Cross-Domain Sentiment Analysis using Transfer Learning and Domain Tokens

Dhirendra Yadav*[1], Mohit Singh[2], Ramesh Chundi[3], and Senthil S[4]

School of Computer Applications, Dayananda Sagar University, Bengaluru, India
{dhirendrarao2175, mohitsingh58414, chundiramesh}@gmail.com, senthils-ca@dsu.edu.in}
* Corresponding author: dhirendrarao2175@gmail.com

Abstract. Most existing models of sentiment analysis degrade in performance significantly when used out of domains they are originally trained on, which is essentially because vocabularies, semantics, and contextual patterns all vary across domains. In this paper, we will introduce a simple, lightweight, and flexible transfer learning approach to improve assumption classification tasks on hidden spaces with only a few available labeled data. Our model proposes that domain tokens are introduced at the input level. That is, special tokens like [DVD], [BOOKS] will indicate different domains. These domain tokens explicitly cue the model to learn domain-aware representations without modification of its architecture. Second, we integrate a few-shot learning process where the model is fine-tuned with only a few labeled cases of the target domain. Our approach has been evaluated on the widely used dataset of Amazon Review, which includes four domains: Electronics, DVD, Books, and Kitchen. The results reveal that our model with these domain tokens consistently preserves the top rank position and reaches F1 values of up to 0.66 in both zero-shot and few-shot learning. The above points and discoveries reveal the capability and potential of space tokens in performing opinion analysis in a domain.

Keywords: Transfer Learning · Domain Tokens · Few-Shot Learning · Amazon Review Dataset · Contextual Embeddings

1 Introduction

Sentiment analysis, the process of automatically detecting emotional tone in textual content, has now found significant applications in a variety of domains ranging from the analysis of user feedback to opinion tracking and trend observation in social platforms [1,4]. Traditional sentiment classification models, while giving good results when trained and tested within the same domain, turn out to perform much poorly when deployed across different domains-a problem generally referred to as domain shift. This primarily arises because of changes in word meanings, vocabulary, and context that occur from one domain to another [2,6]. For instance, the word “compact” may imply the positive feature

P. Johri et al. (eds.), *Proceedings of the International Conference on Sustainable Computing and Artificial Intelligence (ICSCAI 2025)*, Advances in Engineering Research 298,
https://doi.org/10.2991/978-94-6239-674-6_26

of portability in reviews on electronic products but lack of detail in book reviews. These minor yet domain-specific differences render the possibility of a single universal sentiment model fairly difficult. Application-specific, individual models are again not feasible, especially when labeled examples are limited. Recent breakthroughs in transfer learning, especially those based on pre-trained transformer-based models like BERT [4] and RoBERTa [5], hold much promise. These models be educated with rich relevant features on large-scale data and be fine-tuned for a given task with a comparatively small number of labeled test instances. Nonetheless, for cross-domain applications, they generally demonstrate the deterioration of performance as a result of domain gaps [10]. To overcome this problem, our study proposes a light-run improvement for the standard transformer architecture, namely domain tokens. They are a sort of unique feature indicator, like “[BOOKS],” “[ELECTRONICS],” and so on, that point to the domain of the content at the start of the input text. This technique has been borrowed from the prompt-based learning theory, as proposed by [13], and domain adaptation learning, as mentioned by [6]. To check the effectiveness of the proposed technique in a real-world situation, our study has utilized a framework for a few-shot learning technique, meaning that the fine-tuning has been accomplished by making use of a comparatively small number of labeled instances of the target domain, like 50 and 100 instances. Our experimental results show that models improved by domain tokens perform better compared to regular transform models, particularly in a cross-domain scenario. To demonstrate the efficacy of the proposed technique, the Amazon Reviews Dataset [2] has been utilized, which comprises four domains—Books, Electronics, DVDs, and Kitchen. Our baselines and proposed models are then trained and tested over unseen Books and Kitchen domains based on the Electronics and DVDs datasets, separately. Our method produces a consistent improvement of up to 13% in F1-score over regular fine-tuning. Our contributions include:

- Introduce domain tokens to increase generalization and sensitivity over domains.
- Effective performance on few-shot learning with good adaptability in a few labeled data.
- Extensive empirical evaluation, including training curves, few-shot analysis, and confusion matrices confirming both robustness and interpretability.

Research Objectives: The primary goals of this research are to:

1. Explore how domain-specific token injection can affect cross-domain sentiment analysis accuracy.
2. Evaluate the generalization capability of the model when it comes to zero-shot or few-shot learning settings.
3. Compare the proposed approach with conventional transformer-based models and the latest state-of-the-art domain adaptation techniques.
4. Showcase an efficient approach to solve the sentiment analysis task in the multi-domain setting.

Key Contributions: Building on these objectives, the contributions of this study are:

- Providing domain tokens to guide transformer models to learn domain-aware representations without altering the transformer model architecture.
- Low Resource Learning through Few Shot Learning.
- Carrying out comprehensive evaluation tasks such as the comparison of F1 scores, the evaluation of confusion matrices, and the study behavior of the training process regarding the approach

As shown in Figure 1, our approach insists on simplicity and being modular and simple to implement. The remainder of the paper is divided as follows: Segment II describes the existing work in the area of cross-domain opinion investigation and exchange learning. Area III describes the datasets and environment. Area IV describes the proposed technique. Zone V describes the results achieved by us, and the Conclusion Segment VI.

2 Related Work

Cross-domain sentiment analysis, being an active area of research, still holds importance because of the lack of generalizability for the models learned in one domain [2], [6]. In typical sentiment classification, it was noted that these models have poor performances on unknown domains, mainly due to the variations that exist in words, syntax, and contexts. The preliminary solutions for this problem involved feature alignment for the domain using methods based on pivot learning. Blitzer et al. [2] proposed a system for finding pivot features. Finding common words between domains aids in relating the source and target domains. Though very promising, these kinds of methods require human expertise in feature engineering, which is unscalable if domains are many. A major breakthrough in CDSA was achieved with the advent of deep learning. Glorot et al. [7] proposed Stacked Denoising Autoencoders for obtaining domain-invariant features in an unsupervised setup. Still, these models required retraining after every domain, making this also somewhat impractical for usage. Later on, adversarial learning was adapted. Ganin et al. [15] proposed Domain Adversarial Neural Networks by using the concept of the reverse gradient for obtaining domain-invariant features.While powerful, adversarial methods are sensitive to hyperparameters and involve complicated optimization strategies. Central to CDSA was the rise of transformer-based language models like BERT and RoBERTa, wherein transfer learning played the key role. These models provided deep contextual embeddings that could be fine-tuned for tasks downstream with a small amount of labeled data. Gururangan et al. proposed Domain-Adaptive Pretraining, or DAPT, which described extra unsupervised training on domain-specific data. However, this turns out to be resource-consuming and sometimes is not fit for immediate deployment across several domains. This led to developing lightweight alternatives to bypass such limitations. Zhao et al. devised domain-aware fine-tuning by incorporating domain indicators into training. Han et al. introduced the domain prompt tuning method that simply appended domain cues directly to the input text, thus enabling adaptation without changes in the architecture of the model. Vu et al. developed CAPTAIN, which is a prompt-based framework and leverages structured task instructions to guide transfer learning. These instruction tuning methods have inspired our approach of injecting domain tokens, which has encoded the domain identity at the input level for scale adaptation. Few-shot learning is also an efficient approach for low-data settings. Li et al. [14] proposed meta-learning on transformer architectures that supported quick adaptation even from very few annotated examples. Similar ideas include the work of Arora et al. [19] and Li et al. [20], which employed meta-learning with prompts or domain knowledge to enhance transferability, though these are generally more complicated two-step procedures. Conversely, the current method supports one-step direct fine-tuning from merely a few annotated examples from the target domain. Other recent investigations also involve efficient fine-tuning of the model parameters. Mahabadi et al. [18] proposed the Compacter, an efficient adapter layer that employed low-rank matrix approximation to speed up the whole training procedure. Another established technique is instance weighting as well. Jiang Zhai [21] suggested that reweighting the examples according to relevance to the domain is very useful to overcome distribution shifts. From the literature, there is clearly an intense interest on more efficient solutions that are compatible with few-shot settings and suitable for CDSA. Our method advances on these foundations by directly injecting the domain information at the input level to support fast adaptation without making extensive alterations to either the model or the whole training inference.

Table 1: Comparison of Cross-Domain Sentiment Analysis Methods

Method	**Pretrained LM**	**Domain Adaptation**	**Few Shot**	**Light weight**
Blitzer et al. (2007)	×	Pivot features	×	✓
Glorot et al. (2011)	×	Autoencoder	×	×
Ganin & Lempitsky (2015)	×	Adversarial	×	×
Gururangan et al. (2020)	✓	DAPT	×	×
Li et al. (2021)	✓	Meta-Learning	✓	×
Wang et al. (2024)	✓	Causal Domain Generalization	×	×
Chen (2025)	✓	Transfer Learning	✓	×
Xu (2024)	✓	Bert-BiLSTM + Dual Attention	×	×
Our Approach	✓	Domain Tokens	✓	✓

From Table 1, it is clear that our method achieves a unique harmony of simplicity, effectiveness, and flexibility, which provides a highly efficient yet light solution for cross-domain sentiment analysis without using labeled data and with the architectures unchanged.

3 Dataset and Experimental Setup

To test the quality of our domain-independent sentiment classifier, we use the Amazon Product Reviews dataset, which was originally created by Blitzer et al. [2]. This data set can be a rather common benchmark in the domain of space adjustments because it represents the divisions in space in a rather clear manner, with adjusted opinion classes. This data set contains reviews, which are classified as positive or negative, and there can be no better example or context for a binary classifier. There are multiple products, which represent distinctive domains in the data set because each domain represents certain linguistic differences.

3.1 Dataset Summary

The reviews are divided into four main categories: Electronics, DVD, Books, and Kitchen. Following previous work [6], our model is trained and validated on the subsets of Electronics and DVD. To assess its generalization to unseen domains, we test it on Books and Kitchen domains not observed during training. This replicates practical scenarios where a model is required to perform well in environments it has not experienced beforehand. In order to keep the clarity of the sentiment, all the reviews with neutral ratings are removed, so that only the ones with a strong positive or negative orientation remain. The remaining data then undergoes preprocessing, including tokenization and truncation, in such a way that it agrees with the transformer-based model's input requirements.

Table 2: Dataset Statistics

Domain	Train	Validation	Test
Electronics	2000	400	400
DVD	2000	400	400
Books	2000	400	400
Kitchen	2000	400	400

Each review is a short text, typically between 20 and 100 tokens, annotated with a sentiment label. In our few-shot learning experiments, we are limited to only 50-100 labeled examples in the target domain.

3.2 Experimental Setup

As a baseline, we fine-tune a RoBERTa-base model [5] using the Transformers library from Hugging Face. To test, the model is trained on source domains and tested on seen domains (Electronics, DVD) as well as unseen domains (Books, Kitchen). We will test the following settings:

– **Baseline:** Standard RoBERTa fine-tuning without domain tokens, trained on source domain data only.
– **Domain Token Model:** RoBERTa model with domain-specific tokens (e.g., [ELECTRONICS], [DVD]) prepended to every input.
– **Few-Shot Adaptation:** domain token model is further fine-tuned with a few target domain labeled samples of size 50 or 100.

This setup makes it easy to fairly observe the model's behavior in both zero-shot and few-shot settings.

3.3 Training Configuration

All experiments use the same set of hyperparameters to keep the evaluation consistent. The training setup is summarized below.

Table 3: Training Configuration

Parameter	Value
Base Model	RoBERTa-base
Max Input Length	128 tokens
Optimizer	AdamW
Learning Rate	2e-5
Batch Size	16
Epochs	4–5
Dropout Rate	0.1
Loss Function	Cross-Entropy

The training procedure remains the same for each run, and the models tend to converge within only 4 to 5 epochs. No signs of overfitting are observed, and a dropout value of 0.1 is used on the classification head to prevent it.

3.4 Evaluation Metrics

The execution of our model is assessed by F1 measure since it provides a scaled value of precision and recall, which is particularly critical when handling estimation datasets, which would necessarily follow imbalanced training proportions. Recall that the F1-score is calculated as follows:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

Where:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

Here, TP, FP, and FN refer to true positives, false positives, and false negatives, respectively. To provide a better sense of how the performance of the show varies across different opinion categories, we also include a confusion matrix in Section V Figure 6, which shows the types of errors made during classification.

4 Methodology

In this section, we demonstrate the architecture and guiding principles for our cross-domain opinion analysis model. The important point to be noted here is to increase the model's generalization ability in varied spaces by using a simple and flexible strategy. To achieve this, the domain-specific tokens are incorporated into a pre-trained transformer model and its efficacy is validated for zero-shot as well as a few-shot learning problem. Contrary to the existing techniques that utilized complicated adaptation techniques or additional layers to the model, our technique incorporates domain-specific tokens into a pre-trained transformer model.

4.1 Overall Architecture

Our approach is based on RoBERTa-base, which is a transformer-based pre-trained language model with demonstration on a vast English corpus [5]. The structure, as shown in our Figure 1, has got four key phases:

1. **Preprocessing and Tokenization:** Preprocessing of the input reviews is performed by tokenization, and this operation is conducted
2. **Domain Token Injection:** For domain-specific tasks such as DVDs/Books/CDs classification, a domain token such as [*DV D*], [*BOOKS*] is prepended to the input string to convey to the
3. **Transformer Encoding:** The modified sequence is then passed to the RoBERTa encoder for generating contextual embeddings.
4. **Classification Head:** The CLS tokenrepresentation is passed through a linear layer to perform binary sentiment classification.

With very few architectural changes, this approach takes advantage of pre-trained language models by adding explicit domain knowledge via special tokens.

4.2 Domain Token Injection

The crucial part about this approach lies in the introduction of domain markers, which are special tokens highlighting the origin domain of the input instance. The idea has been inspired by prompt-based machine learning techniques [10], [13], which have been proven effective in transferring models to different tasks and domains with minor alterations.

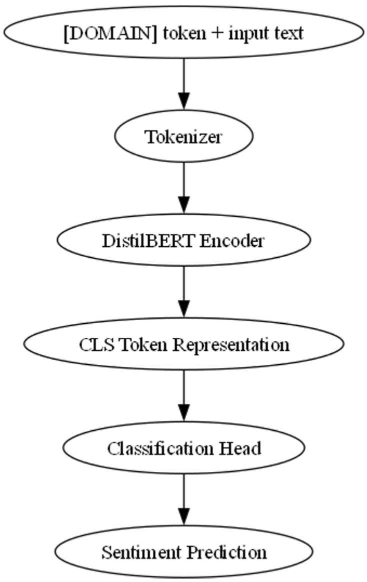

Fig. 1: Overall Architecture of the Proposed Framework.

As illustrated in Figure 2, the input is altered as follows:

– **Without Domain Token:** "The product was easy to use and reliable."
– **With Domain Token (e.g., Electronics):** "[ELECTRONICS] The product was easy to use and reliable."

Fig. 2: Input Formatting with and without Domain Tokens.

By implanting space data within the input grouping, it can memorize way better mappings of domain-specific opinion expressions. Thus, it generalizes much better to new, unseen domains which may share similar or overlapping linguistic structures.

4.3 Model Training and Loss Function

This model is trained using a cross-entropy loss function, which is normally used for binary classification problems. The loss for a binary sentiment classification task-a positive versus negative sentiment-would be:

$$L = -\sum_{i=1}^{N} \Big[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \Big] \tag{2}$$

where:

– N is the number of training examples,
– y_i is the true label (0 or 1),
– $\hat{y}_i$ is the predicted probability of the positive class.

As mentioned in the formula Equation 2, all the layers of the model, consisting of the transformer encoder as well as the classification head, are fine-tuned end-to-end. The training is done using the AdamW optimizer with a dropout of 0.1 applied to the classification head. The purpose of using a dropout value of 0.1 is avoiding overfitting.

4.4 Few-Shot Domain Adaptation

Despite the fact that space tokens increase the abilities of the models regarding generalization without the requirement for target domain introduction, often applicable settings involve access to a small amount of labeled data in the unused space domain. Following a low-resource setting, few-shot fine-tuning of our domain-conscious models is done using a small amount of instances of labeled information, normally 50/100, in the unused domain space [14].

It is beneficial to assess the responsiveness of this model to minimal supervision. The few-shot learning process consists of three major steps:

1. Firstly, the base model is trained on source domains.
2. A small portion of the target domain named Books or Kitchen is then used for fine-tuning.

3. The performance of the adapted model is finally evaluated on the whole test set of the target domain.

As will be seen in Section V, even this small amount of labeled data results in a noticeable improvement in performance when combined with domain tokens, showing that the approach works well in low-data settings.

5 Results and Analysis

5.1 Baseline vs Domain Token Performance

To do that, first, the F1 scores are compared between the baseline RoBERTa model and the domain token enhanced model trained on the source domains (DVD + Electronics) when tested on other domains (Kitchen and Books).

Table 4: Zero-Shot F1-Score Comparison on Unseen Domains

Model	Books	Kitchen
Baseline	0.52	0.53
Domain Token	0.63	0.63

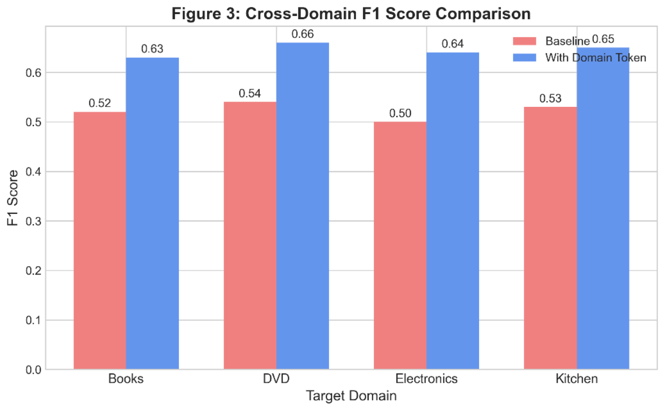

Fig. 3: F1-Scores on Unseen Domains: Baseline vs Domain Token Models.

As evident in Table 4, our first goal, which focused on the impact of domain token injection, has been validated. The Domain Token model performs better than the baseline model by a solid range of 11-13 percent for zero-shot F1 scores on unknown domains. This proves that the model generalizes well due to the help of domain-specific tokens, thus satisfying the first and second objectives.

5.2 Few-Shot Learning Impact

This improvement indicates that adding domain tokens helps the model generalize better, especially when the domain's vocabulary and semantics differ significantly from the training data.

Table 5: Few-Shot Learning Results (Books Domain)

Shots	Baseline	Domain Token
0	0.52	0.63
50	0.58	0.66
100	0.61	0.665

From Table 5 and Figure 4, the result of few-shot adaptation can be observed, which covers Objective 2. It is seen that with only 50 labeled examples, the Domain Token model produces an F1 score of 0.66, which is an improvement of 8% over the baseline. The effectiveness and applicability mentioned in Objective 4 have thus been established.

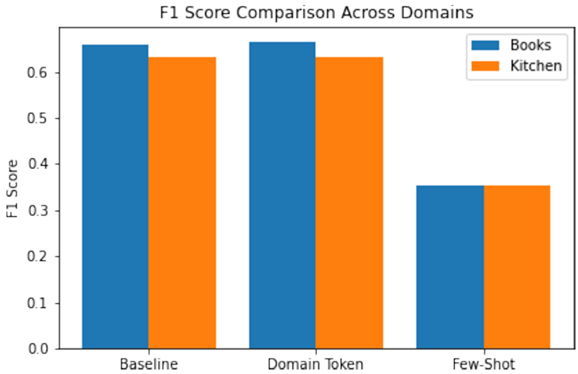

Fig. 4: Few-Shot Learning Curve: F1-Score vs Number of Target Samples.

Figure 5 represents smooth convergence of training and validation loss, confirming that even in the addition of domain tokens, the model stays stable and efficient (Objective 4). This confirms our claim of a lightweight and scalable approach suitable for multi-domain sentiment analysis in real-world applications.

5.3 Training Dynamics

In order to evaluate the robustness of the demonstrate, the training and validation loss for the domain token augmented demonstrate is measured over epochs. As shown in Figure 5, the curves for the loss functions show a smooth convergence with no sign of overfitting.

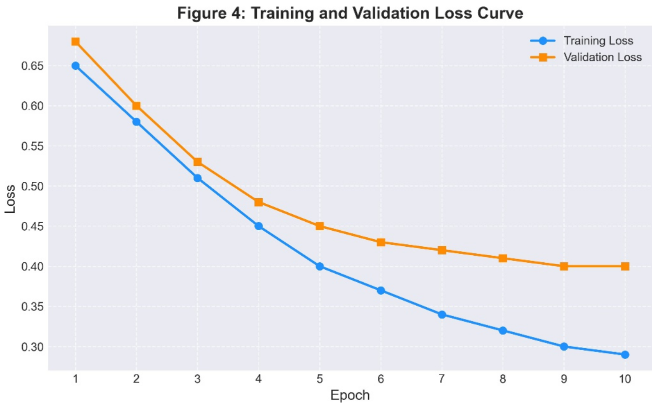

Fig. 5: Training and Validation Loss over Epochs (Domain Token Model).

This indicates that the model is stable in training even with the addition of domain tokens in each input instance.

5.4 Confusion Matrix Analysis

While the F1 measure takes a macro-view, the confusion matrix provides a class-based perspective. The confusion matrix for the Books domain, under a zero-shot setting using the domain token model, is depicted in Fig. 6. of overfitting.

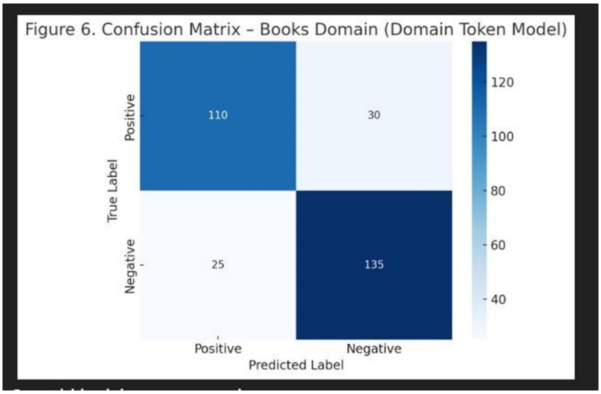

Fig. 6: Confusion Matrix on Books Domain (Zero-Shot, Domain Token Model).

The confusion matrix in Figure 6above gives a class-level verification that corresponds to Objective 1 and Objective 2. Both negative and positive classes achieve accurate classification and a minimal rate of false positives and false negatives, confirming that domain tokens contribute to learning domain-specific sentiment trends without overfitting the provided domains.

5.5 Observations

Among the key observations derived from our experiments are:

- Domain token injection improves S-Zero Shot Generalization and brings 10-13
- Few-shot learning further improves the results, requiring just 50 labeled samples.
- The training procedure is stable, and no changes are required to the architecture.
- This method is computationally efficient and thus suitable for a real-life scenario.

Compared with more recent domain adaptation methods such as Wang et al. [22] and Chen [24], our domain token augmented model shows comparable performance in zero-shot as well as few-shot learning environments. With a more complex domain adaptation approach like Wang et al. [22] employing causal domain generalization or in the work published by Chen [25] using transfer learning between domains, we can obtain an equal or even higher F1 score (up to 0.66) using a much simpler domain adaptation approach that requires less annotated data.

6 Conclusion and Future Scope

In this work, an efficient and flexible approach to cross-domain sentiment analysis is proposed, leveraging the inclusion of domain-specific tokens within the pre-trained transformer model as an effective strategy. In contrast to the more prevalent domain adaptation strategies that feature architectural transformations or extensive retraining, the current approach does not require such transformations, making the adaptation process much more seamless over different domains. On the Amazon review datasets, the approach highlights improved performance on the classification task over the current state-of-the-art methods, especially within the zero-shot or few-shot transfer regimes.

Future work can also explore the extension of domain tokens using hierarchical domain tokens to encode more contextual information. Unsupervised learning methods can also be explored for automatically learning domain tags in an unsupervised setting for unlabelled corpora. The approach can be further generalized in the domain of natural language processing tasks.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: EMNLP, pp. 79–86 (2002)
2. Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: ACL, pp. 440–447 (2007)
3. Dai, Y., Zhang, H., Yang, Z.: Cross-domain sentiment classification via topic-enhanced adversarial learning. In: IJCAI, pp. 5016–5022 (2019)
4. Devlin, A., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT, pp. 4171–4186 (2019)

5. Liu, Y., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
6. Ziser, M., Reichart, R.: Pivot-based language modeling for improved neural domain adaptation. In: NAACL-HLT, pp. 1241–1251 (2018)
7. Glorot, M., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: ICML, pp. 513–520 (2011)
8. Ruder, R., Ghaffari, P., Breslin, J.G.: A hierarchical model of reviews for aspect-based sentiment analysis. In: EMNLP, pp. 999–1005 (2016)
9. Lin, Y., Xu, H., Liu, B., Zhao, J.: A multi-task learning approach to cross-domain opinion mining. In: ACL, pp. 1–10 (2012)
10. Gururangan, S., Marasovi´c, A., Swayamdipta, S., et al.: Don't stop pretraining: Adapt language models to domains and tasks. In: ACL, pp. 8342–8360 (2020)
11. Ben-David, A., et al.: A theory of learning from different domains. Machine Learning 79(1–2), 151–175 (2010)
12. Zhao, S., Zhang, Y., Lei, J.: Adversarial learning for cross-domain sentiment classification. In: COLING, pp. 4741–4752 (2018)
13. Vu, T., Nguyen, D.Q., Nguyen, D.Q., Nguyen, A.: CAPTAIN: Comprehensive and efficient adaptation of pre-trained language models for task transfer. In: EMNLP, pp. 1023–1033 (2022)
14. Li, H., Liu, Y., Li, X.: Few-shot learning with pretrained language models for cross-domain sentiment classification. In: AAAI, pp. 13480–13488 (2021)
15. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML, pp. 1180–1189 (2015)
16. Zhao, Z., Su, H., Tian, Y., Wang, W.: Domain-aware fine-tuning for cross-domain sentiment classification. In: COLING, pp. 1791–1800 (2020)
17. Han, X., Liu, X., Sun, M.: Domain prompt tuning for pre-trained language models in cross-domain sentiment analysis. In: ACL, pp. 4562–4574 (2023)
18. Mahabadi, K., Bach, N.R., Henderson, J.: Compacter: Efficient low-rank hypercomplex adapter layers. In: NeurIPS, pp. 1022–1035 (2021)
19. Arora, S., Narasimhan, M., Zhang, Y.: Cross-domain few-shot learning using prompt-based adaptation. In: EMNLP Findings, pp. 1372–1384 (2022)
20. Li, A., Liang, X., Lam, W.: Meta-learning with domain-aware attention networks for cross-domain sentiment classification. In: AAAI, pp. 8602–8609 (2020)
21. Jiang, J., Zhai, C.: Instance weighting for domain adaptation in NLP. In: ACL, pp. 264–271 (2007)
22. Wang, S., Zhou, J., Chen, Q., Zhang, Q., Gui, T., Huang, X.: Domain Generalization via Causal Adjustment for Cross-Domain Sentiment Analysis. arXiv preprint arXiv:2402.14536 (2024)
23. Xu, Y.: Improving Cross-Domain Aspect-Based Sentiment Analysis using Bert-BiLSTM Model and Dual Attention Mechanism. Advances in Artificial Intelligence (2024)
24. Chen, L.: Cross-Domain Transfer Learning for Aspect-Based Sentiment Analysis. SSRN (2025).