Domain Generalization via Causal Adjustment for Cross-Domain Sentiment Analysis

Siyin Wang[1], **Jie Zhou**[2,*], **Qin Chen**[2], **Qi Zhang**[1], **Tao Gui**[3], **Xuanjing Huang**[1,*]
[1] School of Computer Science, Fudan University, Shanghai, China
[2] School of Computer Science and Technology, East China Normal University, Shanghai, China
[3] Institute of Modern Languages and Linguistics, Fudan University, Shanghai, China

Abstract

Domain adaption has been widely adapted for cross-domain sentiment analysis to transfer knowledge from the source domain to the target domain. Whereas, most methods are proposed under the assumption that the target (test) domain is known, making them fail to generalize well on unknown test data that is not always available in practice. In this paper, we focus on the problem of domain generalization for cross-domain sentiment analysis. Specifically, we propose a backdoor adjustment-based causal model to disentangle the domain-specific and domain-invariant representations that play essential roles in tackling domain shift. First, we rethink the cross-domain sentiment analysis task in a causal view to model the causal-and-effect relationships among different variables. Then, to learn an invariant feature representation, we remove the effect of domain confounders (e.g., domain knowledge) using the backdoor adjustment. A series of experiments over many homologous and diverse datasets show the great performance and robustness of our model by comparing it with the state-of-the-art domain generalization baselines. The codes of our model and baselines are available at https://github.com/sinwang20/DeepDG4nlp.

Keywords: Domain generalization, causal adjustment, cross-domain sentiment analysis

1. Introduction

In the field of cross-domain sentiment analysis (Zhou et al., 2020; Du et al., 2020), domain adaptation (DA) has been extensively studied to transfer the sentiment knowledge from a source (label-rich) domain to a target domain. However, most existing methods in this area assume that the target domain is known during the training phase, which limits their generalization performance when applied to unknown test domain, a scenario commonly encountered in practical applications. In reality it is often, such as in Amazon's products, collecting unlabeled data and fine-tuning (like domain adaptation) is expensive and extravagant, prohibitively impossible. To address this problem, we focus on Domain Generalization (DG) in the field of cross-domain sentiment analysis, which sets a more strict situation, the test domain is unseen (Figure 1).

Recently, domain generalization (Wang et al., 2021b) has attracted increasing interest in the field of computer vision, which aims to learn a model that can generalize to an unseen target domain from some different but related source domains.

The existing methods mainly focus on learning general invariant representations from multiple domains by data manipulation (Adila and Kang, 2021; Volpi et al., 2018), adversarial training (Ganin et al., 2016; Arjovsky et al., 2019), and meta-learning (Chen et al., 2020).

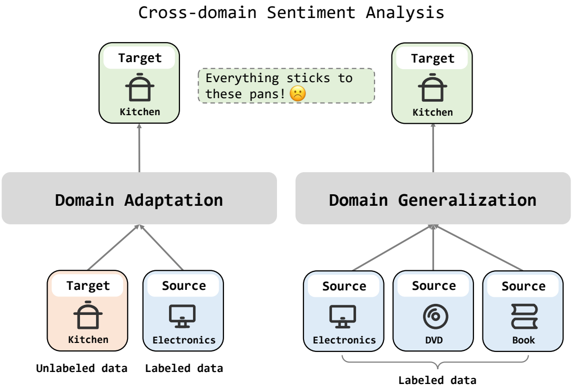


Figure 1: Difference between domain adaptation and domain generalization.

However, there are two significant challenges with current domain generalization methods for cross-domain sentiment analysis. One major challenge (**C1**) is the existence of spurious correlations in domain invariance. It is hard to guarantee that the learned representation is the true cause of the sentiment polarity. For example, the term "hot" indicates popularity in book domain (e.g., "hot-selling books") and delicacy in kitchen domain (e.g., "hot pizza"), leading to the invariant behavior across domains. This spurious correlation fails to hold in other domains, such as "hot CPU," where the sentiment may differ. Such false invariances can degrade the performance of sentiment analysis models in the presence of domain shift.

Furthermore, another challenge (**C2**) is that many existing models focus solely on capturing domain-invariant information, such as general sentiment

[*] Corresponding authors, jzhou@cs.ecnu.edu.cn, xjhuang@fudan.edu.cn.

LREC-COLING 2024, pages 5286–5298

words (e.g., good, bad), while disregarding crucial domain-specific knowledge. This approach results in the loss of valuable domain-specific features that are essential for accurate sentiment analysis. For the previous example, if a model learns that “hot" represents a successful book sale in the book domain and “hot" represents terrible in electronics domain, it can understand that “hot DVD" is positive, while "hot DVD device" is negative, leveraging the domain-specific information.

To address these challenges, we first rethink the task of cross-domain sentiment analysis from a causal perspective, aiming to model the causal relationships among different variables. Then, we propose a causal adjustment-based framework for domain generalization, which learns the generalized representation by disentangling the domain-invariant and domain-specific information. Our model shows great performance over both homologous and diverse datasets. We also explore the robustness of our model on 13 unseen homologous datasets and do ablation studies to verify the effectiveness of components consisting of our model. Representation visualization shows that our model learns a better domain-invariant representation than the baseline.

Our key contributions are summarized as follows.

- We focus on domain generalization for cross-domain sentiment analysis with the assumption that test domains are unseen. Moreover, we rethink this task in a causal view to analyze causal-and-effect relationships among various variables.
- We propose a causal adjustment method to disentangle the domain-invariant and domain-specific representation.
- Extensive experimental results on more than 20 homologous and diverse datasets indicate that our model can remove the influence of confounders to learn a generalized representation.

2. Preliminaries

2.1. Formulation

We define cross-domain sentiment analysis as follows. In the training phase, we have datasets $\mathcal{D}^{train} = \{(x_i^d, y_i^d)\}_{i=1}^{N_d}, d \in \{1, 2, ..\xi\}$, where x_i^d denotes the ith input text(training sample) from the dth source domain, y_i^d is the corresponding sentiment label, and N^d is the number of training samples in domain d. The goal of domain generalization is that given a sample x^T from an unseen domain, we aim to predict its output $\hat{y}^T$ through generalizable representation $\Phi(x)$.

Unlike traditional domain adaptation methods that align representations between source and target domains or other methods that focus on finding invariant representations $\Phi(x_{inv})$, we propose a novel approach that considers both domain-invariant and domain-specific representations based on causal mechanisms and achieve a better generalizable representation $\Phi(x)$.

2.2. Structural Causal Model

Structural Causal Models (SCMs) (Pearl et al., 2000) are widely used to represent causal relationships, such as the causal relationship between the text X and sentiment Y. They are depicted as directed acyclic graphs (DAGs) $G = \{V, E\}$, where V represents the set of variables (e.g., text X, sentiment Y, domain D) and E represents the direct causal connections.

In our work, we utilize SCMs to model the relationships among variables in cross-domain sentiment analysis by specifying how the value of a variable is determined given its parents. These relationships are known as Causal Mechanisms (Peters et al., 2017). Specifically, in our model, the sentiment Y is influenced by its parental variables, which consist of the domain-invariant factor $X_{pa(Y)}^{inv}$ and the domain-specific factor $X_{pa(Y)}^{spc}$. We represent this relationship as follows:

Definition 1 *(Causal Mechanisms)*

$$Y \leftarrow f_Y(X_{pa(Y)}^{inv}, X_{pa(Y)}^{spc}, \epsilon_Y), X_{pa(Y)}^{inv} \perp\!\!\!\perp \epsilon_Y$$

Here, $pa(Y)$ refers to the set of parental variables for sentiment Y. The parental set includes both the domain-invariant factor $X_{pa(Y)}^{inv}$ and domain-specific factor $X_{pa(Y)}^{spc}$ of sentiment Y. ϵ_Y represents the errors due to omitted factors.

In simpler terms, our model captures the causal relationships between the variables. The sentiment Y is influenced by both the domain-invariant aspects of the text X and the domain-specific characteristics. By considering these causal relationships, we can better understand and analyze the sentiment in cross-domain scenarios.

2.3. Backdoor Adjustment

We first consider a simplified setting with only text (X), sentiment labels (Y) and confounders (D). In sentiment analysis, where the goal is to predict sentiment labels (Y) based on text inputs (X), it is important to consider potential confounders that may introduce biases and shortcuts in the causal relationships. One such confounder is the presence of domains (D), which can influence both the text inputs and the sentiment labels.

To tackle the potential confounders existing in the causal inference, one of the regular methods is backdoor adjustment (Pearl et al., 2000) (**Definition 2**). This approach identifies the pure causal effect $P(Y \mid do(X))$ from the total effect

$P(Y \mid X)$ by eliminating the supurios correlation of potential backdoor paths, i.e. $X \leftarrow D \rightarrow Y$.

Definition 2 *(Backdoor Adjustment Formula)*

$$P(Y \mid do(X)) = \sum_D P(Y \mid X, D)P(D)$$

Through this adjustment, we can get the pure relationship between X and Y, $X \rightarrow Y$ without the backdoor path, $X \leftarrow D \rightarrow Y$.

The backdoor adjustment formulation cannot be used arbitrarily, the adjustment variable D should satisfy the backdoor criterion relative to X and Y if:

- No node in D is a descendant of Y.
- Every path between X and Y that contains an arrow pointing to X is blocked by D.

In our SCM, we consider the adjustment between M_{inv} and Y, and the adjustment varivable is the domain D. Fortunately, the domain D satisfies the backdoor criterion: (1) sentiment label D obviously cannot be the factor of domain and therefore is not the descendant of Y. (2) "path between M_{inv} and Y that contains an arrow pointing to M_{inv}" is the backdoor path between M_{inv} and Y. The two backdoor path in our SCM, i.e. $M_{inv} \leftarrow D \rightarrow Y$ and $M_{inv} \leftarrow D \leftarrow M_{spc} \rightarrow Y$ both can be blocked by D. It means that when condition on D as the backdoor adjustment formula, D elimate the correlation between M_{inv} and Y.

So the causal effect between M_{inv} and Y is identifiable and can be formulated as,

$$P(Y \mid do(M_{inv})) = \sum_D P(Y \mid M_{inv}, D)P(D)$$

From the adjustment above, The pure causal effect $P(Y \mid do(M_{inv}))$ is extracted from the total effect $P(Y \mid M_{inv})$ by removing the spurious correlation caused by backdoor path.

3. Our Approach

We propose a backdoor adjustment-based causal model for cross-domain sentiment analysis. We first rethink this task in a causal view (Section 3.1). Then, we introduce the overview of our model (Section 3.2) and integrate backdoor adjustment to learn a better invariant representation (Section 3.3).

3.1. Causal View of Cross-Domain Sentiment Analysis

Despite achieving great performance under the i.i.d condition, the model will fail when encountering the problem of domain shift. That's because the training goal is to minimize $\mathbb{E}_{(x,y)\sim P^{\text{train}}(X,Y)} l(f(x), y)$, where f represents the model, and the objective is to minimize the loss l between the model's predictions and the true labels on the training data. But the empirical distribution of training data $P^{\text{train}}(X, Y)$ is not identical to the test data distribution $P^{\text{test}}(X, Y)$. As the $P(X, Y) = P(Y|X)P(X)$, $P(X)$ is the marginal distribution of X. The period work under the assumption that $P(Y|X)$ remains stable, expected to align the $P^{\text{train}}(X)$ to $P^{\text{test}}(X)$, like many domain adaption methods to align the representation using unlabeled test domain data to achieve the better cross-domain performance (Ganin et al., 2016; Du et al., 2020). Different from DA, the test data is unseen in the DG setting, which makes the prior distribution of target domain $P^{\text{test}}(X)$ cannot be accessed.

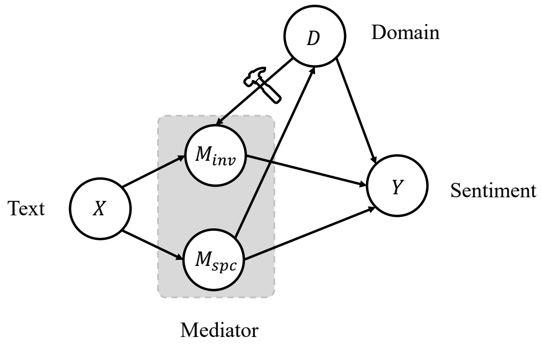


Figure 2: Structural Causal Model of Cross-domain Sentiment analysis

The key challenge for domain generalization is to learn a generalizable representation $\Phi(X)$ without the test domain distribution, which performs well over all domains. Fortunately, the causal mechanisms, $P(Y|\Phi(X))$ has the ability to generalize to the unseen target domain (Scholkopf et al., 2021).

We then construct the Structural Causal Models of cross-domain sentiment analysis in Figure 2 to illustrate the causal relationship with the variables we used. Along with the causal mechanism in section 2.2, we disentangle the text input into domain-invariant and -specific features to model the causal mechanism between text and sentiment. These two features both will cause the sentiment, denoted as the path $M_{\text{inv}} \rightarrow Y$ and $M_{\text{spc}} \rightarrow Y$. Note that we also consider the relationship $M_{\text{spc}} \rightarrow Y$ which is different from the generalization in image classification, because the domain-specific information like the word "hot" in different domains will also affect the sentiment. For domain generalization, another critical variable is the Domain variable D, which represents the domain of the text. Obviously, the domain-specific feature is the cause of the domain. Additionally, the domain variable serves as a confounder that affects the prediction of sentiment Y as well as text. An inexhaustive disentanglement of domain-invariant and -specific may cause the M_{inv} to suffer from the effect of the domain variable.

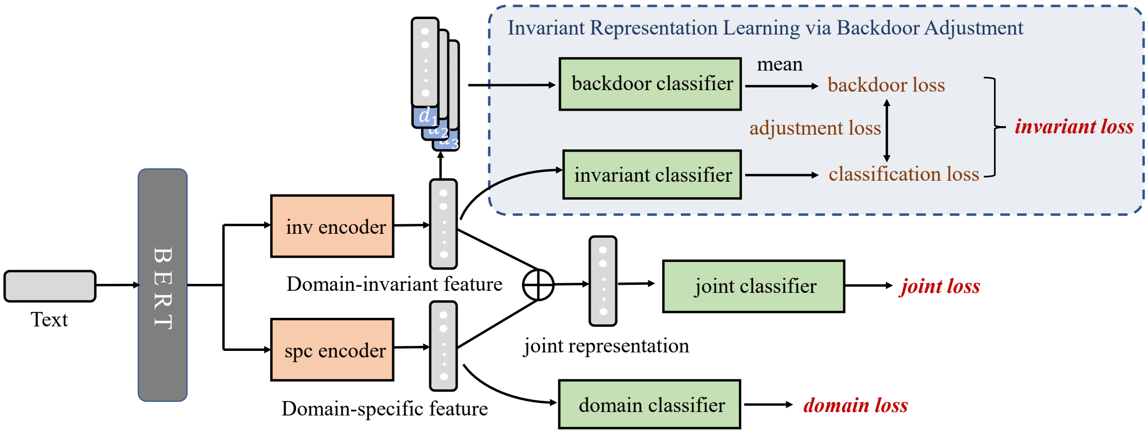


Figure 3: The framework of our model.

3.2. Overview of Proposed Model

We introduce the structure of our causal model for cross-domain sentiment analysis (Figure 3). We use a BERT $\mathcal{M}$ model to obtain the representation of text x using the “[CLS]" representation.

$$h = \mathcal{M}(x) \tag{1}$$

where h is the text representation. Then, to extract the domain-invariant and domain-specific features, we set two independent encoders ϕ (invariant encoder and specific encoder), a three-layer multi-layer perception (MLP) with ReLU activation, $m_{\text{inv}} = \phi_{\text{inv}}(h), m_{\text{spc}} = \phi_{\text{spc}}(h)$.

To help learn the domain-specific information in multiple domains, we constrain its learning with the domain classification loss.

$$\mathcal{L}_{\text{specific}} = \text{CE}(f_{\text{specific}}(m_{\text{spc}}), d) \tag{2}$$

where d denotes the domain of m_{spc}.

As both the domain-invariant and -specific features are the causal factors of the output (polarity), the joint representation (m_{joint}) is formed through element-wise addition of the domain-invariant feature (m_{inv}) and the domain-specific feature (m_{spc}). To ensure accurate polarity prediction, we employ a cross-entropy loss function, which facilitates the alignment between the joint representation and the target sentiment label (y) .

$$\mathcal{L}_{\text{joint}} = \text{CE}(f_{\text{joint}}(m_{\text{inv}} + m_{\text{spc}}), y) \tag{3}$$

where “+" means add by elements.

For a domain-invariant representation, we design the invariant loss $\mathcal{L}_{\text{invariant}}$ from the standpoint of causality and will elaborate on it in Section 3.3.

Overall, our loss function is as follows,

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{joint}} + \mathcal{L}_{\text{invariant}} + \mathcal{L}_{\text{specific}} \tag{4}$$

3.3. Invariant Representation Learning via Backdoor Adjustment

Theoretical Analysis From a causal perspective, if the invariant representation wrongly contains some specific information as the inadequate disentanglement, a path $D \to M_{\text{inv}}$ emerges that should not exist. Which means that the domain variable affects the wrong invariant representation and leads to the spurious correlation between M_{inv} and Y, denoted as $M_{\text{inv}} \leftarrow D \to Y$.

For example, model may wrongly think “hot” is an invariant sentiment word, but the relationship between “hot” and sentiment is caused by the spurious correlation by domain. We are more likely to describe the heating capacity of kitchen application with “hot” instead of “warm" (a common word in the clothes domain), like *“hot enough to boiling water"*, ($D \to M_{\text{inv}}$). Furthermore, in the electronics domain, "hot" is more likely to be associated with negative polarity, such as "a very hot CPU," indicating a relationship between the invariant representation and sentiment, i.e., $M_{\text{inv}} \leftarrow D \to Y$.

To address the issue of spurious correlations, we can adjust them using the Backdoor Adjustment Formula (Definition 2) and identify the backdoor path involving D conventionally. Given the test domain is unknown, we propose incorporating a constraint condition during the training phase to achieve a purer invariant representation instead of adjusting the correlations during inference.

In our specific task, we consider that if the learned invariant features are truly domain-invariant, then the domain should have no effect on these features (i.e., M_{inv} is independent of D). Therefore, the prediction probabilities, with and without backdoor adjustment, should be identical, referred to as "Backdoor Condition" we proposed.

Proposition of Backdoor Condition If a learned representation M is an invariant representation, then the backdoor adjustment is invalid for it, i.e.

$P(Y \mid do(M=m)) = P(Y \mid M=m)$.

Proof of Backdoor Condition The invariant representation obviously should be independent of the domain variable, denoted as $M \perp\!\!\!\perp D$. According to the backdoor adjustment formula, $P(Y \mid do(M=m) = \sum_D P(Y \mid X, D)P(D)$. And the conditional distribution of Y given X can be written as, $P(Y \mid M=m) = \sum_D P(Y \mid X, D)P(D \mid M)$. As the $M \perp\!\!\!\perp D$, so the $P(Y \mid do(M=m) = P(Y \mid M=m)$.

By incorporating the Backdoor Condition into the training phase, we aim to facilitate effective disentanglement between domain-invariant and -specific features. This approach disentangles the backdoor and causal paths during training rather than during inference, as the target domain is unknown.

Loss Design In practice, we design an adjustment loss to achieve the Backdoor Condition to approximate invariant representation. First, we design the after-adjustment distribution of M_{inv} according to the backdoor adjustment (section 2.3), and set the corresponding loss to help the construction.

$$\mathcal{L}_{\text{backdoor}} = \text{CE}(\sum_{d \in D} f_{\text{backdoor}}(m_{\text{inv}} \oplus e^d) \cdot P(d), y)$$

where e^d is set as the learnable embedding of domain d and $\oplus$ means concatenation. We simplify $P(d)$ to the proportion of the domain d in the input data, i.e. $P(d) = \frac{1}{|D|}$.

The $P(Y \mid M_{inv})$ is modeled by the classification loss,

$$\mathcal{L}_{\text{classification}} = \text{CE}(f_{\text{classification}}(m_{\text{inv}}), y) \quad (5)$$

Adjustment loss aligns the $P(Y \mid M_{\text{inv}})$ and $P(Y \mid do(M_{\text{inv}}))$ to achieve the invariant representation under the backdoor condition.

$$\begin{aligned}\mathcal{L}_{\text{adjustment}} &= \mid \mathcal{L}_{\text{classification}} - \mathcal{L}_{\text{backdoor}} \mid^2 \\ &= [\sum_{i=1}^{C} y_i \log(y_i^{\text{inv}}) - \sum_{i=1}^{C} y_i \log(y_i^{\text{back}})]^2 \\ &= [\sum_{i=1}^{C} y_i \log(\frac{y_i^{\text{inv}}}{y_i^{\text{back}}})]\end{aligned}$$

The above derivation shows that such a loss setting makes the $P(Y \mid M_{\text{inv}})$ equals to $P(Y \mid do(M_{\text{inv}}))$.

Finally, we set the α and β to adjust the weight of the backdoor classifier and adjustment. The complete invariant loss is as follows,

$$\begin{aligned}\mathcal{L}_{\text{invariant}} = \mathcal{L}_{\text{classification}} + \alpha \cdot \mathcal{L}_{\text{backdoor}} \\ + \beta \cdot \mid \mathcal{L}_{\text{invariant}} - \mathcal{L}_{\text{backdoor}} \mid^2\end{aligned} \quad (6)$$

4. Experimental Setups

4.1. Datasets and Metrics

Homologous Datasets We use the multi-domain Amazon reviews dataset (Blitzer et al., 2007), a widely-used standard benchmark datasets for domain adaptation. It contains reviews on four domains: Books (B), DVDs (D), Electronics (E), and Kitchen appliances (K). For domain generalization, We follow the experiment settings proposed by (Ziser and Reichart, 2017). Each domain also has 2,000 labeled examples (1,000 positive and 1,000 negative). To further evaluate our model's performance and robustness, we adopt dozens of unseen datasets of Amazon reviews dataset (Blitzer et al., 2007), with 13 types of products.

Diverse Datasets To consider a more challenging setup we experiment with a large gap domain generalization. We randomly sample the 2,000 labeled examples (1,000 positive and 1,000 negative) from four sentiment analysis datasets, including products domain from Amazon reviews (Blitzer et al., 2007), restaurant domain from Yelp reviews (Zhang et al., 2015), airline domain from airline reviews [1], movie domain from IMDb reviews (Maas et al., 2011). In contrast to Homologous Datasets, Diverse Datasets are extracted from different writing platforms and therefore are more diverse in terms of the writing type, navigator, etc.

4.2. Baselines

As most of the past research in cross-domain sentiment analysis concentrates on domain adaptation, which requires the target domain data, we mainly compare our proposed model with several popular and strong domain generalization methods. **(1) MoE, MoEA** (Guo et al., 2018) model the domain relationship with a mixture-of-experts (MoE) approach in non-adversarial and adversarial settings. **(2) BERT-base (ERM)** is a basic BERT model with a binary classification layer at the output and minimizes empirical risk. **(3) DANN** (Ganin et al., 2016) utilizes an adversarial approach to learn features to be domain indiscriminate. **(4) Mixup** (Zhang et al., 2018; Sun et al., 2020) adopts pairs of examples from random domains along with interpolated labels to perform ERM. **(5) GroupDRO** (Sagawa et al., 2019) explicitly minimizes the loss in the worst training environment to tackle the problem that the distribution minority lacks sufficient training. **(6) IRM** (Arjovsky et al., 2019) seeks data representations where the optimal classifier on top of those representations matches across randomly

[1] https://github.com/quankiquanki/skytrax-reviews-dataset

	DEK-B	EKB-D	KBD-E	BDE-K	Avg
MoE	87.55	87.85	89.20	90.45	88.76
MoE-A	87.85	87.65	89.50	90.45	88.86
Bert-base	88.10	89.65	90.00	90.35	89.53
DANN	88.80	89.75	89.80	89.95	89.58
Mixup	87.40	89.20	89.00	89.95	88.89
GroupDRO	89.65	89.65	89.20	89.75	89.56
IRM	88.55	90.05	88.80	91.00	89.60
VREx	88.40	89.80	90.30	90.35	89.71
EQRM	88.55	89.80	90.40	90.85	89.90
Ours	**90.20**	**90.15**	**90.95**	**91.95**	**90.81**

Table 1: The results over Homologous datasets.

partitioned environments. **(7) VREx** (Krueger et al., 2021) reduces the variance of risks in test environments by minimizing the risk variances in training environments. **(8) EQRM** (Eastwood et al., 2022) leverages invariant risk like VREx, but learns predictors that perform well with high probability rather than on-average or in the worst case.

4.3. Implementation Details

In accordance with the commonly used leave-one-domain-out protocol (Li et al., 2017a), one domain will be set aside for testing and the remaining domains will be used for training. The data in each training domain is randomly divided into a training set (80 %) and a validation set (20 %). During training, the learning rate is set as 1e-5 and the batch size is set as 16. Adam optimizer (Kingma and Ba, 2014) is used to update all the parameters. For our $\mathcal{L}_{\text{invariant}}$, α and β are searched in $[0.1, 100]$. For the representation visualization, both settings are identical except for the input variations, with n_components=2 and perplexity=100.

5. Experimental Analysis

In this section, we first evaluate the performance (Section 5.1) and robustness (Section 5.2) of our model by comparing it with baselines. Then, we conduct experiments on diverse domains to further verify the model's effectiveness (Section 5.3). Finally, we report more analysis on ablation studies, representation visualization and comparison with Large Language Models (Section 5.4).

5.1. Main Results

Compared with the current methods, our method outperforms in all settings. Through a comparative analysis of existing methods, we further illustrate the reasons why our approach shows significant advantages (Table 1).

Adversarial training like DANN may fail in some cases, which coincides with the results by (Wright and Augenstein, 2020) in the domain adaptation. Although DANN adversarially trains a domain classifier to make features indistinguishable, the domain confounders may well also remain and the feature suffers from the spurious correlation. Our method of invariant representation via backdoor adjustment well tackles this issue.

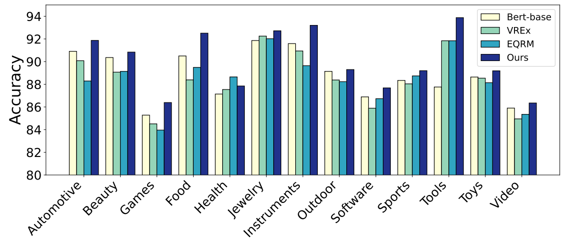


Figure 4: Results on 13 Homologous datasets.

Data augmentation method like Mixup fails in most cases, as the gain in random augmentation does not cover distribution difference triggered by domain shift. Augmented data more closely to the target domain (Yu et al., 2021; Calderon et al., 2022) may be more effective than the universal method, but this is not possible when the target domain is not accessible.

Other methods focusing on learning invariant prediction (like GroupDRO, VREx, EQRM) fall behind ours because they don't consider domain-specific information. It also demonstrates the sufficiency of our consideration of domain-specific information.

5.2. Robustness

To further evaluate the robustness of our model, we conduct experimental results on more homologous datasets of Amazon (Figure 4). Specifically, we train our model on four domains (i.e., Books, DVDs, Electronics, and Kitchen appliances) and test on 13 other unseen domains. It is evident from the results that our method improves the performance of Bert-base in generalizing to multiple unseen domains. Specifically, our model outperforms the Bert-base model over all 13 domains. These denote that our model has a wider generalization capability as it has the causal ability to reason both domain-invariant and domain-specific features.

5.3. Performance on Diverse Domain

Considering the major research in cross-domain sentiment analysis base on the amazon benchmark and only transfer across the products. We consider a more challenging setting on Diverse Dataset, where the distribution gap between the source domain and target domain is much bigger, like the restaurant domain to the airline domain rather than DVDs to Electronics. We maintain the same training setting of homologous datasets, three domain

	Airline	Amazon	Imdb	Yelp	Avg
Bert-base	82.02	87.70	88.05	94.85	88.16
DANN	82.62	87.90	88.45	93.80	88.19
Mixup	82.67	87.85	89.65	94.65	88.71
GroupDRO	83.17	88.65	87.95	94.15	88.48
IRM	83.32	88.60	89.35	92.75	88.51
VREx	83.37	88.75	88.05	94.20	88.59
EQRM	83.47	87.75	89.20	94.35	88.69
Ours	**85.81**	**89.10**	**90.00**	**95.20**	**90.03**
Improvement	+2.34	+1.35	+0.80	+0.85	+1.34

Table 2: The performance over diverse datasets.

data for training and validation, and one domain for testing (Table 2). For example, Airline in Table 2 means training on Amazon, IMDb, and Yelp, and testing on Airline.

In contrast to the homologous scenario, the performance of different test domains shows relatively large differences, reflecting the difference between the fields as well. Due to growing differences between domains and smaller common features, the gain from the original method decreases (like VREx). Owing to our capability to causally model both invariant and specific information, our approach is still able to maintain good performance in the diverse scenario.

5.4. Further Analysis

Ablation Studies To further analyze the effectiveness of the key parts in our model, we provide the ablation studies (Table 3). Specifically, we remove the backdoor loss and adjustment loss that aims to learn a domain-invariant representation (w/o Invariant), the specific loss and joint loss that is designed to learn a domain-specific representation (w/o Specific), and both of them (w/o Both (Bert-base)). From the results, we can obtain the following findings. First, the backdoor adjustment can help our model learn a better domain-invariant representation, which improves the generalization of our model (row 1 and 2). Second, both domain-invariant and domain-specific representations are important for cross-domain sentiment analysis (row 1-3). Removing one of them will reduce the performance of all four datasets.

Representation Visualization To better understand our model, we visualize the representation of the text (Figure 5) over two homologous and diverse datasets. Particularly, we compare our model with Bert-base model by obtaining the representations of samples in the test set. we use t-SNE to translate the 768-dimension representation into a 2-dimension vector. We can observe that it is hard for Bert-base to distinguish the samples with different sentiment polarities. In contrast, the gap between our invariant representations of positive and negative samples is clear. These observations indicate that our backdoor adjustment helps our model learn good generalized representations.

	DEK-B	EKB-D	KBD-E	BDE-K
ours	**90.20**	**90.15**	**90.95**	**91.95**
w/o Invariant	88.30	90.00	88.30	89.85
w/o Specific	89.75	89.65	90.05	89.15
w/o Both (Bert-base)	88.10	89.65	90.00	90.35

Table 3: The ablation results of our model.

	B	M	T
ERM (Bert-base)	88.10	91.59	87.76
ChatGPT (zero-shot)	91.89	89.64	86.73
ChatGPT (3-shot)	**93.86**	92.88	91.84
Ours	90.20	**93.20**	**93.88**

Table 4: Performance Comparison with LLM (ChatGPT)- (B) Book, (M) Musical Instruments, (T) Tools & Hardware

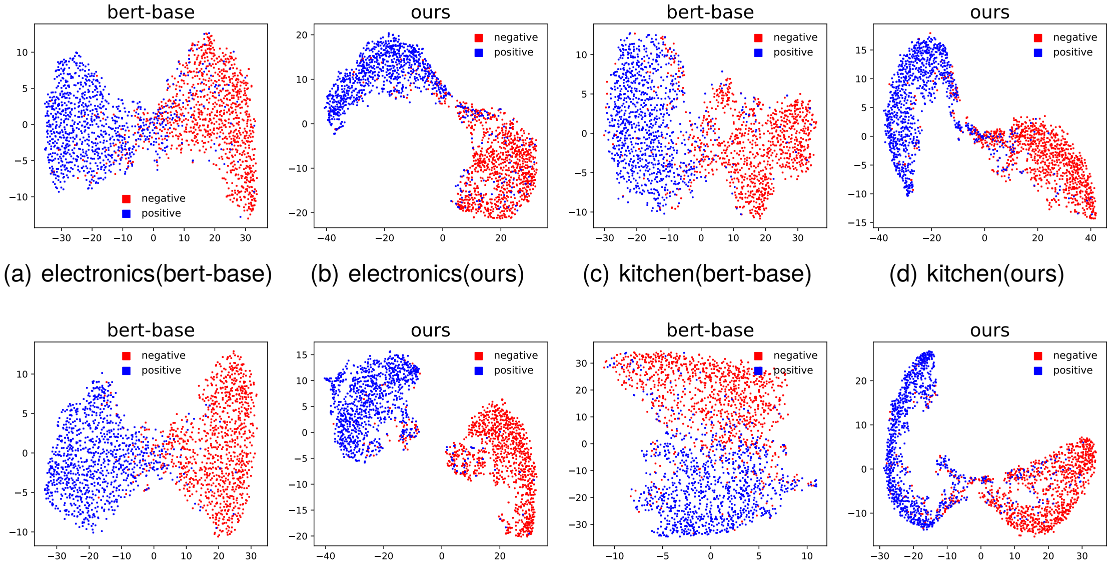
Comparison with Large Language Models (LLMs) To comprehensively assess the implications of our research, we have undertaken a performance evaluation comparing our model with ChatGPT, a well-established Large Language Model (LLM) (Table 4). The experiments involved evaluating both models using zero-shot and few-shot techniques across three diverse datasets: books, musical instruments, and tools & hardware. We gauged the effectiveness of our model in comparison to ChatGPT under various scenarios, shedding light on the relative strengths and weaknesses of our approach against this established LLM.

While ChatGPT exhibits superior performance in the book domain, attributable to its pre-training on extensive datasets encompassing common domains (e.g., books), our model demonstrates competitive efficacy. Notably, our model outperforms ChatGPT in categories such as musical instruments and tools & hardware. This suggests that, although LLMs excel in specific domains due to their pre-training on large-scale datasets, they may encounter challenges in generalizing beyond those domains. The observed differences highlight the intricate nature of language models, underscoring the critical importance of addressing and resolving the fundamental challenge of domain generalization, even for large-scale models.

6. Related Work

6.1. Cross-Domain Sentiment Analysis

Cross-domain sentiment analysis aims to generalize a classifier that is trained on a source domain, for which typically plenty of labeled data is available, to a target domain, for which labeled data is scarce



(a) electronics(bert-base) (b) electronics(ours) (c) kitchen(bert-base) (d) kitchen(ours)

(e) yelp(bert-base) (f) yelp(ours) (g) amazon(bert-base) (h) amazon(ours)

Figure 5: Representation visualization.

(Blitzer et al., 2007; Du et al., 2020). In **single-source domain adaptation**, one line of work employs pivot features to bridge the gap between a source domain and a target domain (Blitzer et al., 2007; Yu and Jiang, 2016; Ziser and Reichart, 2018; Peng et al., 2018; Ben-David et al., 2020). Another branch expects to learn invariant representation, by adversarial training (Ganin et al., 2016; Li et al., 2017b; Du et al., 2020), contrastive learning (Long et al., 2022). Motivated by the success of masked pre-trained language models, some other recent studies base on data augmentation (Calderon et al., 2022; Wang and Wan, 2022), prompt tuning (Wu and Shi, 2022).

There is also a little research on **multi-source domain adaptation**, which uses multiple domains as sources and adapts to one target domain. In this setting, limited works mainly focus on adversarial training (Zhao et al., 2018; Chen and Cardie, 2018; Wu and Guo, 2020) and mixture of expert (Guo et al., 2018). However, most domain adaptation models assume access to unlabeled data from the target domain in-hand during training. For a more reasonable generalization setting, we consider a more challenging and realistic setting, **domain generalization**, where only source domain data can be used during training.

6.2. Domain Generalization

In recent years, domain generalization (DG) has received much attention in ML, which can be divided into three categories (Wang et al., 2021b): (1) **Data Manipulation.** Data manipulation/augmentation methods (Zhang et al., 2018; Sun et al., 2020) aim to increase the diversity of existing training data with operations including randomization, transformation, etc. (2) **Invariant representation learning.** A widely used method is adversarial training (Ganin and Lempitsky, 2015; Li et al., 2018), which adversarially trains the generator (to fool the discriminator) and discriminator (to distinguish the domains). In other works, learning invariant features is approximated by enforcing some invariance conditions across training domains by adding a regularization term to the usual empirical risk minimization (Arjovsky et al., 2019; Krueger et al., 2021). Some group-based works (Sagawa et al., 2019; Liu et al., 2021a) improve the worst group performance. (3) **Learning Strategy.** This line of work focuses on exploiting the general learning strategy to promote the generalization capability, like meta-learning (Chen et al., 2020), and ensemble learning (Mancini et al., 2018; Guo et al., 2018).

There are also several works that consider extending DG to the NLP field, including rumor detection and MNLI (Ben-David et al., 2022), SLU (Shen et al., 2021), Text-to-SQL(Gan et al., 2021), Semantic Parsing (Marzinotto et al., 2019; Wang et al., 2021a), etc. In this work, we consider the DG for cross-domain sentiment analysis. Following the line of invariant representation learning, we propose the backdoor condition for invariant representation and balance the domain-specific features.

6.3. Causality for NLP

Recent years have witnessed the boom of causality, many research combines causal inference with existing machine learning approaches to achieve good results (Feder et al., 2022). This method is used in a wide range of fields, including spurious correlation (Wang and Culotta, 2020; Veitch et al., 2021; Wang et al., 2022), data augmentation (Zmigrod et al., 2019; Liu et al., 2021b), interpretability (Vig et al., 2020; Elazar et al., 2020), etc. The most related to our method is several works utilizing backdoor adjustment to debias in various NLP application, including text classification (Landeiro and Culotta, 2016), distantly supervised named entity recognition (Zhang et al., 2021), court's view generation (Wu et al., 2020). Different from the current method using backdoor adjustment only in the inference period , we design the backdoor adjustment as an invariant prediction condition and add it into the training period to achieve the invariant representation.

7. Conclusion

In this paper, we consider a more challenging scenario, domain generalization for cross-domain sentiment analysis, where the target domain is unseen. Therefore, we propose a framework that disentangles domain-invariant and domain-specific features and leverages both to predict. We rethink the cross-domain sentiment analysis in a causal view and uncover the potential confounders in so-called invariant representations. Taking inspiration from the backdoor adjustment in causal intervention, we propose the backdoor condition to achieve an invariant representation that is not confounded by the domain. Extensive experimental results on more than 20 homologous and diverse datasets demonstrate the great generalization of our model in cross-domain sentiment analysis.

Limitations

Experimental results show that our proposed invariant representation learning does alleviate the problem of potential confounders triggered by domain. Despite giving some examples of domain knowledge as the confounder, it remains impossible for us to enumerate in detail all confounder cases and how variations in performance improvement vary with the differences in confounder between domains. We expect more good interpretable approaches to unveil the potential confounders in cross-domain sentiment analysis and explain the validity of our proposed invariant learning satisfying the backdoor condition.

Moreover, as with other DG studies, the hyperparameters need to be set manually, which limits generalization to some extent. In the future, we expect to eliminate this manual process through self-learning, etc., so that the model is more generalizable.

Acknowledge

The authors wish to thank the reviewers for their helpful comments and suggestions. This work was partially funded by National Natural Science Foundation of China (No.62376061 and No.62307028), the Science and Technology Commission of Shanghai Municipality Grant (No.22511105901 and No.21511100402), Shanghai Science and Technology Innovation Action Plan (No.23ZR1441800 and No.23YF1426100), CURE (Hui-Chun Chin and Tsung-Dao Lee Chinese Undergraduate Research Endowment) (No.23916), and National Undergraduate Training Program on Innovation and Entrepteneurship grant (No.202310246188).

Bibliographical References

Dyah Adila and Dongyeop Kang. 2021. Understanding out-of-distribution: A perspective of data dynamics. In *ICBINB@NeurIPS*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. PADA: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, 10:414–433.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. PERL: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics*, 8:504–521.

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.

Keyu Chen, Di Zhuang, and J. Morris Chang. 2020. Discriminative adversarial domain generalization with meta-learning based cross-domain validation. *Neurocomputing*, 467:418–426.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text 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

1226–1240, New Orleans, Louisiana. 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*, pages 4019–4028, Online. Association for Computational Linguistics.

Cian Eastwood, Alexander Robey, Shashank Singh, Julius von Kügelgen, Hamed Hassani, George J. Pappas, and Bernhard Scholkopf. 2022. Probable domain generalization via quantile risk minimization. *ArXiv*, abs/2207.09944.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2020. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Yujian Gan, Xinyun Chen, and Matthew Purver. 2021. Exploring underexplored limitations of cross-domain text-to-SQL generalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8926–8931, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Tejas Gokhale, Swaroop Mishra, Man Luo, Bhavdeep Sachdeva, and Chitta Baral. 2022. *Generalized but not Robust?* comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2705–2718, Dublin, Ireland. Association for Computational Linguistics.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR.

Virgile Landeiro and Aron Culotta. 2016. Robust text classification in the presence of confounding bias. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. 2017a. Deeper, broader and artier domain generalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5543–5551.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex Chichung Kot. 2018. Domain generalization with adversarial feature learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017b. 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.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021a. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR.

Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2022. Challenges in generalization in open domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2014–2029, Seattle, United States. Association for Computational Linguistics.

Qi Liu, Matt Kusner, and Phil Blunsom. 2021b. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197, Online. Association for Computational Linguistics.

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.

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*, pages 7099–7111, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. 2018. Best sources forward: Domain generalization through source-specific nets. *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1353–1357.

Y. Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2008. Domain adaptation with multiple sources. In *NIPS*.

Gabriel Marzinotto, Géraldine Damnati, Frédéric Béchet, and Benoît Favre. 2019. Robust semantic parsing with adversarial learning for domain generalization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 166–173, Minneapolis, Minnesota. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2).

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.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ArXiv*, abs/1911.08731.

Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109:612–634.

Yilin Shen, Yen-Chang Hsu, Avik Ray, and Hongxia Jin. 2021. Enhancing the generalization for intent classification and out-of-domain detection in SLU. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2443–2453, Online. Association for Computational Linguistics.

Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip S. Yu, and Lifang He. 2020. Mixup-transformer: Dynamic data augmentation for NLP tasks. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3436–3440. International Committee on Computational Linguistics.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *ArXiv*, abs/2106.00545.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *ArXiv*, abs/2004.12265.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. 2018. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Bailin Wang, Mirella Lapata, and Ivan Titov. 2021a. Meta-learning for domain generalization in semantic parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–379, Online. Association for Computational Linguistics.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. 2021b. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35:8052–8072.

Ke Wang and Xiaojun Wan. 2022. Counterfactual representation augmentation for cross-domain sentiment analysis. *IEEE Transactions on Affective Computing*.

Siyin Wang, Jie Zhou, Changzhi Sun, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Causal intervention improves implicit sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6966–6977, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online. Association for Computational Linguistics.

Dustin Wright and Isabelle Augenstein. 2020. Transformer based multi-source domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7963–7974, Online. Association for Computational Linguistics.

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.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780, Online. Association for Computational Linguistics.

Yuan Wu and Yuhong Guo. 2020. Dual adversarial co-learning for multi-domain text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6438–6445.

Jianfei Yu, Chenggong Gong, and Rui Xia. 2021. Cross-domain review generation for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4767–4777, Online. Association for Computational Linguistics.

Jianfei Yu and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 236–246, Austin, Texas. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Lei Zhang and B. Liu. 2012. Sentiment analysis and opinion mining. In *Synthesis Lectures on Human Language Technologies*.

Wenkai Zhang, Hongyu Lin, Xianpei Han, and Le Sun. 2021. De-biasing distantly supervised named entity recognition via causal intervention. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4803–4813, Online. Association for Computational Linguistics.

Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, Joao P Costeira, and Geoffrey J Gordon. 2018. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020. SentiX: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 568–579, Barcelona, Spain

(Online). International Committee on Computational Linguistics.

Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. 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 1241–1251, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

Language Resource References

Blitzer, John and Dredze, Mark and Pereira, Fernando. 2007. *Biographies, Bollywood, Boomboxes and Blenders: Domain Adaptation for Sentiment Classification*. Association for Computational Linguistics.

Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher. 2011. *Learning Word Vectors for Sentiment Analysis*. Association for Computational Linguistics.

Zhang, Xiang and Zhao, Junbo and LeCun, Yann. 2015. *Character-level Convolutional Networks for Text Classification*. Curran Associates, Inc.

Ziser, Yftah and Reichart, Roi. 2017. *Neural Structural Correspondence Learning for Domain Adaptation*. Association for Computational Linguistics.