# Using Review Combination and Pseudo-Tokens for Aspect Sentiment Quad Prediction

**Jiazhou Chen**[1,2,3], **Xu Jia**[1,2,3], **Ruiqiang Guo**[1,2,3]*
[1] College of Computer and Cyber Security, Hebei Normal University
[2] Hebei Provincial Key Laboratory of
Network & Information Security, Hebei Normal University
[3] Hebei Provincial Engineering Research Center for
Supply Chain Big Data Analytics & Data Security, Hebei Normal University
jzchen@stu.hebtu.edu.cn,
{jiaxu,rqguo}@hebtu.edu.cn

## Abstract

Aspect Sentiment Quad Prediction (ASQP) aims to identify quadruples consisting of an aspect term, aspect category, opinion term, and sentiment polarity from a given sentence, which is the most representative and challenging task in aspect-based sentiment analysis. A major challenge arises when implicit sentiment is present, as existing models often confuse implicit and explicit sentiment, making it difficult to extract the quadruples effectively. To tackle this issue, we propose a framework that leverages distinct labeled features from diverse reviews and incorporates pseudo-token prompts to harness the semantic knowledge of pre-trained models, effectively capturing both implicit and explicit sentiment expressions. Our approach begins by categorizing reviews based on the presence of implicit sentiment elements. We then build new samples that combine those with implicit sentiment and those with explicit sentiment. Next, we employ prompts with pseudo-tokens to guide the model in distinguishing between implicit and explicit sentiment expressions. Extensive experimental results show that our proposed method enhances the model's ability across four public datasets, averaging 1.99% F1 improvement, particularly in instances involving implicit sentiment[1].

## 1 Introduction

Aspect-Based Sentiment Analysis (ABSA) focuses on identifying opinions and sentiments related to specific aspects in user-generated content (Pontiki et al., 2014). A key challenge in ABSA is the task of Aspect Sentiment Quad Prediction (ASQP) (Zhang et al., 2021a). ASQP involves extracting four elements from a sentence: aspect terms, aspect category, opinion terms, and sentiment polarity and presenting them as quadruples.

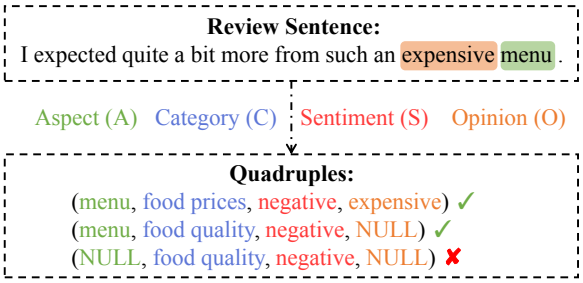


Figure 1: An example of the ASQP task is illustrated, where the aspect term, aspect category, opinion term, and sentiment polarity are highlighted in green, blue, orange, and red, respectively. "NULL" indicates cases where the review contains implicit aspect terms or opinion terms. Current models tend to confuse implicit and explicit sentiment expressions, as seen in the last row of the extracted sentiment quadruples.

For example, in the sentence "I expected quite a bit more from such an expensive menu," two quadruples are identified: (menu, food prices, negative, expensive) and (menu, food quality, negative, NULL), as illustrated in Figure 1. Initially, some works treated ASQP as a classification task, Qiu et al. (2011) proposed the Double Propagation (DP) method, which first identifies aspect-opinion-sentiment triples, then classifies the aspect category. Building on Wan et al. (2020), Cai et al. (2021) introduced TAS-BERT-ACOS, which can simultaneously extract category-sentiment pairs and aspect-opinion pairs. Then, Xiong et al. (2023) proposed the BART-based Contrastive and Retrospective Network (BART-CRN), which uses a machine reading comprehension-based contrastive and retrospective learning module to establish connections between all quadruples. Recently, significant progress has been made in generative methods for aspect sentiment analysis. These approaches involve training models by constructing output targets using different schemas (Zhang et al., 2021a,b; Bao et al., 2022), adopting different data augmentation meth-

* Corresponding author.

[1]We release our code at https://github.com/chienarmor/absa-implicit



*Findings of the Association for Computational Linguistics:*
*NAACL 2025*, pages 3872–3883

ods (Hu et al., 2022; Gou et al., 2023; Zhang et al., 2024b), or employing template-agnostic prompt selection methods (Hu et al., 2023).

However, existing approaches often fail to differentiate between explicit and implicit sentiment expressions in reviews, making it difficult for models to capture implicit sentiment accurately. We reproduce the experiments from Gou et al. (2023) and find that if there is only implicit sentiment in the sample, that is, both the aspect term and the opinion term are marked as NULL, then the aspect term or opinion term in the quaternion extracted by the current model will be the content of the sentence instead of NULL. This phenomenon occurs frequently, and vice versa. When the aspect term or opinion term is implied rather than directly stated, models tend to confuse the two types of sentiment, resulting in inaccurate extraction of sentiment quadruples.

In light of this observation, we propose a novel framework for Aspect Sentiment Quad Prediction (ASQP) that leverages review combination and pseudo-tokens to address the above challenge. Firstly, many reviews feature only explicit or implicit sentiment elements, so our framework merges reviews containing both types to generate more complex sentiment-rich samples. Based on these enhanced reviews, we treat sentiment quadruple extraction as a generation task, guiding model training through prompts that incorporate pseudo-tokens. These pseudo-tokens are tailored to match the specific sentiment expressions (explicit or implicit) present in each review. Finally, we apply constrained decoding to ensure consistency between the model's generated results and the sentiment elements identified in the review. We extensively evaluate our framework across four public datasets, including the ACOS and ASQP benchmarks. The experimental results demonstrate that our method, despite its simplicity, outperforms existing approaches. Detailed analysis reveals several strengths of our framework, such as the use of pseudo-token prompts and the combination of both implicit and explicit sentiment reviews, which significantly enhance the model's capacity to capture complex emotional nuances. Our framework shows substantial improvements over strong baseline models, both in fully supervised and low-resource settings.

Our major contributions are as follows:

(1) We introduce a framework for extracting sentiment quadruples in the ASQP task by combining original reviews and embedding the pseudo-tokens to the prompt.

(2) We introduce a method that classifies reviews based on implicit or explicit sentiment expressions and then creates new samples by merging the classified reviews. Additionally, we use prompts with pseudo-tokens to guide the model in distinguishing between complex sentiment expressions.

(3) We conducted extensive experiments on four ASQP datasets, demonstrating that our approach outperforms current strong baselines.

## 2 Methodology

### 2.1 Formulation and Overview

In this section, we discuss the Aspect Sentiment Quad Prediction (ASQP) task. Given an input review sentence $X = \{x_1, x_2, \cdots, x_l\}$ with length $l$, the objective is to extract all sentiment quadruples $Q = \{(a, c, s, o)\}_{j=1}^{M}$, where $M$ represents the total number of extracted quads. Here, $a$ denotes the aspect term, which can either be a specific term from the review sentence or designated as NULL; $s$ indicates the sentiment polarity, categorized as positive, negative, or neutral; and $o$ refers to the opinion term, also potentially NULL. The variable $c$ represents the aspect category, which is drawn from a predefined set of categories $C$. For the training set $D = \{(X_i, \sum_{j=1}^{M} Q_{ij})\}_{i=1}^{N}$, where $N$ represents the total number of reviews, we aim to maximize the likelihood:

$$L(D) = \prod_{i=1}^{|D|} \prod_{(a,c,s,o) \in Q_i} P((a, c, s, o)|X_i) \quad (1)$$

where $P((a, c, s, o)|X_i)$ represents the conditional probability of observing the quad $(a, c, o, s)$ given review sentence $X_i$.

As illustrated in Figure 2, our proposed framework comprises three components. First, recognizing that individual review sentences may lack sufficient semantic information, we employ a hybrid algorithm to classify original reviews and prioritize the combination of review sentences with significantly differing semantic features. This process combines reviews containing only implicit sentiment expressions with those containing explicit sentiment expressions. Second, based on the combined reviews and their sentiment elements,

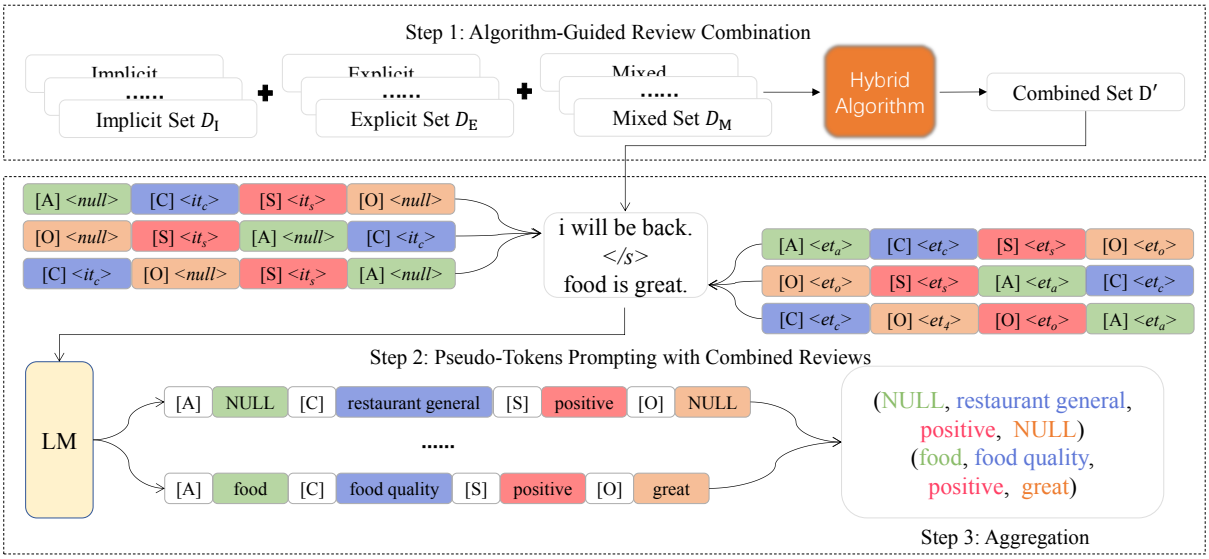


Figure 2: An overview of the proposed framework. The framework combines original reviews and guides model training through prompts with pseudo-tokens. In step 1, three sets, $D_I$, $D_E$, and $D_M$, represent implicit sentiment, explicit sentiment, and mixed sentiment, respectively. Set $D^{'}$ is created by recombining these three sets. In step 2, we use abbreviations to denote pseudo-tokens. For example, $< it_s >$ represents the pseudo-token for sentiment polarity in the implicit sentiment set, and $< et_s >$represents the pseudo-token for sentiment polarity in the explicit sentiment set. Recombined reviews are connected using the $< /s >$ symbol.

we select prompts with varying pseudo-tokens to guide the T5 model during training. Additionally, inspired by Gou et al. (2023), we rearrange the order of sentiment elements in the prompts to further enhance our dataset. During the inference phase, different prompts yield varying results. To aggregate these outcomes, our framework utilizes a voting mechanism. Next, we will describe each section in detail.

## 2.2 Algorithm-Guided Review Combination

Many studies (Zhai et al., 2022; Yu et al., 2023b; Zhang et al., 2023; Wang et al., 2023) are currently exploring data augmentation methods for sentiment analysis. However, these approaches often overlook the impact of missing explicit or implicit sentiment on the data. In our framework, we designed a data augmentation strategy to address this gap. Assuming a review sentence $S_i$ contains $n_i^a$ aspect terms and $n_i^o$ opinion terms, we categorize the sentiment elements accordingly. If all $n_i^a + n_i^o$ elements are implicit (marked as NULL), we label the review as implicit. Conversely, if they are all explicit (words present in the sentence), we label it as explicit. Review sentences containing both types are classified as mixed. This classification results in three sets: $D_I$ for reviews with only implicit sentiment expressions, $D_E$ for those with only explicit expressions, and $D_M$ for reviews containing both. Next, we prioritize combining $D_I$ and $D_E$ by sequentially pairing review sentences from these sets. Since their sizes may differ, the larger set will also be combined with $D_M$. Any remaining review sentences that do not form pairs will stand alone. The method is formally summarized in Algorithm 1. Through these steps, our framework enhances the semantic features of selected reviews in the original dataset, resulting in richer representations for analysis.

**Algorithm 1** Hybrid Algorithm

**Input:** Training set $D = \{S_i\}_{i=1}^{N}$, aspect term $n_i^a \in \{S_i, NULL\}$, opinion term $n_i^o \in \{S_i, NULL\}$
**Output:** Combined dataset $D^{'}$
**Initialize:** $D^{'} = \emptyset, D_I = \emptyset, D_E = \emptyset, D_M = \emptyset$
**// Classification Stage**
**for** each review sentence $S_i \in D$ **do**
  If $n_i^a + n_i^o = NULL$, assign $S_i \rightarrow D_I$
  Else if $n_i^a + n_i^o \in S_i$, assign $S_i \rightarrow D_E$
  Else, assign $S_i \rightarrow D_M$
**end for**
**// Combination Stage**
**while** $|D_I| > 0$ **and** $|D_E| > 0$ **do**
  Select $S_j \in D_I, S_k \in D_E$
  $S^{'} = Combine(S_j, S_k)$ or $Combine(S_k, S_j)$
  Add $S^{'} \rightarrow D^{'}$
**end while**
If $|D_I| \neq |D_E|$, combine remaining review sentences from larger set with $D_M$
Add any unpaired review sentences from $D_I$, $D_E$ or $D_M$ individually to $D^{'}$
Return $D^{'}$

### 2.3 Pseudo-Tokens Prompting with Combined Reviews

Following the methods of Hu et al. (2022) and Gou et al. (2023), our framework models the task of extracting sentiment quadruples as a sequence-to-sequence generation problem. Prompts with pseudo-tokens are appended to the input sequence to guide the model in understanding the relationships among the four sentiment elements and in distinguishing between the semantic features of implicit and explicit sentiment expressions. The model then autoregressively generates all sentiment quads present in the reviews.

Since pseudo-tokens do not carry specific semantic meanings, the model iteratively adjusts their semantic features during training. Over time, the pseudo-tokens align with appropriate vector representations in the model's embedding space. We define two categories of pseudo-tokens to represent sentiment elements in implicit and explicit review sentences. For implicit reviews, we design five pseudo-tokens, '*<implicit_vtoken_a>*', '*<implicit_vtoken_c>*', '*<implicit_vtoken_s>*', '*<implicit_vtoken_o>*' and '*<null>*', while for explicit reviews, we design four pseudo-tokens, '*<explicit_vtoken_a>*', '*<explicit_vtoken_c>*', '*<explicit_vtoken_s>*', '*<explicit_vtoken_o>*'. Thus, we design eight pseudo-tokens corresponding to the four sentiment elements in implicit and explicit reviews, along with one additional token to represent when the aspect or opinion terms are marked as NULL. Unlike Gou et al. (2023), we do not replace implicit aspect terms with "it" but instead use NULL directly. Since our framework first combines the original reviews, we select pseudo-tokens based on the sentiment types of the two sub-clauses in the combined sentence, connecting them with '*</s>*'. For instance, a review sentence might look like this:

[A] *<null>* [C] *<implicit_vtoken_c>* [S] *<implicit_vtoken_s>* [O] *<null>* $X_i$ *</s>* [A] *<explicit_vtoken_a>* [C] *<explicit_vtoken_c>* [S] *<explicit_vtoken_s>* [O] *<explicit_vtoken_o>* $X_j$.

The pseudo-tokens must align with the general sentiment elements prompt format: '[A] [C] [S] [O]'. Similar to previous work (Hu et al., 2022; Gou et al., 2023), we adjust the relative positions of the sentiment elements within the prompt, allowing the model to generate sentiment quads from different templates rather than relying on a fixed order.

### 2.4 Training

We utilize a standard transformer-based encoder-decoder architecture for the text generation process, initializing model's parameters with the pre-trained language model T5 (Raffel et al., 2020). We input $U = (u_1, u_2, ..., u_l, u_{l+p})$ contains review sentence and prompt into the language model to compute the conditional probability $\hat{y}_i$:

$$\hat{y}_i = LM(u_1, u_2, ..., u_l, u_{l+p}) \tag{2}$$

where $p$ represents the length of the prompt. Next, we calculate the cross-entropy loss $L$ between the decoder output and the target sequence $Y_i$:

$$L = -\sum_{i=1}^{N} log\hat{y}_i \tag{3}$$

### 2.5 Inference

For inference, the framework employs an aggregation strategy to combine the results generated from different prompt templates with pseudo-tokens. The final output is selected through a voting mechanism. Given the small size of the training dataset, the model's generated sequences may not always meet the desired criteria, so we apply constrained decoding to regulate token generation.

#### 2.5.1 Multi-Prompts Aggregation

We select the sentiment quads that appear most frequently across templates and include them in the final result set. Specifically, the framework defines a minimum threshold $k$ to filter out quads that occur infrequently, retaining only those that appear more often.

$$\mathcal{P} = \{q | q \in \bigcup_{i=1}^{m} \mathcal{T}_i \text{ and } (\sum_{i=1}^{m} \mathbb{1}_{\mathcal{T}_i}(q) \geq k)\} \tag{4}$$

where $q$ denotes the quad obtained in the template $\mathcal{T}_i$, $m$ is the number of templates and $\mathcal{P}$ is the final prediction of quads.

#### 2.5.2 Constrained Decoding

We employ a constrained decoding strategy (Hokamp and Liu, 2017; Bao et al., 2022; Gou et al., 2023) to ensure that the tokens generated by the model belong to the appropriate set. This method dynamically adjusts the candidate token list based on the previously generated token rather than relying on the entire vocabulary. Appendix A provides a detailed candidate list for the next token following the current token.

## 3 Experimental Setup

### 3.1 Datasets

To validate the effectiveness of our proposed framework, we conducted extensive experiments on four public datasets: Laptop-ACOS, Restaurant-ACOS, Rest15, and Rest16. Cai et al. (2021) created the Laptop-ACOS and Restaurant-ACOS datasets. Laptop-ACOS is derived from Amazon reviews (2017-2018) (Zhang et al., 2024a), while Restaurant-ACOS is an extension of the SemEval 2016 Restaurant dataset (Pontiki et al., 2016). Both quadruples in these two datasets contain implicit aspects or opinions, and Laptop-ACOS has a higher percentage of implicit opinions than Restaurant-ACOS. The Rest15 and Rest16 datasets were created by Zhang et al. (2021a), based on the SemEval shared challenges, with annotations for aspect categories and opinion terms from Peng et al. (2020) and Wan et al. (2020). In these two datasets, only aspect terms contain implicit sentiment, while opinion terms do not. Appendix B provides detailed statistics for these four benchmark datasets.

### 3.2 Implementation Details

We use the T5-BASE model (Raffel et al., 2020) from the Huggingface Transformers library[2] (Wolf et al., 2020) as the pre-trained model for our framework. T5 follows a standard encoder-decoder architecture similar to the Transformer (Vaswani, 2017). The same hyperparameters are applied across all datasets, with detailed settings provided in Appendix C. All experiments were conducted using an NVIDIA RTX 4090 GPU.

In the main experiments, using the aforementioned datasets, the number of orders $m$ varies between 1 and 15. For the low-resource setting experiments, $m$ is fixed at 5. To maintain simplicity, the number of orders during inference is the same as during training. Consistent with prior works (Peng et al., 2020), we evaluate model performance using standard metrics: F1 score (F1), recall (R), and precision (P). All reported results in supervised settings are averaged over 5 runs with different random seeds.

### 3.3 Baselines

To validate the effectiveness of our proposed framework, we compare our results with other strong baseline models:

**EXTRACT-CLASSIFY** is a robust model introduced by Cai et al. (2021), which employs an extraction-based approach. **PARAPHRASE** (Zhang et al., 2021a) is currently the leading ABSA model. This model transforms quadruple extraction into full-text generation with the assistance of Pre-trained Language Models. **GAS** (Zhang et al., 2021b) is an ABSA model that transforms the classification-based scheme into a generative paradigm. **DLO/ILO** (Hu et al., 2022) augments dataset given the order-free property of the quadruplet based on templates. **UAUL** (Hu et al., 2023) is a template-agnostic method based on T5 that effectively handles negative noise and enhances prediction accuracy. **MvP** (Gou et al., 2023) incorporates the DLO method (Hu et al., 2022) and employs element markers to represent the information structure (Paolini et al., 2021). **Mivls** (Nie et al., 2024) adopts a non-autoregressive generative framework and induces a latent variable learning to model the aspect and opinion elements. **ADA-joint** (Zhang et al., 2024b) proposes an Adaptive Data Augmentation (ADA) framework to tackle the data imbalance in the ASQP task.

## 4 Results and Discussions

### 4.1 Main Results

**Supervised settings.** The main results of our experiments are summarized in Table 1, with the F1 score being the most critical evaluation metric. Our proposed framework outperforms the compared methods across all four datasets. Specifically, it achieves improvements in F1 score of 3.96% on ACOS-Laptop, 2.58% on ACOS-Rest, 1.34% on Rest15, and 0.82% on Rest16 over the current state-of-the-art. We attribute the strong performance of our framework to two key factors: (1) The majority of reviews in the original datasets contain either explicit or implicit sentiment expressions. By combining these sentence types, our framework allows the model to learn richer semantic representations. (2) The use of pseudo-token prompts during training effectively leverages the pre-existing knowledge in the pre-trained model. Our framework performs particularly well on the ACOS-Laptop and ACOS-Rest datasets, where both the aspect term and opinion term may be labeled as NULL, indicating implicit sentiment in the sentence. In contrast, in the Rest15 and Rest16 datasets, only the aspect term can be marked as NULL. This suggests that our framework is especially effective in handling

[2] https://github.com/huggingface/transformers

| Model | ACOS-Laptop | | | ACOS-Rest | | | Rest15 | | | Rest16 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| EXTRACT-CLASSIFY | 45.56 | 29.28 | 35.80 | 38.54 | 52.96 | 44.61 | 35.64 | 37.25 | 36.42 | 38.40 | 50.93 | 43.77 |
| PARAPHRASE | - | - | - | - | - | - | 46.16 | 47.72 | 46.93 | 56.63 | 59.30 | 57.93 |
| GAS | 43.46 | 42.69 | 43.07 | 59.81 | 5731 | 58.63 | 47.15 | 46.01 | 46.57 | 57.30 | 57.82 | 57.55 |
| DLO | 43.40 | 43.80 | 43.60 | 60.02 | 59.84 | 59.18 | 47.08 | 49.33 | 48.18 | 57.92 | 61.80 | 59.79 |
| ILO | 44.14 | 44.56 | 44.35 | 58.43 | 58.95 | 58.69 | 47.78 | 50.38 | 49.05 | 57.58 | 61.17 | 59.32 |
| DLO+UAUL | 43.78 | 43.53 | 43.65 | 61.03 | 60.55 | 60.78 | 48.06 | 50.54 | 49.26 | 59.05 | 62.05 | 60.50 |
| PARAPHRASE+UAUL | 44.91 | 44.01 | 44.45 | 60.39 | 60.04 | 60.21 | 48.96 | 49.81 | 49.38 | 58.28 | 60.58 | 59.40 |
| MvP | - | - | 43.92 | - | - | 61.54 | - | - | 51.04 | - | - | 60.39 |
| Mivls | **53.47** | 33.45 | 43.71 | 60.46 | 51.14 | 55.25 | **54.46** | 48.53 | 51.25 | 57.01 | 59.54 | 58.79 |
| ADA-joint | 45.03 | 44.53 | 44.78 | 60.15 | 61.95 | 61.04 | 49.31 | **53.96** | 51.53 | 59.34 | **62.83** | 61.03 |
| ours(5 templates) | 47.57 | 47.91 | 47.74 | 62.84 | 63.19 | 63.02 | 52.55 | 53.21 | **52.87** | 62.60 | 60.95 | 61.76 |
| ours(15 templates) | 49.42 | **48.08** | **48.74** | **63.85** | **64.41** | **64.12** | 52.83 | 50.44 | 51.61 | **62.64** | 61.08 | **61.85** |

Table 1: Results for supervised settings on four datasets of ASQP tasks. The best and the second best results are in **bold** and underlined, respectively. In this experiment, mm is set to 1-15, and we choose mm equal to 5 and 15.

| Dataset | Model | 1% | 5% | 10% | AVG |
|---|---|---|---|---|---|
| ACOS-Laptop | MvP† | **14.63** | 27.01 | 32.01 | 24.55 |
| | ours | 11.01 | **33.54** | **38.21** | **27.59** |
| ACOS-Rest | MvP† | 16.70 | 32.31 | 42.38 | 30.46 |
| | ours | **18.88** | **44.77** | **46.27** | **36.64** |
| Rest15 | MvP† | 14.02 | 26.57 | 31.48 | 24.02 |
| | ours | **16.16** | **30.44** | **37.99** | **28.20** |
| Rest16 | MvP† | 17.88 | 37.67 | 42.91 | 32.82 |
| | ours | **20.42** | **38.87** | **48.26** | **35.85** |

Table 2: Results for low-resource settings. The results with "†" are reproduced by their released code.

reviews with implicit sentiment expressions.

**Low-resource settings.** To further evaluate the performance of our framework in low-resource settings, we trained both our model and MvP using only 1%, 5%, and 10% of the data from the four datasets. The F1 scores on the test sets are shown in Table 2. Our framework consistently outperforms MvP, even with limited training samples. Notably, our framework achieves better results than MvP across all resource settings. In low-resource scenarios, the reduction in sample size makes it crucial to increase the complexity of individual samples. Since our proposed framework combines reviews with different sentiment categories, a single review sentence may contain both implicit and explicit sentiment expressions, allowing the model to learn more complex representations. This clearly demonstrates the ability of our framework to adapt quickly in low-resource scenarios, highlighting its robustness and efficiency when sample sizes are limited.

### 4.2 Ablation Study

In this section, we conducted ablation experiments to assess the effectiveness of our proposed framework, specifically focusing on the **Algorithm-Guided Sample Mixing** and **Pseudo-Tokens Prompting with Mixed Samples** components. All experiments were carried out in a supervised setting, and the results are summarized in Table 3.

First, we removed the review combination module and used the original dataset for experiments. The results show that removing this component led to a decrease in F1 scores on the ACOS-Laptop, ACOS-Rest, Rest15, and Rest16 datasets by 2.72%, 1.97%, 2.08%, and 2.85%, respectively. This demonstrates the critical role review combination plays in our framework, which will be further discussed in Section 4.2.1. Next, we excluded the pseudo-tokens from the prompt sequences, resulting in performance drops across all datasets. The F1 scores for ACOS-Laptop, ACOS-Rest, Rest15, and Rest16 declined by 1.74%, 1.88%, 3.53%, and 0.83%, respectively. This indicates that pseudo-tokens significantly enhance the effectiveness of the prompts, which will be thoroughly analyzed in Section 4.2.2. Finally, we removed the constrained decoding module, resulting in a slight decrease in

| Model | Laptop | Rest | Rest15 | Rest16 |
|---|---|---|---|---|
| ours (5 templates) | **47.74** | **63.02** | **52.87** | **61.76** |
| w/o sample combination | 45.02 | 61.05 | 50.79 | 58.91 |
| w/o pseudo-tokens | 46.00 | 61.14 | 49.34 | 60.93 |
| w/o cd | 47.17 | 62.72 | 51.28 | 60.95 |

Table 3: Results from the ablation study under supervised settings for our framework with five templates.

| Model | ACOS-Laptop | | | | ACOS-Rest | | | |
|---|---|---|---|---|---|---|---|---|
| | EA & EO | IA & EO | EA & IO | IA & IO | EA & EO | IA & EO | EA & IO | IA & IO |
| TAS-BERT-ACOS | 26.1 | 41.5 | 10.9 | 21.2 | 33.6 | 31.8 | 14.0 | 39.8 |
| Extract-Classify-ACOS | 35.4 | 39.0 | 16.8 | 18.6 | 45.0 | 34.7 | 23.9 | 33.7 |
| PARAPHRASE | 45.7 | 51.0 | 33.0 | 39.6 | 65.4 | 53.5 | 45.6 | 49.2 |
| GEN-SCL-NAT | 45.8 | 54.0 | 34.3 | 39.6 | **66.5** | 56.5 | 46.2 | 50.7 |
| ours | **50.3** | **57.1** | **47.1** | **45.1** | 65.7 | **65.1** | **66.7** | **71.6** |

Table 4: Breakdown of F1 performance per example split, with each split comprising reviews containing that quadruple type. E: explicit, I: implicit, A: aspect, O: opinion. The best and the second best results are in **bold** and underlined, respectively.

model performance.

#### 4.2.1 Effect of Review Combination

The results shown in Table 3 clearly demonstrate that removing the review combination component leads to a significant drop in F1 scores, highlighting the effectiveness of our proposed review combination approach. This method serves as a form of data augmentation, combining reviews that contain only implicit sentiment expressions with those that have only explicit sentiment expressions. By merging these types of reviews, the model is able to capture richer sentiment features from a single sentence, improving its ability to handle more complex sentiment scenarios. Importantly, this augmentation technique relies solely on the original dataset without introducing any external data.

To further analyze the framework's performance, we evaluated its effectiveness across specific sentiment quadruples, as shown in Table 4. The sentiment quads were categorized into four groups based on whether the aspect term and opinion term are implicit: EAEO, IAEO, EAIO, and IAIO. We tested the framework on the ACOS-Laptop and ACOS-Rest datasets, and the results indicate that our framework outperforms previous state-of-the-art models (Wan et al., 2020; Cai et al., 2021; Zhang et al., 2021a; Peper and Wang, 2022) across all categories of sentiment quads, particularly in cases where either the aspect term or opinion term is implicit. Additionally, our experiments reveal that existing models often confuse implicit and explicit sentiment expressions when extracting sentiment quadruples. In summary, the review combination approach we introduced is both effective and impactful.

#### 4.2.2 Effect of Pseudo-Tokens

We designed two types of pseudo-tokens: one to represent implicit sentiment and the other for explicit sentiment. The framework selects different pseudo-tokens based on the types of sentiment quads in the sentence. Each of the four sentiment elements in a quad is associated with a corresponding pseudo-token. Additionally, we created a specific pseudo-token, '*<null>*', to represent implicit aspect terms and opinion terms. These pseudo-tokens help capture relationships between different sentiment elements and leverage the prior knowledge of pre-trained models, making the prompts more adaptable across contexts.

To explore the impact of the '*<null>*' token, we conducted ablation experiments on the ACOS-Laptop and ACOS-Rest datasets, as shown in Table 5. When the '*<null>*' token was removed, the performance of the framework decreased, confirming that having a dedicated pseudo-token for implicit sentiment elements is indeed effective.

| Model | ACOS-Laptop | | | ACOS-Rest | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| w/o *<null>* | 47.05 | 46.60 | 46.82 | 64.17 | 63.75 | 63.96 |
| ours | **49.42** | **48.08** | **48.74** | **63.85** | **64.41** | **64.12** |

Table 5: Comparative experiment after removing '*<null>*' from Pseudo-tokens.

#### 4.2.3 Effect of the number of templates

Inspired by the success of multi-order template-based data augmentation in the MvP model, we further investigate the impact of using prompts with different sentiment element orders for data augmentation. We evaluate the effect of multi-order templates across all four datasets. As shown in Figure 3, the F1 score improves consistently across all datasets as the value of $m$ increases, with the

model reaching near-optimal performance around $m = 5$. This indicates that prompts with pseudo-tokens, when combined with multi-order template augmentation, can effectively help the model learn semantic relationships between sentiment elements. For more analysis on this section, see Appendix C.

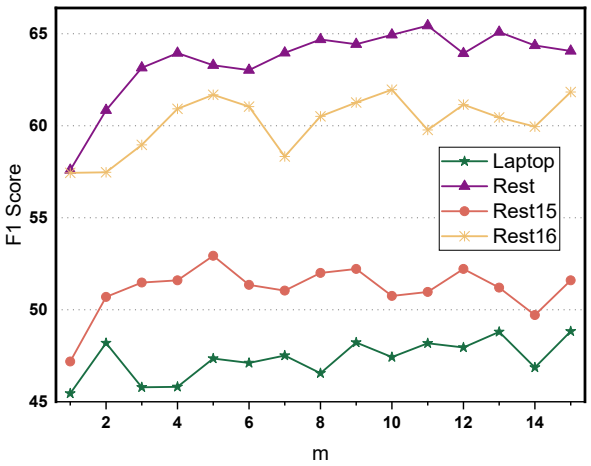


Figure 3: F1 scores of variants by setting different numbers of templates in all datasets.

## 5 Related Work

**Task Evolution.** According to Liu (2022), Sentiment Analysis (SA) is a field that uses computational methods to analyze opinions and emotions expressed in text. Aspect-Based Sentiment Analysis (ABSA), a subtask of SA, focuses on sentiment analysis at the aspect level, identifying four key elements in a sentence: aspect term, aspect category, sentiment polarity, and opinion term. The aspect term refers to the entity or specific feature being evaluated, which can be explicitly or implicitly mentioned. The aspect category represents the broader domain to which the aspect belongs. Sentiment polarity indicates the sentiment (positive, negative, or neutral) toward the aspect term, inferred from the opinion term and its context. The opinion term expresses an attitude toward the aspect term, which may also be explicit or implicit. Two tasks have been defined to extract all four sentiment elements simultaneously: the ACOS (Aspect-Category-Opinion-Sentiment) task introduced by Cai et al. (2021), and Aspect-Based Sentiment Quadruple Prediction (ASQP) defined by Zhang et al. (2021a), both aiming to extract sentiment quadruples from review sentences.

**Generative Methods.** Instead of separate or pipeline methods (Phan and Ogunbona, 2020), recent research has increasingly adopted generative approaches to address various ABSA challenges. These generative methods excel by minimizing the error propagation common in pipeline models and leveraging rich semantic label information (Paolini et al., 2021; Zhang et al., 2022; Yu et al., 2023a). Moreover, Hu et al. (2022) and Gou et al. (2023) explored the impact of element ordering and proposed methods to enhance data on the target side by selecting optimal element orders for ABSA tasks. Additionally, Zhang et al. (2024b) introduced an adaptive data augmentation framework to enhance data quality within a generative approach further. Another study (Zhang et al., 2024c) proposes a self-training framework with a pseudo label scorer, which constructs human and AI annotated comparison datasets, uses a generative model as the scorer, conducts two stage filtering in self-training, and applies the scorer as a reranker to improve model performance.

## 6 Conclusion

In this paper, we present a framework designed to address the challenge of distinguishing between implicit and explicit sentiment expressions when extracting sentiment quadruples. Our approach first uses a hybrid algorithm to reorganize the original dataset based on sentiment expression types. We then introduce the prompts with pseudo-tokens through multiple templates to guide the model's training. We design nine pseudo-tokens, each representing the sentiment elements under different types of sentiment expression. Depending on the sentiment expression present in the reorganized review sentences, the appropriate pseudo-tokens are embedded in the prompt. Finally, a voting strategy is applied to aggregate the results generated by the model from different prompts, yielding the final output. Extensive experiments demonstrate that this framework enhances the model's ability to extract sentiment quads, particularly in cases involving implicit sentiment expressions.

## 7 Limitations

In our proposed framework, we use a straightforward hybrid algorithm to reorganize the original reviews. This serves as a form of data augmentation; however, it focuses solely on combining data without considering its quality. While this approach enriches the original reviews with more diverse sentiment expressions, it overlooks the semantic relationships between the two reviews being

merged. When there is a semantic connection between the combined reviews, the resulting samples are of higher quality.

## Acknowledgments

We want to thank the anonymous reviewers for their insightful comments.

## References

Xiaoyi Bao, Wang Zhongqing, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4044–4050. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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, Online. Association for Computational Linguistics.

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.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Mengting Hu, Yinhao Bai, Yike Wu, Zhen Zhang, Liqi Zhang, Hang Gao, Shiwan Zhao, and Minlie Huang. 2023. Uncertainty-aware unlikelihood learning improves generative aspect sentiment quad prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13481–13494, Toronto, Canada. Association for Computational Linguistics.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Yu Nie, Jianming Fu, Yilai Zhang, and Chao Li. 2024. Modeling implicit variable and latent structure for aspect-based sentiment quadruple extraction. *Neurocomputing*, 586:127642.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Ma Jie, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, Stefano Soatto, et al. 2021. Structured prediction as translation between augmented natural languages. In *ICLR 2021-9th International Conference on Learning Representations*, pages 1–26. International Conference on Learning Representations, ICLR.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8600–8607.

Joseph Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6089–6095, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minh Hieu Phan and Philip O. Ogunbona. 2020. Modelling context and syntactical features for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3220, Online. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9122–9129.

An Wang, Junfeng Jiang, Youmi Ma, Ao Liu, and Naoaki Okazaki. 2023. Generative data augmentation for aspect sentiment quad prediction. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 128–140, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haoliang Xiong, Zehao Yan, Chuhan Wu, Guojun Lu, Shiguan Pang, Yun Xue, and Qianhua Cai. 2023. Bart-based contrastive and retrospective network for aspect-category-opinion-sentiment quadruple extraction. *International Journal of Machine Learning and Cybernetics*, 14(9):3243–3255.

Chengze Yu, Taiqiang Wu, Jiayi Li, Xingyu Bai, and Yujiu Yang. 2023a. Syngen: A syntactic plug-and-play module for generative aspect-based sentiment analysis. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jianfei Yu, Qiankun Zhao, and Rui Xia. 2023b. Cross-domain data augmentation with domain-adaptive language modeling for aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1456–1470, Toronto, Canada. Association for Computational Linguistics.

Zepeng Zhai, Hao Chen, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. COM-MRC: A COntext-masked machine reading comprehension framework for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3230–3241, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hao Zhang, Yu-N Cheah, Osamah Mohammed Alyasiri, and Jieyu An. 2024a. Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and chatgpt: a comprehensive survey. *Artificial Intelligence Review*, 57(2):17.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. 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.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Wenyuan Zhang, Xinghua Zhang, Shiyao Cui, Kun Huang, Xuebin Wang, and Tingwen Liu. 2024b. Adaptive data augmentation for aspect sentiment quad prediction. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11176–11180. IEEE.

Yice Zhang, Yifan Yang, Meng Li, Bin Liang, Shiwei Chen, and Ruifeng Xu. 2023. Target-to-source augmentation for aspect sentiment triplet extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12165–12177, Singapore. Association for Computational Linguistics.

Yice Zhang, Yifan Yang, Yihui Li, Bin Liang, Shiwei Chen, Yixue Dang, Min Yang, and Ruifeng Xu. 2022. Boundary-driven table-filling for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6485–6498, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yice Zhang, Jie Zeng, Weiming Hu, Ziyi Wang, Shiwei Chen, and Ruifeng Xu. 2024c. Self-training with pseudo-label scorer for aspect sentiment quad prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11862–11875, Bangkok, Thailand. Association for Computational Linguistics.

## A Constrained Decoding and Experiment settings

The specific constraint decoding strategy and experimental settings are shown in Table 6 and Table 7

## B Data Statistics

Table 8 shows the data statistics of all datasets of the ASQP and ACOS task.

| Current Token | Candidate tokens |
|---|---|
| [A] | Input Sentence, NULL, [SSEP] |
| [C] | Aspect Category, [SSEP] |
| [S] | Positive, Negative, Neutral, NULL, [SSEP] |
| [O] | Input Sentence, NULL, [SSEP] |

Table 6: Candidate tokens of different current token. The [SSEP] token is used to separate multiple sentiment quads within a single input sentence.

| Hyperparameters | Ours | Ours (Low Resource) | | |
|---|---|---|---|---|
| | | 1% | 5% | 10% |
| Epoch | 20 | 200 | 100 | 50 |
| Batch Size | 16 | 8 | | |
| Learning Rate | 1e-4 | | | |

Table 7: Hyper-parameters for all supervised and low-resource settings.

## C Analysis of Multi-prompts

### C.1 Effect of different aggregation strategies

The multi-order template data augmentation strategy can lead the model to generate different sentiment quads. Therefore, selecting the final quads from these varied results becomes crucial. To address this, we employ a voting strategy. During result aggregation, a threshold value $k$ is set. Only quads that appear more than $k$ times in the result set are included in the final output. As seen in Figure 4 and Figure 5, increasing $k$ improves precision but reduces recall, with minimal fluctuation in the F1 score. This occurs because a smaller $k$ allows more quads to be selected, boosting recall but introducing potential errors that lower precision. Conversely, a larger $k$ reduces recall but increases precision. To balance model performance, we recommend setting $k$ to half the number of generated prompts.

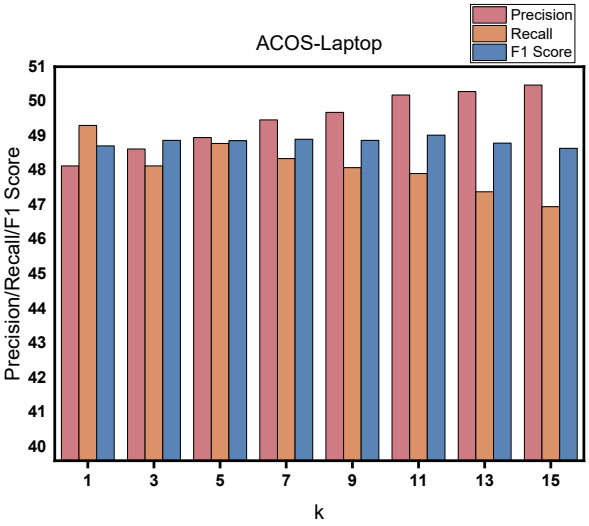


Figure 4: The precision, recall and F1 score for ACOS-Laptop training on aggregation strategies

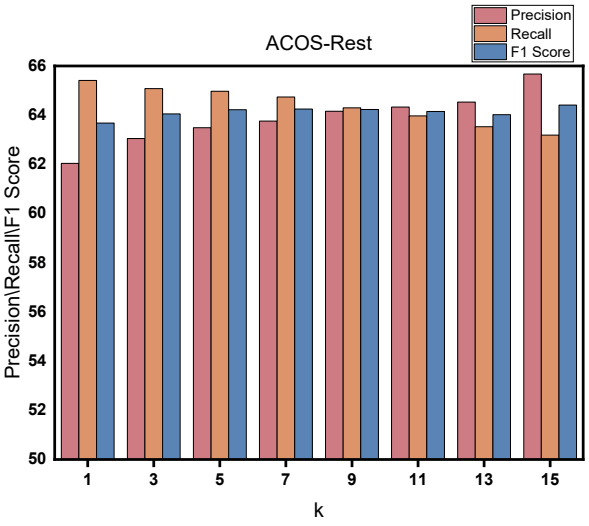


Figure 5: The precision, recall and F1 score for ACOS-Rest training on aggregation strategies

### C.2 Effect of prompt input position

We explored the impact of placing the prompt at different positions within the input sequence, considering two configurations: placing the prompt at the beginning (pre) or at the end (post) of the sequence. Experiments on the ACOS-Laptop and ACOS-Rest datasets, as shown in Table 9, reveal that both settings yield similar F1 scores. However, the pre setting results in higher precision but lower recall, while the post setting produces the opposite effect. Since T5 is a generative model, placing the prompt at the end likely encourages the model to generate more sentiment quads based on the sentiment elements present in the prompt, which could explain this outcome.

| Dataset | ACOS-Laptop | | | ACOS-Rest | | | Rest15 | | | Rest16 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | #S | #Q | #IS | #S | #Q | #IS | #S | #Q | #IS | #S | #Q | #IS |
| Train | 2934 | 4172 | 1527 | 1530 | 2484 | 589 | 834 | 1354 | 227 | 1264 | 1989 | 382 |
| Dev | 326 | 440 | 155 | 171 | 261 | 74 | 209 | 347 | 58 | 316 | 507 | 87 |
| Test | 816 | 1161 | 407 | 581 | 913 | 267 | 537 | 795 | 184 | 544 | 799 | 146 |

Table 8: Statistics of the datasets. #S and #Q are the numbers of review sentences and sentiment quads. #IS are the numbers of review sentences which contain implicit sentiment expression.

| Model | ACOS-Laptop | | | ACOS-Rest | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| post | 47.57 | **47.91** | 47.74 | **62.84** | **63.19** | **63.02** |
| pre | **48.57** | 47.29 | **47.92** | 62.31 | 62.86 | 62.58 |

Table 9: The impact of prompt at different positions in the input sequence.