HITSZ-HLT at SIGHAN-2024 dimABSA Task: Integrating BERT and LLM for Chinese Dimensional Aspect-Based Sentiment Analysis

Hongling Xu[1,3*], **Delong Zhang**[1,3*], **Yice Zhang**[1,3*], **Ruifeng Xu**[1,2,3†]
[1] Harbin Institute of Technology, Shenzhen, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
xuhongling@stu.hit.edu.cn, xuruifeng@hit.edu.cn

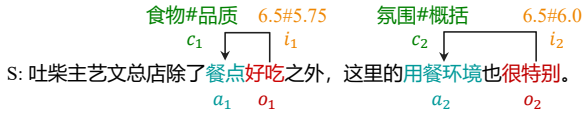
Abstract

This paper presents the winning system participating in the ACL 2024 workshop SIGHAN-10 shared task: Chinese dimensional aspect-based sentiment analysis (dimABSA). This task aims to identify four sentiment elements in restaurant reviews: aspect, category, opinion, and sentiment intensity evaluated in valence-arousal dimensions, providing a concise yet fine-grained sentiment description for user opinions. To tackle this task, we introduce a system that integrates BERT and large language models (LLM) to leverage their strengths. First, we explore their performance in entity extraction, relation classification, and intensity prediction. Based on preliminary experiments, we develop an integrated approach to fully utilize their advantages in different scenarios. Our system achieves first place in all subtasks and obtains a 41.7% F1-score in quadruple extraction.

1 Introduction

Aspect-based sentiment analysis (ABSA) is a fine-grained problem that aims to recognize aspect-level sentiments and opinions of users (Pontiki et al., 2016). ABSA generally involves four fundamental elements: (1) aspect term (a), the mention of the reviewed entity in the text; (2) aspect category (c), a predefined category of the evaluated aspect; (3) opinion term (o), the sentiment word or phrase towards the aspect; and (4) sentiment (Cai et al., 2021; Zhang et al., 2021). For example, in the review "*the sushi was delicious but the staff was unfriendly*", the quadruples are (*sushi*, food#quality, *delicious*, positive) and (*staff*, services#general, *unfriendly*, negative).

Existing ABSA works have typically treated sentiment as coarse-grained polarities, overlooking the complexity of sentiment dimensions. Pioneeringly, the SIGHAN-2024 dimABSA task (Lee et al., 2024) proposes to represent sentiment states as continuous real-valued scores in valence-arousal dimensions, referred to as intensity (i). Valence measures the positivity or negativity, and arousal evaluates the degree of emotional activation (Russell, 1980). As depicted in Figure 1, dimABSA consists of three subtasks: (1) Intensity Prediction, predicting the intensity of the given aspect; (2) Triplet Extraction, extracting the triplets composed of (a, o, i) from the given sentence; (3) Quadruple Extraction, extracting the quadruples composed of (a, c, o, i) from the given sentence.



Subtask	Input	Output	Task Type
Intensity Prediction	S + a_1	$\{i_1\}$	Regression
	S + a_2	$\{i_2\}$	
Triplet Extraction	S	$\{a_1, o_1, i_1\}$ $\{a_2, o_2, i_2\}$	Extraction & Regression
Quadruple Extraction	S	$\{a_1, c_1, o_1, i_1\}$ $\{a_2, c_2, o_2, i_2\}$	Extraction & Classification & Regression

Figure 1: Illustration of three dimABSA subtasks.

To tackle these subtasks, we develop a system that integrates BERT and large language models (LLM), representing two leading paradigms for natural language understanding tasks. Specifically, we devise both BERT-based and LLM-based methods and evaluate them to highlight their respective advantages. The **BERT-based method** employs a pipeline approach that sequentially performs aspect-opinion extraction, pairing and classification, and intensity prediction. We implement three improvements to enhance performance: domain-adaptive pre-training (Gururangan et al., 2020), negative pairs construction, and removing dropout in intensity prediction. The **LLM-based method** transforms the three subtasks into text generation tasks and then fine-tunes a unified model using a multi-task learning strategy. We craft

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), pages 175–185

code-style prompts (Li et al., 2023) to enhance the extraction capabilities of LLMs and employ QLoRA (Dettmers et al., 2024) to reduce memory usage during training.

Through preliminary experiments, we make two observations: (1) in structure extraction (aspect-opinion extraction and pairing), the BERT-based method outperforms the LLM-based method; (2) in intensity prediction, the BERT-based method performs better with continuous values, while the LLM-based method excels in integer-level predictions. Therefore, for Subtask 1, we employ the BERT-based method. For Subtask 2 and 3, we utilize the BERT-based method to derive the aspect, category, and opinion, which are then fed into LLM to generate integer-level intensity predictions.

Our contributions are summarized as follows:

- We propose both BERT-based and LLM-based methods to address the dimABSA tasks and devise various strategies to enhance their performance.
- We analyze the strengths of BERT-based and LLM-based methods in different scenarios and develop an ensemble solution.
- Extensive experimental results demonstrate that our system achieves superior performance and validate the effectiveness of each module. Additionally, we conduct several discussions to provide further insights.

2 Related Work

2.1 Aspect-Based Sentiment Analysis

Aspect-level Sentiment Classification (ASC) is the most fundamental task in ABSA, aiming to identify the sentiment of specific aspect terms in a sentence (Pontiki et al., 2016). Early methods utilized LSTM with attention mechanisms to capture the interaction between aspects and their contextual relationships (Wang et al., 2016b; Ma et al., 2017). With the development of the fine-tuning paradigm, it became mainstream for ASC. Strategies such as interaction mechanism designs (Wu and Ong, 2021; Zhang et al., 2022b), post-training (Xu et al., 2019; Li et al., 2021; Zhang et al., 2023), graph neural networks (Wang et al., 2020; Chen et al., 2022), and contrastive learning (Liang et al., 2021; Cao et al., 2022) have been used to enhance fine-grained sentiment classification. With the advent of LLMs, recent work has explored the effect of LLMs, including in-context learning (Wang et al., 2023b; Xu et al., 2024), chain-of-thought prompting (Fei et al., 2023), and sentiment explanation (Wang et al., 2023a).

Aspect Sentiment Quad Prediction (ASQP) is the most comprehensive task in ABSA, aiming to extract all ABSA quadruples in a review (Cai et al., 2021; Zhang et al., 2021). Research can be categorized into three main types: discriminative methods, generative methods, and LLM-based methods. In the first stream, Cai et al. (2021) applied extract-classify techniques, and Zhou et al. (2023) involved table-based methods to extract aspect-category and opinion-sentiment pairs via simultaneous training. Generative methods, like Zhang et al. (2021), converted quad prediction into paraphrase generation, while Gou et al. (2023) used different permutations as prompts to generate quadruples in various orders for voting. Additionally, some methods enhanced ASQP performance through tree generation designs (Bao et al., 2022; Mao et al., 2022). In the third stream, LLM-based approaches mainly leveraged the rationale of LLMs to improve quad prediction (Kim et al., 2024).

However, early ABSA work solely modeled sentiment with three-class polarities. Our system predicts sentiment in valence-arousal dimensions, providing more fine-grained sentiment information.

2.2 Dimensional Sentiment Analysis

This task focuses on the multiple dimensions of emotional states, such as valence-arousal space (Russell, 1980). Valence measures positivity or negativity, while arousal evaluates excitement or calmness. Previous studies provided various multi-dimensional affective resources, such as lexicons (Warriner et al., 2013) and sentence-level corpora (Preoţiuc-Pietro et al., 2016; Buechel and Hahn, 2017). Meanwhile, some works developed multi-granularity Chinese dimensional sentiment resources, filling the gap in Chinese resources (Yu et al., 2016; Lee et al., 2022). To effectively predict dimensional scores, early approaches mainly used LSTM for modeling, including Densely Connected LSTM for phrase-level predictions (Wu et al., 2017), a relation interaction model for sentence-level predictions (Xie et al., 2021), and a Regional CNN-LSTM model for text-level predictions (Wang et al., 2016a, 2019). With the advancement of Transformer (Vaswani et al.,

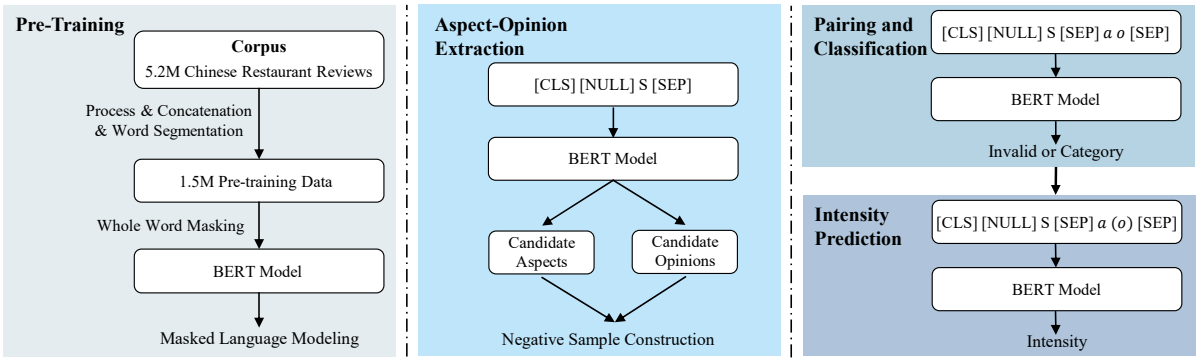


Figure 2: Overview of our BERT framework.

2017), researchers began leveraging PLMs for improvement. For instance, Deng et al. (2023) proposed a multi-granularity BERT fusion framework, and Wang et al. (2024) introduced soft momentum contrastive learning for pre-training. Different from them, our work further evaluates LLMs for dimensional score prediction, providing advanced exploration and analysis.

3 Methods

3.1 Task Definition

Given a sentence $S = [w_1, \cdots, w_T]$ and a predefined aspect term a (a substring of S), the objective of Subtask 1 is to predict the sentiment intensity val, aro, which are continuous values ranging from 1 to 9. For Subtasks 2&3, the input consists only of the sentence S, and the output includes all triplets $(a, o, val\text{-}aro)$ and quadruples $(a, c, o, val\text{-}aro)$, where c and o denote the aspect category and opinion term, respectively. In Subtasks 2&3, (1) the aspect term a and opinion term o can either be a substring of S or be implicit, in which case they are represented by 'NULL'; (2) the aspect category belongs to a pre-defined category set C.

3.2 BERT-based Method

As shown in Figure 2, our BERT framework is structured into four main steps: (i) Domain-adaptive Pre-training, (ii) Aspect-opinion Extraction, (iii) Pairing and Classification, and (iv) Intensity Prediction.

Domain-adaptive Pre-training. Pre-training on sentiment-dense corpus has been proven to enhance downstream sentiment analysis tasks (Xu et al., 2019; Zhang et al., 2023). We first collect 5.2 million open-source Chinese restaurant reviews and conduct data cleaning to remove duplicates and excessively short entries. Subsequently, we concatenate all data and split it according to the maximum length, resulting in 1.5 million pre-training corpora.[1] Moreover, we employ LTP (Che et al., 2010) for Chinese word segmentation and implement a dynamic whole-word masking strategy for masked language modeling (Cui et al., 2021), aiming at enhancing BERT's contextual understanding in the restaurant domain.

Aspect-Opinion Extraction. This step utilizes the pre-trained BERT model to extract aspect and opinion terms. To identify implicit terms, we augment the given sentence by prepending a special [NULL] token. We add this token to the vocabulary and initialize its embedding. Subsequently, we transform the extraction task into a `BIO` sequence labeling task. Using BERT, we predict the category of each token as follows:

$$\boldsymbol{h}_0, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_T = \text{BERT}(S'), \quad (1)$$

$$P(y_t) = \text{softmax}(\text{Linear}(\boldsymbol{h}_t)), \quad (2)$$

where $S' = [[\text{NULL}], w_1, \cdots, w_T]$ denotes the augmented sentence, and y_t represents the tag for the t-th token in the sentence, belonging to {`BA`, `IA`, `BO`, `IO`, `O`}.

Pairing and Classification. This step pairs aspect and opinion terms and determines the corresponding aspect categories. In the BERT-based method, we frame the aspect-opinion pairing and category classification as a unified multi-class classification task. To achieve this, we input the sentence S' along with the aspect term a and opinion term o into BERT and feed the hidden vector at the

[1] Here, we set the maximum sequence length to 512 after adding [CLS] and [SEP] tokens.

[CLS] token position to a classifier, formulated as follows:

$$\boldsymbol{h}_{[CLS]} = \text{BERT}(S', a, o), \quad (3)$$

$$P(c) = \text{softmax}(\text{Linear}(\boldsymbol{h}_{[\text{CLS}]})), \quad (4)$$

where $c \in \{\text{Invalid}\} \cup C$.

In this step, we introduce **negative pairs construction** to mitigate error propagation. During training, the input aspect and opinion terms are true values. However, at the inference stage, these terms are predicted values obtained from the previous step. This discrepancy can lead the classifier to fail in rejecting those aspect and opinion terms with minor boundary errors, resulting in error propagation. To address this issue, we train the extraction model using k-fold cross-validation and incorporate incorrectly extracted aspect and opinion terms into the negative pairs, labeling them as invalid. These negative pairs, along with the true aspect and opinion terms, are then fed into the relation model during training to enhance its robustness against such errors.

Intensity Prediction. This step predicts the valence-arousal scores of an aspect term (for Subtask 1) or an aspect-opinion pair (for Subtask 2&3). We exploit two models for intensity prediction: a regression model and a classification model.

- The regression model obtains the valence and arousal scores s_{val}, s_{aro} by feeding the hidden vector at the [CLS] token position to two separate linear layers:

$$\boldsymbol{h}_{[CLS]} = \text{BERT}(S', a, o), \quad (5)$$

$$\hat{s}_{val} = \text{Linear}(\boldsymbol{h}_{[\text{CLS}]}), \quad (6)$$

$$\hat{s}_{aro} = \text{Linear}(\boldsymbol{h}_{[\text{CLS}]}). \quad (7)$$

  We then compute two losses by mean squared error (MSE) and average them as the regression loss.

- The classification model first converts continuous scores into categories c_{val}, c_{aro} at fixed intervals and then predicts these two categories using two classifiers:

$$\hat{c}_{val} = \text{softmax}(\text{Linear}(\boldsymbol{h}_{[\text{CLS}]})), \quad (8)$$

$$\hat{c}_{aro} = \text{softmax}(\text{Linear}(\boldsymbol{h}_{[\text{CLS}]})). \quad (9)$$

  We use the cross-entropy function to compute two losses and average them as the classification loss.

Furthermore, for the regression model, we adopt the strategy of **removing BERT's internal dropout**. This approach was discussed in a Kaggle forum[2]. The rationale behind this strategy is that BERT's internal dropout may lead to inconsistencies in the variance of neuron activations between the training and inference phases, potentially affecting the numerical stability of the regression.

3.3 LLM-based Method

We transform the dimABSA tasks into text generation tasks and fine-tune a unified LLM using a multi-task learning strategy. To augment the extraction capabilities of the LLM, we employ code-style prompts, as suggested by Li et al. (2023). Additionally, we utilize QLoRA (Dettmers et al., 2024) to reduce memory usage during training. Our framework is illustrated in Figure 3.

Multi-task Learning. Recent work shows that LLMs exhibit excellent task generalization capabilities (Touvron et al., 2023). Inspired by this, we design a multi-task learning strategy for dimABSA to enable the LLM to acquire diverse sentimental knowledge across different tasks. Specifically, we manually construct 6 typical tasks from existing data and labels, including three target subtasks. These are aspect extraction, aspect intensity prediction, aspect-opinion-intensity triplet extraction, aspect-category-opinion triplet extraction, quadruple extraction, and aspect-opinion intensity prediction. These tasks encompass a variety of extraction, classification, and regression task types, thus allowing for a comprehensive learning of aspect-related sentiment knowledge.

Code-style Prompt. LLMs are general-purpose text generation models. To adapt them for specific tasks, it is necessary to craft prompts that direct their output to align with the specific requirements of these tasks. Following Li et al. (2023), we design code-style instructions as prompts. As shown in Figure 3, we formalize each task as Python code, explaining necessary information through comments and standardizing the output format or content via specific code to serve a more instructive role.

Optimization with QLoRA. After completing task selection and prompt design, we construct the

[2] https://www.kaggle.com/competitions/commonlitreadabilityprize/discussion/260729

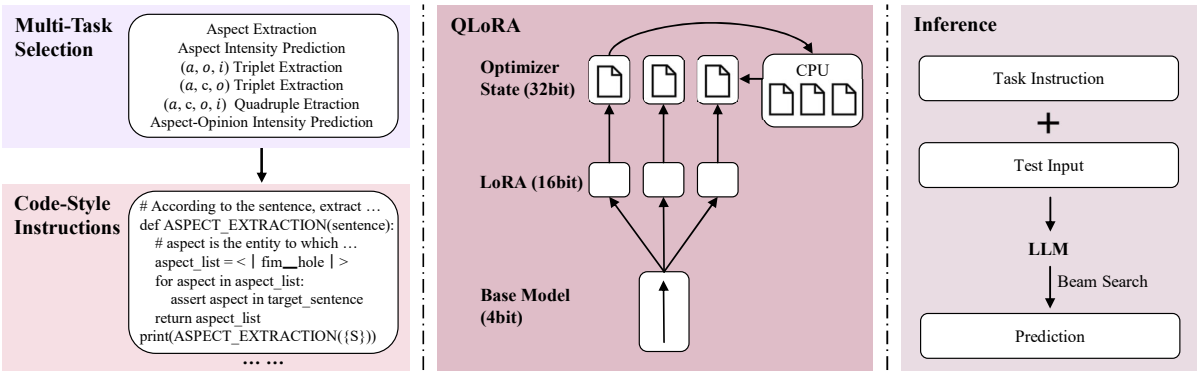


Figure 3: Overview of our LLM framework.

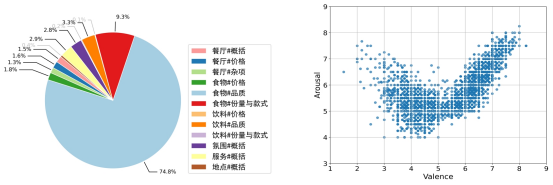
training data for fine-tuning the LLM. This fine-tuning approach is QLoRA (Dettmers et al., 2024). QLoRA is a typical parameter-efficient fine-tuning approach that integrates low-rank matrices into the architecture of LLMs and further quantizes the base model to 4-bit. QLoRA enables us to fine-tune most LLMs on a single 40G A100 GPU.

Inference. We load the parameters of the base model along with those obtained during the fine-tuning phase to perform inference. Utilizing code-style instructions as prompts for each task, we integrate these prompts with the test text inputs for model decoding. During decoding, we set the temperature coefficient to 1 and utilize the beam search strategy (Freitag and Al-Onaizan, 2017) with 'num_beams=2'.

3.4 Ensemble

We conduct preliminary experiments to compare the BERT-based and LLM-based methods. The results indicate that BERT performs better in continuous intensity predicting and aspect-opinion extraction. Conversely, the LLM shows superior performance in integer-level intensity prediction tasks. We suppose this difference arises because LLMs, constrained by their natural language generation output format, may not ensure an accurate understanding of continuous values and extraction, but exhibit better results in coarse-grained predictions due to the larger parameter size.

To fully leverage the strengths of both models, we develop an integrated method. For Subtask 1, we average the predictions of the regression and classification models in the BERT-based method. For Subtasks 2 and 3, we use the BERT-based method to extract (a, c, o) tuples. Then, we input all valid aspect-opinion pairs into the LLM, employing the aspect-opinion intensity prediction



(a) Category distribution. (b) Intensity distribution.

Figure 4: Visualization of training data distribution.

prompt to output integer-level predictions of valence and arousal.

4 Experiments

4.1 Experimental Settings

Datasets. In experiments, we use the Chinese restaurant review dataset provided by the organizer, which includes 6,050 sentences for training, 2,000 sentences for Subtask 1 testing, and 2,000 sentences for Subtasks 2&3 testing. Specifically, the average sentence length, aspect length, and opinion length in the training set are 14.12, 3.14, and 3.07, respectively. Additionally, the training set contains 8,523 quadruples, with 22.81% of quadruples sharing the same aspect in one sentence, 6.10% sharing the same opinion, and 1.98% being implicit aspects. As depicted in Figure 4a, there are 12 predefined categories, with their specific distribution. The training set also includes valence-arousal annotations for aspect-opinion pairs, with real values ranging from 1 to 9. The distribution of valence-arousal annotations is visualized in Figure 4b.

Evaluation Metrics. For Subtask 1, the performance of sentiment intensity prediction is assessed using Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC). These metrics evaluate the difference between model-predicted results and human-annotated scores for valence and

arousal dimensions, respectively.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

where y_i is the actual value and $\hat{y}_i$ is the prediction.

$$\text{PCC} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{11}$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the means of the actual and predicted values, respectively. Lower MAE values indicate more accurate predictions, while PCC ranges from -1 to 1, with higher values indicating a stronger linear correlation. To evaluate Subtasks 2 and 3, the Precision (P), Recall (R), and F1-score (F1) are employed. Meanwhile, valence and arousal values are rounded to the nearest integer. A tuple is correct only if all elements and their combinations match the gold standard.

$$\text{F1} = \frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}} \tag{12}$$

where P denotes the number of correct tuples divided by the total number of extracted tuples, and R denotes the number of correct tuples divided by the total number of standard tuples. Higher values of F1 indicate better performance. Additionally, each metric is calculated independently for valence and arousal dimensions or in combination.

Implementation Details. For BERT, we use *ernie-3.0-xbase-zh* (Sun et al., 2021) as the backbone encoder. The pre-training settings are as follows: batch size of 32, gradient accumulation steps of 12, bf16 mixed precision, 5 training epochs, initial learning rate of 1e-4, and a maximum sequence length of 512. During fine-tuning, we set the learning rate to 2e-5 and the batch size to 32. The fine-tuning epochs are 7 for aspect-opinion extraction, pairing-and-classification, and BERT_{CLS} models, and 6 for the BERT_{REG} model, using the AdamW optimizer. Besides, the interval l for classification is set to 0.25. All models are fine-tuned on five different random seeds and results are aggregated by voting. For LLM, we use *deepseek-7b-instruct-v1.5* (Guo et al., 2024) as the backbone. The training settings include the learning rate of 1e-4, 5 epochs, batch size of 4, bf16 mixed precision, and maximum sequence length of 2048. Besides, the rank of QLoRA fine-tuning is set to 8, and the scaler factor is set to 16. All implementations are based on the PyTorch framework, using NVIDIA A6000 GPUs.

Comparison. We apply different BERT models, LLMs, and the pipeline ensemble method for comparison, including: (1) $\textbf{BERT}_{\textbf{REG}}$, which utilizes regression method for intensity prediction; (2) $\textbf{BERT}_{\textbf{CLS}}$, which employs the interval-based classification approach to predict intensity scores; (3) $\textbf{LLM}_{\textbf{INT}}$, which trains LLM with integer-level intensity; (4) $\textbf{LLM}_{\textbf{DEC}}$, which uses one decimal place intensity and corresponding prompts for training; and (5) **Ensemble**, referring to the ensemble method described in Section 3.4.

4.2 Main Results

The main results are presented in Table 1, from which we can draw the following conclusions:

Firstly, the proposed ensemble method demonstrates obvious superiority, achieving the best results on the majority of metrics. For instance, in Subtasks 2 and 3, the Ensemble method shows improvements of 0.8% and 0.6% in VA-T-F1 and VA-Q-F1 compared to BERT_{CLS}. Compared to LLM_{INT}, these improvements even more achieve 4.1% and 3.8%. These results indicate that our ensemble method effectively leverages the respective strengths of both BERT and LLM in different scenarios, achieving better performance than single-model approaches.

Secondly, we find that the performance of LLM across various metrics is generally inferior to that of BERT. For example, the BERT_{CLS} outperforms LLM_{INT} by 1.1% on V-PCC and surpasses the LLM_{DEC} model by 3.2% on VA-Q-F1. This indicates that BERT is more suitable for predicting the intensity of continuous numerical scores. Additionally, further exploration reveals that although LLM underperforms in Subtasks 2&3, the performance is primarily constrained by aspect-opinion extraction. Conversely, LLM excels in predicting valence-arousal at integer levels, the superiority of Ensemble also supports this viewpoint.

Lastly, we compare different training methods within the same model. We observe that BERT_{CLS} significantly outperforms BERT_{REG} in Subtasks 2 and 3, indicating that the classification model is more suitable for coarse-grained evaluation. Furthermore, comparing LLM_{INT} and LLM_{DEC}, we find that LLM_{DEC} performs better in Subtask 1, whereas LLM_{INT} excels in Subtasks 2 and 3. We assume that in Subtasks 2 and 3, the joint extraction

Methods	Subtask 1				Subtask 2			Subtask 3		
	V-MAE	V-PCC	A-MAE	A-PCC	V-T-F1	A-T-F1	VA-T-F1	V-Q-F1	A-Q-F1	VA-Q-F1
BERT_{REG}	0.287	0.930	0.311	0.773	0.574	0.526	0.405	0.555	0.511	0.393
BERT_{CLS}	**0.279**	0.930	0.316	0.766	0.583	0.543	0.425	0.564	**0.527**	0.411
LLM_{INT}	0.367	0.884	0.394	0.683	0.530	0.498	0.392	0.512	0.482	0.379
LLM_{DEC}	0.294	0.919	0.331	0.738	0.457	0.437	0.312	0.443	0.426	0.302
Ensemble	**0.279**	**0.933**	**0.309**	**0.777**	**0.589**	**0.545**	**0.433**	**0.567**	0.526	**0.417**

Table 1: Main experimental results of our dimABSA system across three Subtasks. V for valence, A for arousal, T for Triplet, and Q for Quadruple. The best scores of each metric are in bold.

Methods	Type	V-Q-F1	A-Q-F1	VA-Q-F1
Voting	BERT	0.557	0.509	0.393
Co-Voting	BERT&LLM	0.563	0.526	0.413
Replace	BERT&LLM	0.565	**0.526**	0.416
Pipeline	BERT&LLM	**0.567**	**0.526**	**0.417**

Table 2: Results of different ensemble strategies for BERT and LLM on Subtask 3.

tasks require generating multiple tuples at once and generating more complex decimals may impact the overall extraction result.

4.3 Analysis of Ensemble

To further verify the effectiveness of the proposed ensemble method, we compare several different ensemble approaches on Subtask 3, including (1) Voting, where results from both types of BERT models are averaged; (2) Co-Voting, where votes are cast only for (a, c, o) tuples that are consistent between LLM and BERT while retaining BERT results for all other tuples; (3) Replace, using intensity results from LLM to replace those of BERT for consistent tuples; (4) Pipeline (ours), where extracted tuples from BERT are input into LLM for intensity prediction. Results are shown in Table 2. We observe that Voting performs poorest, highlighting the importance of combining LLM with BERT. Furthermore, when comparing the last three methods, we find that both Co-Voting and Replace underperform Pipeline. Since LLM excels in coarse-grained intensity prediction, the Pipeline method can more effectively leverage this advantage and achieve superior results.

4.4 Ablation Study

Ablation of BERT. To investigate the effectiveness of various components in BERT, we conduct ablation studies on BERT_{REG}, as shown in Table 3. We observe that removing pre-training (w/o pre-training) leads to a slight decline across all metrics, validating the effectiveness of domain-specific pre-training. Furthermore, eliminating the no-dropout strategy (w/o no-dropout) results in a substantial decrease in most metrics, confirming that dropout can introduce biases in the numerical outputs of regression models. Lastly, omitting the negative sample construction strategy during aspect-opinion pairing training (w/o construction) also degrades performance, proving that this strategy effectively reduces error propagation in the pipeline model.

Ablation of LLM. To explore the effectiveness of various strategies within the LLM framework, we conduct ablation studies on LLM_{INT}, specifically targeting code-style prompts, multi-task learning, and beam search. These modifications are denoted as w/o code prompt, w/o multi-task, and w/o beam search, respectively. The results, as shown in Table 4, indicate that replacing code-style prompts with standard natural language instructions significantly reduces performance in Subtasks 2&3, confirming the effectiveness of this method. Additionally, removing multi-task learning leads to a decline in all metrics, suggesting that LLM benefits from learning generalized emotional knowledge across tasks. Lastly, the performance also declines upon removing the beam search, highlighting the importance of decoding strategy design in LLM inference.

4.5 Effect of Pre-Trained Language Models

To compare the effectiveness of different PLMs on the dimABSA tasks, we conduct experiments on Subtask 1 using several types of models with varying parameter sizes. The results are presented in Table 5. Specifically, we employ our ensemble method to test five different Chinese language models, including *chinese-roberta-wwm-ext* and *chinese-roberta-wwm-ext-large* (Cui et al., 2021), *ernie-3.0-base-zh* and *ernie-3.0-xbase-zh* (Sun et al., 2021), and *erlangshen-deberta-v2-*

Methods	Subtask 1				Subtask 2			Subtask 3		
	V-MAE	V-PCC	A-MAE	A-PCC	V-T-F1	A-T-F1	VA-T-F1	V-Q-F1	A-Q-F1	VA-Q-F1
BERT_{REG}	**0.287**	**0.930**	**0.311**	0.773	**0.574**	**0.526**	**0.405**	**0.555**	**0.511**	**0.393**
w/o pre-training	0.294	0.924	0.313	0.771	0.565	0.520	0.401	0.544	0.502	0.386
w/o no-dropout	0.337	0.933	0.348	**0.779**	0.537	0.503	0.365	0.521	0.487	0.354
w/o construction	-	-	-	-	0.567	0.518	0.399	0.549	0.502	0.387

Table 3: Ablation study of BERT_{REG}.

Methods	Subtask 1				Subtask 2			Subtask 3		
	V-MAE	V-PCC	A-MAE	A-PCC	V-T-F1	A-T-F1	VA-T-F1	V-Q-F1	A-Q-F1	VA-Q-F1
LLM_{INT}	**0.367**	**0.884**	0.394	**0.683**	0.530	**0.498**	**0.392**	0.512	**0.482**	**0.379**
w/o code prompt	0.367	0.882	0.394	0.672	0.515	0.472	0.373	0.495	0.454	0.358
w/o multi-task	0.381	0.876	0.406	0.632	**0.535**	0.481	0.381	**0.514**	0.464	0.367
w/o beam search	0.377	0.880	**0.391**	0.670	0.531	0.489	0.388	0.511	0.472	0.374

Table 4: Ablation study of LLM_{INT}.

Model (Params)	Valence		Arousal	
	MAE	PCC	MAE	PCC
roberta-base (102M)	0.300	0.918	0.310	0.766
ernie-base (118M)	0.300	0.915	0.313	0.762
ernie-xbase (296M)	0.286	0.926	**0.309**	**0.776**
deberta-large (320M)	**0.284**	**0.930**	0.310	0.774
roberta-large (326M)	0.289	0.923	0.314	0.769

Table 5: Results of different pre-trained language models on Subtask 1 (using Ensemble strategy).

320m-chinese (Zhang et al., 2022a). The results indicate that larger models with more parameters tend to perform better than base models. Additionally, our backbone, ernie-xbase, with a moderate parameter size, demonstrates superior performance, ensuring both training efficiency and excellent results for our system.

5 Conclusions

In this paper, we describe our winning system in the SIGHAN-2024 dimABSA task, which involves identifying fundamental sentiment elements in restaurant reviews: aspect, category, opinion, and intensity. Our system integrates BERT and LLM, utilizing their strengths in entity extraction and intensity prediction across three subtasks. The experimental results not only validate the effectiveness of our methods but also underscore the potential of BERT-LLM ensemble strategies in advanced sentiment analysis, providing technical insights and a solid foundation for future research.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (62176076), the Natural Science Foundation of Guangdong (2023A1515012922), the Shenzhen Foundational Research Funding (JCYJ20220818102415032), and the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

Limitations

Despite proposing a novel approach that integrates BERT and LLM for the dimABSA task and achieves promising performance, our study has several limitations. Firstly, our exploration is confined to ensemble methods such as voting and pipeline approaches, leaving deeper integration strategies between BERT and LLMs unexplored. Methods such as knowledge distillation and designing hybrid architectures could potentially enhance performance by capturing more respect advantages. Secondly, our research is constrained by limited computational resources, preventing us from investigating the application of more advanced LLMs to this task. These advanced models might offer better performance in terms of both accuracy and generalization. Lastly, our work does not leverage existing dimensional sentiment resources, such as sentiment lexicons and annotated datasets, which we believe could further improve the prediction of sentiment dimensions. Future work should consider incorporating these resources to enhance the robustness and accuracy of sentiment predictions.

References

Xiaoyi Bao, Zhongqing Wang, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *IJCAI*, volume 2022, pages 4044–4050.

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: Volume 2, Short Papers*, 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 (Volume 1: Long Papers)*, pages 340–350.

Jiahao Cao, Rui Liu, Huailiang Peng, Lei Jiang, and Xu Bai. 2022. Aspect is not you need: No-aspect differential sentiment framework for aspect-based sentiment analysis. In *Proceedings of NAACL-HLT*, pages 1599–1609.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *Coling 2010: demonstrations*, pages 13–16.

Chenhua Chen, Zhiyang Teng, Zhongqing Wang, and Yue Zhang. 2022. Discrete opinion tree induction for aspect-based sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2064, Dublin, Ireland. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yu-Chih Deng, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2023. Towards transformer fusions for chínese sentiment intensity prediction in valence-arousal dimensions. *IEEE Access*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Jieyong Kim, Ryang Heo, Yongsik Seo, SeongKu Kang, Jinyoung Yeo, and Dongha Lee. 2024. Self-consistent reasoning-based aspect-sentiment quad prediction with extract-then-assign strategy. *arXiv preprint arXiv:2403.00354*.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–18.

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.*

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023. CodeIE: Large code generation models are better few-shot information extractors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Toronto, Canada. Association for Computational Linguistics.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of EMNLP*, pages 246–256.

Bin Liang, Wangda Luo, Xiang Li, Lin Gui, Min Yang, Xiaoqi Yu, and Ruifeng Xu. 2021. Enhancing aspect-based sentiment analysis with supervised contrastive learning. In *Proceedings of CIKM*, pages 3242–3247.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4068–4074. AAAI Press.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2Path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225, Dublin, Ireland. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

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.

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

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2016a. Dimensional sentiment analysis using a regional cnn-lstm model. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 225–230.

Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. 2019. Tree-structured regional cnn-lstm model for dimensional sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:581–591.

Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2024. Softmcl: Soft momentum contrastive learning for fine-grained sentiment-aware pre-training. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15012–15023.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of ACL*, pages 3229–3238.

Qianlong Wang, Keyang Ding, Bin Liang, Min Yang, and Ruifeng Xu. 2023a. Reducing spurious correlations in aspect-based sentiment analysis with explanation from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2930–2941, Singapore. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016b. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2023b. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

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

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, Sixing Wu, and Zhigang Yuan. 2017. Thu_ngn at ijcnlp-2017 task 2: Dimensional sentiment analysis for chinese phrases with deep lstm. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 47–52.

Zhengxuan Wu and Desmond C Ong. 2021. Context-guided bert for targeted aspect-based sentiment analysis. In *Proceedings of AAAI*, volume 35, pages 14094–14102.

Housheng Xie, Wei Lin, Shuying Lin, Jin Wang, and Liang-Chih Yu. 2021. A multi-dimensional relation model for dimensional sentiment analysis. *Information Sciences*, 579:832–844.

Hongling Xu, Qianlong Wang, Yice Zhang, Min Yang, Xi Zeng, Bing Qin, and Ruifeng Xu. 2024. Improving in-context learning with prediction feedback for sentiment analysis. *arXiv e-prints*, pages arXiv–2406.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of NAACL-HLT*, pages 2324–2335.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K Robert Lai, and Xuejie Zhang. 2016. Building chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022a. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Kai Zhang, Kun Zhang, Mengdi Zhang, Hongke Zhao, Qi Liu, Wei Wu, and Enhong Chen. 2022b. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. In *Findings of ACL*, pages 3599–3610.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yice Zhang, Yifan Yang, Bin Liang, Shiwei Chen, Bing Qin, and Ruifeng Xu. 2023. An empirical study of sentiment-enhanced pre-training for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9633–9651, Toronto, Canada. Association for Computational Linguistics.

Junxian Zhou, Haiqin Yang, Yuxuan He, Hao Mou, and JunBo Yang. 2023. A unified one-step solution for aspect sentiment quad prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12249–12265, Toronto, Canada. Association for Computational Linguistics.