Master's Thesis Presentation for the 141th Dissertation Evaluation

# Efficient Fine-tuning Approaches for Large Language Models in Aspect-based Sentiment Analysis

# 측면 기반 감정 분석에서 대규모 언어 모델을 위한 효율적인 미세 조정 접근 방식

Chaelyn Lee

mylynchae@gmail.com

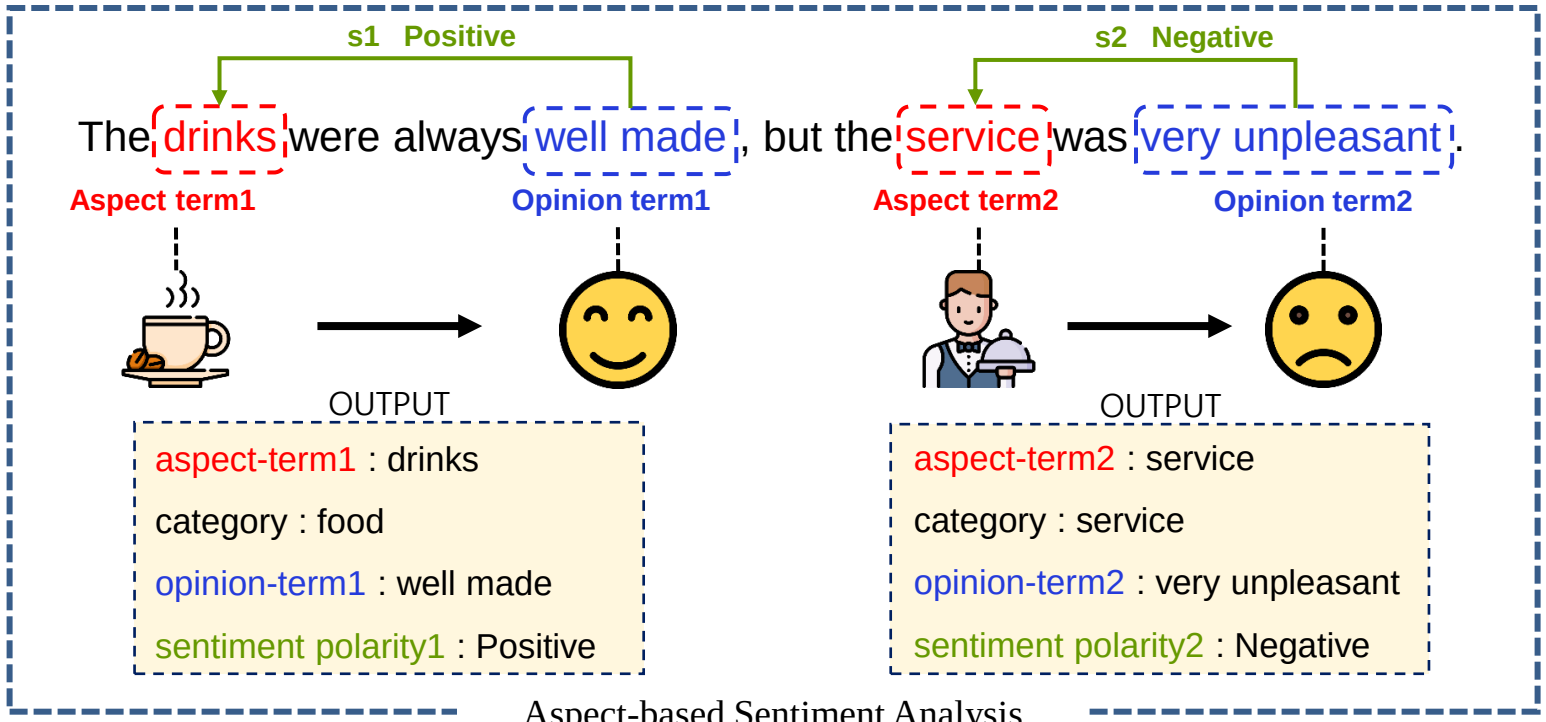
21th May, 2024

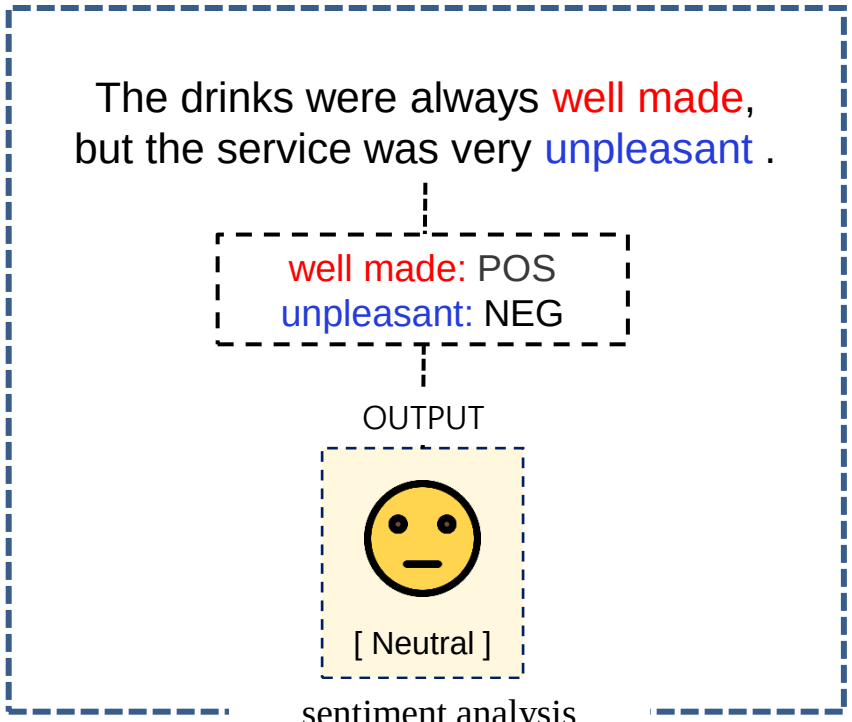
# Contents

# Introduction

## Aspect-based Sentiment Analysis (ABSA)

- Marketing, product sales, and service sectors need to focus on specific aspects through emotional analysis of customer feedback to accurately understand and analyze opinions, thereby continuously enhancing the quality of their products and services.
- In traditional sentiment analysis, the goal is to output a single sentiment for the entire given sentence, but unlike this, the main task of aspect-based sentiment analysis is to find each sentiment for multiple aspects that appear in a sentence.

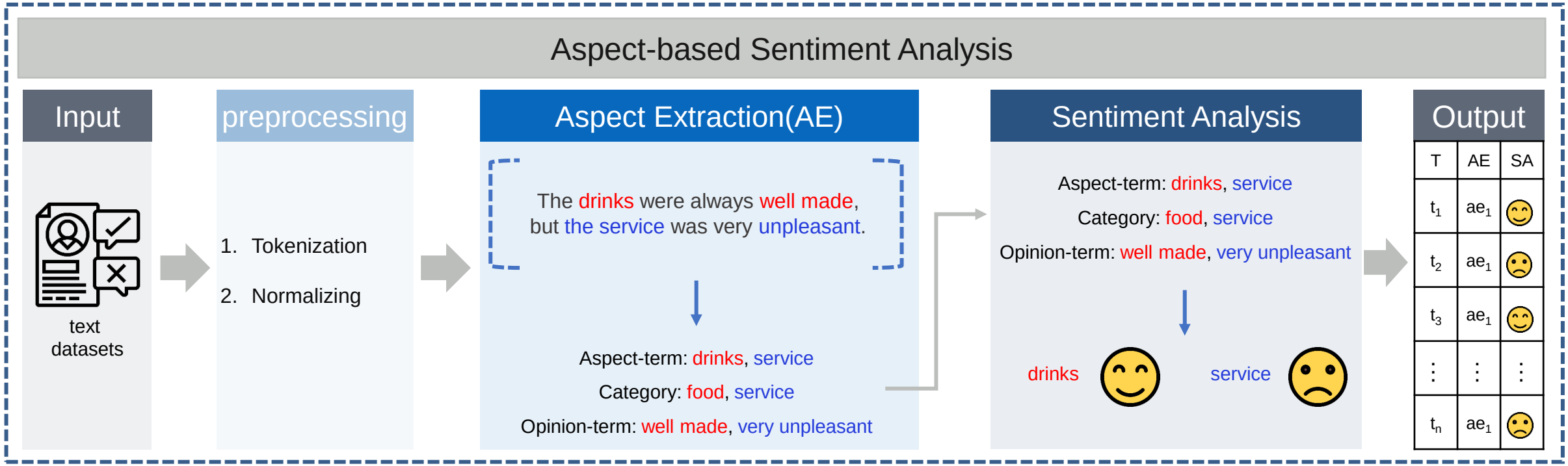


# Introduction

## Objective and Procedure

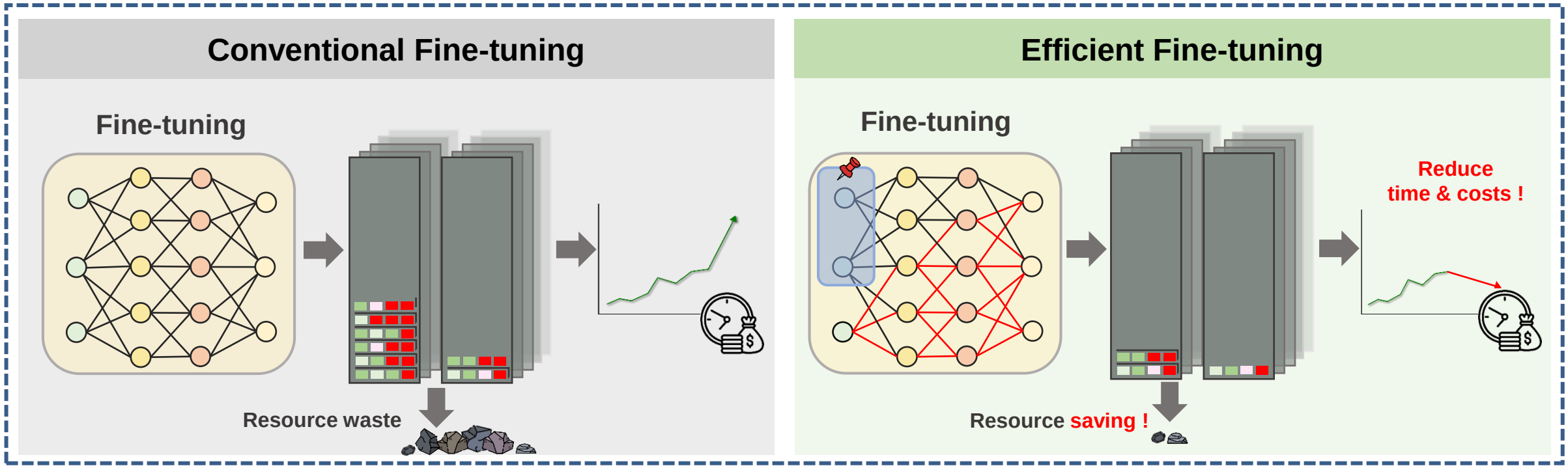
- ABSA identifies and categorizes opinions about specific "aspects" of products, services, or topics in a text, where the input is the text and the included aspect terms, and the output is the sentiment label that classifies the sentiment towards each aspect in the text.
- Conventional methods in ABSA research employ large language models to better understand the context between words in sentences and fine-tune the layers for each task, which results in time delays due to repetitive fine-tuning whenever new information is updated.



# Introduction

## Problem and Strategy

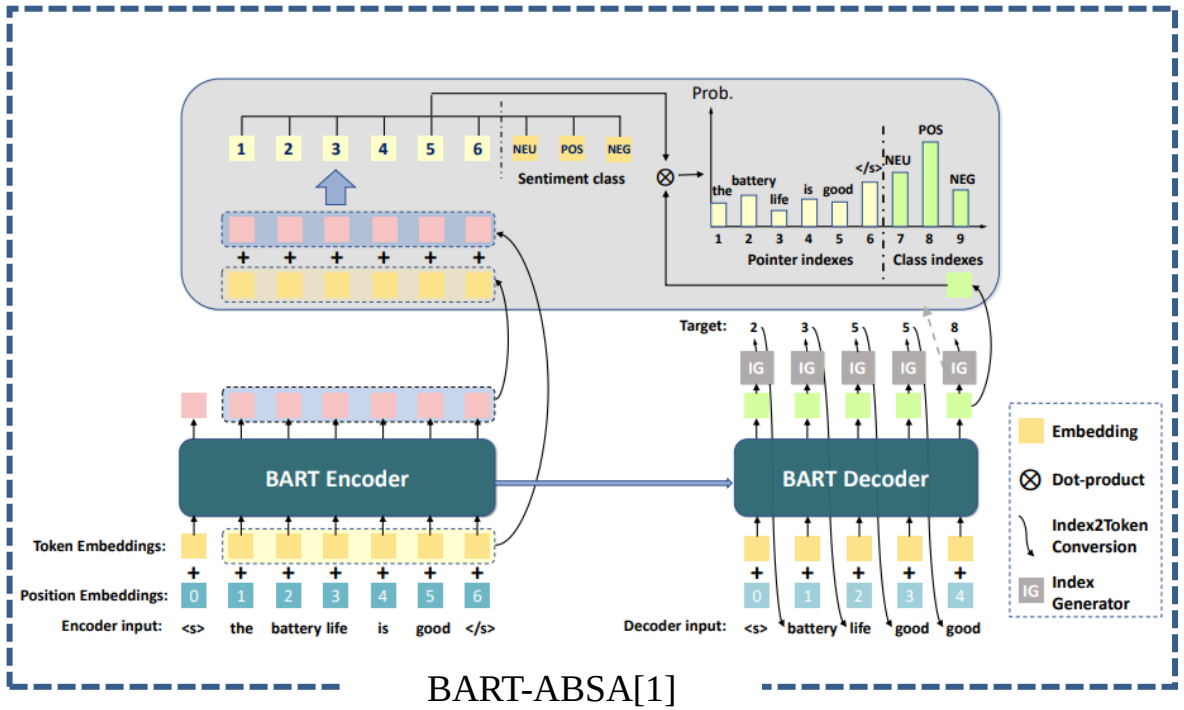
- Customer feedback about new products may include new product-related words, requiring repeated fine-tuning, existing LLM-based ABSA ignores this issue, and repeated fine-tuning may result in unnecessary resources, resulting in poor commercial performance.
- In order to accelerate fine tuning of LLM, this study was the first to apply a strategy to greatly improve the learning speed by using low-rank additional parameters without directly modifying the model parameters, and proved its effectiveness in the ABSA field.



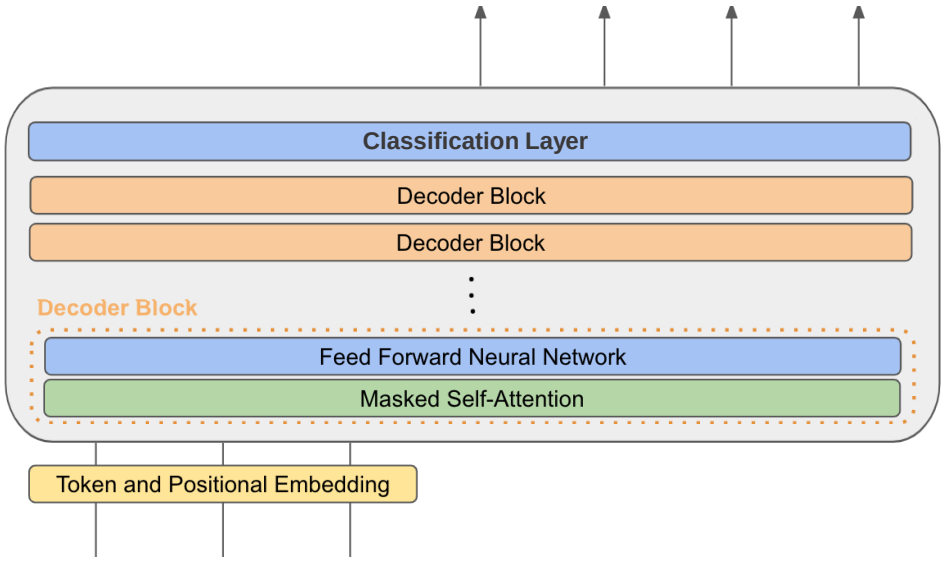
# Related Work

## Conventional Methods

- BART-ABSA uses pre-trained BART models to generate end-to-end target sequences without task-specific layers or sub-models. However, unnecessary parameter adjustments may occur due to the structure specialized for complex denoising tasks.
- GPT-ABSA uses a generative language model to reorganize the task of extracting side terms and polarities into a sequence generation task. However, there are limitations in providing accurate analysis of the latest information due to slow response to the latest data.




BART-ABSA[1]




GPT-ABSA[2]

[1] X. Wang, F. Li, Z. Zhang, G. Xu, J. Zhang, and X. Sun, “A unified position-aware convolutional neural network for aspect based sentiment analysis,” Neurocomputing, vol. 450, pp. 91–103, 2021.
[2] E. Hosseini-Asl, W. Liu, and C. Xiong, “A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis,” in Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 770–787.

# Related Work

## Common Drawback of Conventional Methods

- Large language models such as BART or GPT have problems such as the absence of up-to-date information due to slow information updates and the need to repeat the process of fine-tuning all large amounts of parameters every time, making fast fine tuning difficult.
- The company periodically releases new products, and whenever customer feedback is reflected in the model, unnecessary resources are wasted and a lot of time is consumed, making it impossible to work efficiently and commercial performance may also decline.

| Model | Reference (Author, Year) | Reference (Conference/Journal) | Summary | Drawbacks |
|---|---|---|---|---|
| BART-ABSA | Yan, Hang, et al. 2021 | In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* | Use pre-trained BART models to generate end-to-end target sequences without task-specific layers or sub-models. | Fine-tuning the entire parameters of BART wastes significant time, including memory and processing power. |
| GPT-ABSA | Hosseini-Asl, E., Liu, W., & Xiong, C. 2022 | In *Findings of the Association for Computational Linguistics: NAACL 2022* | Using a GPT, the task of extracting aspect terms and their polarities is reorganized into a sequence generation task. | GPT's information update delay problem limits its performance in areas where real-time updated information is important. |

# Proposed Method

## Notation and Definition

- Text dataset D consists of tuples, each containing a sequence of text sentences S, a set of aspects A, and sentiment polarity sequences Y for each sentence and its aspects. This structured format facilitates nuanced sentiment analysis at both the sentence and aspect levels.
- ABSA is a task that extracts a series of aspect A from each sentence of sequence S included in dataset D, accurately determines the emotional polarity sequence Y for each aspect within the sentence, and effectively classifies them into corresponding categories.

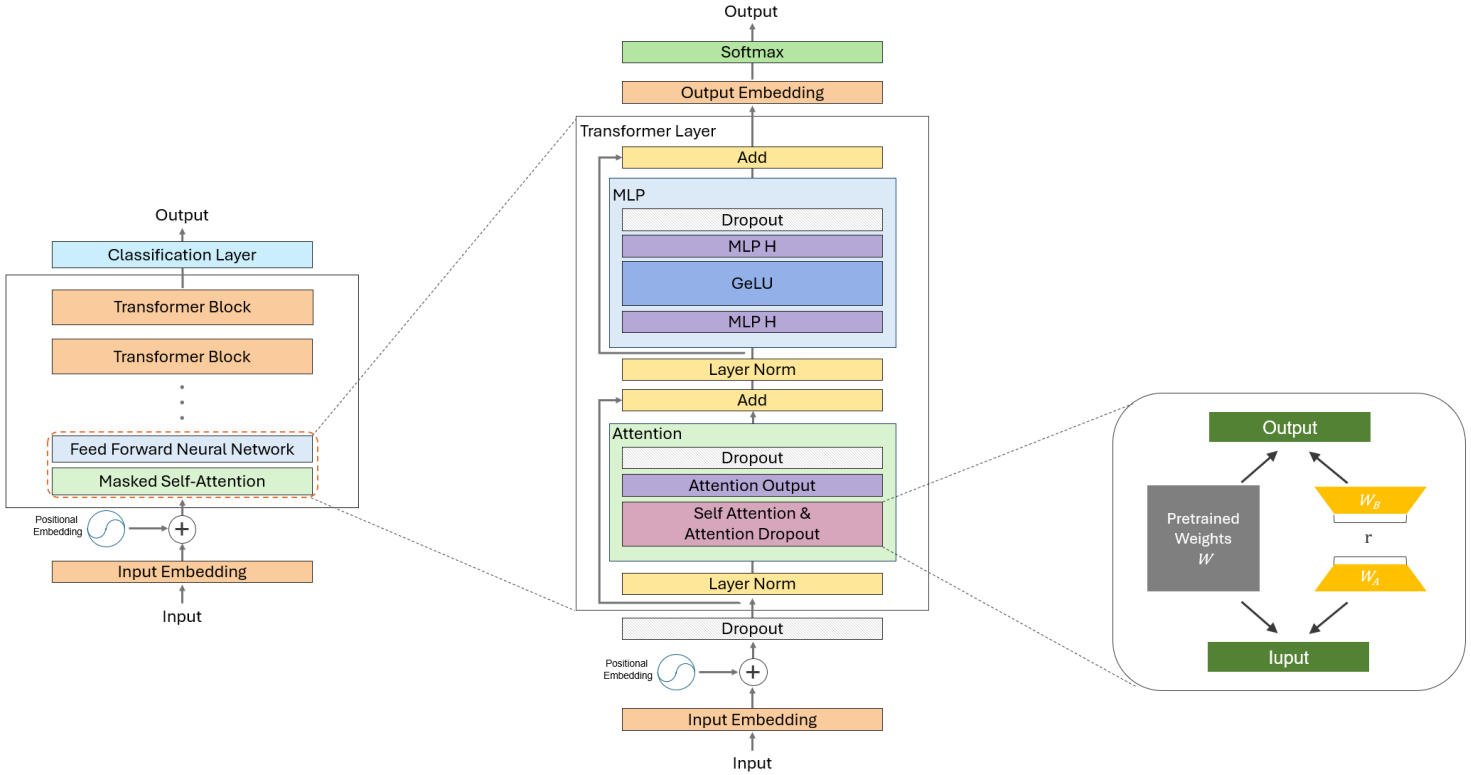
| character | name | Definition |
|---|---|---|
| $S$ | Sentences | Sequence of text sentences $S = \langle s_1, s_2, \dots, s_{\lvert S \rvert} \rangle$ |
| $A$ | Aspects terms | Set of aspects mentioned in each sentence $\mathrm{A} = \{a_1, \dots, a_{\lvert A \rvert}\}$ |
| $Y$ | Sentiment polarity | Sentiment polarity sequence for each sentence and its aspects $Y = \langle y_1, y_2, \dots, y_{\lvert Y \rvert} \rangle$ |
| $D$ | Dataset | Text dataset $D = \{(S_1, A_1, Y_1), \dots, (S_{\lvert D \rvert}, A_{\lvert D \rvert}, Y_{\lvert D \rvert})\}$ |

Notations used in a proposed method

# Proposed Method

## Procedural Overview

- This study leverages the decoder-based architecture of GPT to develop an efficient fine-tuning approach for ABSA tasks. Specifically, within the self-attention blocks, we apply additional low-dimensional projection matrices to the Q, K, and V matrices within each block.
- After fixing the existing method weights, some optimize the rank decomposition matrix [5], which is a special type of weight matrix. As a result, the computational amount of downstream tasks is reduced, enabling efficient training with small memory and computational cost.




Proposed method overview

# Proposed Method

## Algorithm

**Algorithm 2** Fast Fine-tuning for ABSA

**Require:** Text dataset $D = \{(S_1, A_1, Y_1), \ldots, (S_{|D|}, A_{|D|}, Y_{|D|})\}$, Pre-trained generative model $\mathcal{M}_{\text{GPT}}$, Tokenizer $\mathcal{T}_{\text{GPT}}$, Rank decomposition matrices $\Delta\Phi$, Initial weight matrices $\Phi_0$

**Ensure:** Polarity-classified dataset $D' = \{(S_1, A_1, Y_1'), \ldots, (S_{|D|}, A_{|D|}, Y_{|D|}')\}$

**for** $i = 1$ to $|D|$ **do**

$(S, A, Y) \leftarrow D[i]$

$S_t \leftarrow \mathcal{T}_{\text{GPT}}.\text{encode}(S)$

$S_t' \leftarrow \mathcal{M}_{\text{GPT}}(S_t, A)$ with $\Phi_0 + \Delta\Phi$

Compute loss using predicted $S_t'$ and true $Y$

Perform backpropagation to update $\Delta\Phi$ using the computed loss

$Y_i' \leftarrow \arg\max(\text{softmax}(S_t'))$

$D'[i] \leftarrow (S, A, Y_i')$

**end for**

**return** $D'$

Pseudocode of the proposed method

# Proposed Method

## Details

- Lists the inputs needed by the algorithm, which include the text dataset *D*, a pre-trained GPT model $\mathcal{M}_{GPT}$, a tokenizer $T_{GPT}$ , and rank decomposition matrices $\Delta\Phi$, Initial weight matrices $\Phi_0$.

**Require:** Text dataset $D = \{(S_1, A_1, Y_1), \ldots, (S_{|D|}, A_{|D|}, Y_{|D|})\}$, Pre-trained generative model $\mathcal{M}_{\text{GPT}}$, Tokenizer $\mathcal{T}_{\text{GPT}}$, Rank decomposition matrices $\Delta\Phi$, Initial weight matrices $\Phi_0$

- Defines the output of the algorithm, which is a polarity-classified dataset *D* , where each item in the dataset has sentiment predictions.

**Ensure:** Polarity-classified dataset $D' = \{(S_1, A_1, Y_1'), \ldots, (S_{|D|}, A_{|D|}, Y_{|D|}')\}$

# Proposed Method

## Details

- ***For* loop**: Iterates over each example in the dataset *D*

$$\textbf{for } i = 1 \text{ to } |D| \textbf{ do}$$

- $(\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{Y}) \leftarrow \boldsymbol{D}[\boldsymbol{i}]$: Extracts the sentence sequence *S*, aspects *A*, and true sentiment labels *Y* for each data point

$$(S, A, Y) \leftarrow D[i]$$

  - $\boldsymbol{S_t} \leftarrow \mathbf{T}_{\boldsymbol{GPT}}.\boldsymbol{encode}(\boldsymbol{S})$: Uses the GPT tokenizer to convert the sentence sequence *S* into a format suitable for model processing, resulting in tokenized sub tokens $S_t$.

$$S_t \leftarrow \mathcal{T}_{\text{GPT}}.\text{encode}(S)$$

# Proposed Method

## Details

- Feeds the tokenized sentence $S_t$ and aspects $A$ into the GPT model, using the existing pre-trained weights $\Phi_0$ augmented by the low-rank decomposition matrices $\Delta\Phi$ to adapt the model to the ABSA task.

$$S'_t \leftarrow \mathcal{M}_{\text{GPT}}(S_t, A) \text{ with } \Phi_0 + \Delta\Phi$$

- Calculates a loss function to measure the difference between the predicted sentiments and the actual sentiments $Y$. Backpropagation is used to update the $\Delta\Phi$ matrices based on this loss, enhancing the model's accuracy for sentiment prediction.

Compute loss using predicted $S'_t$ and true $Y$
Perform backpropagation to update $\Delta\Phi$ using the computed loss

- Applies a softmax function to the model's outputs $S'_t$ to obtain a probability distribution over potential sentiment classes. The sentiment class with the highest probability is chosen as the predicted sentiment $Y'_t$ for that data point.

$$Y'_i \leftarrow \arg\max(\text{softmax}(S'_t))$$

- Updates the output dataset $D$ with the new tuple containing the original sentence $S$, aspects $A$, and the predicted sentiment $Y'_i$ .

$$D'[i] \leftarrow (S, A, Y'_i)$$

**end for**

**return** $D'$

# Experimental Settings

## Dataset

- The experiments evaluate the model's performance using the Semeval-2014 Task 4 [3] and Semeval-2016 Task 5 [4] datasets constructed for ABSA. This dataset is based on consumer reviews and includes ratings for the laptop and restaurant domains.
- These datasets aim to identify specific aspects of a given target entity and the emotions expressed about those aspects. SemEval-2016 Task 5 is an ABSA-focused dataset that, like SemEval-2014, uses restaurant review data, but does not include conflict labels.

| | | Train | | | | | Test | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Total | Positive | Negative | Neutral | Conflict | Total | Positive | Negative | Neutral | Conflict |
| Semeval2014 | Laptop | 1,914 | 818 | 690 | 369 | 37 | 654 | 341 | 128 | 169 | 16 |
| | Restaurant | 2,980 | 1744 | 643 | 520 | 73 | 1,130 | 726 | 14 | 195 | 195 |
| Semeval2016 | Restaurant | 1,289 | 933 | 307 | 49 | x | 465 | 362 | 27 | 76 | x |

Datasets

```
<sentence id="11351725#582163#9">
   <text>Our waiter was friendly and it is a shame that he didnt have a supportive
staff to work with.</text>
        <aspectTerms>
            <aspectTerm term="waiter" polarity="positive" from="4" to="10"/>
            <aspectTerm term="staff" polarity="negative" from="74" to="79"/>
        </aspectTerms>
        <aspectCategories>
            <aspectCategory category="service" polarity="conflict"/>
        </aspectCategories>
</sentence>
```

Format of the Semeval data set

# Experimental Settings

## Hyperparameter Settings

- Batch size: 16
- Epochs : 20
- Learning Rate: 3e-2
- Cross Validation: 80 : 20
- Repeat experiment: 10

## Evaluation Measures

- Mean Average Precision (Accuracy) $\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$
- Mean Reciprocal Rank (Macro F1) $\frac{\sum_{i=1}^{n} \text{F1 Score}_i}{n}$
- Precision $\frac{TP}{TP+FP}$
- Recall $\frac{TP}{TP+FN}$

# Experimental Results

## Comparison to Conventional Methods

- For the 14lap, the proposed method showed high performance in all evaluation metrics compared to the comparison method. This means better understanding and processing of text data overall. It also means that we can effectively identify relevant emotions within a dataset.
- For the 16res dataset, the strength of the proposed method is particularly noticeable, achieving 91.56% and 75.11% in accuracy and F1 score, respectively. This shows that the proposed method adapts better to newer data and more complex emotional expressions.

| | | **Proposed** | GPT-ABSA | BART-ABSA |
|---|---|---|---|---|
| 14lap | Accuracy | **79.37±0.0055** | 78.65±0.0077▼ | 70.09±0.1191▼ |
| | f1-score | **63.57±0.0362** | 63.29±0.0209▼ | 48.52±0.2099▼ |
| | Precision | **71.00±0.0702** | 70.74±0.0391▼ | 48.51±0.2369▼ |
| | recall | **63.29±0.0251** | 63.15±0.0171▼ | 51.60±0.1786▼ |
| 14res | Accuracy | **84.84±0.0077** | 84.31±0.0073▼ | 79.23±0.0818▼ |
| | f1-score | 65.54±0.0181 | **66.92±0.0139** | 60.62±0.1656▼ |
| | Precision | 68.95±0.1436 | **72.79±0.0268** | 60.88±0.1853▼ |
| | recall | **64.01±0.0209** | 63.64±0.0143▼ | 62.68±0.1495▼ |
| 16res | Accuracy | **91.56±0.0030** | 89.87±0.0122▼ | 84.79±0.0612▼ |
| | f1-score | **75.11±0.0120** | 70.45±0.0371▼ | 51.77±0.1954▼ |
| | Precision | 78.48±0.0183 | **81.08±0.0435** | 52.15±0.2284▼ |
| | recall | **74.65±0.0132** | 67.51±0.0314▼ | 52.91±0.1697▼ |

Experimental Results of method performance

# Experimental Results

## Friedman Test

- If the significance level is set to 0.05, the test considers a statistically significant difference if the p-value is less than or equal to this value. The p-values for each dataset are 0.049787, 0.017644. The p-value is less than 0.05 and is therefore statistically significant.

| Evaluation Measure | Friedman statistic | p-value (α = 0.05) |
|---|---|---|
| Accuracy | 6.0 | 0.049787 |

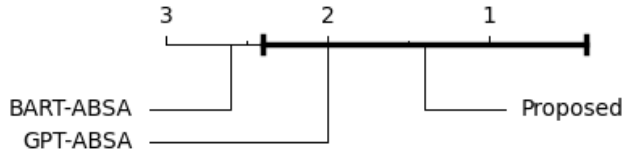
Semeval-2014 Laptop dataset

| Evaluation Measure | Friedman statistic | p-value (α = 0.05) |
|---|---|---|
| Accuracy | 20.04 | 0.017644 |

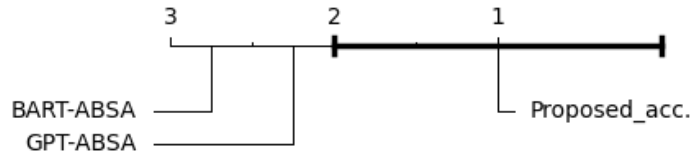
Semeval-2016 Restaurant dataset

## Bonferroni-Dunn test (α = 0.05, CD = 1.002205)

- The test is run at a significance level of α = 0.05, and the critical difference (CD) is set to 1.002205. CD value is the minimum rank difference required between two methods before the performance difference is considered statistically significant at a given alpha level.
- An alpha level of 0.05 shows that the proposed method is a superior method for the given task. It also exceeds the critical difference threshold of 1.002205, showing statistically significant superiority over other methods.




Semeval-2014 Laptop dataset



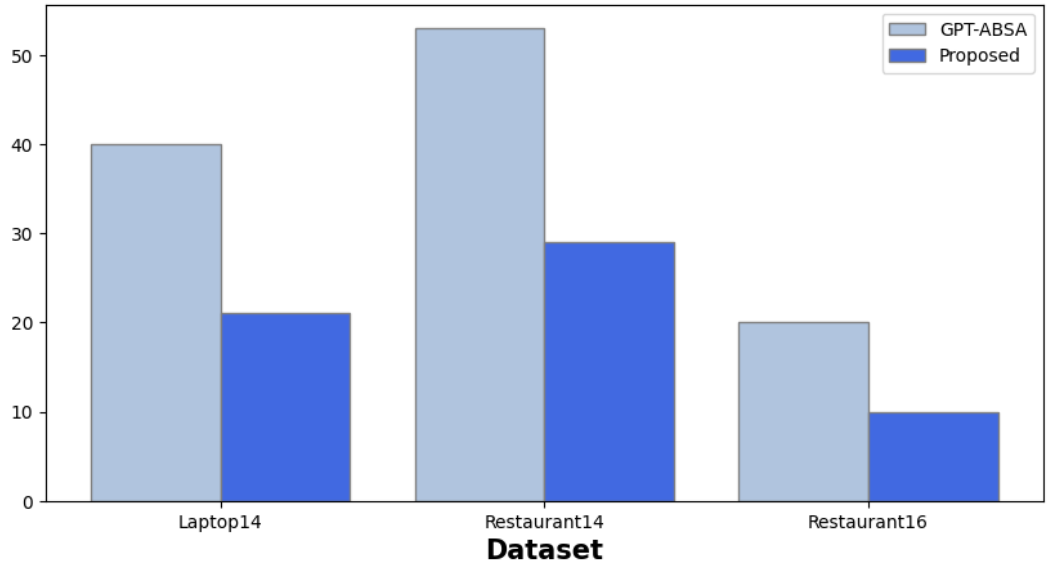

Semeval-2016 Restaurant dataset

Comparison of Bonferroni-Dunn test results between the proposed method and the comparative method

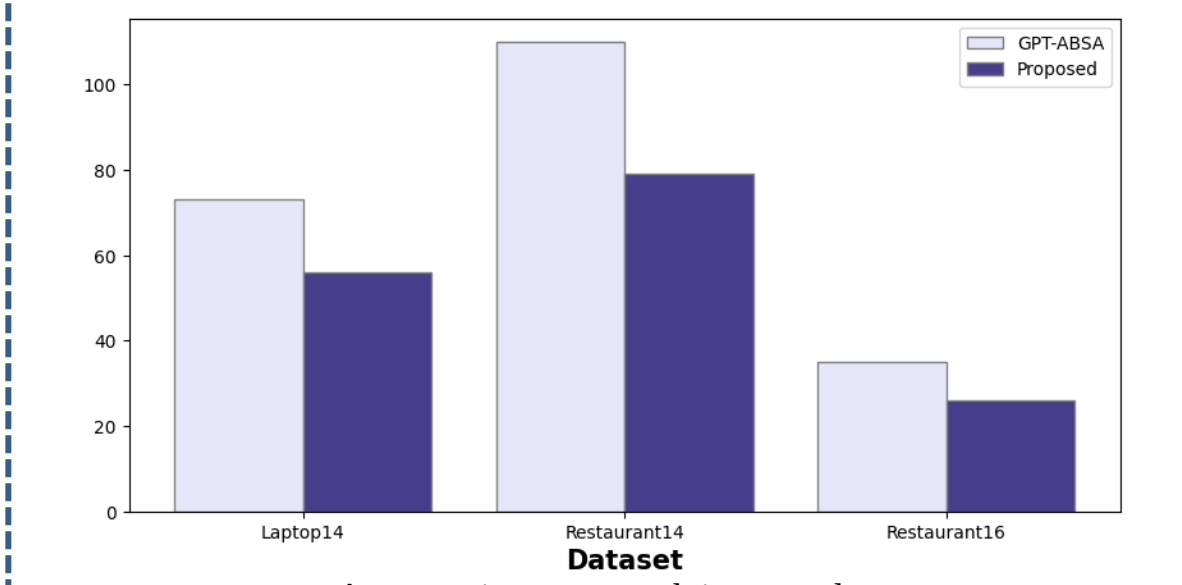
# Experimental Results

## In-depth Analysis

- The number of epochs applied in the experiment is 20, and the total training time is calculated using Python's time module. The results show that the proposed method reduces the overall learning time by up to 50% by reducing the number of learnable parameters.
- The result is the average of the sum of the training times of the epochs divided by the number of epochs. The proposed method enables efficient learning by reducing the learning time for each epoch by up to 30 seconds or more compared to the conventional method.




Training Time in minutes




Average time per epoch in seconds

# Conclusion and Contribution

## Conclusion

- This study proposed an efficient and fast fine-tuning method that reduces the execution time required for fine-tuning while maintaining analysis accuracy when learning new information in a large-scale language model in aspect-based sentiment analysis.
- Conventional methods repeatedly fine-tune already learned parameters, resulting in poor commercial performance. The proposed method solved this problem by fixing the existing weights and optimizing the low-rank matrix to adapt the model to a specific task.
- The experimental results showed that the learning time was noticeably reduced with the proposed fine-tuning method while the proposed method maintained the accuracy of polarity prediction, and the significance was verified through the Bonferroni-Dunn statistical test.

## Contribution

- In International Conferences
  - A Crux on Current Neural Collaborative Filtering Methods, ***ICEIC***, 2023
- In Domestic Conferences
  - An Efficient Fine-Tuning of Generative Language Model for Aspect-Based Sentiment Analysis, ***ICCE***, 2024
- Training and Event
  - ELTE Budapest Summer University, Eötvös Loránd University, Budapest, **Hungary**, 2023
  - 2023 Singapore K-Tech Festival - AI/ML Innovation Research Center Booth, Singapore, **Singapore**, 2023

# Thank you

# Reference

[1] X. Wang, F. Li, Z. Zhang, G. Xu, J. Zhang, and X. Sun, “A unified position-aware convolutional neural network for aspect based sentiment analysis,” Neurocomputing, vol. 450, pp. 91–103, 2021.

[2] E. Hosseini-Asl, W. Liu, and C. Xiong, “A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis,” in Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 770–787.

[3] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, “SemEval-2014 Task 4: Aspect Based Sentiment Analysis,” in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Aug. 2014, pp. 27–35. doi: 10.3115/v1/S14-2004.

[4] Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., ... & Eryiğit, G. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)* (pp. 19-30). Association for Computational Linguistics.

[5] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.