Enhancing Dimensional Aspect-Based Sentiment Analysis: A Comparative Study of Few-Shot and Fine-Tuned LLM Approaches with Semantic Evaluation Metrics

Balbo Enrico, Cutrone Benedetto, Grillo Giovanni, Magliano Dylan, Perno Marco

Politecnico di Torino
{s338793, s333990, s347941, s337915, s339450}@studenti.polito.it

Abstract

Dimensional Aspect-Based Sentiment Analysis (DimABSA) extends traditional ABSA by replacing categorical sentiment labels with continuous valence-arousal (VA) scores, enabling more nuanced emotional representations. In this work, we address the DimASQP task, which requires extracting quadruplets consisting of aspect terms, categories, opinion terms, and VA scores from restaurant reviews. We compare two approaches: a few-shot Llama model and a LoRA fine-tuned Llama model for aspect-opinion-category extraction, both combined with a BERT model for VA prediction. Our experiments show that fine-tuning improves performance significantly, achieving 53.68% cF1 with exact matching compared to 30.51% for few-shot learning (75.9% improvement). Additionally, we propose a novel semantic evaluation metric based on embedding similarity and Hungarian matching that ensures 1-to-1 assignment while better capturing semantic equivalences, further improving cF1 to 63.37% for the fine-tuned approach versus 46.94% for few-shot (35.0% improvement). Our results demonstrate the effectiveness of fine-tuning and the importance of semantic-aware evaluation metrics for DimABSA tasks.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) is a fundamental task in opinion mining that aims to extract fine-grained sentiment information from text. Traditional ABSA approaches classify sentiment into discrete categories (positive, negative, neutral), which limits their ability to capture the nuanced emotional expressions present in natural language. To address this limitation, Dimensional ABSA (DimABSA) has been proposed as a framework that represents sentiment along continuous dimensions of valence (ranging from negative to positive) and arousal (from low to high intensity), following established psychological theories of emotion.

The DimABSA shared task introduces three subtasks of increasing complexity. This work focuses on Subtask 3: Dimensional Aspect Sentiment Quad Prediction (DimASQP), which requires systems to extract quadruplets (Aspect, Category, Opinion, VA) from text. Each quadruplet consists of an aspect term (e.g., "salads"), an aspect category following the Entity#Attribute format (e.g., "FOOD#QUALITY"), an opinion term (e.g., "fantastic"), and a valence-arousal score pair (e.g., "7.88#7.75"), where each score ranges from 1.00 to 9.00.

Large Language Models (LLMs) have demonstrated remarkable capabilities in various NLP tasks, including information extraction and sentiment analysis. However, their application to DimABSA presents unique challenges due to the hybrid nature of the task, which combines discrete element extraction with continuous value prediction. This work investigates the effectiveness of different LLM-based approaches for DimASQP and proposes improvements to both the methodology and evaluation metrics.

1.1 Research Questions

Our research addresses the following questions:

1. **RQ1:** How do few-shot and fine-tuned LLM approaches compare for aspect-opinion-category (AOC) extraction in the DimASQP task?

2. **RQ2:** Is the standard continuous F1 (cF1) metric based on exact matching appropriate for evaluating DimASQP, or can a semantic similarity-based metric provide better assessment of model performance?

3. **RQ3:** What are the error patterns in VA prediction and how do they affect overall system performance?

2 Background

2.1 Aspect-Based Sentiment Analysis

Traditional ABSA tasks focus on extracting sentiment elements from text. The seminal work in this area includes Aspect Term Extraction, Aspect Sentiment Classification, and more recently, joint tasks such as Aspect Sentiment Triplet Extraction (ASTE) and Aspect Sentiment Quad Prediction (ASQP). These tasks typically use categorical sentiment labels (positive, negative, neutral), which provide coarse-grained sentiment information.

2.2 Dimensional Sentiment Analysis

Dimensional sentiment analysis represents emotions in a continuous space defined by valence and arousal dimensions, based on Russell's circumplex model of affect (Russell, 1980). This representation enables more fine-grained distinctions between emotional states compared to categorical approaches. Recent work has begun integrating dimensional sentiment representations into ABSA frameworks, leading to the DimABSA paradigm.

2.3 Large Language Models for Information Extraction

Recent advances in LLMs have shown their effectiveness in various extraction tasks. Few-shot learning with LLMs leverages in-context learning, where models are provided with task examples in the prompt. Fine-tuning approaches, particularly parameter-efficient methods like LoRA (Low-Rank Adaptation), enable adaptation to specific tasks while maintaining computational efficiency and avoiding catastrophic forgetting.

2.4 Evaluation Metrics for Hybrid Tasks

Traditional F1-score for extraction tasks relies on exact string matching, which may be overly strict for semantically equivalent predictions. Recent work has explored semantic similarity-based metrics using embeddings to provide more robust evaluation, particularly for tasks involving paraphrasing or lexical variation.

3 System Overview

We developed two pipeline-based systems for the DimASQP task, both following a two-stage architecture: (1) AOC extraction using Llama, and (2) VA prediction using BERT. The two systems differ in their approach to the first stage: Pipeline 1 uses few-shot learning, while Pipeline 2 employs LoRA fine-tuning. Figure 1 illustrates the architectures of both pipelines.

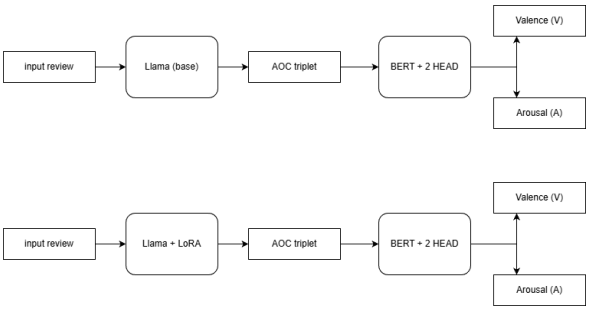

Figure 1: System architecture for both pipelines. Top: Pipeline 1 using few-shot Llama for AOC extraction. Bottom: Pipeline 2 using LoRA fine-tuned Llama for AOC extraction. Both pipelines share the same BERT-based VA prediction stage.

3.1 Dataset

We used the DimABSA Track A Subtask 3 English restaurant dataset, which contains customer reviews annotated with quadruplets. The dataset was split into training, development, and test sets following standard practices. Each instance consists of a text and zero or more quadruplets containing aspect terms, categories (from a predefined ontology), opinion terms, and VA scores rounded to two decimal places.

3.2 Pipeline 1: Few-Shot Llama + BERT

3.2.1 AOC Extraction with Few-Shot Llama

The first pipeline employs Llama in a few-shot setting for extracting aspect terms, categories, and opinion terms. We designed prompts that include:

- Task description explaining the DimASQP objective
- Format specifications for triplet output
- Few-shot examples from the training set
- The input text to be analyzed

The model outputs structured predictions in JSON format containing the extracted AOC elements. We post-process the outputs to ensure consistency with the expected format and validate category values against the predefined ontology.

3.2.2 VA Prediction with BERT

For the second stage, we fine-tuned a BERT model on the training dataset to predict VA scores. The model takes as input the concatenation of the text,

aspect term, opinion term, and category term, and outputs two continuous values representing valence and arousal. The model was trained using Mean Squared Error (MSE) loss to minimize prediction error for both dimensions.

3.3 Pipeline 2: LoRA Fine-Tuned Llama + BERT

3.3.1 AOC Extraction with LoRA Fine-Tuned Llama

The second pipeline uses LoRA to fine-tune Llama on the DimASQP training data. LoRA introduces trainable low-rank matrices into the transformer layers, enabling efficient adaptation while keeping most parameters frozen. This approach allows the model to learn task-specific patterns for AOC extraction while maintaining the general capabilities of the pre-trained LLM.

We configured LoRA with the following hyperparameters:

- Rank: 16
- Alpha: 32
- Dropout: 0.1
- Target modules: query key value projection layers

The training objective was to generate correctly formatted triplets given input text, using teacher forcing and cross-entropy loss.

3.3.2 VA Prediction with BERT

We used the same BERT-based VA prediction training as in Pipeline 1, ensuring a fair comparison focused on the AOC extraction component.

4 Proposed Semantic Evaluation Metric

The standard evaluation metric for DimASQP is continuous F1 (cF1), which combines categorical matching for AOC elements with a penalty based on VA prediction error. Specifically, a prediction is considered a categorical true positive only if the aspect, category, and opinion terms exactly match a gold annotation. This strict matching criterion may fail to recognize semantically equivalent predictions that use different but synonymous expressions.

4.1 Motivation

Consider a gold quadruplet with aspect "tiger roll" and opinion "good", and a prediction with aspect "the tiger roll" and opinion "good". Under exact matching, this prediction would be considered a false positive, despite being semantically very similar. To address this limitation, we propose a semantic similarity-based matching approach.

4.2 Semantic cTP Formulation

Instead of binary categorical matching, we compute continuous similarity scores for each element using sentence embeddings. We use the all-MiniLM-L6-v2 model from sentence-transformers to encode aspect, category, and opinion terms into dense vector representations.

For a predicted triplet t and its candidate gold match g, we compute:

$$\text{sim}_{\text{aspect}} = \cos(\mathbf{e}_A^{(t)}, \mathbf{e}_A^{(g)}) \tag{1}$$

$$\text{sim}_{\text{category}} = \cos(\mathbf{e}_C^{(t)}, \mathbf{e}_C^{(g)}) \tag{2}$$

$$\text{sim}_{\text{opinion}} = \cos(\mathbf{e}_O^{(t)}, \mathbf{e}_O^{(g)}) \tag{3}$$

where $\mathbf{e}_A$, $\mathbf{e}_C$, and $\mathbf{e}_O$ denote the embeddings for aspect, category, and opinion respectively, and $\cos$ is the cosine similarity.

The semantic categorical true positive score is then:

$$cTP_{\text{semantic}}^{(t)} = \frac{\text{sim}_{\text{aspect}} + \text{sim}_{\text{category}} + \text{sim}_{\text{opinion}}}{3} \tag{4}$$

To ensure 1-to-1 matching between predictions and gold annotations, we employ the Hungarian algorithm. For each review, we construct a cost matrix where entry (i, j) represents the negative semantic similarity score $-cTP_{\text{semantic}}^{(i,j)}$ between gold triplet i and predicted triplet j. We apply a semantic threshold (0.9 for Pipeline 1, 0.925 for Pipeline 2) to determine valid pairs: if $cTP_{\text{semantic}}^{(t)}$ is below this threshold, the pair is considered invalid and assigned zero cost (equivalent to a dummy match).

The Hungarian algorithm finds the optimal 1-to-1 assignment that maximizes total semantic similarity (minimizes negative cost). After assignment, we identify:

- **True Positives (TP)**: Valid matched pairs (above semantic threshold)

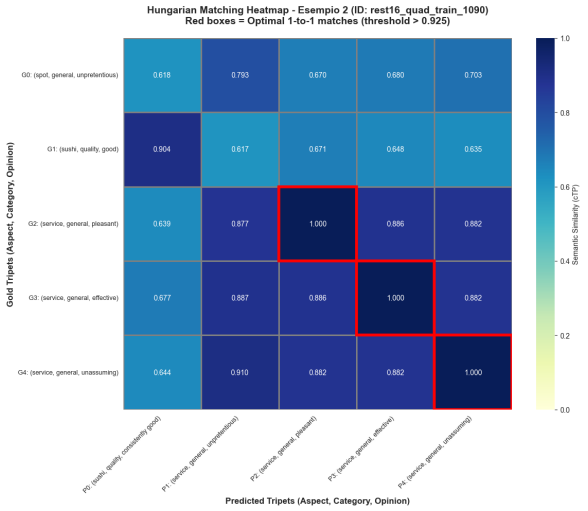

Figure 2: Hungarian Matching heatmap example showing semantic similarity scores between gold triplets (rows) and predicted triplets (columns). Red boxes indicate optimal 1-to-1 matches selected by the Hungarian algorithm with similarity threshold > 0.925. Darker blue indicates higher semantic similarity.

- **False Positives (FP)**: Unmatched predictions
- **False Negatives (FN)**: Unmatched gold annotations

This approach guarantees that each gold triplet matches at most one prediction and vice versa, while properly handling cases where no valid match exists. Importantly, the VA distance penalty is applied only after the matching is established, affecting the final cF1 computation but not the matching itself.

Figure 2 illustrates an example of the Hungarian matching process, showing the semantic similarity matrix between gold and predicted triplets. Red boxes indicate the optimal 1-to-1 matches selected by the algorithm based on similarity scores above the threshold.

The final semantic continuous recall and precision are:

$$cRecall_{\text{sem}} = \frac{\sum_{t \in P_{\text{cat}}} cTP_{\text{sem}}^{(t)} - \text{dist}(VA_p^{(t)}, VA_g^{(t)})}{TP_{\text{cat}} + FN_{\text{cat}}} \tag{5}$$

$$cPrecision_{\text{sem}} = \frac{\sum_{t \in P_{\text{cat}}} cTP_{\text{sem}}^{(t)} - \text{dist}(VA_p^{(t)}, VA_g^{(t)})}{TP_{\text{cat}} + FP_{\text{cat}}} \tag{6}$$

where $\text{dist}(VA_p, VA_g) = \sqrt{(V_p - V_g)^2 + (A_p - A_g)^2}/\sqrt{128}$ as defined in the official evaluation metric.

5 Experimental Results

We evaluated both pipelines on the test split of the English restaurant dataset. Results are reported for both the standard exact matching cF1 and our proposed semantic cF1.

Metric	Pipeline 1 (Few-shot)	Pipeline 2 (LoRA)
Exact Matching (VA-based)		
TP_{cat}	183	314
FP_{cat}	380	227
FN_{cat}	368	237
cPrecision	0.3018	0.5417
cRecall	0.3084	0.5319
cF1	**0.3051**	**0.5368**
Semantic Matching		
TP_{cat}	288	375
FP_{cat}	275	166
FN_{cat}	263	176
cPrecision	0.4644	0.6395
cRecall	0.4745	0.6279
cF1	**0.4694**	**0.6337**
Improvement	+53.87%	+18.06%

Table 1: Performance comparison of Pipeline 1 (few-shot) and Pipeline 2 (LoRA fine-tuned) using exact matching and semantic matching evaluation metrics.

5.1 Few-Shot vs Fine-Tuned Comparison (RQ1)

Table 1 presents the main results. Pipeline 2 (LoRA fine-tuned) significantly outperforms Pipeline 1 (few-shot) under both evaluation metrics. With exact matching, Pipeline 2 achieves 53.68% cF1 compared to 30.51% for Pipeline 1, representing a 75.9% relative improvement. This substantial gain demonstrates the effectiveness of fine-tuning for the AOC extraction component of DimASQP.

The fine-tuned model shows improvements across all metrics:

- TP_{cat} increases from 183 to 314 (+71.6%)
- FP_{cat} decreases from 380 to 227 (-40.3%)
- FN_{cat} decreases from 368 to 237 (-35.6%)

With semantic matching, the improvements are still significantly important:

- TP_{cat} increases from 288 to 375 (+30.2%)
- FP_{cat} decreases from 275 to 166 (-39.6%)
- FN_{cat} decreases from 263 to 176 (-33.1%)

These results indicate that fine-tuning enables the model to better learn the extraction patterns

and category taxonomy specific to the restaurant domain, leading to more accurate predictions with fewer false positives and false negatives.

5.2 Impact of Semantic Evaluation (RQ2)

The semantic evaluation metric reveals higher performance for both pipelines, indicating that exact matching underestimates actual model capabilities. Pipeline 1 improves from 30.51% to 46.94% cF1 (53.87% relative improvement), while Pipeline 2 improves from 53.68% to 63.37% cF1 (18.06% relative improvement).

The larger improvement for Pipeline 1 suggests that the few-shot model produces more lexical variations that are semantically correct but fail exact matching. The fine-tuned model, having learned from specific training examples, produces predictions more aligned with the exact wording in annotations, thus benefiting less from semantic matching.

Analysis of the semantic similarity scores shows that many predictions have cosine similarities between 0.85 and 0.95, indicating they are semantically close but lexically different from gold annotations. This validates our hypothesis that exact matching is overly strict for this task. Figure 3 illustrates the performance differences between VA-based and semantic-based metrics for both pipelines, while also showing the distribution of cTP scores.

Metric	Exact	Semantic
Pipeline 1 (Few-shot)		
cF1	30.51%	46.94%
Improvement	-	+53.87%
Pipeline 2 (LoRA)		
cF1	53.68%	63.37%
Improvement	-	+18.06%

Table 2: Impact of semantic evaluation on cF1 scores for both pipelines.

5.3 VA Prediction Error Analysis (RQ3)

We analyzed the distribution of VA prediction errors to understand their impact on the overall cF1 scores. The VA distance penalty contributes to reducing cTP values even for categorically correct predictions. Figure 4 shows the detailed VA prediction analysis for both pipelines.

The BERT model achieves reasonable VA prediction accuracy, with most errors concentrated in the 0–2 range on the normalized distance scale (where maximum distance is $\sqrt{128} \approx 11.3$). Larger errors (distance > 3) are relatively rare, occurring primarily for examples with ambiguous or complex sentiment expressions.

Comparing the VA distance penalty contribution between the two pipelines, we observe that Pipeline 2 suffers slightly lower VA penalties on average. This may be because the more accurate AOC extraction provides better context for VA prediction, as the BERT model receives correct aspect and opinion terms as input.

5.4 Error Analysis by Category

Breaking down performance by aspect category reveals varying difficulty levels. Categories like FOOD#QUALITY and SERVICE#GENERAL are more common and achieve higher F1 scores, while rarer categories like AMBIENCE#GENERAL show lower performance due to limited training examples.

Common error patterns include:

- Boundary errors: incorrect aspect span extraction (e.g., “great service” vs “service”)
- Category confusion: selecting semantically related but incorrect categories
- Implicit aspects: missing aspects that are implied but not explicitly mentioned
- Multi-word expressions: difficulty capturing complete opinion phrases

6 Conclusion

This work presents a comprehensive study of LLM-based approaches for Dimensional Aspect Sentiment Quad Prediction. Our key findings are:

Main Outcomes:

1. Fine-tuning with LoRA substantially improves AOC extraction performance compared to few-shot learning, achieving 75.9% relative improvement in cF1 with exact matching (30.51% $\rightarrow$ 53.68%) and 35.0% improvement with semantic matching (46.94% $\rightarrow$ 63.37%).

2. The proposed semantic evaluation metric based on embedding similarity and Hungarian matching provides a more robust assessment of model performance, capturing semantically correct predictions that fail exact matching while ensuring 1-to-1 assignment. Both pipelines benefit significantly, with few-shot

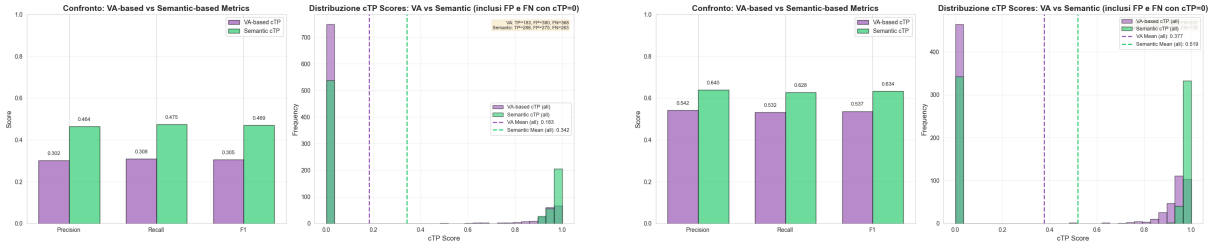

Figure 3: Performance comparison between Pipeline 1 (Few-shot, left) and Pipeline 2 (LoRA, right). Each panel shows: (top) Precision, Recall, and F1 scores comparing VA-based and semantic-based metrics; (bottom) Distribution of cTP scores for both evaluation methods. The fine-tuned model achieves higher cTP scores overall, while semantic matching reveals additional valid predictions for both pipelines.

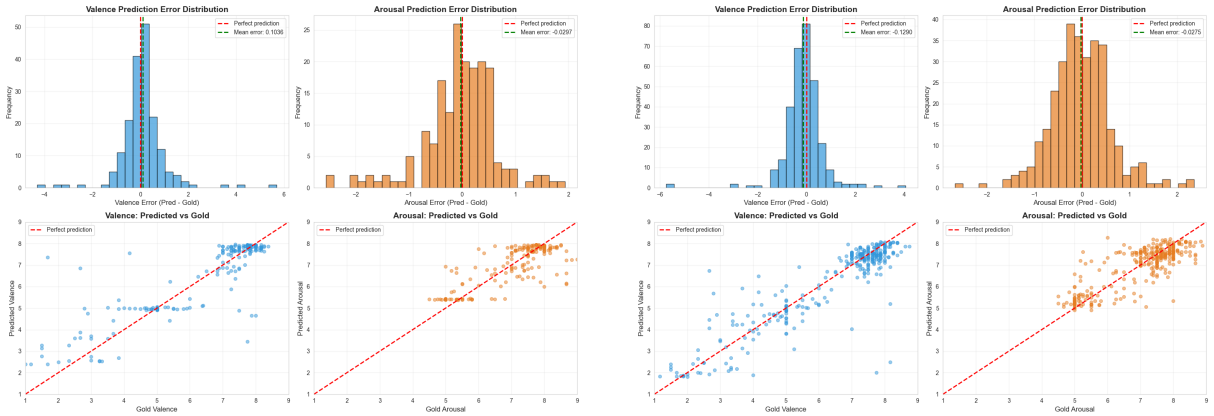

Figure 4: VA prediction error analysis. Left: Pipeline 1 (Few-shot) VA distance distribution. Right: Pipeline 2 (LoRA) VA distance distribution. Both show the distribution of normalized VA distances, with most errors concentrated in the lower range.

showing a 53.87% improvement and LoRA showing a 18.06% improvement.

3. The two-stage pipeline architecture combining LLM for AOC extraction and BERT for VA prediction is effective for the DimASQP task.

Limitations: Our study has several limitations. First, experiments were conducted only on the English restaurant domain, limiting generalizability to other domains and languages. Second, the semantic similarity thresholds were chosen empirically based on development set performance, trying to be as close as possible to exact matching and may not be optimal. Third, the computational cost of fine-tuning limits accessibility for resource-constrained settings. Finally, the two-stage pipeline may propagate errors from AOC extraction to VA prediction.

Future Work: Several directions could extend this work. Evaluating on multi-domain and multilingual datasets would assess generalizability. Developing end-to-end models that jointly predict AOC and VA could reduce error propagation. Finally, studying optimal threshold selection methods for semantic matching would improve the metric's robustness.

References

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