DS-Group at SIGHAN-2024 dimABSA Task: Constructing In-context Learning Structure for Dimensional Aspect-Based Sentiment Analysis

Ling-ang Meng, Tianyu Zhao and Dawei Song*
School of Computer Science & Technology,
Beijing Institute of Technology, China
{ling.ang.meng, tyzhao, dwsong}@bit.edu.cn

Abstract

Aspect-Based Sentiment Analysis (ABSA) is an important subtask in Natural Language Processing (NLP). More recent research within ABSA have consistently focused on conducting more precise sentiment analysis on aspects, i.e., dimensional Aspect-Based Sentiment Analysis (dimABSA). However, previous approaches have not systematically explored the use of Large Language Models (LLMs) in dimABSA. To fill the gap, we propose a novel In-Context Learning (ICL) structure with a novel aspect-aware ICL example selection method, to enhance the performance of LLMs in dimABSA. Experiments show that our proposed ICL structure significantly improves the fine-grained sentiment analysis abilities of LLMs. Our code is publicly available at: https://github.com/Maydayflower/dimABSA-ICL.

1 Introduction

Aspect-based Sentiment Analysis (ABSA) has been a significant research topic in Natural Language Processing (NLP). The goal of ABSA is to identify specific aspects within a sentence and determine the corresponding sentiment polarity (positive, neutral, or negative) for each aspect (Zhang et al., 2023b). This is different from traditional sentiment analysis (SA) that provides an overall sentiment prediction for the sentence. ABSA has been extensively studied, resulting in numerous effective algorithms.

However, human emotions are inherently continuous rather than discrete, involving two finer-grained dimensions of sentiment, including valence and arousal Russell (1980). As illustrated in Figure 1, the valence dimension represents the degree of pleasure or displeasure sentiment, while the arousal dimension indicates the intensity of the sentiment. In this two-dimensional space, all emotions can be precisely represented. For instance, an emotion with a valence of 7 and an arousal of 7 would be closer to delighted, whereas a valence of 1 and an arousal of 9 would signify a very intense negative sentiment. Extending the traditional SA and ABSA to the two-dimensional space of sentiment has led to Dimensional Sentiment Analysis (DSA) and Dimensional Aspect-Based Sentiment Analysis (dimABSA).

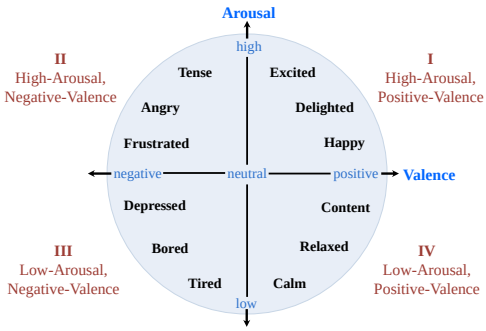

Figure 1: Valence-Arousal space. The picture is originally from (Yu et al., 2016).

As a recently emerging yet largely under-investigated task, dimABSA aims to conduct finer-grained sentiment analysis by assigning corresponding valence and arousal values to each aspect in a sentence, as illustrated in Figure 2. Despite various DSA methods have been developed, which are mainly based on lexicons at the word-level or phrase-level, there is a lack of extensive and systematic studies on the aspect-level dimABSA. This paper aims to fill the gap. Inspired by the success of Large Language Models (LLMs) on the aspect-level sentiment analysis tasks (Wang et al., 2023; Zhang et al., 2023a; Yang et al., 2024), we propose an in-context learning (ICL) framework for dimABSA and evaluates its effectiveness on three mainstream LLMs: qwen-plus (Bai et al., 2023), GPT-3.5 (OpenAI, 2023) and GPT-4 (Achiam et al., 2023). The main contributions of this paper are as follows:

*Corresponding author.

Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), pages 127–132

(1) This paper is the first to explore the performance of LLMs on the dimABSA task. (2) This paper proposes an ICL framework for the dimABSA task, which is facilitated by a novel sample selection method. Experimental results demonstrate that our method significantly improves the performance of LLMs on the dimABSA task.

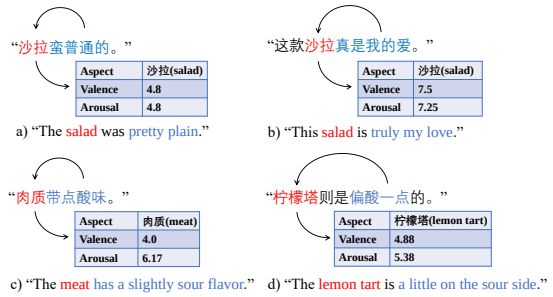

Figure 2: Some examples of dimABSA task demonstrate that when the same aspect is described differently, the aspect can have different valence and arousal. Similarly, when different aspects receive similar evaluations, they can also have different valence and arousal.

2 Related work

2.1 Aspect-based Sentiment Analysis

Early ABSA research is focused on the assessment of single sentiment elements. However, with advancements in the field, ABSA has evolved to include a growing number of sub-tasks, such as Aspect Sentiment Triplet Extraction (ASTE) (Zhang et al., 2020, 2022) and Aspect Sentiment Quad Prediction (ASQP) (Cai et al., 2021; Mao et al., 2022). A more recently emerged area is Dimensional Aspect-Based Sentiment Analysis (dimABSA), which introduces two scalar dimensions to more accurately describe sentiment.

2.2 Dimensional Aspect-Based Sentiment Analysis

Russell proposed a two-dimensional space for more precise emotion modeling, as illustrated in Figure 1. One dimension describes the intensity ranging from pleasant to unpleasant (i.e., Valence), while the other captures the intensity from calm to excited (i.e., Arousal). Based on this model, human emotional states can be represented in a more accurate manner (Bradley and Lang, 1999; Malandrakis et al., 2011). Researchers have incorporated this two-dimensional VA space into sentiment analysis, leading to the development of Dimensional Sentiment Analysis (DSA).

Existing research on DSA is heavily based on lexicons. In the field of Chinese research, the most commonly used lexicon is the Chinese EmoBank proposed by Lee et al., which includes 5,512 single words, 2,998 multi-word phrases, 2,582 single sentences, and 4,969 multi-sentence texts. Consequently, current DSA methods have primarily focused on the word-level (Wei et al., 2011) and phrase-level (Wu et al., 2017), neglecting high-level emotional features. Lee et al. proposed the task of dimensional aspect-based sentiment analysis (dimABSA), extending DSA to the aspect-level. Our work primarily focuses on this task.

2.3 In-Context Learning

In recent years, Large Language Models (LLMs) have demonstrated remarkable performance across various NLP downstream tasks, and have shown excellent In-Context Learning (ICL) capabilities. ICL refers to the ability of LLMs to be applied directly to downstream tasks by adding a few examples to the prompt, without the need for parameter updates (Dong et al., 2023). Demonstration designing is a crucial component in constructing an in-context learning structure. Although the capabilities of LLMs in sentiment analysis have been widely studied (Lian et al., 2023; Wang et al., 2023; Yang et al., 2024), there has been no research exploring the impact of ICL on the dimensional ABSA abilities of LLMs. In this paper, we propose an ICL framework, demonstrating through experiments that our ICL framework significantly enhances the sentiment analysis capabilities of LLMs.

3 Methodology

In this section, we introduce the task definition and the components of our proposed ICL structure.

3.1 Task definition

Given a n-word sentence $s = \{w_1, w_2, ..., w_n\}$, the output of dimABSA is $y = \{(A_1, v_1\#a_1), (A_2, v_2\#a_2), ..., (A_x, v_x\#a_x)\}$, where A_i denotes the representation of an aspect and x represents the number of aspects in the sentence. v_i denotes the valence value of the aspect A_i, ranging from 1 to 9, with 1 representing unpleasant and 9 representing pleasant. a_i denotes the arousal value of A_i, also ranging from 1 to 9, with 1 representing calm and 9 representing excited.

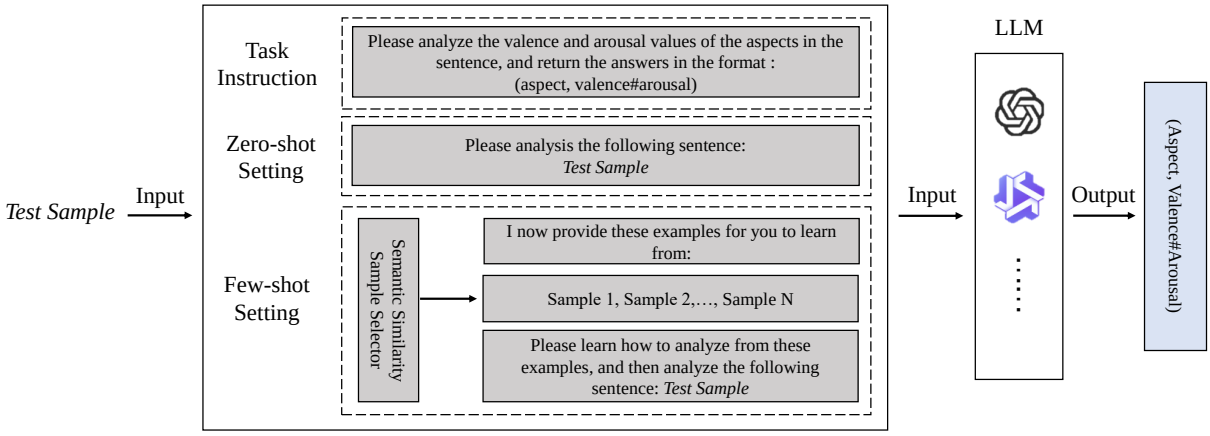

Figure 3: The framework of our proposed ICL method.

3.2 Semantic Similarity Sample Selector (S4)

To select the most helpful samples for the dimABSA task, inspired by (Liu et al., 2021), we choose samples from the training set that are semantically closest to the test samples and use them as examples in the prompt. Given the specific nature of the dimABSA task, we believe that directly computing the semantic similarity of two sentences is inadequate. Instead, the aspect present in the samples should also be considered. The same description can represent different sentiment orientations for different aspects. For instance, the word "sour" typically does not convey negative sentiment when describing a lemon, but when referring to spoiled meat, it strongly indicates a negative sentiment.

Therefore, in our approach, we consider the aspect's presence in the sentence when calculating similarity, leading to an aspect-aware semantic similarity measure.

First, we use BERT (Devlin et al., 2019) to obtain the representation $\mathcal{T}_i$ of the text for calculating semantic similarity. BERT is also used to obtain the representation $\mathcal{A}_i$ of the aspect. The process of obtaining $\mathcal{T}_i$ and $\mathcal{A}_i$ from a sentence S_i and an aspect a is as follows:

$$\mathcal{T}_i = BERT(S_i), \mathcal{A}_i = BERT(a) \quad (1)$$

Then we use Cosine similarity to calculate the semantic similarity between two sentences S_p and S_q. The formula is as follows:

$$sim_t = \text{cosine}(\mathcal{T}_p, \mathcal{T}_q) = \frac{\mathcal{T}_p \cdot \mathcal{T}_q}{\|\mathcal{T}_p\| \cdot \|\mathcal{T}_q\|} \quad (2)$$

Next, we take into consideration the aspects contained in the sentences. Assuming the test sample and the target sample in the training dataset contain m and n aspects respectively, we will calculate the similarity between each pair of aspects across the test sample and the target sample using Equation 3, and ultimately select the highest similarity value for the final computation.

$$sim_a = \max_{\substack{i\in\{1,2,\ldots,m\} \\ j\in\{1,2,\ldots,n\}}} \frac{\mathcal{A}_i \mathcal{A}_j}{\|\mathcal{A}_i\| \|\mathcal{A}_j\|} \quad (3)$$

We believe that the presence of semantically similar aspects in two samples indicates that these reviews were likely given in similar contexts to some extent.

The overall aspect-aware similarity between two samples is ultimately computed, as shown in Equation 4, where α_0 and α_1 are trade-off parameters[1].

$$sim(S_t, S_i) = \alpha_0 \cdot sim_t(\mathcal{T}_t, \mathcal{T}_i) + \alpha_1 \cdot sim_a(\mathcal{A}_t, \mathcal{A}_i) \quad (4)$$

We select N samples from the training set with the highest aspect-aware similarity scores to the test sample for use in subsequent prompt construction. Considering the impact of context length on the performance of LLMs, we set N to 10 in this paper.

3.3 In-context Learning Structure

The prompt we construct comprises a detailed description of the task, including the meanings of Valence, Arousal and Aspect, the input format, the required processing of the input, and the output format. Additionally, depending on various settings, the prompt may also include different sample examples for the LLMs to learn from. Figure 3 illustrates the prompt construction process of our proposed ICL framework.

[1]In our experiment, α_0 and α_1 are both set to 0.5.

Zero-shot setting. To demonstrate the effectiveness of our method, we first test the sentiment analysis capabilities of LLMs in a zero-shot setting. In the zero-shot setting, the prompt does not include additional examples for the LLMs to learn from. The prompt content is: [*Please analyze the following sentence: test sample*].

Few-shot setting with Random Selection. In NLP downstream tasks, the zero-shot setting often fails to achieve satisfactory results. Consequently, a common approach is to randomly select some examples for prompt construction. However, the samples chosen through this method often lack task representativeness, leading to limited improvements in the capabilities of LLMs.

Few-shot setting with S4. To address the lack of effective ICL frameworks in the dimABSA domain, we have proposed a Semantic Similarity Sample Selector (S4) for sample selection, detailed in Section 3.2. After obtaining samples through the S4, we construct the prompt in the following format: [*I now provide these examples for you to learn from: Sample 1, Label 1; Sample 2, Label 2;...; Sample N, Label N. Please learn how to analyze from these examples, and then analyze the following sentence: Test Sample*].

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset

In this task we use the dataset provided by the organizer which contains 3000 sentences. Each sentence contains one or more aspects, and each of these aspects is annotated with corresponding valence and arousal values from 1-9.

4.1.2 Evaluation Metrics

To compare the sentiment analysis capabilities of different models, we use two metrics, Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PCC), to indicate the performance of different models. The formulas for these two metrics are shown in Equation 5 and Equation 6, where $a_i \in A$ represents the ground truth value, and $p_i \in P$ represents the model prediction result. μ_A and μ_P denote the arithmetic mean of **A** and **P**, respectively. σ denotes the standard deviation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i| \quad (5)$$

The smaller MAE value, the better quality of the model's predictions.

$$PCC = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{a_i - \mu_A}{\sigma_A})(\frac{p_i - \mu_P}{\sigma_P}) \quad (6)$$

A larger PCC value indicates a better quality of the model's predictions.

4.2 Main Results

The evaluation results of our proposed ICL structure are presented in Table 1. Among all the results, the GPT-4o model utilizing our proposed ICL framework achieved the best performance on the V-MAE, V-PCC, and A-PCC metrics. The qwen-plus model with our proposed ICL framework, slightly outperformed GPT-4o on the A-MAE metric. Experimental results show that our proposed method significantly improves the sentiment analysis capability of LLMs.

	V-MAE	V-PCC	A-MAE	A-PCC
qwen-plus†	0.697	0.713	0.911	0.300
qwen-plus w. RS	0.541	0.845	0.718	0.345
qwen-plus w. S4	0.542	0.891‡	**0.480**‡	0.495‡
GPT3.5†	0.600	0.882	0.524	0.515
GPT3.5 w. RS	0.460	0.858	0.501	0.490
GPT3.5 w. S4	0.392‡	0.890‡	0.500‡	0.528‡
GPT4o†	0.552	0.838	0.676	0.453
GPT4o w. RS	0.409	0.870	0.500	0.510
GPT4o w. S4	**0.391**‡	**0.900**‡	0.485‡	**0.606**‡

Table 1: Comparison between LLMs with different settings, where † indicates that the LLM is using a zero-shot setting, RS denotes Random Select and S4 denotes our proposed ICL Structure. ‡indicates that our method is significantly better than zero-shot setting and Random Select with p-value < 0.05 based on t-test.

5 Conclusions

This paper explores the enhancement of LLMs for the dimABSA task through ICL. We have designed a sample selection method called Semantic Similarity Sample Selector (S4) and used it to select samples for prompt construction. Experimental results indicate that our proposed ICL framework significantly improves the performance of LLMs for the dimABSA task.

Limitations

The primary limitation of our proposed approach lies in its reliance on proprietary LLMs, which may pose challenges for reproducibility. To achieve optimal results, we did not conduct experiments on

mainstream open-source LLMs such as LLaMA2 and LLaMA3. However, the experimental results on proprietary LLMs demonstrate that our proposed method is significantly effective. In future work, we plan to extend our experiments to include a broader range of LLMs to develop a more performant and generalized approach.

Acknowledgements

This work is funded by the Natural Science Foundation of China (grant no: 62376027) and Beijing Municipal Natural Science Foundation (grant no: 4222036 and IS23061).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. arXiv preprint arXiv:2309.16609.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint, arXiv:1810.04805.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning. Preprint, arXiv:2301.00234.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. ACM Trans. Asian Low-Resour. Lang. Inf. Process., 21(4).

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.

Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Shun Chen, Bin Liu, and Jianhua Tao. 2023. Gpt-4v with emotion: A zero-shot benchmark for multimodal emotion understanding. arXiv preprint arXiv:2312.04293.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? Preprint, arXiv:2101.06804.

Nikos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2011. Kernel models for affective lexicon creation. In Twelfth annual conference of the international speech communication association.

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.

OpenAI. 2023. Chatgpt.

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

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

Wen-Li Wei, Chung-Hsien Wu, and Jen-Chun Lin. 2011. A regression approach to affective rating of chinese words from anew. In Affective Computing and Intelligent Interaction, pages 121–131, Berlin, Heidelberg. Springer Berlin Heidelberg.

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 International Joint Conference on Natural Language Processing.

Li Yang, Zengzhi Wang, Ziyan Li, Jin-Cheon Na, and Jianfei Yu. 2024. An empirical study of multimodal entity-based sentiment analysis with chatgpt: Improving in-context learning via entity-aware contrastive learning. Information Processing & Management, 61(4):103724.

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.

Chen Zhang, Qiuchi Li, Dawei Song, and Benyou Wang. 2020. A multi-task learning framework for opinion triplet extraction. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 819–828, Online. Association for Computational Linguistics.

Chen Zhang, Lei Ren, Fang Ma, Jingang Wang, Wei Wu, and Dawei Song. 2022. Structural bias for aspect sentiment triplet extraction. In COLING.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023a. Sentiment analysis in the era of large language models: A reality check. arXiv preprint arXiv:2305.15005.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023b. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. IEEE Transactions on Knowledge and Data Engineering, 35(11):11019–11038.