information

MDPI

Article

Aspect-Enhanced Prompting Method for Unsupervised Domain Adaptation in Aspect-Based Sentiment Analysis

Binghan Lu *, Kiyoaki Shirai and Natthawut Kertkeidkachorn

Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan; kshirai@jaist.ac.jp (K.S.); natt@jaist.ac.jp (N.K.)
* Correspondence: s2320043@jaist.ac.jp

Abstract: This study proposes an Aspect-Enhanced Prompting (AEP) method for unsupervised Multi-Source Domain Adaptation in Aspect Sentiment Classification, where data from the target domain are completely unavailable for model training. The proposed AEP is based on two generative language models: one generates a prompt from a given review, while the other follows the prompt and classifies the sentiment of an aspect. The first model extracts Aspect-Related Features (ARFs), which are words closely related to the aspect, from the review and incorporates them into the prompt in a domain-agnostic manner, thereby directing the second model to identify the sentiment accurately. Our framework incorporates an innovative rescoring mechanism and a cluster-based prompt expansion strategy. Both are intended to enhance the robustness of the generation of the prompt and the adaptability of the model to diverse domains. The results of experiments conducted on five datasets (Restaurant, Laptop, Device, Service, and Location) demonstrate that our method outperforms the baselines, including a state-of-the-art unsupervised domain adaptation method. The effectiveness of both the rescoring mechanism and the cluster-based prompt expansion is also validated through an ablation study.

Keywords: aspect-based sentiment analysis; unsupervised domain adaptation; multi-source domain adaptation; text generation model; prompt engineering

check for updates

Academic Editor: Diego Reforgiato Recupero

Received: 6 April 2025
Revised: 6 May 2025
Accepted: 14 May 2025
Published: 16 May 2025

Citation: Lu, B.; Shirai, K.; Kertkeidkachorn, N. Aspect-Enhanced Prompting Method for Unsupervised Domain Adaptation in Aspect-Based Sentiment Analysis. *Information* **2025**, *16*, 411. https://doi.org/10.3390/info16050411

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Sentiment Analysis (SA) has become a vital task in Natural Language Processing (NLP), aiming to identify the polarity (positive, negative, or neutral) of an opinion expressed in a text [1]. Traditional sentiment analysis primarily focuses on categorizing a text with predefined labels, which is inadequate when different aspects of an entity are discussed in the same text. To address this limitation, Aspect-Based Sentiment Analysis (ABSA) has emerged as a more fine-grained task, seeking to identify sentiments tied to specific aspects within a text. While many NLP algorithms for SA and ABSA have undergone remarkable advances [2,3], a significant challenge remains: the deterioration of the performance of the model when a trained model is applied to an out-of-distribution domain. This deterioration can be attributed to the fact that most NLP algorithms assume that the training and test data are drawn from the same underlying distribution [4].

Domain Adaptation (DA) addresses this challenge by enabling NLP algorithms to adapt to out-of-distribution data. Traditional DA research in NLP has primarily focused on Supervised Domain Adaptation, which makes use of a small amount of labeled data from the target domain alongside a large quantity of labeled data from the source domain. However, acquiring labeled data in the target domain requires substantial manual effort.

To alleviate this, Unsupervised Domain Adaptation (UDA) has emerged, allowing the use of unlabeled target domain data to improve the generalization to out-of-distribution data. Nevertheless, in most UDA settings, even though the target domain is unlabeled, it is still known. This does not take into account all possible real-world scenarios, in some of which the target domain is entirely unknown. For example, when a new product or service is released by a company, customer reviews for it are not accumulated (i.e., a large amount of even unlabeled data in the target domain are unavailable), but the company wishes to automatically analyze reviews posted by pioneer customers. The model's generalization ability may be restricted when faced with unseen target domains. A more challenging and yet underexplored area of DA is to adapt models to any possible unknown target domains that are not available during the training.

This paper proposes a novel method of DA for ABSA, called the Aspect-Enhanced Prompting (AEP) method, which adapts a classification model from multiple source domains to an unknown target domain. ABSA is performed by giving a prompt to a text generation model, which identifies the sentiment of an aspect in a review and produces a corresponding sentiment word as output. A prompt is automatically generated by another text generation model. This model extracts several words from the review that are closely related to the aspect and includes them in the generated prompt so that these extracted important words can be highlighted in the sentiment classification. The prompt generation model is trained on multiple source domains, so it can generate appropriate aspect-enhanced prompts in any domain. Therefore our AEP is an effective method for sentiment classification when applied to a completely unseen target domain.

The main contributions of this paper can be summarized as follows:

1. We explore a severe domain adaptation setting for ABSA, where labeled data of multiple source domains are available but the data of a target domain are completely unavailable during training. This setting has not been actively studied in terms of ABSA.
2. We propose a novel domain-agnostic prompt generation framework that automatically generates prompts highlighting important keywords for sentiment classification.
3. We evaluate the effectiveness of our proposed method through experiments on diverse sentiment analysis datasets.

2. Related Work

2.1. Unsupervised Domain Adaptation

Approaches to Unsupervised Domain Adaptation (UDA) can be broadly categorized into two types: feature-based methods and Generative Data Augmentation methods [4].

The feature-based approach can be divided into two key categories: feature augmentation and feature generalization. Feature augmentation methods use pivots to construct an aligned feature space, intending to expand the feature space to more efficiently capture domain-invariant information by identifying common features across domains [5,6]. Recently, deep learning techniques have shown strong performance in UDA tasks. These methods often involve aligning feature representations between the source and target domains through adversarial learning or statistical measures to minimize the distance between domains [7,8]. Feature generalization methods have also been widely explored. Early approaches were often based on autoencoders, which learn latent representations by minimizing an input reconstruction loss in an unsupervised manner. Inspired by denoising autoencoders [9], initial work in this area introduced the stacked denoising autoencoder (SDA) [10], which learns unified feature representations by stacking multiple layers and corrupting the inputs with Gaussian noise. A more efficient approach was introduced with the marginalized stacked denoising autoencoder (mSDA) [11], which marginalizes the

noise, leading to faster and more scalable training while maintaining the robustness of the learned representations. Despite these advances, a notable drawback of autoencoder-based methods is their lack of integration with linguistic information, which is crucial for many NLP tasks. More recent approaches for feature generalization have shifted toward models that leverage pre-trained language models, such as BERT [12], which capture linguistic and contextual features across domains [13,14]. A representative feature generalization approach under the source-free setting is Source-Free ABSA, which addresses ABSA by combining feature-based adaptation and pseudo-labeling strategies [15]. It effectively learns transferable representations and achieves superior performance through iterative self-training in the target domain.

Generative Data Augmentation techniques have also gained considerable attention in UDA. The core idea behind these methods is to generate data similar to the target dataset by leveraging labeled source data and low-resource, unlabeled target data. For example, Mixup, a data augmentation technique originally designed for image classification, has been adapted for NLP tasks by performing interpolations on word and sentence embeddings. This approach has achieved notable improvements in sentence classification across various datasets [16]. In the context of SA, Generative Data Augmentation methods have been employed to generate synthetic reviews that resemble ones in the target domain. These target-like synthetic reviews help to bridge the domain gap, facilitating better generalization on the target domain [17]. Refining and Synthesis Data Augmentation (RSDA) introduces a two-stage framework that refines pseudo-labeled target data using natural language inference filtering and synthesizes diverse labeled data through label composition and paraphrasing for cross-domain ABSA [18].

Our proposed method differs from existing approaches in two key ways. First, it does not rely on target awareness, as it operates solely on knowledge derived from the source domains without reference to knowledge in the target domain. Second, rather than the transfer of knowledge between domains, our approach classifies sentiment by using the representations learned from examples within the source domains. The model is exclusively trained on data from the source domain and applies this learned knowledge to classify examples from the target domain, without direct adaptation to the target domain. This enables the effective classification of data from an unseen target domain, based on training on the source domain.

2.2. Multi-Source Domain Adaptation

Multi-Source Domain Adaptation (MSDA) is an extension of UDA in which labeled data are available from multiple source domains, while the target domain remains unlabeled. The challenge in MSDA is to effectively use data from diverse source domains while mitigating negative transfer and enhancing generalization to the target domain.

A key concept in MSDA is the measurement of the divergence of two domains, which plays a critical role in transferring knowledge. For instance, the DistanceNet-Bandit approach [19] explores distance-based metrics to quantify the dissimilarity between domains. The model incorporates these metrics as an additional loss function to enhance domain adaptation and dynamically selects the most relevant source domain through a multi-armed bandit controller.

Another line of research employs a mixture of experts (MoE) approach [20] to handle the multiple source domains. This method explicitly models the relationship between the target domain and various source domains by learning how to combine predictions from domain-specific experts.

Transformer-based MSDA models [21] have demonstrated that large pre-trained models like BERT are naturally robust to a change of domain. However, these models

benefit from MoE strategies, where attention-based mixing functions assist in combining predictions from domain-specific experts. Despite these advances, the challenge of learning effective functions for combining domain expert predictions remains, as transformer-based experts tend to produce homogeneous predictions across different domains. In another approach, BERT-based domain classification and data selection tackle domain shifts by fine-tuning BERT with domain-specific data selection [22]. The model employs curriculum learning and domain-discriminative data selection to adapt BERT to new domains. While this method significantly mitigates the problems arising from domain shift, it relies on the availability of small validation sets for the target domain in order to tune its hyperparameters, which may not always be feasible in fully unsupervised settings.

Recent advances in MSDA have explored various techniques to enhance cross-domain generalization. Feature Structure Matching (FSM) introduces a dynamic parameter fusion module and a feature structure matching constraint that improves the robustness of adaptation across multiple source domains through effective feature correlation alignment [23]. FaiMA focuses on MSDA for ABSA and leverages feature-aware in-context learning by optimizing linguistic, domain, and sentiment features through a graph attention network and contrastive learning [24].

Our method introduces several key differences. First, we propose a domain-agnostic prompt generation framework that generates prompts only from source domain data, thereby avoiding the reliance on explicit domain-specific features or domain divergence metrics, in contrast to DistanceNet-Bandit. Second, unlike traditional mixture-of-experts approaches, our model directly classifies target data based on source domain prompts, thereby reducing the complexity of the model. Third, our approach is highly efficient in low-resource scenarios, as it requires no validation data for the target domain for its fine-tuning.

2.3. Example-Based Prompt Learning

Recent research in Example-Based Prompt Learning has shown significant potential in improving the performance of DA, especially in scenarios where the target domain is unseen during training. Instead of relying solely on parametric representations, models in this framework use specific examples from the training data to make predictions. Notably, recent work demonstrates that by tuning soft prompts, large models like T5 [25] can be adapted to various tasks with far fewer parameters while maintaining competitive performance compared to fully fine-tuned models [26].

Prompt learning for on-the-fly Any-Domain Adaptation (PADA) introduces a dynamic framework where example-based prompts are generated for each target instance, facilitating adaptation to unseen domains [27]. By leveraging the representations learned from multiple source domains, the T5 model generates prompts for unseen target domains, guided by examples from the source domains. Prompt Tuning with Domain Knowledge (PTDK) introduces a hybrid strategy that combines hard prompts and trainable soft prompts [28]. Although PTDK effectively integrates external domain knowledge to enrich prompt representations, it relies primarily on structured knowledge bases.

Our approach is an advance on Example-Based Prompt Learning with a particular focus on ABSA. Specifically, we propose the use of Aspect-Related Features (ARFs) to enhance both the selection and generation of the prompt; this enables the model to effectively capture aspect-based sentiments in the unseen target domain.

3. Proposed Method

3.1. Problem Settings

In this study, we focus on the task of Aspect Sentiment Classification (ASC), which is a subtask of ABSA, in a Multi-Source Domain Adaptation setting. Specifically, we

aim to classify the sentiment of a given aspect in a review text within an unseen target domain, where neither labeled nor unlabeled target domain data are available. Our method leverages only labeled data from multiple source domains and adapts the model to perform well in the target domain.

Let $D_s = \{D_1, D_2, \ldots, D_n\}$ be a set of datasets from several source domains. The dataset, D_i, consists of labeled data in the form of (x, a, y), where x is a text containing an aspect a and y is the sentiment label (positive, negative, or neutral) associated with the aspect a. Furthermore, D_{ut} denotes test samples in the unseen target domain that are unavailable during the training phase. The objective is to accurately classify the sentiment of the aspects of the texts in D_{ut} by transferring knowledge from the source domain data D_s.

3.2. Overview of the Method

In our Aspect-Enhanced Prompting (AEP) method, the ASC is defined as a generation task. Two generation models are employed to perform the ASC as illustrated in Figure 1. The first model is the "Prompt Generation Model", which takes a given review and aspect as input and generates a prompt as output. The second model is the "Sentiment Classification Model", which generates a sentiment word from a given review and a prompt obtained by the first model. The prompt is a sentence that asks a text generation model about the polarity of the user's opinion, such as "What is the sentiment of [aspect] considering [ARF] in the review?". Here, ARF stands for "Aspect-Related Feature", which is defined as a word that appears in a review and is closely related to an aspect. ARFs are indicated in blue in Figure 1.

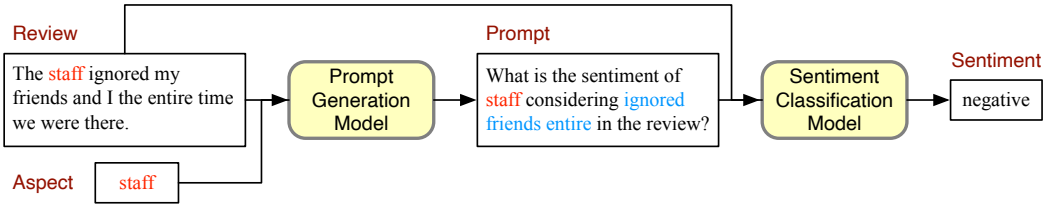


Figure 1. Overview of Aspect-Enhanced Prompting Method.

The Prompt Generation Model plays a key role in our proposed method. It learns how to extract ARFs from a review, which must be an important clue for classifying the sentiment of the aspect. Even when it is applied to reviews in an unseen target domain, it can identify several useful linguistic features and generate them in the prompt, thus improving the classification performance. In addition, since the Prompt Generation Model is trained using samples from multiple source domains, general and domain-independent linguistic features can be incorporated into the model. Furthermore, although sentiment classification based on the text generation model is sensitive to the input prompt, our Prompt Generation Model can automatically generate the most appropriate prompt for a target review.

3.3. Training

Figure 2 shows an overview of the training of two text generation models. In Step 1, several ARFs are extracted from each training sample in source domains. In Step 2, candidates for the prompts are produced by filling an aspect and ARFs into a template. In Step 3, scores of the prompts are calculated to choose the best prompt. In Step 4, the Prompt Generation Model is obtained by fine-tuning a pre-trained generative language model using pairs of reviews and chosen prompts. Also, the Sentiment Classification Model is fine-tuned using reviews, prompts, and ground-truth sentiment labels. In Step 5, the prompts are

updated by rescoring based on the initially trained Sentiment Classification Model. In Step 6, the Prompt Generation Model and the Sentiment Classification Model are retrained by fine-tuning with the updated prompts.

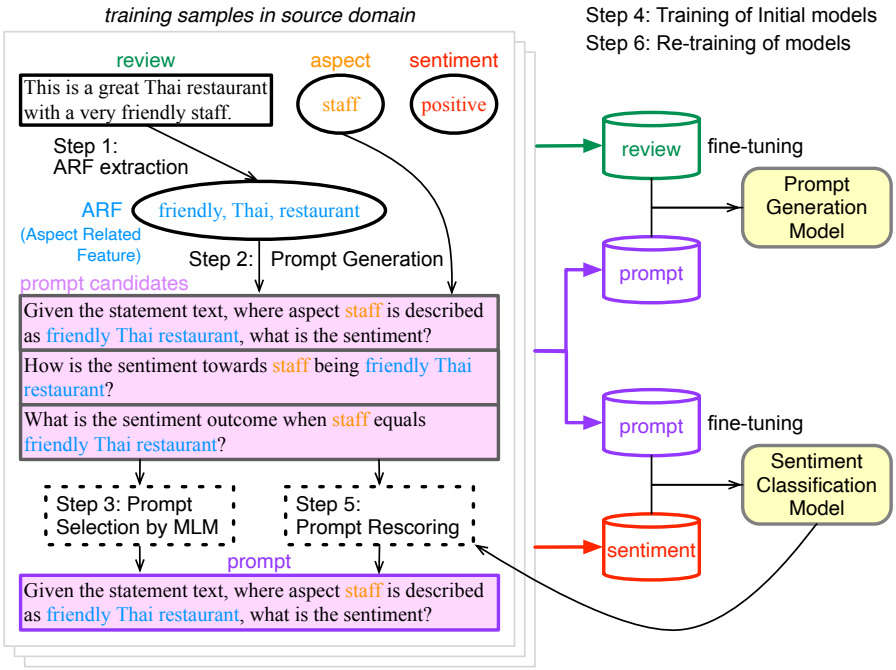


Figure 2. Overview of model training.

3.3.1. Extraction of Aspect-Related Features

The first step of the model training is to extract ARFs from a review sentence. ARFs play an important role in the prompt-based learning of the ASC task since they are supposed to be highly correlated with the target aspect.

Candidates of ARFs are obtained by (1) tokenizing sentences that contain the target aspect, (2) removing stopwords, punctuation marks, and other non-informative words, and (3) extracting all remaining words. We measure the Pointwise Mutual Information (PMI) [29] between an aspect and each candidate feature as in Equation (1).

$$PMI(a; f) = \log \frac{P(a, f)}{P(a)P(f)}, \tag{1}$$

where $P(a, f)$ denotes the joint probability between an aspect a and a feature f, which measures how likely a and f are to co-occur; $P(a)$ and $P(f)$ denote the probabilities of appearance of a and f. Note that $PMI(a; f)$ is measured separately in each of the source domains. Then, for a given review sentence and an aspect, the several words with the highest $PMI(a; f)$ are extracted as ARFs. We define the parameter N_{arf} as the number of extracted ARFs.

Table 1 shows examples of extracted ARFs for different domains. We can find a strong relationship between ARFs and their corresponding aspects. For example, in the first example in Table 1, “friendly”, “Thai”, and “restaurant” are highly correlated with the aspect “staff”. It is also confirmed that appropriate ARFs are extracted from different domains.

Table 1. Extracted Aspect-Related Features from different domains.

Review	Domain	Extracted ARFs
This is a great Thai restaurant with a very friendly staff.	Restaurant	friendly, Thai, restaurant
The battery was completely dead, in fact it had grown about a quarter inch thick lump on the underside.	Laptop	thick, grown, quarter
It is a perfect phone in such a small and appealing package.	Device	perfect, small, appealing
I have heard of people having problems receiving email out of order or not receiving some of their messages at all.	Service	heard, people, problems
Most are very expensive to rent or buy even the grotty little flats and bedsits in LOCATION1.	Location	expensive, rent, buy

Note: N_{arf} is set to 3. The aspects are underlined.

3.3.2. Generation of Candidate Prompt

Once the ARFs are extracted, the next step involves generating ARF-based prompts to effectively guide the sentiment classification process. We have manually designed a collection of prompt templates. These templates are divided into two categories: (1) question templates that probe the sentiment associated with a given aspect and its ARFs and (2) description templates that explicitly describe the relationship between the aspect and its corresponding ARFs and instruct a language model to guess the sentiment of the aspect. An example of each type of template is shown below.

(1) What is the sentiment of *[aspect]* considering *[ARFs]* in the review?
(2) Predict the sentiment for *[aspect]* described as *[ARFs]*.

There are 20 templates. All templates are shown in Appendix A. The candidates for the prompts are obtained by filling an aspect and ARFs into a template.

3.3.3. Prompt Scoring

Next, the best prompt is chosen from among the multiple prompts derived from different templates. The score of a prompt is calculated by the Masked Language Model (MLM) of BERT [12]. The MLM can estimate the probability of filling a word at a [MASK] position in a sentence. Here, the probabilities of filling ARFs at the masked positions of the template are estimated by the MLM. Specifically, the score of prompt p_i is defined by Equation (2).

$$Score(p_i) = \sum_j \text{MLM}(p_i, f_j) = \sum_j P(f_j \mid t_i) = \sum_j \text{softmax}(\text{Logits}_{\text{masked_position}(f_j)}), \quad (2)$$

where t_i is the ith template filled with the aspect, f_j is the jth ARF, and $MLM(t_i, f_j)$ is the MLM-estimated probability that f_j is filled into t_i. Finally, the prompt p_i with the highest score is selected.

3.3.4. Training of Initial Models

After selecting the best prompt, we obtain the prompt p_i for each training sample (x, a, y). These prompt-augmented training samples are used to train the Prompt Generation Model and the Sentiment Classification Model. The Text-to-Text Transfer Transformer

(T5) is chosen as the backbone model for both models. We chose the T5-base [30] as the pre-trained T5 model. This T5 model is fine-tuned using training samples where the input is the review x and the aspect a and the output is the prompt p_i, which yields the Prompt Generation Model. The loss function used for fine-tuning is the token-level cross-entropy loss as follows:

$$\mathcal{L}_{\text{prompt}} = -\sum_{t=1}^{T} \log P(p_t \mid p_{<t}, [x; a]), \tag{3}$$

where p_t denotes the t-th token of the ground-truth prompt, $p_{<t}$ represents all previously generated tokens, and $P(\cdot \mid p_{<t}, [x; a])$ is the probability distribution over the vocabulary output by the model at step t given the input context. Similarly, the Sentiment Classification Model is trained by fine-tuning the T5 model using the samples of the original review x and prompt p as input and the ground-truth sentiment label y as output. It also uses the cross-entropy loss as the loss function as follows,

$$\mathcal{L}_{\text{sentiment}} = -\log P(y \mid [x; p]), \tag{4}$$

where y is the ground-truth sentiment label, and $P(y \mid [x; p])$ is the probability that the model generates y given the input.

3.3.5. Rescoring of the Prompts and Re-Training of the Models

The prompt selected by the method in Section 3.3.3 may not be the most optimal choice. While the MLM primarily assesses the fluency of the prompt, it is uncertain whether such a prompt can perform the sentiment classification accurately. Therefore, we propose a mechanism to evaluate the appropriateness of the prompts based on their performance in the ASC task.

All prompt templates are rescored using the initially trained Sentiment Classification Model. Specifically, the prediction probability of the model is used as the new score of a template. Compared to the initial scoring using the MLM, this scoring allows us to select the template that can best contribute to improving the performance of sentiment classification.

Algorithm 1 gives the pseudocode for the rescoring. It shows how the prompt is renewed for a training sample (x, a, y), the chosen best prompt (p_{best}), and all prompts obtained by filling the ARFs into the templates (P). In Line 3, Prediction_Prob is the probability that label y be predicted by the initial Sentiment Classification Model M_{sc} using the best prompt p_{best} where the review x and the aspect a are given as input. If this probability is sufficiently high, the prompt is left unchanged. If not, a new prompt is chosen from the remaining prompts (P'), where the new prompt has the highest prediction probability (Lines 8–9). Note that θ is the hyperparameter that controls whether the initially chosen prompt is kept or replaced by the another promising prompt.

Algorithm 1 Prompt Rescoring Algorithm.

1: **Input:** Training sample (x, a, y), Selected prompt p_{best}, Set of all prompts P, Initial Sentiment Classfication Model M_{sc}
2: **Output:** New prompt p_{new}
3: $prob \leftarrow$ Prediction_Prob$(M_{sc}, x, a, p_{best}, y)$
4: **if** $prob > \theta$ **then**
5: $\quad p_{new} \leftarrow p_{best}$
6: **else**
7: $\quad P' \leftarrow P \setminus \{p_{best}\}$
8: $\quad p_{new} \leftarrow \arg\max_{p_i \in P'}$ Prediction_Prob(M_{sc}, x, a, p_i, y)
9: **end if**
10: **return** p_{new}

After updating the prompts for all training samples, the Prompt Generation Model and the Sentiment Classification Model are fine-tuned using the updated dataset.

3.4. Cluster-Based Prompt Expansion

To further enhance the diversity and robustness of the sentiment classification process, we introduce the *cluster-based prompt expansion* method at the inference phase.

The use of one prompt may suffer from prompt bias, which is the tendency of the prompt to favor a certain sentiment class. To alleviate this bias and improve the robustness of the classification model, the set of prompts is expanded by clustering the prompt, and then this expanded ensemble of prompts is used to determine the sentiment class more accurately.

Figure 3 shows an overview of cluster-based prompt expansion, while Algorithm 2 presents its pseudocode. For a given review x and an aspect a in the target domain, the prompt is generated by the Prompt Generation Model M_{pg} (Line 3). Then, the ARFs in the prompt p_{orig} are extracted as F (Line 4). Next, a set of prompts is constructed by filling the aspect a and the ARFs F in the templates t_i (Lines 5–8).

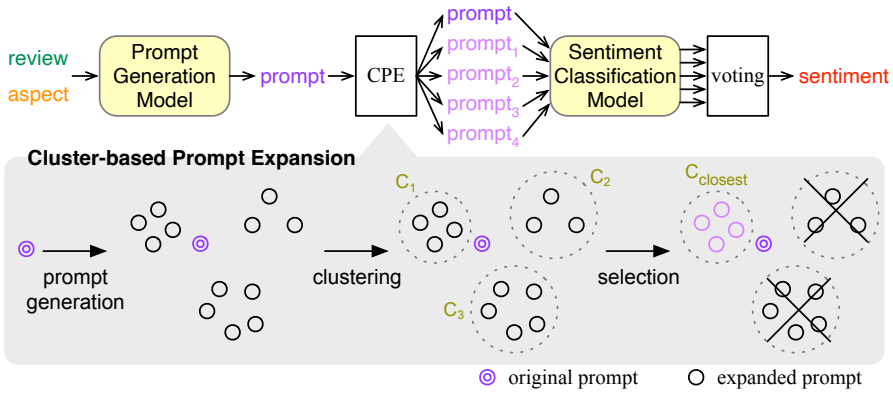


Figure 3. Overview of cluster-based prompt expansion.

Algorithm 2 Cluster-Based Prompt Expansion Algorithm.

1: **Input:** Review text and aspect (x, a), Set of templates T, Prompt Generation Model M_{pg}, Sentiment Classification Model M_{sc}
2: **Output:** Sentiment label y
3: $p_{orig} \leftarrow$ Prompt_Generation(M_{pg}, x, a)
4: $F \leftarrow$ Extract_ARF(p_{orig})
5: $P \leftarrow \emptyset$
6: **for** $t_i \in T$ **do**
7: $\quad p_i \leftarrow$ Fill_Template(t_i, a, F) ; $P \leftarrow P \cup \{p_i\}$
8: **end for**
9: $C \leftarrow$ Kmeans_Clustering(P, k)
10: $C_{closest} \leftarrow \arg\max_{C_i \in C}$ Similarity(p_{orig}, C_i)
11: $P_{all} \leftarrow \{p_{orig}\} \cup C_{closest}$
12: $\hat{y} \leftarrow$ Voting(M_{sc}, P_{all})
13: **return** $\hat{y}$

We perform K-means clustering [31] to make clusters of the prompts (Line 9). The number of clusters, k, is set to five so that four prompts from a total of twenty prompts can be included in each cluster. Each prompt is converted into an embedding (a dense vector) in order to carry out the clustering. In this study, the pre-trained BERT model is used to obtain the sentence embeddings. Then, the cluster most similar to p_{orig} ($C_{closest}$) is chosen (Line 10). The similarity between the embedding of the p_{orig} and the centroid of the cluster

C_i is measured. The original prompt and the prompts in $C_{closest}$, denoted by P_{all} in Line 11, are used for the sentiment classification.

The sentiment label is determined by voting on the classification labels predicted by the Sentiment Classification Model M_{sc} using each prompt $p_i \in P_{all}$. Two aggregation mechanisms are employed to finalize the predicted sentiment: *majority voting* and *weighted voting*. In the majority voting, the sentiment class with the highest frequency from among the predictions from all prompts in P_{all} is selected:

$$\hat{y} = \arg\max_{y} \sum_{p_i \in P_{all}} \mathbf{1}(M_{sc}(x, a, p_i) = y), \tag{5}$$

where $\mathbf{1}(\cdot)$ is an indicator function that returns 1 if $M_{sc}(x, a, p_i)$ (the sentiment class predicted by M_{sc} with the prompt p_i for a given review x and an aspect a) is equal to y, and 0 otherwise. In the weighted voting, the chosen sentiment class is

$$\hat{y} = \arg\max_{y} \left(w_{orig} \cdot \mathbf{1}(M_{sc}(x, a, p_{orig}) = y) + \sum_{p_i \in C_{closest}} w_{sup} \cdot \mathbf{1}(M_{sc}(x, a, p_i) = y) \right), \tag{6}$$

where w_{orig}, the weight for the original prompt, is set higher than those for the support prompts in the selected cluster, with all weights for the support prompts set equal to each other. That is, $w_{orig} = 1$ and $w_{sup} = \frac{1}{|C_{closest}|}$.

4. Evaluation

4.1. Datasets

Our evaluation experiments use datasets from five distinct domains: Restaurant (R), Laptop (La), Device (D), Service (S), and Location (Lo). Each dataset consists of triples of a review, an aspect, and a sentiment class. The details of the five datasets are shown below.

Restaurant and Laptop: The datasets of the SemEval-2014 ABSA task [32], which contain reviews of various aspects of restaurants and laptops, are used.
Device: The device domain dataset is taken from Toprak et al. [33]. It contains user reviews of electronic devices such as smartphones, tablets, and home appliances.
Service: The service domain dataset comes from Hu and Liu [34]. It includes users' feedback on service quality, staff responsiveness, and customer satisfaction.
Location: The Sentihood [35], which focuses on location-based sentiment analysis, is used. This dataset consists of reviews that evaluate the perceptions of specific places or neighborhoods.

These datasets were selected to represent a diverse range of application scenarios, with significant differences in domain-specific features and contents. While the Laptop and Device domains exhibit some similarities, other domains differ significantly from one another. Experiments on these datasets can assess the robustness and generalizability of the proposed method.

The numbers of positive, negative, and neutral samples in the datasets are shown in Table 2. A neutral label is not used in the datasets of Device and Location. The experiment employs the leave-one-domain-out strategy, where one dataset is used as an unseen target domain and the rest are used as multiple source domains. Binary classification is performed when the target dataset is Device or Location, while trinary classification is carried out for the remaining datasets.

Table 2. Statistics of datasets.

Domain	Positive	Negative	Neutral	Total
Restaurant	795	287	217	1299
Laptop	482	505	177	1164
Device	589	375	–	964
Service	920	629	135	1684
Location	954	480	–	1434

4.2. Experimental Settings

The settings of the hyperparameters are summarized in Table 3. The number of extracted ARFs per review, N_{arf}, is set to three. As described in Section 3.3.4, the T5-base model is used for both the Prompt Generation Model and the Sentiment Classification Model. We set the different learning rates and number of epochs to fine-tune these two models. Also, the different hyperparameters are used for additional fine-tuning of the models after rescoring the prompts.

Table 3. Hyperparameter settings. M_{pg} and M_{sc} stand for the Prompt Generation Model and the Sentiment Classification Model, respectively.

Process	Parameter	Value
Extraction of ARFs	N_{arf}	3
Fine-tuning of M_{pg}	Learning Rate	1×10^{-4}
	Epochs	10
Fine-tuning of M_{sc}	Learning Rate	3×10^{-5}
	Epochs	2
Additinal fine-tuning of M_{pg} & M_{sc}	Learning Rate	3×10^{-6}
	Epochs	1

The evaluation criteria for the ASC task are the accuracy and the macro average of the F1-scores for two (when the target domain is Device or Location) or three (otherwise) sentiment classes.

4.3. Models

This subsection describes baseline methods and our proposed method that are evaluated and compared in this experiment.

- **T5-base:** It is a T5-base model fine-tuned using the training data from multiple source domains. It serves as a benchmark for comparison. Unlike our framework, this model does not use any domain adaptation techniques. Instead, it directly processes the review and the aspect to generate a sentiment word as a classification result.
- **AutoPrompt [36]:** AutoPrompt (AP) is a prompt-based method for automatically constructing prompts using gradient-guided search. It is applied to multi-source unseen domain adaptation of the ASC task. Two pre-trained models, BERT and RoBERTa, are employed in this experiment. (RoBERTa is used in the original paper [36]).
- **LM-BFF [37]:** LM-BFF (Better Few-Shot Fine-Tuning of Language Models) is a prompt-based fine-tuning approach designed to enhance model performance with a limited number of annotated training examples by using automatically generated prompts and demonstrations. Although this method is designed for few-shot learning, it is

extended for multi-source unseen domain adaptation by using the entire training dataset. RoBERTa is used as the base pre-trained language model.

- **PADA:** The PADA model is an important previous model: it is a state-of-the-art method of multi-source unseen domain adaptation. It adapts source domain knowledge to the target domain by incorporating domain-specific and general features. The comparison with PADA allows us to evaluate the generalization ability of our method in the domain adaptation of the multi-source domain setting. In this experiment, we reimplemented the PADA model and applied it to our dataset to ensure a fair and consistent comparison.
- **AEP+RS+CE:** Aspect-Enhanced Prompting method, which is our proposed method. "+RS" and "+CE" indicate that the prompt rescoring module and the cluster-based prompt expansion module are employed, respectively.

4.4. Results

Table 4 shows the accuracy and macro F1-score of the six models in five cases of domain adaptation, where one of the five datasets is chosen as the unseen target domain. The "Average" column shows the macro average over these five cases.

Table 4. Performance of Aspect Sentiment Classification.

(a) accuracy						
Model	$\mathcal{MS}\rightarrow$R	$\mathcal{MS}\rightarrow$La	$\mathcal{MS}\rightarrow$D	$\mathcal{MS}\rightarrow$S	$\mathcal{MS}\rightarrow$Lo	Average
T5-base	0.746	0.754	0.882	0.798	0.809	0.798
AP(BERT)	0.743	0.758	0.893	0.793	0.818	0.801
AP(RoBERTa)	0.767	0.770	0.901	0.819	0.857	0.823
LM-BFF	0.752	0.764	0.891	0.802	0.806	0.803
PADA	**0.756**	**0.780**	0.907	0.832	0.815	0.818
AEP+RS+CE	0.754	0.769	**0.928**	**0.835**	**0.875**	**0.832**
(b) macro F1-score						
Model	$\mathcal{MS}\rightarrow$R	$\mathcal{MS}\rightarrow$La	$\mathcal{MS}\rightarrow$D	$\mathcal{MS}\rightarrow$S	$\mathcal{MS}\rightarrow$Lo	Average
T5-base	0.522	0.545	0.892	0.550	0.796	0.661
AP(BERT)	0.519	0.552	0.895	0.546	0.807	0.664
AP(RoBERTa)	**0.551**	0.582	0.897	0.561	0.842	0.687
LM-BFF	0.539	0.558	0.893	0.552	0.796	0.668
PADA	0.544	0.575	0.908	**0.577**	0.805	0.682
AEP+RS+CE	0.527	**0.596**	**0.925**	0.576	**0.862**	**0.697**

Note: $\mathcal{MS}$ stands for "multi-source domains". Bold indicates the best among all models.

Our AEP model significantly outperforms the T5-base. On average, the accuracy of the AEP+RS+CE is 0.832, which is 0.034 points higher than the T5-base. A significant improvement is observed when the target domain is Location (+0.066) and Device (+0.047). A similar trend is confirmed in the comparison of the F1-score. These results indicate that our Aspect-Enhanced Prompting method is effective for domain adaptation, even when the target domain is completely unseen. Furthermore, the superiority of the AEP+RS+CE over the T5-base is observed for all five target domains, demonstrating the robustness of our method.

Among the other baselines, the performance of AP(BERT) and LM-BFF is somewhat constrained. Given that LM-BFF is designed for few-shot learning, it may not achieve sufficient performance when it is applied to domain adaptation. The performance of AP(RoBERTa) is enhanced compared to AP(BERT) by using the better pre-trained language model. When the target domain is Restaurant or Laptop, i.e., the datasets of the SemEval 2014 ABSA task, the accuracy of AP(RoBERTa) and PADA is better than our method.

Nevertheless, for three out of the five domains, the accuracy of the AEP+RS+CE is better than these strong baselines. In terms of the F1-score, our method is less effective than the AP(RoBERTa) and PADA for the Restaurant domain, but better or comparable for the other domains. On average across the five domains, the AEP+RS+CE outperforms all baselines in terms of both the accuracy and F1-score.

These results indicate that the AEP has an excellent ability to adapt the sentiment classification model to different types of domains. The mechanism for automatically generating the prompts including ARFs could contribute to preventing the model from performing worse on an unseen target domain.

4.5. Detailed Evaluation of the Components

Several additional experiments have been conducted to evaluate the contribution of the components in the AEP model. As the results for accuracy and F1-score are nearly the same, only the F1-scores will be presented in this subsection.

4.5.1. Ablation Study

Table 5 shows the results of the ablation study of the rescoring of the prompts and the cluster-based prompt expansion. The symbol × in the columns "RS" and 'CE" indicate that the prompt scoring and the cluster-based prompt expansion modules were not used.

Table 5. Ablation study of the AEP model. Δ indicates the difference in the F1-score from the full model AEP+RS+CE.

Model	RS	CE	$\mathcal{MS}$→R	$\mathcal{MS}$→La	$\mathcal{MS}$→D	$\mathcal{MS}$→S	$\mathcal{MS}$→Lo	Average
AEP+RS+CE			0.527	0.596	0.925	0.576	0.862	0.697
AEP+CE (Δ)	×		0.526 (−0.001)	0.596 (0.000)	0.921 (−0.004)	0.575 (−0.001)	0.853 (−0.009)	0.694 (−0.003)
AEP+RS (Δ)		×	0.525 (−0.002)	0.583 (−0.013)	0.923 (−0.002)	0.572 (−0.004)	0.857 (−0.005)	0.692 (−0.005)
AEP (Δ)	×	×	0.521 (−0.006)	0.549 (−0.047)	0.916 (−0.009)	0.572 (−0.004)	0.846 (−0.016)	0.681 (−0.016)
T5-base	×	×	0.522	0.545	0.892	0.550	0.796	0.661

When the rescoring of the prompts is not performed, the F1-score of the model decreases to 0.694. The most notable decline is observed in the Location domain, where the F1-score drops from 0.862 to 0.853 (a decrease of 0.009). Although the F1-score remains unchanged for the Laptop domain, this outcome confirms the effectiveness of the method of selecting the prompt by rescoring based on the initial Sentiment Classification Model.

When the cluster-based prompt expansion is not carried out, the average F1-score drops to 0.692. While the prompt expansion generally improves the sentiment classification, its influence depends on the domain. For instance, not carrying it out leads to a slight reduction in the F1-score for the Laptop domain, but has minimal impact on the Restaurant and Device domains.

A comparison of the two components reveals that the cluster-based prompt expansion is more effective than prompt rescoring in terms of improving the F1-score. Not carrying out the prompt expansion results in a more pronounced decline in the average F1-score (0.692) than does omitting rescoring the prompts (0.694). The AEP model without both modules shows a substantial decline in performance, with its average F1-score dropping to 0.681, especially in the Laptop domain (−0.047). Furthermore, the difference in F1-scores between AEP and the full model (AEP+RS+CE) is greater than the sum of the differences

of each component (i.e., (AEP+RS+CE − APE+CE) plus (AEP+RS+CE − AEP+RS)). This shows that rescoring the prompts and the cluster-based prompt expansion complement each other in improving the performance of the sentiment classification.

In addition, the comparison between AEP and T5-base, which does not use the AEP, demonstrates that only the AEP can improve the performance of ABSA for the most domains. On average, AEP improves the macro F1-score by 2.0 points.

To address the additional computational cost associated with the cluster-based prompt expansion, the supplementary processing time incurred per sample for each target domain is measured using an NVIDIA RTX A6000 GPU server. The range of the average of this supplementary time was from 0.014 s (Service domain) to 0.044 s (Device domain). These results indicate that the cluster-based prompt expansion imposes a relatively moderate computational burden.

4.5.2. Impact of Parameters on Prompt Rescoring

As described in Section 3.3.5, our rescoring method refines the prompt by filtering out prompts with low sentiment prediction probabilities. By adjusting the threshold θ, the model can either retain the initially selected prompts or replace them with alternatives that deserve higher confidence. To investigate the effect of this threshold, we have evaluated the performance of the models with different values of θ. In this experiment, the cluster-based prompt expansion is not used, this is so as to focus on evaluating the prompt rescoring module. Table 6 shows the results of the experiment.

Table 6. F1-score of models with different parameters for rescoring.

Model	$\mathcal{MS}\rightarrow$R	$\mathcal{MS}\rightarrow$La	$\mathcal{MS}\rightarrow$D	$\mathcal{MS}\rightarrow$S	$\mathcal{MS}\rightarrow$Lo	Ave.
AEP	0.521	0.549	0.916	0.572	0.846	0.681
AEP+RS (θ = 0.99)	0.522	0.581	0.920	**0.574**	0.853	0.690
AEP+RS (θ = 0.98)	**0.525**	**0.583**	**0.923**	0.572	**0.857**	**0.692**
AEP+RS (θ = 0.95)	0.524	0.576	0.920	**0.574**	0.856	0.690

The models with the prompt rescoring using arbitrary values of θ consistently outperform the model without rescoring (AEP), except in the case of the Service domain. This implies that the contribution of the prompt rescoring is less sensitive to the threshold for any target domain. A comparison of models using three different thresholds reveals that the AEP+RS (θ = 0.98) model achieves the best F1-score for the Restaurant, Laptop, Device, and Location domains, as well as on average. The scores of the prompts are calculated based on the probability that the MLM fills the template with ARFs in the initial prompt generation phase, while the prompts are re-evaluated based on the reliability of the sentiment classification in the rescoring phase. High values of the parameters prefer the former scoring schemes and low values of the parameters the latter. The moderate threshold (i.e., θ = 0.98) could adequately consider these two scoring criteria, resulting in a good performance.

4.5.3. Investigation of Voting Strategy in Cluster-Based Prompt Expansion

In the cluster-based prompt expansion, the results of the sentiment classification using the expanded multiple prompts are combined in two ways: (simple) majority voting and weighted majority voting. We have carried out an experiment to compare the effectiveness of these two methods. Table 7 shows the results of the models using the two different strategies; "CE(ma)" and "CE(we)" indicate the cluster-based prompt expansion using majority voting and weighted voting, respectively. AEP+CE is the model without prompt

rescoring. Since the weighted voting is applied in our full model, AEP+RS+CE(we) is equivalent to AEP+RS+CE in Tables 4 and 5.

Table 7. Macro F1 of models with two voting methods in cluster-based prompt expansion.

Model	$\mathcal{MS}\rightarrow$R	$\mathcal{MS}\rightarrow$La	$\mathcal{MS}\rightarrow$D	$\mathcal{MS}\rightarrow$S	$\mathcal{MS}\rightarrow$Lo	Ave.
AEP+CE (ma)	0.523	**0.603**	0.910	0.573	**0.853**	0.692
AEP+CE (we)	**0.526**	0.596	**0.921**	**0.575**	**0.853**	**0.694**
AEP+RS+CE (ma)	**0.528**	**0.597**	0.918	0.571	**0.863**	0.695
AEP+RS+CE (we)	0.527	0.596	**0.925**	**0.576**	0.862	**0.697**

Comparing AEP+RS+CE (ma) and AEP+RS+CE (we), the weighted voting achieves a better performance in the Device and the Service domains, as well as on average. This demonstrates the significance of the weights in Equation (6) for the aggregation of the results derived from different prompts. When the prompt rescoring is not carried out, the weighted voting achieves better results than the majority voting in three domains, as well as on average. Although the average accuracy of the model using weighted voting is higher than when it uses the majority voting, the superiority of weighted voting is relatively modest. Overall, however, we can conclude that weighted voting is a more effective approach for cluster-based prompt expansion.

When either majority voting or weighted voting is applied, the models with the prompt rescoring are better than the models without it. These results support the effectiveness of the prompt rescoring, which is consistent with the results of the ablation study.

4.5.4. Investigation of Input Format of Sentiment Classification Model

In our two-stage method of domain adaptation for sentiment classification, a prompt is generated in the first stage, and then the review and the generated prompt are fed into the classification model in the second stage. Here, we will present and evaluate two strategies for feeding a review and a prompt into the Sentiment Classification Model.

AEP-Separate is a method that simply concatenates a review and a prompt to form an input to the Sentiment Classification Model. In contrast, the AEP-Insert method inserts a review text into a prompt so that a single long sentence is fed into the classification model. A template of a prompt is designed to fill an aspect, ARFs, and a review text, such as "Consider the text: [TEXT], what sentiment does [ARF] convey about [ASPECT]?". See Table 8 for examples of the formatted texts in these two methods.

Table 8. Two input formats of Sentiment Classification Model. The review text is underlined.

Method	Example
AEP-Separate	*This is a great Thai restaurant with a very friendly staff.* With *staff* being *great Thai restaurant*, how is the sentiment?
AEP-Insert	Consider the text: *This is a great Thai restaurant with a very friendly staff.* what sentiment does *great Thai restaurant* convey about *staff*?

Table 9 compares between AEP-Separate and AEP-Insert. In this experiment, neither the prompt rescoring nor the cluster-based prompt expansion was applied. This was in order to enable a focused evaluation of these two different strategies. The results indicate that AEP-Separate outperforms AEP-Insert for all target domains except for the Restaurant domain. A significant difference in the average F1-scores of the two methods is also found. In AEP-Insert, it seems difficult for the model to identify the sentiment of an aspect in a

complicated single prompt that includes a review text. The separation of a review and a prompt could help a model to classify the sentiment more precisely.

Table 9. F1-scores of AEP-Separate and AEP-Insert.

Model	$\mathcal{MS}\rightarrow$R	$\mathcal{MS}\rightarrow$La	$\mathcal{MS}\rightarrow$D	$\mathcal{MS}\rightarrow$S	$\mathcal{MS}\rightarrow$Lo	Ave.
AEP-Separate	0.521	**0.549**	**0.916**	**0.572**	**0.846**	**0.681**
AEP-Insert	**0.525**	0.548	0.910	0.520	0.806	0.662

4.5.5. Impact of Prompt Templates

To investigate how sensitive the performance of the AEP is to different prompts, several AEP models with different prompt templates are compared. Specifically, different sets of prompt templates are created by decreasing the number of templates. Figure 4 shows the comparison of three models: "AEP" using all 20 templates, "AEP-Q-only" using the question templates (11 templates), and "AEP-D-only" using the description templates (9 templates). It is found that the macro-F1 scores of the models using fewer templates are worse than the model using all templates, although the differences are not noteworthy. This suggests that the preparation of a wide variety of templates plays an important role in the AEP framework.

In the future, the dynamic generation of templates using pre-trained language models, such as T5 [25] and GPT [38], holds considerable potential for investigation in the future, since it could address the limitation of linguistic variety inherent to a few hand-crafted templates.

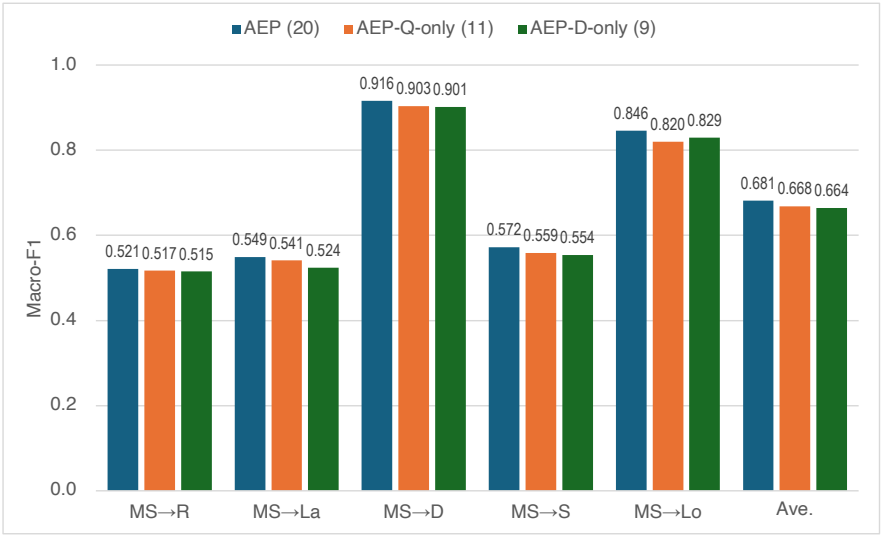


Figure 4. Comparison of different sets of prompt templates. The number of used templates is shown in parentheses.

4.6. Error Analysis

An error analysis is conducted to gain a deeper insight into the weaknesses of our model. Since both the accuracy and the F1-score of our full model are worse than PADA when the target domain is the restaurant, as shown in Table 4, we manually investigate the misclassification made by the AEP+RS+CE model in the Restaurant domain.

Table 10 shows examples of misclassification. The first example shows neutral sentiment for the aspect "food" and negative sentiment for the aspect "ambience". Since the negative word "annoying" is generated as an ARF in the prompt, the model may incorrectly predict the sentiment of the aspect "food" as negative. Similarly, the error in the second example may be caused by the positive ARF "friendly", which is not related to

the target aspect. In the third example, the generated ARFs ("much", "bring", and "back") are unrelated to the sentiment for the aspect "food". Overall, these cases illustrate the limitation of our model, which may be caused by a failure to generate appropriate ARFs. Additionally, PADA also fails to generate words related to the sentiment of the aspect in the prompt. Although it correctly predicts the sentiment for the first and second examples, it often causes misclassifications for other samples, resulting in a decrease in accuracy compared to AEP.

Table 10. Misclassified examples in the Restaurant domain.

Input Text (Aspect, Gold Label)	Prompt		Prediction	
	AEP	PADA	AEP	PADA
The food is decent at best, and the ambience, well, it's a matter of opinion, some may consider it to be a sweet thing, I thought it was just annoying. (*food*, *neutral*)	With food being decent ambience annoying, how is the sentiment?	food egroups overall quickly table	neg	neu
The service was a bit slow, but they were very friendly. (*service*, *negative*)	Predict the sentiment for service described as bit slow friendly.	service slow egroups dinner toshiba	pos	neg
As much as I like the food there, I cant bring myself to go back. (*food*, *positive*)	With food being much bring back, how is the sentiment?	food egroups week extremely simple	neg	neg

Figure 5 shows the number of samples, which are correctly predicted by the PADA model but misclassified by our model, for each of the sentiment classes. (While there is no neutral sample in the Device and Location datasets (see Table 2), no neutral sample that only our model misclassified is found in the Service domain.) It was found that our AEP model struggled significantly with neutral reviews in the Restaurant domain. While the PADA model generates general domain-specific features in a prompt, which are not always extracted from a review sentence, ARFs are usually extracted from a review sentence. By inappropriately extracting sentiment words as ARFs, our model incorrectly predicts a neutral sample as positive or negative. In addition, the poor performance in the Restaurant domain can be attributed to the diverse and descriptive nature of reviews about restaurants, where aspects with neutral and non-neutral sentiments can appear in a sentence.

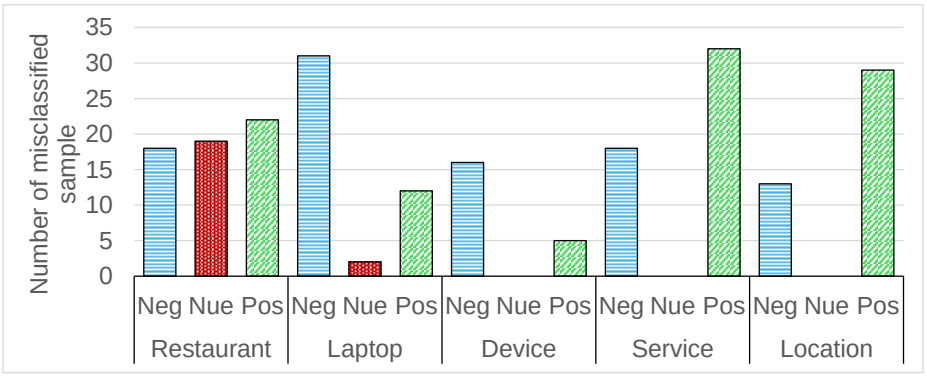


Figure 5. Number of errors for each sentiment class.

5. Conclusions

In this study, we proposed the Aspect-Enhanced Prompting (AEP) framework for Aspect Sentiment Classification in a Multi-Source Domain Adaptation setting. Our AEP employed Aspect-Related Features (ARFs), which were crucial for the sentiment classification. The generation of a prompt containing ARFs helped the generative language model to classify the sentiment of an aspect in a review in an unseen target domain. The generation of the prompts was enhanced by rescoring the prompts, taking into consideration the reliability of the sentiment classification. Our classification procedure using aspect-enhanced prompts was further refined by a clustering-based prompt expansion, which determined the sentiment class using the resulting expanded ensemble of multiple prompts. Experiments on five datasets from different domains demonstrated the effectiveness of our AEP framework. The contributions of the rescoring of the prompts and the cluster-based prompt expansion were confirmed by an ablation study.

However, some limitations still remain. This study has focused primarily on single-aspect reviews, limiting its relevance for the evaluation of reviews containing multiple or implicit aspects. Although the AEP framework can be extended to handle multiple aspects and implicit aspects, further experiments are necessary to validate its efficacy in these scenarios. Another important direction for future research is to enhance the efficiency of the process of selecting the prompts. Currently, this process relies on MLM scoring and a subsequent rescoring. We will investigate the potential of introducing a prompt selection model, which could be implemented using Multi-Layer Perceptrons and so on, and jointly training the prompt selection model and the sentiment classification model to optimize the process of selecting the prompts. It is also important to explore the integration of the AEP framework with Large Language Models (LLMs). Instead of T5, experiments with larger and more recent models, such as the LLaMA [39] and the GPT models (e.g., GPT-3, GPT-4) [38,40], could provide further insight into the effectiveness of AEP.

Author Contributions: Conceptualization, B.L. and K.S.; methodology, B.L. and K.S.; software, B.L.; validation, B.L.; formal analysis, B.L. and K.S.; investigation, B.L. and K.S.; resources, B.L. and K.S.; data curation, B.L.; writing—original draft preparation, B.L., K.S., and N.K.; writing—review and editing, B.L., K.S., and N.K.; visualization, B.L., K.S., and N.K.; supervision, K.S.; project administration, B.L. and K.S.; B.L. is the first author (Binghan Lu), K.S. is the second author (Kiyoaki Shirai), and N.K. is the third author (Natthawut Kertkeidkachorn). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used for training and evaluation is available at https://github.com/lbh-nlp/AEP (accessed on 28 March 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. List of Templates

Table A1 shows all the templates of the prompt. “Q” and “D” indicate the question and description templates, respectively. Table A2 shows the insertion templates used in the experiment in Section 4.5.4, where a review is filled into [TEXT].

Table A1. List of prompt templates.

Template	Type
Given the statement text, where aspect [ASPECT] is described as [ARF], what is the sentiment?	Q
For the statement text and focusing on [ASPECT] being [ARF], what is the sentiment?	Q
Analyze the sentiment of text with emphasis on [ASPECT] being [ARF].	D
How does text portray [ASPECT] as [ARF] in terms of sentiment?	Q
Considering the text, what sentiment does [ARF] convey about [ASPECT]?	Q
Evaluate the sentiment towards [ASPECT] being [ARF].	D
What is the emotional tone when [ASPECT] is [ARF]?	Q
In the text, [ASPECT] is described as [ARF]. What sentiment does this reflect?	Q
Assess the feeling towards [ASPECT] being [ARF].	D
Determine the sentiment of [ASPECT] is characterized as [ARF].	D
What emotion is associated with [ASPECT] being [ARF]?	Q
Identify the sentiment when [ASPECT] is mentioned as [ARF].	D
Predict the sentiment for [ASPECT] described as [ARF].	D
In the text, [ASPECT] is [ARF]. How does this make the sentiment?	Q
Sentiment analysis with [ASPECT] as [ARF].	D
How is the sentiment towards [ASPECT] being [ARF]?	Q
What is the sentiment outcome when [ASPECT] equals [ARF]?	Q
Review the sentiment with [ASPECT] as [ARF].	D
With [ASPECT] being [ARF], how is the sentiment?	Q
Analyze for sentiment with a focus on [ASPECT] as [ARF].	D

Table A2. List of prompt templates (insertion templates).

Template	Type
Given the statement [TEXT], where aspect [ASPECT] is described as [ARF], what is the sentiment?	Q
For the statement [TEXT] and focusing on [ASPECT] being [ARF], what is the sentiment?	Q
Analyze the sentiment of [TEXT] with emphasis on [ASPECT] being [ARF].	D
How does [TEXT] portray [ASPECT] as [ARF] in terms of sentiment?	Q
Considering the [TEXT], what sentiment does [ARF] convey about [ASPECT]?	Q
In the [TEXT], evaluate the sentiment towards [ASPECT] being [ARF].	D
In the [TEXT], what is the emotional tone when [ASPECT] is [ARF]?	Q
In the [TEXT], [ASPECT] is described as [ARF]. What sentiment does this reflect?	Q
In the [TEXT], assess the feeling towards [ASPECT] being [ARF].	D
In the [TEXT], determine the sentiment of [ASPECT] is characterized as [ARF].	D

Table A2. *Cont.*

Template	Type
In the [TEXT], what emotion is associated with [ASPECT] being [ARF]?	Q
In the [TEXT], identify the sentiment when [ASPECT] is mentioned as [ARF].	D
In the [TEXT], predict the sentiment for [ASPECT] described as [ARF].	D
In the [TEXT], [ASPECT] is [ARF]. How does this make the sentiment?	Q
In the [TEXT], sentiment analysis with [ASPECT] as [ARF].	D
In the [TEXT], how is the sentiment towards [ASPECT] being [ARF]?	Q
In the [TEXT], what is the sentiment outcome when [ASPECT] equals [ARF]?	Q
In the [TEXT], review the sentiment with [ASPECT] as [ARF].	D
In the [TEXT], with [ASPECT] being [ARF], how is the sentiment?	Q
In the [TEXT], analyze for sentiment with a focus on [ASPECT] as [ARF].	D

References

1. Pang, B.; Lee, L. Opinion Mining and Sentiment Analysis. *Found. Trends® Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
2. Blitzer, J.; Dredze, M.; Pereira, F. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 25–27 June 2007; pp. 440–447.
3. Gong, C.; Yu, J.; Xia, R. Unified Feature and Instance Based Domain Adaptation for Aspect-Based Sentiment Analysis. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 7035–7045. [CrossRef]
4. Ramponi, A.; Plank, B. Neural Unsupervised Domain Adaptation in NLP—A Survey. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 6838–6855. [CrossRef]
5. Blitzer, J.; McDonald, R.; Pereira, F. Domain Adaptation with Structural Correspondence Learning. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006; pp. 120–128.
6. Pan, S.J.; Ni, X.; Sun, J.T.; Yang, Q.; Chen, Z. Cross-domain sentiment classification via spectral feature alignment. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 751–760.
7. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *arXiv* **2016**, arXiv:1505.07818. [CrossRef]
8. Long, M.; Cao, Y.; Wang, J.; Jordan, M.I. Learning Transferable Features with Deep Adaptation Networks. *arXiv* **2015**, arXiv:1502.02791 [CrossRef]
9. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008.
10. Glorot, X.; Bordes, A.; Bengio, Y. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In Proceedings of the International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011.
11. Chen, M.; Xu, Z.; Weinberger, K.; Sha, F. Marginalized Denoising Autoencoders for Domain Adaptation. *arXiv* **2012**, arXiv:1206.4683. [CrossRef]
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [CrossRef]
13. Han, X.; Eisenstein, J. Unsupervised Domain Adaptation of Contextualized Embeddings for Sequence Labeling. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 4238–4248. [CrossRef]
14. Karouzos, C.; Paraskevopoulos, G.; Potamianos, A. UDALM: Unsupervised Domain Adaptation through Language Modeling. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 2579–2590. [CrossRef]
15. Zhao, Z.; Ma, Z.; Lin, Z.; Xie, J.; Li, Y.; Shen, Y. Source-free Domain Adaptation for Aspect-based Sentiment Analysis. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, 20–25 May 2024; pp. 15076–15086.
16. Guo, H.; Mao, Y.; Zhang, R. Augmenting Data with Mixup for Sentence Classification: An Empirical Study. *arXiv* **2019**, arXiv:1905.08941. [CrossRef]

17. Yu, J.; Gong, C.; Xia, R. Cross-Domain Review Generation for Aspect-Based Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 4767–4777. [CrossRef]
18. Wang, H.; He, K.; Li, B.; Chen, L.; Li, F.; Han, X.; Teng, C.; Ji, D. Refining and Synthesis: A Simple yet Effective Data Augmentation Framework for Cross-Domain Aspect-based Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 10318–10329. [CrossRef]
19. Guo, H.; Pasunuru, R.; Bansal, M. Multi-Source Domain Adaptation for Text Classification via DistanceNet-Bandits. *arXiv* **2020**, arXiv:2001.04362. [CrossRef]
20. Guo, J.; Shah, D.; Barzilay, R. Multi-Source Domain Adaptation with Mixture of Experts. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 4694–4703. [CrossRef]
21. Wright, D.; Augenstein, I. Transformer Based Multi-Source Domain Adaptation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 7963–7974. [CrossRef]
22. Ma, X.; Xu, P.; Wang, Z.; Nallapati, R.; Xiang, B. Domain Adaptation with BERT-based Domain Classification and Data Selection. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), Hong Kong, China, 3 November 2019; pp. 76–83. [CrossRef]
23. Li, R.; Liu, C.; Tong, Y.; Dazhi, J. Feature Structure Matching for Multi-source Sentiment Analysis with Efficient Adaptive Tuning. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, 20–25 May 2024; pp. 7153–7162.
24. Yang, S.; Jiang, X.; Zhao, H.; Zeng, W.; Liu, H.; Jia, Y. FaiMA: Feature-aware In-context Learning for Multi-domain Aspect-based Sentiment Analysis. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, 20–25 May 2024; pp. 7089–7100.
25. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* **2023**, arXiv:1910.10683. [CrossRef]
26. Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online/Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3045–3059. [CrossRef]
27. Ben-David, E.; Oved, N.; Reichart, R. PADA: Example-based Prompt Learning for on-the-fly Adaptation to Unseen Domains. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 414–433. [CrossRef]
28. Sun, X.; Zhang, K.; Liu, Q.; Bao, M.; Chen, Y. Harnessing domain insights: A prompt knowledge tuning method for aspect-based sentiment analysis. *Knowl.-Based Syst.* **2024**, *298*, 111975. [CrossRef]
29. Church, K.W.; Hanks, P. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.* **1990**, *16*, 22–29.
30. Google. T5-Base Model. 2024. Available online: https://huggingface.co/google-t5/t5-base (accessed on 15 May 2025).
31. MacQueen, J.B. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, 21 June–18 July 1965; Volume 1, pp. 281–297.
32. Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 27–35. [CrossRef]
33. Toprak, C.; Jakob, N.; Gurevych, I. Sentence and Expression Level Annotation of Opinions in User-Generated Discourse. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 13 July 2010; pp. 575–584.
34. Hu, T.; Liu, B. Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; pp. 168–177.
35. Saeidi, M.; Bouchard, G.; Liakata, M.; Riedel, S. SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 1546–1556.
36. Shin, T.; Razeghi, Y.; IV, R.L.L.; Wallace, E.; Singh, S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020.
37. Gao, T.; Fisch, A.; Chen, D. Making Pre-trained Language Models Better Few-shot Learners. In Proceedings of the Association for Computational Linguistics (ACL), Online, 5–10 July 2021.
38. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165. [CrossRef]

39. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971. [CrossRef]
40. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2024**, arXiv:2303.08774. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.