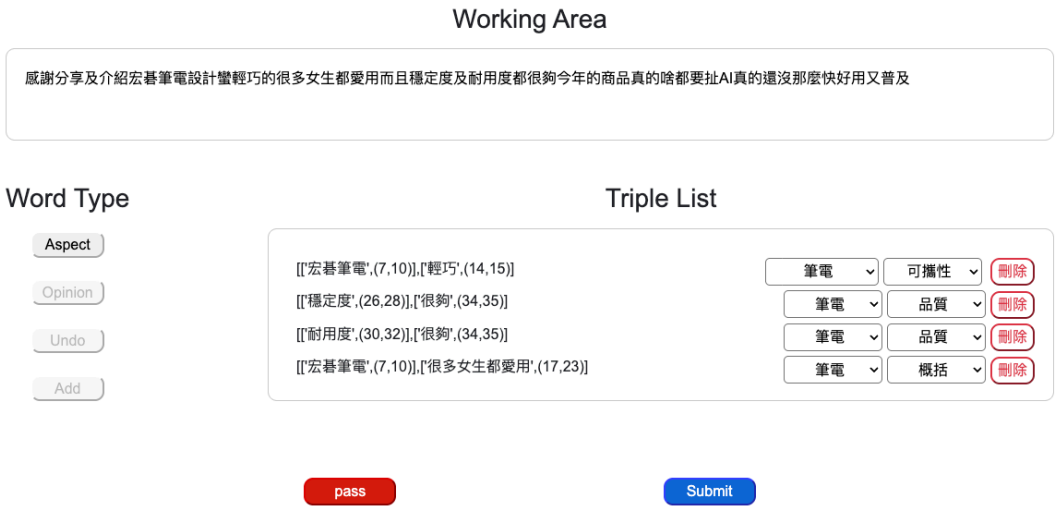
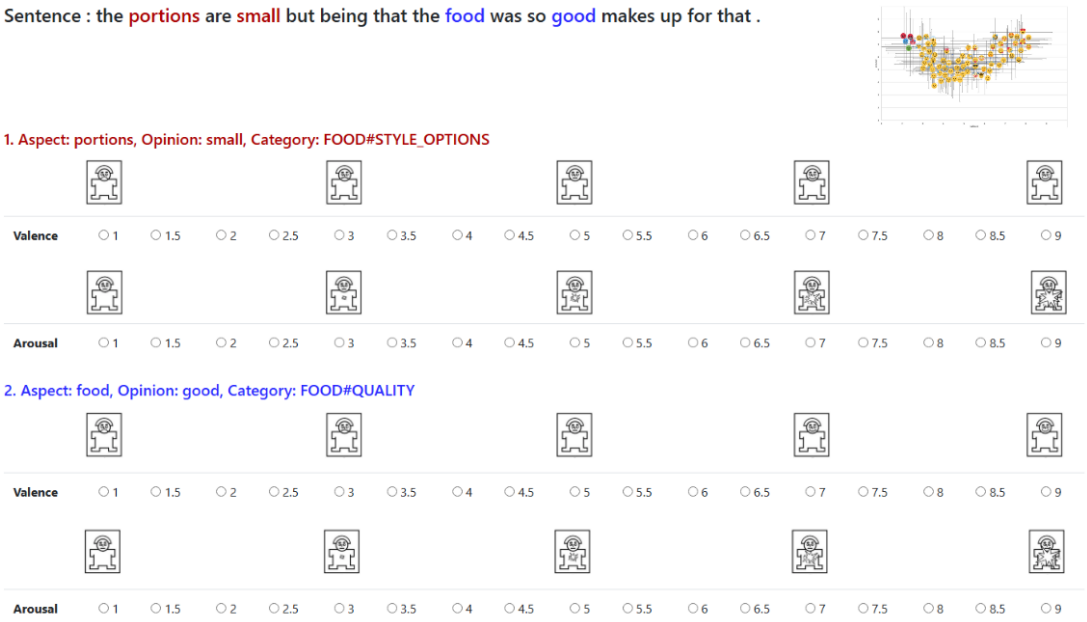


Figure 6: Facial emoji’s emotional states on the valence-arousal axes (Kutsuzawa et al., 2022).

B.4 Annotation Screens

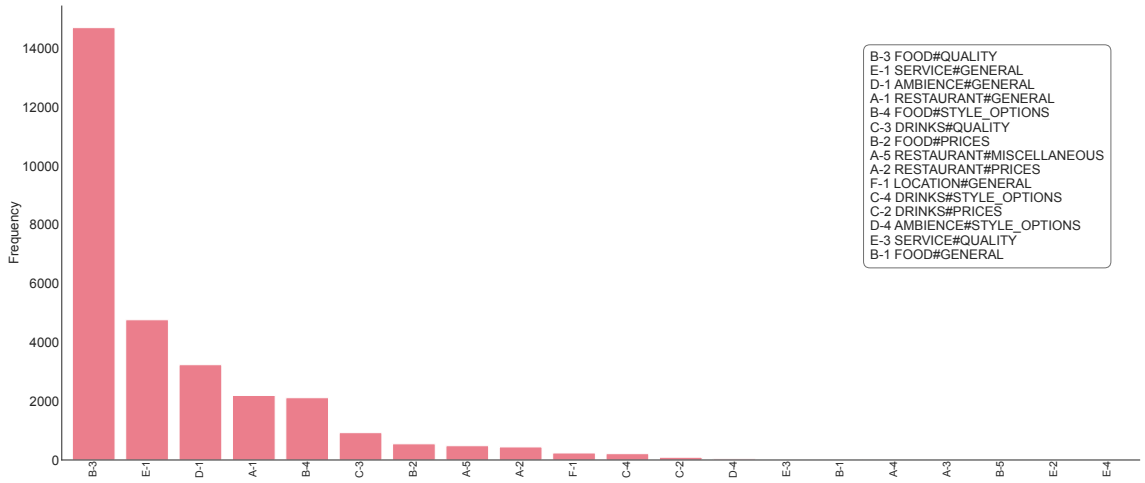


(a) Triplet (A, C, O) annotation interface

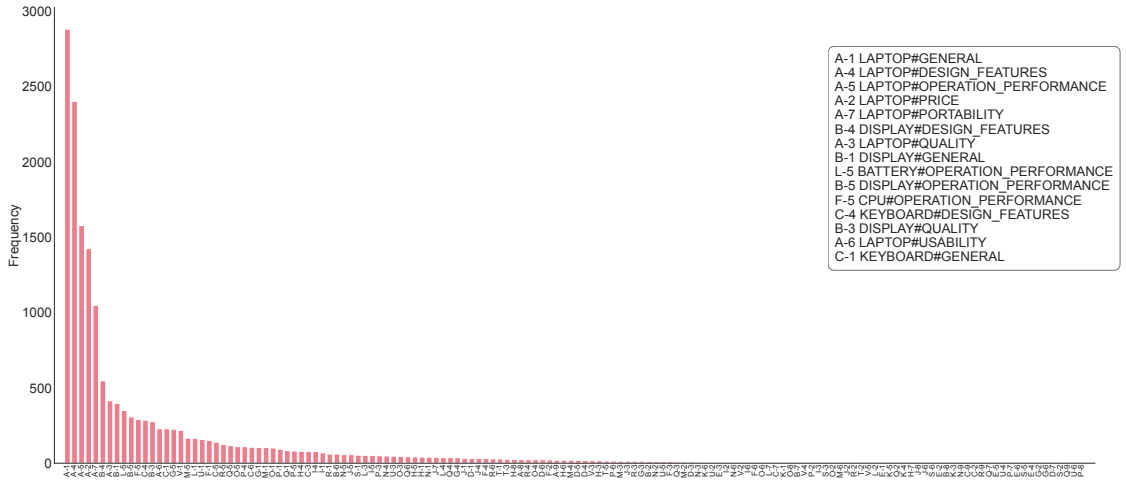


(b) VA annotation interface

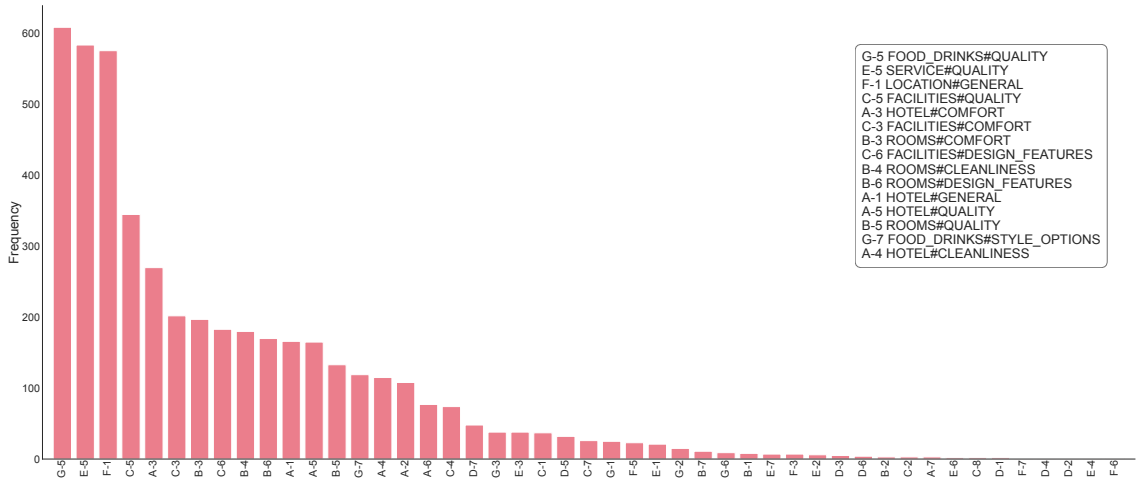
C Category Distribution



(a) Restaurant Category Distribution



(b) Laptop Category Distribution



(c) Hotel Category Distribution

D Prompts used for Few-Shot Learning

DimASR

Definition: The output will be a pair of real numbers [valence#arousal] representing the sentiment toward the identified aspect in the sentence, where both values are on a scale from 1 to 9.
Valence: 1 = most negative, 9 = most positive;
Arousal: 1 = calm/low intensity, 9 = excited/high intensity.
<Examples>
Now complete the following example
Input: <sentence> The aspect is <aspect>.
Output:

DimASQP
(Notes: without "aspect category" and its corresponding part is used for dimASTE)

According to the following sentiment elements definition:
- The "aspect term" refers to a specific feature, attribute, or aspect of a product or service on which a user can express an opinion. Explicit aspect terms appear explicitly as a substring of the given text.
- The "aspect category" refers to the category that aspect belongs to, and the available categories include:
<Aspect_Category_Lists>
- The "opinion term" refers to the sentiment or attitude expressed by a user towards a particular aspect or feature of a product or service. Explicit opinion terms appear explicitly as a substring of the given text.
- The "sentiment score" refers to the emotional response toward the aspect, represented as a Valence-Arousal pair, where Valence indicates positivity/negativity (1.0 = most negative, 9.0 = most positive) and Arousal indicates emotional intensity (1.0 = calm, 9.0 = excited). Both values are real numbers in the range [1.0, 9.0].
Please carefully follow the instructions. Ensure that aspect terms are recognized as exact matches in the review. Ensure that opinion terms are recognized as exact matches in the review.
Recognize all sentiment elements with their corresponding aspect terms, aspect categories, opinion terms, and sentiment score in the given input text (review). Provide your response in the format of a Python list of tuples: 'Sentiment elements: [("aspect term", "aspect category", "opinion term", [valence#arousal]), ...]'. Note that ", ..." indicates that there might be more tuples in the list if applicable and must not occur in the answer. Ensure there is no additional text in the response.
<Examples>
Input: <sentence>
Output:

E Prompts used for Supervised Fine-Tuning

DimASR

System Prompt:

Instruction: The output will be a pair of real numbers [valence#arousal] representing the sentiment toward the identified aspect in the sentence, where both values are on a scale from 1 to 9. Valence: 1 = most negative, 9 = most positive; Arousal: 1 = calm/low intensity, 9 = excited/high intensity.

Provide your response in the format of a Python list: [valence#arousal]

User Prompt:

Text: <text>

Aspect Term: <aspect>

Assistant Prompt:

< valence#arousal >

DimASQP

(Notes: without "aspect category" and its corresponding part is used for DimASTE)

System Prompt:

Instruction: According to the following sentiment elements definition. - The "aspect term" refers to a specific feature, attribute, or aspect of a product or service on which a user can express an opinion. Explicit aspect terms appear explicitly as a substring of the given text.

- The "aspect category" refers to the category that aspect belongs to, and the available categories include: **<Aspect_Category_Lists>**

- The "opinion term" refers to the sentiment or attitude expressed by a user towards a particular aspect or feature of a product or service. Explicit opinion terms appear explicitly as a substring of the given text.

- The "sentiment score" refers to the emotional response toward the aspect, represented as a Valence-Arousal pair, where Valence indicates positivity/negativity (1.0 = most negative, 9.0 = most positive) and Arousal indicates emotional intensity (1.0 = calm, 9.0 = excited). Both values are real numbers in the range [1.0, 9.0].

Please carefully follow the instructions. Ensure that aspect terms are recognized as exact matches in the review. Ensure that opinion terms are recognized as exact matches in the review.

User Prompt:

Text: <text>

Assistant Prompt:

Sentiment elements: <tuple>

F Example Calculation of cF1

Prediction/Gold	TP_{cat} (A)	**VA error distance**		**cTP** (A)-(C)
		Raw (B)	Normalized $(C) = (B)/\sqrt{128}$	
P: (food, good, 8.00#8.00) G: (food, good, 7.00#7.00)	1	$\sqrt{2}$	$\frac{\sqrt{2}}{\sqrt{128}} = 0.125$	0.875
P: (soup, spicy, 7.50#7.50) G: (soup, spicy, 3.50#3.50)	1	$\sqrt{32}$	$\frac{\sqrt{32}}{\sqrt{128}} = 0.5$	0.5
P: (staff, friendly, 7.00#7.00) G: (staff, always friendly, 7.50#7.50)	0	-	-	0
P: (staff, good, 7.00#7.00) G: N/A	0	-	-	0
			Total cTP	1.375
$cRecall = 1.375/3 = 0.458$				
$cPrecision = 1.375/4 = 0.344$				
$cF1 = (2 * 0.458 * 0.344)/(0.458 + 0.344) = 0.393$				

Note: The VA scores lie in the range [1, 9]. When the VA prediction is perfect (i.e., dist=0), cRecall/cPrecision reduces to the standard recall/precision.

G DimASTE and DimASQP Results

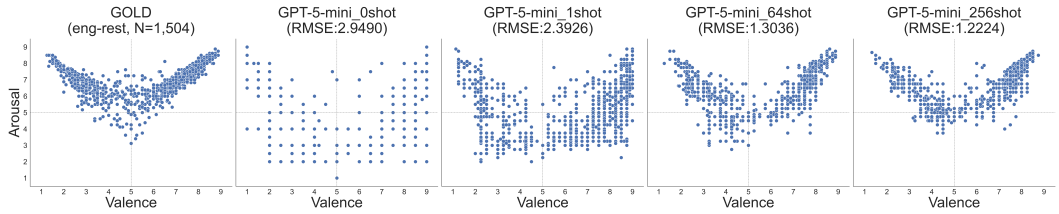
Subtask	Dataset	Zero-Shot Learning						One-Shot Learning					
		GPT-5 mini			Kimi K2 Thinking			GPT-5 mini			Kimi K2 Thinking		
		cP	cR	cF1	cP	cR	cF1	cP	cR	cF1	cP	cR	cF1
DimASTE	eng-rest	0.4719	0.5301	0.4993	0.5043	0.5159	**0.5101**	0.4667	0.5465	**0.5034**	0.4902	0.4938	0.4920
	eng-lap	0.4224	0.4793	0.4491	0.4458	0.4582	**0.4519**	0.4443	0.5399	**0.4874**	0.4265	0.4595	0.4424
	jpn-hot	0.1552	0.1947	0.1727	0.2975	0.3342	**0.3148**	0.2259	0.2767	0.2487	0.3409	0.3520	**0.3464**
	rus-rest	0.3575	0.4580	0.4016	0.3863	0.4214	**0.4031**	0.3380	0.4159	0.3730	0.4287	0.4199	**0.4242**
	tat-rest	0.3038	0.3910	0.3419	0.3363	0.3656	**0.3503**	0.2868	0.3516	0.3159	0.3594	0.3561	**0.3577**
	ukr-rest	0.3571	0.4583	0.4014	0.3875	0.4313	**0.4082**	0.3164	0.3505	0.3326	0.4229	0.4212	**0.4220**
	zho-rest	0.2737	0.3823	0.3190	0.3436	0.4076	**0.3728**	0.2115	0.2842	0.2425	0.3357	0.3719	**0.3529**
	zho-lap	0.1934	0.3065	**0.2372**	0.1975	0.2553	0.2227	0.2271	0.3702	**0.2815**	0.2233	0.2823	0.2494
	AVG	0.3169	0.4000	0.3528	0.3624	0.3987	**0.3792**	0.3146	0.3919	0.3481	0.3785	0.3946	**0.3859**
DimASQP	eng-rest	0.3814	0.4285	**0.4036**	0.3698	0.3783	0.3740	0.3538	0.4143	**0.3816**	0.3732	0.3760	0.3746
	eng-lap	0.2167	0.2459	0.2304	0.2767	0.2844	**0.2805**	0.2590	0.3148	**0.2842**	0.2695	0.2903	0.2795
	jpn-hot	0.0815	0.1022	0.0907	0.1237	0.1389	**0.1309**	0.1400	0.1715	0.1542	0.1912	0.1974	**0.1943**
	rus-rest	0.2233	0.2860	0.2508	0.2546	0.2777	**0.2656**	0.2271	0.2794	0.2505	0.2994	0.2933	**0.2963**
	tat-rest	0.1754	0.2257	0.1974	0.2217	0.2410	**0.2310**	0.1839	0.2254	0.2025	0.2391	0.2369	**0.2380**
	ukr-rest	0.2193	0.2814	0.2465	0.2823	0.3142	**0.2974**	0.2074	0.2297	0.2180	0.2977	0.2966	**0.2971**
	zho-rest	0.2129	0.2974	0.2481	0.2741	0.3252	**0.2975**	0.1649	0.2217	0.1891	0.2720	0.3013	**0.2859**
	zho-lap	0.1105	0.1752	0.1356	0.1391	0.1799	**0.1569**	0.1549	0.2526	**0.1921**	0.1701	0.2150	0.1900
	AVG	0.2026	0.2553	0.2254	0.2428	0.2675	**0.2542**	0.2114	0.2637	0.2340	0.2640	0.2759	**0.2695**

Subtask	Dataset	Supervised Fine-Tuning											
		Qwen3 (14B)			Ministral-3 (14B)			Llama-3.3 (70B)			GPT-OSS (120B)		
		cP	cR	cF1	cP	cR	cF1	cP	cR	cF1	cP	cR	cF1
DimASTE	eng-rest	0.4634	0.4341	0.4483	0.2913	0.2948	0.2930	0.5547	0.5294	0.5418	0.5644	0.5254	**0.5442**
	eng-lap	0.3944	0.3716	0.3827	0.2673	0.2803	0.2736	0.4939	0.4418	**0.4664**	0.4854	0.4220	0.4515
	jpn-hot	0.1685	0.1563	0.1622	0.1404	0.1517	0.1458	0.4663	0.4724	0.4694	0.5549	0.5253	**0.5397**
	rus-rest	0.3310	0.3373	0.3341	0.1742	0.1808	0.1774	0.4454	0.4736	**0.4590**	0.4217	0.4307	0.4262
	tat-rest	0.2033	0.2008	0.2020	0.1170	0.1139	0.1154	0.3946	0.4268	**0.4101**	0.3406	0.3768	0.3578
	ukr-rest	0.3077	0.3121	0.3099	0.1555	0.1637	0.1595	0.4466	0.4568	**0.4517**	0.4189	0.4314	0.4250
	zho-rest	0.2528	0.2490	0.2509	0.1373	0.1527	0.1446	0.4837	0.4741	**0.4789**	0.4674	0.4848	0.4759
	zho-lap	0.2021	0.2182	0.2099	0.1052	0.1348	0.1182	0.4432	0.4260	0.4344	0.4468	0.4268	**0.4366**
	AVG	0.2904	0.2849	0.2875	0.1735	0.1841	0.1784	0.4661	0.4626	**0.4640**	0.4625	0.4529	0.4571
DimASQP	eng-rest	0.2763	0.2588	0.2673	0.2046	0.2070	0.2058	0.5169	0.4933	**0.5048**	0.5199	0.4840	0.5013
	eng-lap	0.1576	0.1485	0.1529	0.1263	0.1325	0.1293	0.2630	0.2353	**0.2483**	0.2592	0.2253	0.2411
	jpn-hot	0.0416	0.0386	0.0400	0.0299	0.0323	0.0311	0.3554	0.3601	0.3577	0.4268	0.4040	**0.4151**
	rus-rest	0.1666	0.1698	0.1682	0.0981	0.1019	0.1000	0.3995	0.4248	**0.4118**	0.3644	0.3722	0.3683
	tat-rest	0.0959	0.0948	0.0954	0.0619	0.0603	0.0611	0.3562	0.3853	**0.3702**	0.2945	0.3258	0.3094
	ukr-rest	0.1631	0.1653	0.1641	0.0951	0.1001	0.0975	0.4024	0.4117	**0.4070**	0.3610	0.3718	0.3663
	zho-rest	0.1617	0.1593	0.1605	0.0887	0.0986	0.0934	0.4436	0.4348	**0.4391**	0.4173	0.4329	0.4249
	zho-lap	0.1083	0.1169	0.1124	0.0648	0.0831	0.0728	0.3577	0.3438	0.3506	0.3634	0.3471	**0.3551**
	AVG	0.1464	0.1440	0.1451	0.0962	0.1020	0.0989	0.3868	0.3961	**0.3862**	0.3758	0.3704	0.3727

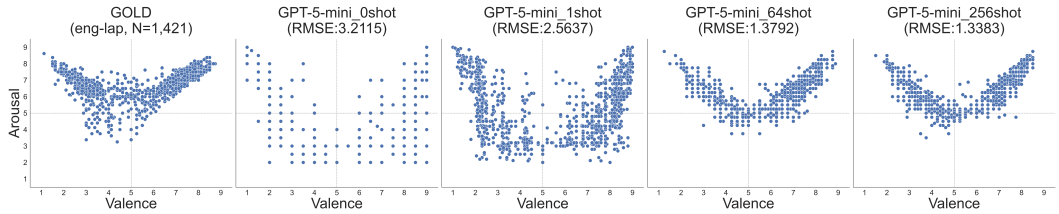
H Few-Shot Results

Subtask	Dataset	Shots									
		0	1	2	4	8	16	32	64	128	256
DimASR	eng-rest	2.9490	2.3926	2.1832	2.1351	1.7992	1.5631	1.4263	1.3036	1.2443	**1.2224**
	eng-lap	3.2115	2.5637	2.5269	2.0482	2.1466	1.9426	1.5950	1.3792	**1.3251**	1.3383
	jpn-hot	3.1406	2.1607	2.0899	2.0094	1.1898	0.8968	0.8478	0.8047	0.7924	**0.7498**
	jpn-fin	2.6760	1.9243	1.5710	1.6804	1.2386	1.1088	1.1425	1.0645	1.0192	**0.9380**
	rus-rest	2.5447	2.0390	2.0192	2.0296	1.8490	1.7170	1.5912	**1.5740**	1.6682	1.6385
	tat-rest	2.6645	2.2308	2.2187	2.2111	1.9360	1.8583	**1.6913**	1.6990	1.7692	1.7318
	ukr-rest	2.5628	2.0438	2.0222	2.0141	1.8340	1.7152	1.6259	**1.5723**	1.6515	1.6383
	zho-rest	2.7125	2.2467	1.8937	1.9228	1.5810	1.3957	1.2485	1.2191	1.1436	**1.1169**
	zho-lap	2.4790	1.9380	1.7699	1.4945	1.3179	1.0741	0.9765	0.9144	0.8802	**0.8781**
	zho-fin	2.6547	2.0094	1.6099	1.3295	1.1311	0.8401	0.7343	0.7253	0.6825	**0.6532**
	AVG	2.7595	2.1549	1.9904	1.8875	1.6023	1.4112	1.2879	1.2256	1.2176	**1.1905**
DimASTE	eng-rest	0.4993	0.5034	0.5117	0.5234	0.5425	0.607	0.6201	**0.6274**	0.6212	0.6247
	eng-lap	0.4491	0.4874	0.4936	0.5210	0.5259	0.5302	0.5664	0.5646	0.5758	**0.5769**
	jpn-hot	0.1727	0.2487	0.2721	0.2923	0.2975	0.3078	0.3204	0.3120	0.3085	**0.3261**
	rus-rest	0.4016	0.3730	0.3466	0.4037	0.4375	0.4295	0.4247	**0.4360**	0.4217	0.4220
	tat-rest	0.3419	0.3159	0.3120	0.3599	0.3901	0.4055	0.3789	0.3779	**0.3823**	0.3677
	ukr-rest	0.4014	0.3326	0.3152	0.3834	**0.4117**	0.3994	0.3973	0.4081	0.3951	0.3890
	zho-rest	0.3190	0.2425	0.3166	0.3403	0.3705	0.3862	**0.4091**	0.4033	0.4066	0.3995
	zho-lap	0.2372	0.2815	0.2787	0.2785	0.2835	0.2758	0.2851	0.2891	0.2905	**0.2929**
	AVG	0.3528	0.3481	0.3558	0.3878	0.4074	0.4177	0.4253	**0.4273**	0.4252	0.4249
DimASQP	eng-rest	0.4036	0.3816	0.4290	0.4585	0.4767	0.5453	0.5525	0.5689	0.5634	**0.5749**
	eng-lap	0.2304	0.2842	0.3038	0.3170	0.3193	0.3255	0.3490	**0.3667**	0.3640	0.3473
	jpn-hot	0.0907	0.1542	0.1720	0.1912	0.1946	0.2032	**0.2301**	0.2251	0.2176	0.2264
	rus-rest	0.2508	0.2505	0.2496	0.3187	0.3516	0.3679	0.3726	**0.3977**	0.3861	0.3878
	tat-rest	0.1974	0.2025	0.2049	0.2748	0.3081	**0.3494**	0.3311	0.3413	0.3437	0.3328
	ukr-rest	0.2465	0.2180	0.2168	0.2878	0.3234	0.3366	0.3441	**0.3644**	0.3561	0.3489
	zho-rest	0.2481	0.1891	0.2529	0.2760	0.3172	0.3381	0.3685	0.3688	**0.3744**	0.3694
	zho-lap	0.1356	0.1921	0.2001	0.2051	0.2057	0.2068	0.2118	0.2192	0.2219	**0.2313**
	AVG	0.2254	0.2340	0.2536	0.2911	0.3121	0.3341	0.3450	**0.3565**	0.3534	0.3524

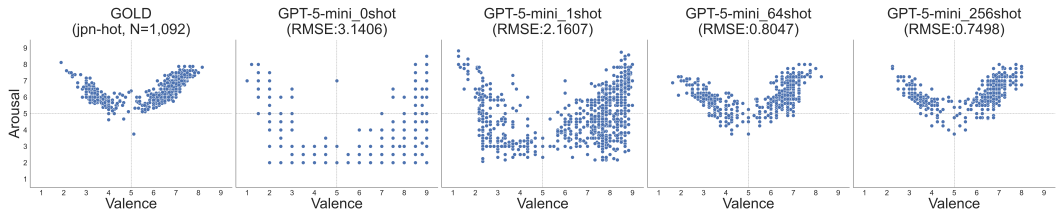
I Few-shot VA scatter



(a) eng-rest

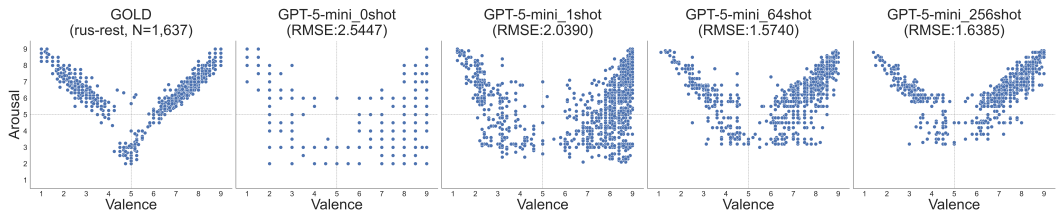


(b) eng-lap

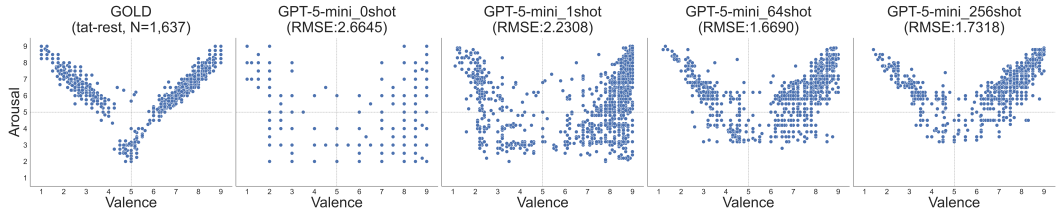


(c) jpn-hot

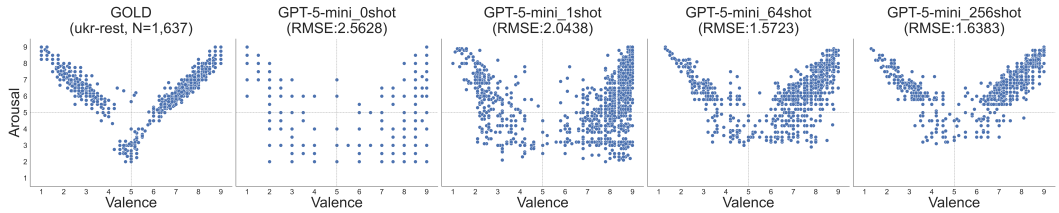
(d) jpn-fin



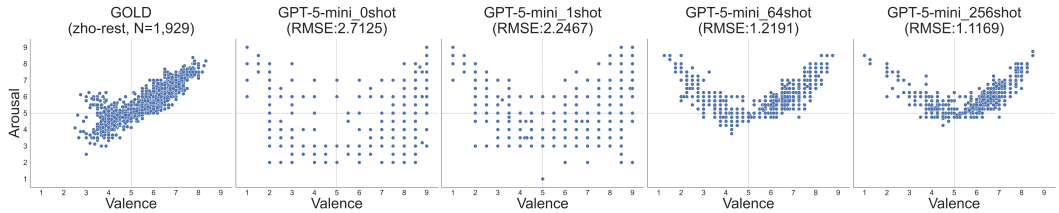
(e) rus-rest



(f) tat-rest

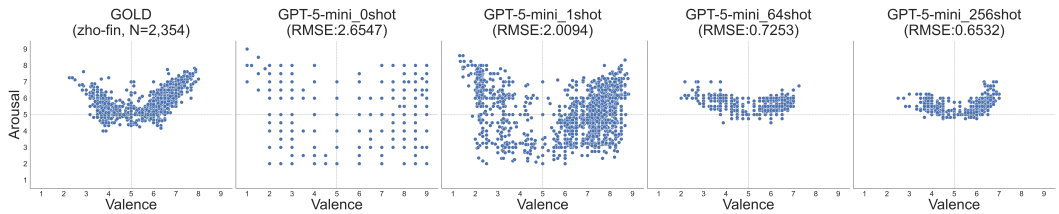


(g) ukr-rest



(h) zho-rest

(i) zho-lap



(j) zho-fin

J Categorical ABSA

J.1 Data Statistics

We split up all data in positive (V>5.5), neutral (4.5<=V<=5.5), and negative (V<4.5) samples. We report the average valence and arousal values for each split of the data. Counts (n) and standard deviations (SD) are shown in tiny.

DimABSA	Positive ($V > 5.5$)			Neutral ($4.5 \leq V \leq 5.5$)			Negative ($V < 4.5$)		
Dataset	%(n)	$V_{mean\ (SD)}$	$A_{mean\ (SD)}$	%(n)	$V_{mean\ (SD)}$	$A_{mean\ (SD)}$	%(n)	$V_{mean\ (SD)}$	$A_{mean\ (SD)}$
eng-rest	71.14% (5,720)	7.327 (0.645)	7.241 (0.692)	9.15% (736)	5.016 (0.290)	5.220 (0.508)	19.7% (1,584)	3.007 (0.804)	6.444 (1.083)
eng-lap	65.58% (6,401)	7.215 (0.625)	7.151 (0.670)	9.42% (919)	4.948 (0.320)	5.269 (0.485)	25.00% (2,440)	3.234 (0.720)	6.284 (0.967)
jpn-hot	83.46% (5,032)	6.803 (0.464)	6.486 (0.559)	1.19% (72)	5.089 (0.332)	5.418 (0.581)	15.34% (925)	3.498 (0.494)	6.015 (0.555)
jpn-fin	59.10% (1,946)	6.189 (0.426)	5.497 (0.497)	5.25% (173)	5.130 (0.342)	4.685 (0.692)	35.65% (1,174)	3.632 (0.480)	5.518 (0.532)
rus-rest	77.69% (4,364)	7.248 (0.636)	6.652 (0.842)	4.2% (236)	4.976 (0.238)	3.616 (1.121)	18.11% (1,017)	2.801 (0.677)	6.532 (0.961)
tat-rest	77.69% (4,364)	7.248 (0.636)	6.652 (0.842)	4.2% (236)	4.976 (0.238)	3.616 (1.121)	18.11% (1,017)	2.801 (0.677)	6.532 (0.961)
ukr-rest	77.69% (4,364)	7.248 (0.636)	6.652 (0.842)	4.2% (236)	4.976 (0.238)	3.616 (1.121)	18.11% (1,017)	2.801 (0.677)	6.532 (0.961)
zho-rest	73.40% (13,295)	6.406 (0.463)	6.016 (0.626)	12.47% (2,258)	4.976 (0.354)	5.092 (0.359)	14.13% (2,559)	3.909 (0.400)	5.071 (0.696)
zho-lap	71.04% (9,630)	6.405 (0.456)	5.769 (0.674)	8.19% (1,111)	5.028 (0.338)	5.185 (0.438)	20.78% (2,817)	3.807 (0.399)	5.402 (0.602)
zho-fin	65.10% (3,613)	6.025 (0.302)	5.466 (0.407)	23.35% (1,296)	5.222 (0.320)	5.165 (0.281)	11.55% (641)	4.042 (0.274)	5.326 (0.421)

J.2 Aspect Sentiment Classification (ASC) Results

Subtask	Dataset	Zero-Shot Learning						One-Shot Learning					
		GPT-5 mini			Kimi K2 Thinking			GPT-5 mini			Kimi K2 Thinking		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
ASC	eng-rest	0.6792	0.7045	0.6885	0.6751	0.7104	**0.6902**	0.6798	0.7067	**0.6904**	0.6692	0.7124	0.6870
	eng-lap	0.6666	0.6723	**0.6626**	0.6424	0.6681	0.6501	0.6491	0.6762	**0.6580**	0.6370	0.6716	0.6458
	jpn-hot	0.7027	0.7044	0.7035	0.6972	0.7152	**0.7053**	0.6953	0.7045	0.6998	0.6978	0.7183	**0.7069**
	jpn-fin	0.7238	0.7233	**0.7235**	0.6757	0.6897	0.6792	0.7427	0.7240	**0.7308**	0.7113	0.7183	0.7146
	rus-rest	0.7737	0.7283	**0.7452**	0.7494	0.7198	0.7295	0.7747	0.7324	**0.7478**	0.7426	0.7153	0.7241
	tat-rest	0.6830	0.6755	0.6785	0.7002	0.6973	**0.6975**	0.6981	0.6872	0.6921	0.6959	0.6960	**0.6956**
	ukr-rest	0.7320	0.7028	**0.7132**	0.7254	0.7047	0.7116	0.7655	0.7286	**0.7427**	0.7118	0.6935	0.6995
	zho-rest	0.7622	0.7207	0.7156	0.7630	0.7325	**0.7295**	0.7457	0.7177	0.7086	0.7538	0.7329	**0.7249**
	zho-lap	0.7920	0.7464	0.7540	0.7805	0.7602	**0.7612**	0.7803	0.7416	0.7464	0.7897	0.7658	**0.7682**
	zho-fin	0.7010	0.6797	**0.6407**	0.6303	0.6642	0.6329	0.7010	0.6716	**0.6302**	0.6274	0.6619	0.6254
	AVG	0.7216	0.7058	**0.7025**	0.7039	0.7062	0.6987	0.7232	0.7091	**0.7047**	0.7037	0.7086	0.6992

Subtask	Dataset	Supervised Fine-Tuning											
		Qwen3 (14B)			Ministral-3 (14B)			Llama-3.3 (70B)			GPT-OSS (120B)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
ASC	eng-rest	0.6796	0.7487	**0.7019**	0.6121	0.6311	0.5915	0.6140	0.6612	0.6197	0.6739	0.7050	0.6603
	eng-lap	0.6395	0.6807	0.6496	0.6206	0.6014	0.5690	0.5896	0.5922	0.5597	0.6654	0.7066	**0.6548**
	jpn-hot	0.6949	0.7732	**0.7080**	0.6503	0.6551	0.6431	0.6769	0.7567	0.6892	0.7245	0.6808	0.6928
	jpn-fin	0.6797	0.7051	0.6768	0.6838	0.6855	0.6831	0.6367	0.6695	0.6402	0.7189	0.7236	**0.7212**
	rus-rest	0.7165	0.7204	0.7184	0.7123	0.6605	0.6421	0.6908	0.6934	0.6732	0.8063	0.7320	**0.7549**
	tat-rest	0.5730	0.6229	0.5915	0.5471	0.5760	0.5060	0.4504	0.4864	0.3712	0.7398	0.6835	**0.7020**
	ukr-rest	0.7226	0.7303	0.7260	0.6640	0.6457	0.6217	0.6882	0.6995	0.6779	0.8008	0.7148	**0.7376**
	zho-rest	0.7198	0.7166	**0.7133**	0.7033	0.6291	0.5614	0.6798	0.6923	0.6718	0.7408	0.6999	0.7060
	zho-lap	0.7465	0.7533	0.7487	0.6989	0.5984	0.5284	0.6802	0.6947	0.6714	0.7676	0.7503	**0.7578**
	zho-fin	0.6827	0.6738	**0.6677**	0.6206	0.6454	0.5525	0.6011	0.6759	0.6136	0.6478	0.6239	0.6348
	AVG	0.6855	0.7125	0.6902	0.6513	0.6328	0.5899	0.6308	0.6622	0.6188	0.7286	0.7020	**0.7022**

J.3 ASTE Results

Subtask	Dataset	Zero-Shot Learning						One-Shot Learning					
		GPT-5 mini			Kimi K2 Thinking			GPT-5 mini			Kimi K2 Thinking		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
ASTE	eng-rest	0.5654	0.6618	**0.6098**	0.5842	0.5932	0.5887	0.5703	0.6501	**0.6076**	0.5526	0.5627	0.5576
	eng-lap	0.5141	0.6015	**0.5544**	0.5396	0.5554	0.5474	0.5402	0.6395	**0.5857**	0.4752	0.5094	0.4917
	jpn-hot	0.2009	0.2509	0.2231	0.3431	0.3805	**0.3608**	0.2728	0.3153	0.2925	0.3827	0.3867	**0.3847**
	rus-rest	0.4152	0.5435	0.4707	0.4554	0.4908	**0.4724**	0.3768	0.4565	0.4128	0.4630	0.4443	**0.4534**
	tat-rest	0.3526	0.4565	0.3979	0.3858	0.4153	**0.4000**	0.3166	0.3840	0.3470	0.4062	0.4084	**0.4073**
	ukr-rest	0.4125	0.5275	**0.4630**	0.4365	0.4725	0.4538	0.3624	0.3901	0.3757	0.4692	0.4534	**0.4612**
	zho-rest	0.3009	0.4352	0.3558	0.3668	0.4352	**0.3981**	0.2301	0.3114	0.2646	0.3485	0.3827	**0.3648**
	zho-lap	0.2265	0.3751	**0.2825**	0.2452	0.3096	0.2736	0.2385	0.3545	0.2851	0.2588	0.3356	**0.2922**
	AVG	0.3735	0.4815	0.4197	0.4196	0.4566	**0.4369**	0.3635	0.4377	0.3964	0.4195	0.4354	**0.4266**

Subtask	Dataset	Supervised Fine-Tuning											
		Qwen3 (14B)			Ministral-3 (14B)			Llama-3.3 (70B)			GPT-OSS (120B)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
ASTE	eng-rest	0.5123	0.4871	0.4994	0.3524	0.3593	0.3558	0.6259	0.6059	**0.6158**	0.5992	0.5871	0.5931
	eng-lap	0.4302	0.4448	0.4374	0.3486	0.3651	0.3567	0.4972	0.4430	**0.4685**	0.4911	0.4446	0.4666
	jpn-hot	0.2003	0.2010	0.2006	0.1749	0.1968	0.1852	0.5686	0.5627	**0.5657**	0.5573	0.5530	0.5551
	rus-rest	0.3210	0.3496	0.3347	0.2107	0.2229	0.2166	0.5695	0.5565	**0.5629**	0.5025	0.5313	0.5165
	tat-rest	0.1744	0.1832	0.1787	0.1414	0.1458	0.1436	0.5276	0.4954	**0.5110**	0.4353	0.4626	0.4486
	ukr-rest	0.3161	0.3359	0.3257	0.2004	0.2153	0.2076	0.5655	0.5763	**0.5709**	0.5065	0.5321	0.5190
	zho-rest	0.2466	0.2587	0.2525	0.1695	0.1943	0.1811	0.4695	0.4708	**0.4702**	0.4526	0.4425	0.4475
	zho-lap	0.1924	0.2431	0.2148	0.1485	0.2005	0.1706	0.4984	0.4743	**0.4860**	0.4386	0.4249	0.4317
	AVG	0.2992	0.3129	0.3055	0.2183	0.2375	0.2272	0.5403	0.5231	**0.5314**	0.4979	0.4973	0.4973

J.4 ASQP Results

Subtask	Dataset	Zero-Shot learning						One-Shot Learning					
		GPT-5 mini			Kimi K2 Thinking			GPT-5 mini			Kimi K2 Thinking		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
ASQP	eng-rest	0.4446	0.5204	**0.4795**	0.4468	0.4537	0.4502	0.4681	0.5336	**0.4987**	0.4437	0.4519	0.4478
	eng-lap	0.2678	0.3134	0.2888	0.3463	0.3565	**0.3513**	0.3272	0.3873	**0.3547**	0.2872	0.3078	0.2972
	jpn-hotel	0.1071	0.1337	0.1190	0.1487	0.1649	**0.1564**	0.1667	0.1927	0.1787	0.2188	0.2211	**0.2199**
	rus-rest	0.2758	0.3611	0.3127	0.3081	0.3321	**0.3196**	0.2703	0.3275	0.2962	0.3190	0.3061	**0.3124**
	tat-rest	0.2205	0.2855	0.2488	0.2752	0.2962	**0.2853**	0.2102	0.2550	0.2304	0.2840	0.2855	**0.2847**
	ukr-rest	0.2549	0.3260	0.2861	0.3103	0.3359	**0.3226**	0.2617	0.2817	0.2713	0.3262	0.3153	**0.3207**
	zho-rest	0.2369	0.3425	0.2801	0.2858	0.3390	**0.3102**	0.1833	0.2482	0.2109	0.2791	0.3065	**0.2922**
	zho-lap	0.1255	0.2078	0.1565	0.1748	0.2208	**0.1951**	0.1640	0.2431	0.1959	0.1999	0.2592	**0.2257**
	AVG	0.2416	0.3113	0.2714	0.2870	0.3124	**0.2988**	0.2564	0.3086	0.2796	0.2947	0.3067	**0.3001**

Subtask	Dataset	Supervised Fine-tuning											
		Qwen3 (14B)			Ministral-3 (14B)			Llama-3.3 (70B)			GPT-OSS (120B)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
ASQP	eng-rest	0.2594	0.2466	0.2528	0.2450	0.2499	0.2474	0.5929	0.5740	**0.5833**	0.5652	0.5538	0.5594
	eng-lap	0.1584	0.1544	0.1564	0.1552	0.1625	0.1588	0.3011	0.2684	**0.2838**	0.2925	0.2648	0.2780
	jpn-hotel	0.0407	0.0381	0.0393	0.0363	0.0409	0.0385	0.4335	0.4290	**0.4312**	0.4204	0.4172	0.4188
	rus-rest	0.1650	0.1748	0.1698	0.1162	0.1229	0.1194	0.5227	0.5107	**0.5166**	0.4542	0.4802	0.4668
	tat-rest	0.0828	0.0870	0.0849	0.0711	0.0733	0.0722	0.4837	0.4542	**0.4685**	0.3958	0.4206	0.4078
	ukr-rest	0.1474	0.1527	0.1500	0.1180	0.1267	0.1222	0.5236	0.5336	**0.5285**	0.4629	0.4863	0.4743
	zho-rest	0.1246	0.1307	0.1276	0.1134	0.1300	0.1212	0.4371	0.4383	**0.4377**	0.4219	0.4124	0.4171
	zho-lap	0.0995	0.1143	0.1064	0.0927	0.1252	0.1065	0.4001	0.3808	**0.3902**	0.3501	0.3392	0.3446
	AVG	0.1347	0.1373	0.1359	0.1185	0.1289	0.1233	0.4618	0.4486	**0.4550**	0.4204	0.4218	0.4209

CiteAssist

Citation Sheet

Generated with citeassist.uni-goettingen.de

(Kaesberg et al., 2024)

BibTeX Entry

```
@inproceedings{lee-etal-2026-dimabsa,
  author={Lee, Lung-Hao and Yu, Liang-Chih and Loukachevitch, Natalia V and Alimova,
      Ilseyar and Panchenko, Alexander and Lin, Tzu-Mi and Xu, Zhe-Yu and Zhou, Jian-Yu
      and Zheng, Guangmin and Wang, Jin and Awasthi, Sharanya and Becker, Jonas and Wahle,
       Jan Philip and Ruas, Terry and Muhammad, Shamsuddeen Hassan and Mohammad, Saif M.},
  title={{D}im{ABSA}: Building Multilingual and Multidomain Datasets for Dimensional
      Aspect-Based Sentiment Analysis},
  abstract={Aspect-Based Sentiment Analysis (ABSA) focuses on extracting sentiment at a
      fine-grained aspect level and has been widely applied across real-world domains.
      However, existing ABSA research relies on coarse-grained categorical labels (e.g.,
      positive, negative), which limits its ability to capture nuanced affective states.
      To address this limitation, we adopt a dimensional approach that represents
      sentiment with continuous valence{--}arousal (VA) scores, enabling fine-grained
      analysis at both the aspect and sentiment levels. To this end, we introduce DimABSA,
       the first multilingual, dimensional ABSA resource annotated with both traditional
      ABSA elements (aspect terms, aspect categories, and opinion terms) and newly
      introduced VA scores. This resource contains 76,958 aspect instances across 42,590
      sentences, spanning six languages and four domains. We further introduce three
      subtasks that combine VA scores with different ABSA elements, providing a bridge
      from traditional ABSA to dimensional ABSA. Given that these subtasks involve both
      categorical and continuous outputs, we propose a new unified metric, continuous F1 (
      cF1), which incorporates VA prediction error into standard F1. We provide a
      comprehensive benchmark using both prompted and fine-tuned large language models
      across all subtasks. Our results show that DimABSA is a challenging benchmark and
      provides a foundation for advancing multilingual dimensional ABSA. We publicly
      released the DimABSA dataset, which was used for Track A of SemEval-2026 Task 3,
      attracting over 300 participants.},
  address={San Diego, California, United States},
  booktitle={Proceedings of the 64th Annual Meeting of the {A}ssociation for {C}
      omputational {L}inguistics (Volume 1: Long Papers)},
  editor={Liakata, Maria and Moreira, Viviane P. and Zhang, Jiajun and Jurgens, David},
  isbn={979-8-89176-390-6},
  pages={40492--40518},
  publisher={Association for Computational Linguistics},
  year={2026},
  month={07}
}
```

Online Access

ACL Anthology https://aclanthology.org/2026.acl-long.1881/