CDA^2: Counterfactual Diffusion Augmentation for Cross-Domain Adaptation in Low-Resource Sentiment Analysis

Dancheng Xin[1†], Kaiqi Zhao[2‡], Jingyun Sun[1†] and Yang Li[1†*]
[1] Northeast Forestry University, China
[2] University of Auckland, New Zealand
[†]{dcxin, sunjingyun, yli}@nefu.edu.cn
[‡]kaiqi.zhao@auckland.ac.nz

Abstract

Domain adaptation is widely employed in cross-domain sentiment analysis, enabling the transfer of models from label-rich source domains to target domains with fewer or no labels. However, concerns have been raised about their robustness and sensitivity to distribution shifts, particularly when significant disparities exist between domains. To address this problem, we propose CDA^2, a framework for **c**ross-**d**omain **a**daptation in low-resource sentiment analysis that leverages **c**ounterfactual **d**iffusion **a**ugmentation. Specifically, it employs samples derived from domain-relevant word substitutions in source domain samples to guide the diffusion model for generating high-quality counterfactual target domain samples. During the training stage, we employ a soft absorbing state and MMD loss, while using an advanced ODE solver to accelerate the sampling process. Our experiments demonstrate that CDA^2 generates high-quality target samples and achieves state-of-the-art performance in cross-domain sentiment analysis.

1 Introduction

Sentiment analysis is a crucial task in Natural Language Processing (NLP), primarily focuses on extracting the underlying emotion or sentiment expressed within textual data. It has surged in popularity in recent years, due to its wide-ranging applications in the real-world(Kertkeidkachorn and Shirai, 2023; Nzeyimana, 2023). In recent years, deep learning technology has experienced significant growth and achieved remarkable success in sentiment analysis(Zhang et al., 2015; Yadav and Vishwakarma, 2020). However, when operating under low-resource conditions or encountering a data distribution shift between the training domain and the target domain, traditional sentiment analysis methods that rely on labeled data to train models in the target domain experience a significant decline in performance(Ben-David et al., 2020).

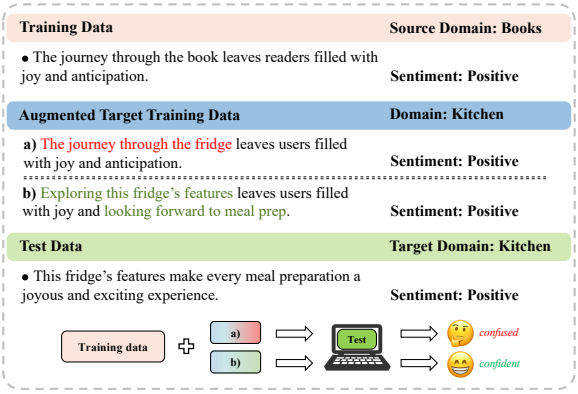

Figure 1: An Illustration of the Cross-Domain SA Task. If the augmented training data exhibit semantic disruptions and spurious associations with the source domain, the model will become confused due to the failure of semantic transfer.

To alleviate the reliance on labeled data, cross-domain sentiment analysis(SA) has garnered the attention from researchers. Many previous works resort to unsupervised domain adaptation techniques, which aim to transfer knowledge from a resource-rich source domain to a target domain with unlabeled data(Blitzer et al., 2007; Pan et al., 2010; Zhuang et al., 2015). In cross-domain sentiment analysis tasks, most existing domain adaptation methods employ adversarial training to prevent models from distinguishing samples from specific domains, thereby transferring knowledge from the source domain to the target domain(Liu et al., 2018; Wang et al., 2019) and some attempts to learn domain-specific knowledge(Du et al., 2020; Qu et al., 2019; Yang et al., 2022). Although these methods achieve promising results, their models are trained only on in-domain labeled data from the source domain, thereby limiting their ability to handle out-of-domain data.

*Corresponding Author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 61–72
January 19–24, 2025.

To address the aforementioned limitations, researchers have attempted to design cross-domain data augmentation methods. The key objective is to generate a large number of labeled target domain samples based on the labeled source domain samples, thereby achieving knowledge transfer. The research within this framework primarily includes two approaches: masked language models (MLM)(Yang et al., 2022) and sequence-to-sequence (Seq2Seq)(Li et al., 2022a) models. While word substitution-based data augmentation methods have demonstrated advancement over feature adaptation methods, they still have some drawbacks: (i) semantic disruptions, (ii) the fixed syntactic structure from the source domain, (iii) the lack of diversity in generated samples.

Taking the cross-domain sentiment analysis(SA) task in Figure 1 as an example, Training models with logically inconsistent augmented data can lead to confusion, especially in context-aware language models. Conversely, incorporating the augmented target domain data can enhance the reliability of the predictive model.

To generate high-quality labeled target domain data for cross-domain sentiment analysis, we propose a framework called CDA^2 for **C**ross-**D**omain **A**daptation in low-resource sentiment analysis, which utilizes **C**ounterfactual **D**iffusion **A**ugmentation. CDA^2 is designed to mitigate semantic disruptions and spurious associations caused by fixed syntactic structures from the source domain. Firstly, we provide the diffusion generator with high-quality raw target samples through domain corruption and domain reconstruction. Next, we design a learnable soft absorbing state by introducing additional discrete noise into the continuous diffusion process to better fit the inherently discrete nature of text. Additionally, we incorporate Maximum Mean Discrepancy loss, utilizing real target domain unlabeled samples to supervise the generation process, thereby facilitating better data distribution shift. During the sampling phase, we employ an advanced Ordinary Differential Equation solver to accelerate sampling while minimizing the sacrifice of sample quality, resulting in the generation of high-quality counterfactual target samples.

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

- We propose a novel diffusion-based cross-domain data augmentation framework, CDA^2, which can generate a large amount of labeled target domain data for cross-domain sentiment analysis tasks.
- Within this framework, we conditionally guide the diffusion model to generate high-quality counterfactual target samples from source samples and raw target samples.
- We conduct experiments on various sentiment analysis datasets , demonstrating that our model achieves state-of-the-art performance.

2 Related Work

Cross-Domain Sentiment Analysis Cross-domain sentiment analysis aims to generalize models trained on a source domain to a target domain. Typically, the source domain has abundant labeled data, while the target domain has scarce or no labeled data(Du et al., 2020). Researchers address this by bridging data distribution differences through shared feature representations(Ziser and Reichart, 2017; Ben-David et al., 2020; Peng et al., 2018) and learning invariant features via adversarial training(Ganin et al., 2017; Du et al., 2020; Li et al., 2017) and contrastive learning(Long et al., 2022). Influenced by masked generation methods, recent works have explored data augmentation(Calderon et al., 2022; Wang and Wan, 2023) and prompt tuning(Wu and Shi, 2022).

Domain Adaptation Unsupervised adaptation is a practical setup that assumes access to unlabeled data from both domains and labeled data from the source domain(Blitzer et al., 2007). A more challenging setup, Any Domain Adaptation(Ben-David et al., 2020), assumes the target domain is unseen during training. Methods include representation learning(Ziser and Reichart, 2017), instance reweighting, and self-training(Rotman and Reichart, 2019). Deep neural networks have focused on the two approaches mentioned in cross-domain sentiment analysis.

Data Augmentation Data augmentation aims to improve model generalization by generating more training data. Synonym-based augmentation methods replace words with synonyms, hypernyms, or hyponyms(Xu et al., 2019; Kobayashi, 2018), but these methods can create spurious associations. To address this, Kaushik et al. (2020) introduced minimal modifications using human annotators for label inversion, though costly and time-consuming. Chen et al. (2021) used automated antonym replacement. Recently, diffusion models have been

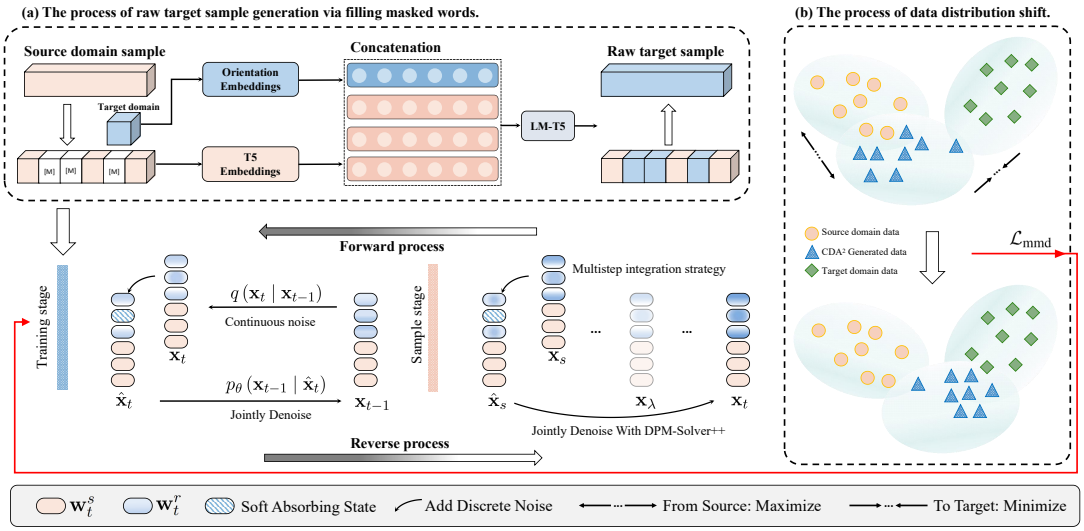

Figure 2: The architecture of counterfactual diffusion augmentation (CDA^2) framework for cross-domain adaptation.

applied for controlled text generation(Li et al., 2022b; Gong et al., 2023a,b), offering stable training and diverse content generation compared to GANs(Goodfellow et al., 2014). Our goal is to use diffusion models to generate high-quality target samples guided by raw target samples, rather than through manual or rule-based efforts.

3 Methodology

In this section, we first define the task of cross-domain sentiment analysis. Subsequently, we present the proposed counterfactual diffusion augmentation framework for cross-domain adaptation (CDA^2 for short). The overall structure of CDA^2 is shown in Figure 2, which comprises three parts: (i) generation of raw target samples, (ii) diffusion-based generator (including training stage and sample stage), and (iii) data filtering mechanism.

3.1 Problem Formulation

In this paper, we focus on cross-domain sentiment classification in low-resource scenarios. Following previous studies(Zhang et al., 2019; Li et al., 2018), we consider two domains: Source and Target. The source domain $\mathcal{D}^s$ contains labeled data $\mathcal{D}_l^s = \{(\mathbf{w}_i^s, y_i^s)\}_{i=1}^{N_l^s}$ and unlabeled data $\mathcal{D}_u^s = \{(\mathbf{w}_i^s)\}_{i=N_l^s+1}^{N_u^s}$, where $\mathcal{D}^s = \mathcal{D}_l^s \bigcup \mathcal{D}_u^s$. Additionally, $N_l^s \ll N_u^s$. The target domain $\mathcal{D}^t$ includes a set of unlabeled data $\mathcal{D}_u^t = \left\{\left(\mathbf{w}_j^t\right)\right\}_{j=1}^{N_u^t}$, where $\mathcal{D}^t = \mathcal{D}_u^t$. The goal of cross-domain sentiment classification is to utilize $\mathcal{D}^s$ and $\mathcal{D}^t$ to predict the labels of test samples from the target domain.

3.2 Generation of Raw Target Samples

To meet the requirements for conditional guidance of the diffusion model, we aim to generate raw target samples that are contextually relevant and sentimentally aligned. We adopt a strategy of corruption and reconstruction on given source domain samples through a masking generation approach, as illustrated in Figure 2a.

Domain Corruption The first step in generating raw target samples $\mathbf{w}^r$ is to mask specific domain-relevant terms from the source domain $\mathcal{D}^s$. Let $\mathbf{w} = \{w_1, w_2, \ldots, w_m\}$ represent a sample, with m denoting the sample length. We mask all uni-grams w for which $\mathrm{M}(w, \mathcal{D}^s, \mathcal{D}^t) > \tau$, with τ being a masking threshold parameter and M representing a function that returns the masking score of a the uni-gram. For bi-grams, we mask those terms that have an overall score exceeding τ, provided that none of their constituent uni-grams have been masked. Similarly, this strategy can be extended to tri-grams. For example, "*paper*" and "*towel*" as uni-grams have weak relevance to the Kitchen domain and are not masked. However, the bi-gram "*paper towel*" has high relevance to the Kitchen domain as a combined term and a score above the τ threshold, so it is masked. This method provides more contextual information and proves our strategy effective in identifying domain-specific terms.

The rationale behind this higher-order n-gram masking approach is to capture the context more accurately. Higher-order terms like bi-grams and tri-

grams provide richer contextual information compared to uni-grams. By masking bi-grams and tri-grams, we ensure that domain-specific phrases are identified, while still allowing the individual words to be used in other contexts where they may not be as relevant. This approach prevents the loss of useful words that might be masked unnecessarily if only higher-dimensional terms were considered.

To clarify the masking score $\mathrm{M}(\cdot)$, we assume equal prior probabilities for each domain and utilizing the Bayes' rule, the probability that an n-gram term w belongs to a domain $\mathcal{D}$ with $n^{\mathcal{D}}$ unlabeled samples is estimated by:

$$P(D = \mathcal{D} \mid W = w) \propto \frac{n_w^{\mathcal{D}} + \alpha}{n^{\mathcal{D}} + \alpha \cdot V} \quad (1)$$

where $n_w^{\mathcal{D}}$ represents the number of samples in $\mathcal{D}$ that include the term w, α is a smoothing hyper-parameter and V represents the total number of unique terms. To effectively identify domain-specific terms, we need a measure that captures both the likelihood of a term belonging to a domain and its specificity to that domain. Therefore, we define the association between w and $\mathcal{D}$ as:

$$\rho(w, \mathcal{D}) = P(\mathcal{D} \mid w) \cdot \left(1 - \frac{H(D \mid w)}{\log N}\right) \quad (2)$$

where N is the number of unlabeled domains, and $\log N$ is the upper bound of the entropy $H(D|w)$. Higher entropy values indicate that the term w is not particularly related to any specific domain. Based on the above, we derive the masking scores for n-gram terms under the source domain $\mathcal{D}^s$ and the target domain $\mathcal{D}^t$.

$$\mathrm{M}\left(w, \mathcal{D}^s, \mathcal{D}^t\right) = \rho(w, \mathcal{D}^s) - \rho\left(w, \mathcal{D}^t\right) \quad (3)$$

where the masking scores $\mathrm{M}(\cdot)$ range from -1 to 1. $\mathrm{M}(\cdot)$ can take negative values to prevent the inadvertent masking of n-grams that should be included in the raw target samples.

Domain Reconstruction The second step in generating raw target samples $\mathbf{w}^r$ involves predicting the masked source domain data using information from the target domain. To incorporate target domain information, we introduce an orientation vector $\mathbf{v}^t$ that encodes the target domain's features. We utilize a T5 (Raffel et al., 2020) generation model based on an encoder-decoder architecture. Given a masked sample of $\mathbf{w}^r$, denoted as $\mathrm{M}(\mathbf{w}^r)$, and a target domain $\mathcal{D}^t$, we concatenate the domain orientation vector $\mathbf{v}^t$ representing $\mathcal{D}^t$ with the embedding vector $\mathbf{v}^r$ of $\mathrm{M}(\mathbf{w}^r)$ along the feature dimension. Then, this concatenated matrix is fed into T5 to generate $\mathbf{w}^r$.

Specifically, we equip the model with a learnable embedding matrix that contains $K \cdot N$ orientation vectors, allowing each domain to be represented by a K different vectors. We initialize the orientation vectors using the embedding vectors of the domain names and the top $K-1$ representative words. For each domain $\mathcal{D}$, representative words are selected based on $\log(n_w^{\mathcal{D}}+1)\rho(w, \mathcal{D})$. Based on the above, we obtain multiple raw target samples $\mathbf{w}^r$ for the specified target domain $\mathcal{D}^t$, each corresponding to a single source domain sample $\mathbf{w}^s$ and sharing the same label. These samples are used to conditionally guide the diffusion model. It is noteworthy that these initialized orientation vectors gradually converge to different effective values over the course of training, according to the requirements of this work.

3.3 Diffusion based Generator

To address the semantic disruptions and spurious associations that arise from the fixed syntactic structure of the source domain. We train a diffusion generator using the raw target sample $\mathbf{w}^r$ generated in Section 3.2. to produce additional high-quality counterfactual target samples $\mathbf{w}^c \in \mathcal{D}^c$. Inspired by Gong et al.(Gong et al., 2023a) and Lu et al.(Lu et al., 2022, 2023), we will detail the diffusion generation process used in this study in the following discussion.

Preliminaries Diffusion models(Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Li et al., 2022b; Gong et al., 2023a) are a type of latent variable model initially designed for continuous domains. These models comprise two processes: a forward diffusion process and a reverse diffusion process. In the forward process, given a sample $\mathbf{x}_0$ drawn from $q(\mathbf{x}_0)$, a Markov chain of latent variables $\mathbf{x_1} \ldots \mathbf{x_T}$ is generated by progressively adding Gaussian noise:

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \quad (4)$$

where β_t is a noise schedule controlling the noise addition step size. Eventually, $\mathbf{x}_T$ approximates an isotropic Gaussian distribution. If β_t is sufficiently small, the reverse process $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ also follows a Gaussian distribution and can be modeled

by:

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, t\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; \mu_\theta\left(\mathbf{x}_t, t\right), \Sigma_\theta\left(\mathbf{x}_t, t\right)\right) \tag{5}$$

where $\mu_\theta(\cdot)$ and $\Sigma_\theta(\cdot)$ can be implemented using a U-Net or a Transformer. By conditioning on $\mathbf{x}_0$, $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)$ has a closed form, allowing the variational lower bound to be minimized to optimize $\log p_\theta(\mathbf{x}_0)$.

When compared to traditional generative models, such as Generative Adversarial Networks (GANs)(Goodfellow et al., 2014), diffusion models have emerged as a novel paradigm for generative models. They come with several potential advantages, particularly in the generation of high-quality text and images. However, most current diffusion works face challenges during training convergence and generation speed, particularly given that these models require the use of a Minimum Bayes Risk(MBR) strategy(Koehn, 2004) for decoding and generation, resulting in significant computational overhead during training. Additionally, in domain adaptation, there are concerns about the quality of generated target domain samples in low-resource settings, especially due to failures in data distribution shift.

Training Stage To ensure the quality of the generated samples, we introduce a Soft Absorbing State(SAS) and Maximum Mean Discrepancy(MMD) loss during the training stage, which facilitates the diffusion model's ability to learn to reconstruct discrete mutations based on the underlying Gaussian space, thereby enhancing its capacity to recover conditional signals. At the same time, under the supervision of real target domain data $\mathcal{D}^t$, the MMD loss can promote the transition of generated samples $\mathbf{w}^c$ from the source domain $\mathcal{D}^s$ to the target domain $\mathcal{D}^t$, as shown in Figure 2(b).

Let $\mathbf{x}$ represent the latent representations of the data from the source domain ($\mathbf{w}^s$). At the initial step of the forward noise-adding process, we follow the Diffusion-LM proposed by Li et al. (2022b) to map the discrete sample $\mathbf{w}^s$ into a continuous space. Specifically, we concatenate the source domain sample $\mathbf{w}^s$ and raw target sample $\mathbf{w}^r$ to embed them into a continuous feature space, denoted as $\operatorname{Emb}\left(\mathbf{w}^{s \oplus r}\right)$.

$$q_\phi\left(\mathbf{x}_0 \mid \mathbf{w}^{s \oplus r}\right)=\mathcal{N}\left(\operatorname{Emb}\left(\mathbf{w}^{s \oplus r}\right), \beta_0 \mathbf{I}\right) \tag{6}$$

where $\mathbf{I}$ is an identity matrix. As shown in Eq. (4), the structure of the perturbed data $\mathbf{x}_0$ during the forward noising process is detailed. From this, we can derive the latent variable $\mathbf{x}_t$ as follow:

$$\mathbf{x}_t=\sqrt{\bar{\alpha}_t} \mathbf{x}_0+\sqrt{1-\bar{\alpha}_t} \epsilon \tag{7}$$

where ϵ is defined at each time step with $\alpha_t = 1-\beta_t$,$\bar{\alpha}_t=\Pi_{i=1}^t \alpha_i$. Additionally, $\mathbf{x}_t=\mathbf{w}_t^s \oplus \mathbf{w}_t^r$, with $\mathbf{w}_t^s$ and $\mathbf{w}_t^r$ representing the latent states of $\mathbf{w}^s$ and $\mathbf{w}^r$, respectively. During this process, we replace the i-th token in $\mathbf{x}_t$ with the the soft absorbing state $\mathbf{n}$ with a certain probability. The SAS $\mathbf{n}$ has the same dimension as the word embeddings and is learnable during the diffusion process.

$$\hat{\mathbf{x}}_t^i=\begin{cases}\mathbf{n} & \text{if } \eta=1 \\ \mathbf{x}_t^i & \text{if } \eta=0\end{cases} \tag{8}$$

where $\eta=\operatorname{Bernoulli}(\beta_t * \gamma)$, and γ is the [MASK] ratio when $t=T$. The introduction of the SAS enhances the model's ability to handle discrete data during continuous diffusion. Simultaneously, it provides a soft constraint in the high-dimensional feature space, which enhances the stability and reliability of the model. Also, in contrast to conventional diffusion models, which perturb $\mathbf{x}_t$ in its entirety, we introduce partial noise solely to $\mathbf{w}_t^r$, by replacing $\mathbf{w}_t^s$ with $\mathbf{w}_0^s$. This is a crucial aspect for enabling the diffusion model to conduct conditional language modeling.

In the reverse process, the objective is to recover the initial $\mathbf{x}_0$ from the partially Gaussian-noised $\hat{\mathbf{x}}_T$ by jointly denoising both continuous and discrete noise, as shown in Eq. (5). Thereby, we compute the variational lower bound following the diffusion process:

$$\begin{aligned}\mathcal{L}_{\mathrm{vlb}} & =\mathbb{E}_q\left[D_{\mathrm{KL}}\left(q\left(\hat{\mathbf{x}}_T \mid \mathbf{x}_0\right) \| p_\theta\left(\hat{\mathbf{x}}_T\right)\right)\right. \\ & +\sum_{t=2}^T D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1} \mid \hat{\mathbf{x}}_t, \mathbf{x}_0\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \hat{\mathbf{x}}_t, t\right)\right) \\ & +D_{\mathrm{KL}}\left(q_\phi\left(\mathbf{x}_0 \mid \mathbf{w}^{s \oplus r}\right) \| p_\theta\left(\mathbf{x}_0 \mid \hat{\mathbf{x}}_1\right)\right) \\ & \left.-\log p_\theta\left(\mathbf{w}^{s \oplus r} \mid \mathbf{x}_0\right)\right]\end{aligned} \tag{9}$$

To ensure the transition of the data distribution from counterfactual target samples $\mathbf{w}^c$ in $\mathcal{D}^c$ to real target domain samples $\mathbf{w}^t$ in $\mathcal{D}^t$, we propose a sentence-level MMD loss as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{mmd}} & =\mathrm{d}_k^2\left(\mathcal{D}^c, \mathcal{D}^t\right)=\frac{1}{\left(N^c\right)^2} \sum_{i, j}^{N^c} k\left(\mathbf{w}_i^c, \mathbf{w}_j^c\right)+ \\ & \frac{1}{\left(N^t\right)^2} \sum_{i, j}^{N^t} k\left(\mathbf{w}_i^t, \mathbf{w}_j^t\right)-\frac{2}{N^c N^t} \sum_i^{N^c} \sum_j^{N^t} k\left(\mathbf{w}_i^c, \mathbf{w}_j^t\right)\end{aligned} \tag{10}$$

where N^c and N^t represent the number of samples in each domain, respectively, and $k(\cdot)$ denotes a Gaussian kernel function. When the MMD loss is minimized, the distribution of $\mathcal{D}^c$ approaches that of $\mathcal{D}^t$, thereby improving the quality of the generated samples.

In conclusion, we derive the overall objective function by summing up the two components:

$$\mathcal{L} = \mathcal{L}_{\text{vlb}} + \varphi \mathcal{L}_{\text{mmd}} \tag{11}$$

where φ is a weight parameter that starts at zero and gradually increases during model training to ensure a balance between reconstruction ability and distribution shift capability throughout the training process.

Sample Stage Previously, diffusion models employed clamping operations during the sampling phase to predict vectors and reduce rounding errors. However, the discrepancy between training and sampling(Tang et al., 2023) can lead to the accumulation of prediction errors and a reduction in sampling speed.

To improve the sampling speed of the diffusion model, we employ the advanced DPM-Solver++(Lu et al., 2023) as a sampling accelerator in the continuous space during the sample stage. This accelerator does not require MBR decoding during the sampling process, thereby saving a substantial amount of time. Importantly, it enhances the sampling speed while also ensuring the quality of the generated samples.

Specifically, as described in Eq. (8), discrete noise is added to the continuous Gaussian noise, which bridges training and inference in the discrete space. Utilizing the precise solution of the diffusion ODEs proposed by DPM-Solver++, given an initial value $\mathbf{x}_s$ at time $s > 0$, the solution $\mathbf{x}_t$ at time $t \in [0, s]$:

$$\mathbf{x}_t = \frac{\sigma_t}{\sigma_s}\mathbf{x}_s + \sigma_t \int_{\lambda_s}^{\lambda_t} e^{\lambda} f_\theta\left(\hat{\mathbf{x}}_\lambda, \lambda\right) d\lambda \tag{12}$$

where the λ_t is a strictly decreasing function of t with an inverse function $t_\lambda(\cdot)$. The term σ_t is monotonic with respect to β_t, and f_θ serves as the data prediction model that recover the corrupt data $\mathbf{x}_t$ to $\mathbf{x}_0$.

Furthermore, Eq. (12) requires an approximation of $\int e^{\lambda} f_\theta d\lambda$. The integral can be analytically computed by repeatedly applying integration by parts n times, and we can approximate only the first few terms while discarding higher-order error terms. In our experiments, we use the second order. After discrete denoising in our method, this algorithm remains applicable since our $f_\theta(\hat{\mathbf{x}}_\lambda, \lambda)$ aligns with the training objectives. Based on the above, we train a classifier using the source domain dataset $\mathcal{D}^s$ and the corresponding generated counterfactual target domain dataset $\mathcal{D}^c$, where the sample labels in $\mathcal{D}^c$ are consistent with those in $\mathcal{D}^s$ and $\mathcal{D}^r$ due to the paired correspondence, to predict the labels of test samples from the target domain.

3.4 Data Filtering Mechanism

Since the counterfactual target samples are generated based on the raw target samples' corresponding labels and domains, the generation process may introduce uncertainties and inconsistencies. To better utilize the counterfactual target domain data $\mathbf{w}^c$ in cross-domain SA tasks, we introduce a data filtering mechanism that eliminates noisy data. Specifically, our filtering mechanism consists of two parts: sentiment label filtering and domain adaptability assessment. (i) For sentiment label filtering, we use the sentiment from $\mathbf{w}^s$ as supervisory information to ensure consistency with the corresponding sentiment labels of $\mathbf{w}^c$. This step helps us eliminate samples with mismatched sentiment labels, thus ensuring the accuracy and reliability of sentiment analysis. (ii) Additionally, we train an extra classifier to assess the domain adaptability of the generated counterfactual target domain samples $\mathbf{w}^c$, benefiting from the access to unlabeled target domain data. This ensures that $\mathbf{w}^c$ is not only consistent in sentiment with the target domain but also closer in semantics and style. We name this enhanced version with the filtering mechanism CDA^2-F.

4 Experiments

In this section, we conduct experiments to explore the following research questions: (i) Does our proposed data augmentation approach have the capability to substantially improve the cross-domain SA performance of the model? If so, how does the enhancement achieved by our approach compare to other baseline methods? (ii) Do the individual components of our framework contribute positively to the overall effectiveness of the model? (iii) Is the proposed CDA^2 framework effective in addressing the problem of semantic disruptions and spurious associations with the source domain while generating high-quality samples?

4.1 Datasets

We follow prior domain adaptation research, concentrating on binary cross-domain sentiment classification. Our experiments utilize the multi-domain Amazon reviews dataset(Blitzer et al., 2007), containing reviews from four domains: Books (B), DVD (D), Electronics (E), and Kitchen appliances (K). A five-fold cross-validation protocol is used, with 20% of samples randomly selected as the development set, and the best model on this set is used for target domain generalization testing. Since we focus on cross-domain generation in low-resource settings where the target domain lacks labeled data, we only utilize unlabeled reviews during the training stage. We initially train on a labeled source domain dataset and an unlabeled target domain dataset, and then evaluate the models on the remaining three datasets, resulting in a total of 12 tasks. Furthermore, to create a more challenging setting, we select labeled reviews along with corresponding unlabeled reviews from various platforms, including the products domain from Amazon reviews the airline domain and the blog domain.

4.2 Experimental Settings

In the generation process of raw target samples, we truncate each example to 100 tokens. The hyper-parameter was chosen based on the length of labeled samples and computational requirements. We apply the NLTK Snowball stemmer to each word in the n-grams. The smoothing hyper-parameters for calculating $P(\mathcal{D}|w)$ are set to 1, 5, and 7 for uni-grams, bi-grams, and tri-grams, respectively. A threshold of $\tau = 0.08$ is used. We use $K = 4$ orientation vectors for each unlabeled domain. The controllable model is built upon a T5-base model and trained on the unlabeled data for 60 epochs with a learning rate of 5e-5 and a weight decay of 1e-5. In the generation process of the diffusion model, we set the embedding dimension d to 300. We set γ to 0.5. We train using NVIDIA A100 80G Tensor Core GPUs with a batch size of 425 and a sampling batch size of 100. All parameters within our experiments are optimized using the AdamW optimizer(Loshchilov and Hutter, 2019).

4.3 Baselines

We compare our model with the several state-of-the-art baselines, including **R-PERL**(Ben-David et al., 2020) enhances Bert by incorporating a pivot-based adaptation, **SAIM^2**(Rostami et al., 2023) employs domain adaptation to bridge the domain gap in sentiment analysis by creating large margins between class representations in an embedding space, **HATN$_{\text{-Bert}}$**(Li et al., 2018) proposes a transfer network that effectively captures both domain-specific and domain-shared sentiment words, **DAAT**(Du et al., 2020) utilizes domain-adversarial training to prompt Bert to identify features that are invariant across domains, **COBE**(Luo et al., 2022) refines the contrastive learning loss for negative samples in batches, separating class representations further in potential space, **CFd**(Ye et al., 2020) implements class-aware feature self-distillation by integrating PLM's features into a feature adaptation module, **TACIT**(Song et al., 2024) use VAE to disentangle robust and unrobust features using VAE, **UDALM**(Karouzos et al., 2021) extends Bert's pretraining on unlabeled target domain data via the MLM task. In addition, we explore three specific Bert variants for baseline comparisons: **Vanilla$_{\text{-Bert}}$**, fine-tuned on the fundamental Bert(Devlin et al., 2019) and RoBERTa(Liu et al., 2019) models; **AT$_{\text{-Bert}}$**, which incorporates adversarial training to enhance robustness against attacks; and **DA$_{\text{-Bert}}$**, leveraging domain-aware training with source domain labeled data.

5 Results

5.1 Main Experimental Results

In Table 1, we compare our model CDA^2 using Bert as text encoders with baseline methods on 12 cross-domain tasks, and we also compare their average performances. As expected, CDA^2 demonstrates a significant performance advantage over the competitive baselines. Moreover, compared to the current most advanced domain adaptation method, UDALM, our approach achieves competitive performance overall from the perspective of generating reliable target domain data, and it has achieved the best accuracy in multiple tasks.

Specifically, (i) compared to the most basic baseline R-PEAL, our CDA^2 model has an average accuracy improvement of 4.08%, and CDA^2-F has an improvement of 4.24%. Moreover, CDA^2 and CDA^2-F have surpassed all baseline methods in 12 tasks, with the exception of TACIT and UDALM. (ii) CDA^2 outperforms the recent TACIT model proposed by Song et al. (2024) in most of the 12 tasks, achieving an average accuracy improvement of 0.26%, and has reached the state-of-the-art from the Electronics to Books domain. CDA^2-F

S→T	(a) Books →			(b) DVD →			(c) Electronics →			(d) Kitchen →			All
	D	E	K	B	E	K	B	D	K	B	D	E	
R-PERL	87.80	87.20	90.20	85.60	89.30	90.40	83.90	84.80	91.20	83.00	85.60	91.20	87.50
Vanilla$_{\text{-Bert}}$	88.96	86.15	89.05	89.40	86.55	87.53	86.50	87.95	91.60	87.55	87.30	90.45	88.25
SAIM^2	87.50	88.30	88.00	90.50	87.30	88.50	89.00	85.50	90.80	88.00	84.50	91.30	88.27
AT$_{\text{-Bert}}$	89.70	87.30	89.55	89.55	86.05	87.69	87.15	88.20	91.91	87.65	87.72	90.25	88.56
HATN$_{\text{-Bert}}$	89.36	87.21	89.41	89.81	86.99	87.59	87.10	88.81	92.01	87.88	87.89	90.31	88.69
DA$_{\text{-Bert}}$	89.75	88.11	90.65	90.40	88.15	88.55	88.31	89.03	92.75	87.90	88.35	90.59	89.37
DAAT	89.70	89.57	90.75	90.86	89.30	90.50	88.91	90.13	93.18	87.98	88.81	91.72	90.12
COBE	90.05	90.45	92.90	90.98	90.67	92.00	87.90	87.87	93.33	88.38	87.43	92.58	90.38
CFd	87.65	91.30	92.45	91.50	91.55	92.45	88.65	88.20	93.60	89.75	87.80	92.60	90.63
TACIT	91.42	91.68	92.73	91.33	91.83	91.55	89.62	89.25	94.18	89.70	89.20	93.40	91.32
UDALM	90.97	**91.69**	93.21	91.00	**92.30**	**93.66**	90.61	88.83	**94.43**	90.29	**89.54**	**94.34**	**91.74**
CDA^2	91.18	91.43	93.01	91.29	92.02	92.51	**90.62**	89.65	94.11	90.24	89.10	93.74	91.58
CDA^2-F	**91.62**	91.41	**93.22**	**91.35**	91.84	92.78	90.35	**90.02**	94.13	**90.65**	89.42	94.04	**91.74**

Table 1: Classification accuracy (%) for the cross-domain sentiment analysis tasks for the Amazon Reviews dataset.

Model	B → D	B → E	B → K	Avg
CDA^2-F	**91.62**	91.41	**93.22**	**92.08**
CDA^2	91.18	91.43	93.01	91.87
- w/o DS++	91.34	**91.45**	93.18	91.99
- w/o MMD	88.72	89.94	91.61	90.09
- w/o SAS	90.11	90.98	92.43	91.17
- w/o Diff	88.35	89.67	91.17	89.73

Table 2: Ablation experimental results using the Books domain as an example for the cross-domain SA task.

shows even better performance relative to these outcomes. (iii) CDA^2-F, which incorporates a data filtering mechanism, achieves performance competitive with the current SOTA method, UDALM, in this task. Moreover, it attains SOTA performance in multiple tasks among the twelve evaluated. It is worth considering that, compared to traditional domain adaptation methods, we have explored a new generative paradigm to more effectively match the tasks. The results clearly demonstrate the consistent superiority of our method across various domain adaptation tasks compared to baseline methods, highlighting its effectiveness in enhancing cross-domain sentiment analysis performance.

5.2 Ablation Study

We conduct ablation studies, using Books as the source domain, to validate the effectiveness of each component in CDA^2.

In Table 2, the "w/o DS++" indicates that we do not utilize DPM-Solver++ for acceleration. The performance demonstrates that our method effectively balances the relationship between sampling speed and quality maintenance.

Additionally, it further proves the effectiveness of our data filtering mechanism in enhancing the quality of the generated samples. "w/o MMD" means that we do not incorporate MMD loss. The results show the effectiveness of the MMD strategy in managing data distribution shift. "w/o SAS" indicates that the model operates solely in continuous diffusion. Experimental results indicate that the flexible and learnable state enhances the quality of generated models to a certain extent. "w/o Diff" scenario indicates that we do not utilize the diffusion-based generator and instead generate samples directly using a word substitution strategy. This omission leads to a comprehensive decline in experimental results. Based on this analysis, it is evident that the absence of any single component leads to a decline in the performance of CDA^2.

5.3 Robustness Analysis

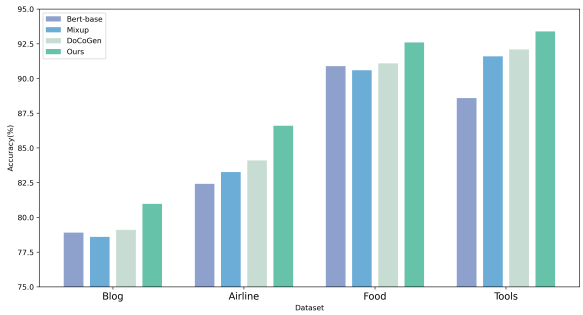

Figure 3: Results on Bert-base and three generation methods for homogeneous and heterogeneous datasets.

To further evaluate the robustness of CDA^2, we conduct comparative experiments on Amazon's homogeneous datasets as well as across-platform datasets. Specifically, we train our model on four

domains and use unlabeled target domain data as supervisory signals for domain adaptation, where the test data remain unseen. Moreover, we compare our method with Bert-base and other generative approaches such as Mixup and DoCoGen. Due to the inconsistent performance of previous generative methods, which lack competitiveness with SOTA, we chose to conduct a separate analysis here. As shown in Figure 3, our method outperforms other generative approaches in the homogeneous Food and Tools datasets, enhancing cross-domain SA performance. In the heterogeneous datasets of Blog and Airline, the large data distribution differences across platforms pose greater challenges; experimental results indicate that our CDA^2 achieves more substantial improvements compared to other methods.

5.4 Data Visualization

To further explore the effectiveness of our method in addressing semantic disruptions and spurious associations with the source domain, we visualize the intermediate representation vectors of text samples using the t-SNE. Figure 4 displays the visualization results for cross-domain pairs from DVD to Kitchen. Although the data distribution produced by DoCoGen exhibits some deviation, it largely remains similar to the source domain because these methods retain many source domain attributes, including context and syntactic structure. In contrast, CDA^2 shows a more similar distribution between the generated data and the target domain data.

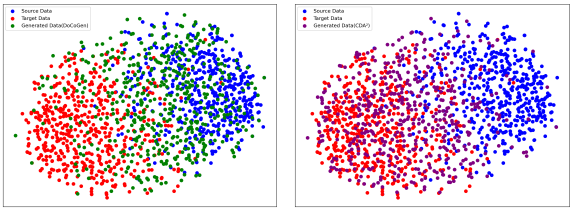

Figure 4: Visualization of discrepancy in distribution.

Additionally, we provide two distinct case studies to analyze the diversity of the text, as shown in Table 3. Specifically, there is a conflict between the action "clean" and "the meal time". While they do include basic domain adaptation operations, there are also instances of unclear and illogical expressions. Therefore, a further understanding of data distribution transfer and mastery of contextual logic are necessary. The analysis above proves that our method not only captures relevant features of domain migration but also exhibits superior expressive capabilities.

	D → K
Original Sample	Sadly, most of the debunking occurs towards the end of the show, in brief statements, before quickly moving on to the next topic. *Negative*
Generated Sample (word substitution)	Sadly, most of the cleaning occurs towards the end of the meal, in brief efforts, before quickly moving on to the next course. *Negative*
Generated Sample (ours)	Unfortunately, the real cleanup only happens at the meal's end, with quick wipes before the next use. *Negative*

Table 3: Cross-domain sentences generated by word substitution strategies and CDA^2 model.

6 Conclusion

In this article, we introduce a Counterfactual Diffusion Augmentation framework for Cross-Domain Adaptation, to address semantic disruptions and spurious associations with the source domain in cross-domain sentiment analysis. CDA^2 excels in generating diverse and realistic counterfactual samples by employing domain-relevant word substitutions from source domain samples to guide a diffusion model. Experiments on benchmark datasets demonstrated that CDA^2 achieves state-of-the-art performance. Through qualitative analysis and visualization, we demonstrate that CDA^2 generates high-quality counterfactual samples that improve domain transfer, effectively alleviating semantic disruptions as well as spurious associations with the source domain.

Limitations

While our study has performed well in cross-domain sentiment analysis, it still has the following limitations.

Firstly, although CDA^2 can generate high-quality text aligned with the target domain, it still relies on unlabeled target-domain data. We should explore how to eliminate this reliance, even when labels are unknown, to generalize the method to unforeseen test data.

Secondly, CDA^2 improves classification by expanding the training set in the target domain but doesn't adjust the classifier's sensitivity to domain knowledge transfer from a causal perspective. Designing causal classification models with augmented data is a promising direction.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work has been supported by the National Natural Science Foundation of China (NSFC) via Grant 62276059 and the Heilongjiang Provincial Natural Science Foundation of China via Grant YQ2023F001. Corresponding author: Yang Li, E-mail: yli@nefu.edu.cn.

References

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. PERL: pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Trans. Assoc. Comput. Linguistics*, 8:504–521.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. DoCoGen: Domain counterfactual generation for low resource domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7727–7746, Dublin, Ireland. Association for Computational Linguistics.

Hao Chen, Rui Xia, and Jianfei Yu. 2021. Reinforced counterfactual data augmentation for dual sentiment classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 269–278, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4019–4028. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2017. Domain-adversarial training of neural networks. In *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pages 189–209. Springer.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023a. Diffuseq: Sequence to sequence text generation with diffusion models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023b. Diffuseq-v2: Bridging discrete and continuous text spaces for accelerated seq2seq diffusion models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 9868–9875. Association for Computational Linguistics.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potamianos. 2021. 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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2579–2590. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Natthawut Kertkeidkachorn and Kiyoaki Shirai. 2023. Sentiment analysis using the relationship between users and products. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8611–8618, Toronto, Canada. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 388–395. ACL.

Junjie Li, Jianfei Yu, and Rui Xia. 2022a. Generative cross-domain data augmentation for aspect and opinion co-extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4219–4229. Association for Computational Linguistics.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022b. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5852–5859. AAAI Press.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2237–2243.

Qi Liu, Yue Zhang, and Jiangming Liu. 2018. Learning domain representation for multi-domain sentiment classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 541–550, New Orleans, Louisiana. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Quanyu Long, Tianze Luo, Wenya Wang, and Sinno Pan. 2022. Domain confused contrastive learning for unsupervised domain adaptation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2982–2995, Seattle, United States. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2023. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Preprint*, arXiv:2211.01095.

Yun Luo, Fang Guo, Zihan Liu, and Yue Zhang. 2022. Mere contrastive learning for cross-domain sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 7099–7111. International Committee on Computational Linguistics.

Antoine Nzeyimana. 2023. KINLP at SemEval-2023 task 12: Kinyarwanda tweet sentiment analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 718–723, Toronto, Canada. Association for Computational Linguistics.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 751–760, New York, NY, USA. Association for Computing Machinery.

Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuanjing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513, Melbourne, Australia. Association for Computational Linguistics.

Xiaoye Qu, Zhikang Zou, Yu Cheng, Yang Yang, and Pan Zhou. 2019. Adversarial category alignment network for cross-domain sentiment classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2496–2508. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text

transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Mohammad Rostami, Digbalay Bose, Shrikanth Narayanan, and Aram Galstyan. 2023. Domain adaptation for sentiment analysis using robust internal representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11484–11498. Association for Computational Linguistics.

Guy Rotman and Roi Reichart. 2019. Deep contextualized self-training for low resource dependency parsing. *Trans. Assoc. Comput. Linguistics*, 7:695–713.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Rui Song, Fausto Giunchiglia, Yingji Li, Mingjie Tian, and Hao Xu. 2024. Tacit: A target-agnostic feature disentanglement framework for cross-domain text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence,Volume 38,17*, pages 18999–19007.

Zecheng Tang, Pinzheng Wang, Keyan Zhou, Juntao Li, Ziqiang Cao, and Min Zhang. 2023. Can diffusion model achieve better performance in text generation ? bridging the gap between training and inference ! In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11359–11386. Association for Computational Linguistics.

Guoyin Wang, Yan Song, Yue Zhang, and Dong Yu. 2019. Learning word embeddings with domain awareness. *CoRR*, abs/1906.03249.

Ke Wang and Xiaojun Wan. 2023. Counterfactual representation augmentation for cross-domain sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):1979–1990.

Hui Wu and Xiaodong Shi. 2022. Adversarial soft prompt tuning for cross-domain sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2438–2447, Dublin, Ireland. Association for Computational Linguistics.

Jingjing Xu, Liang Zhao, Hanqi Yan, Qi Zeng, Yun Liang, and Xu Sun. 2019. LexicalAT: Lexical-based adversarial reinforcement training for robust sentiment classification. 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)*, pages 5518–5527, Hong Kong, China. Association for Computational Linguistics.

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artif. Intell. Rev.*, 53(6):4335–4385.

Linyi Yang, Lifan Yuan, Leyang Cui, Wenyang Gao, and Yue Zhang. 2022. FactMix: Using a few labeled in-domain examples to generalize to cross-domain named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5360–5371, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7386–7399. Association for Computational Linguistics.

Kai Zhang, Hefu Zhang, Qi Liu, Hongke Zhao, Hengshu Zhu, and Enhong Chen. 2019. Interactive attention transfer network for cross-domain sentiment classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5773–5780. AAAI Press.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. 2015. Supervised representation learning: Transfer learning with deep autoencoders. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 4119–4125. AAAI Press.

Yftah Ziser and Roi Reichart. 2017. Neural structural correspondence learning for domain adaptation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pages 400–410. Association for Computational Linguistics.