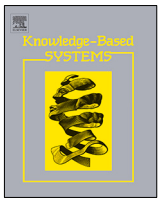
Knowledge-Based Systems 245 (2022) 108649

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



# Contrastive transformer based domain adaptation for multi-source cross-domain sentiment classification

Yanping Fu, Yun Liu *

*School of Electronic and Information Engineering, Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing 100044, China*

ARTICLE INFO

*Article history:*
Received 31 August 2021
Received in revised form 17 February 2022
Accepted 23 March 2022
Available online 29 March 2022

*Keywords:*
Sentiment classification
Contrastive learning
Domain adaptation
Multi-source domains

ABSTRACT

Cross-domain sentiment classification aims to predict the sentiment tendency in unlabeled target domain data using labeled source-domain data. The wide range of data sources has motivated research into multi-source cross-domain sentiment classification tasks. Conventional domain adaptation methods focus on reducing the domain difference between the source and target domains to realize sentiment migration, which ignores the selection of effective sources and fails to deal with negative transfer, leading to limited performance. To address these problems, we propose a contrastive transformer-based domain adaptation (CTDA) method, which not only develops a multi-source domain selection strategy, but also improves the problem of negative transfer from the perspective of data quality. Specifically, the proposed CTDA includes four stages: (1) designing a mixed selector to weight all related sources or pick out the Top-K sources according to the spatial similarity between both domains, (2) building an adaptor to extract domain-invariant information of features by minimizing the Wasserstein distance between both domains, (3) constructing a discriminator to capture the domain-private information of features by contrastive learning, and (4) performing a weighted classifier to predict the sentiment tendency of the target domain according to multiple trained source classifiers. Extensive experiments were performed on two public benchmarks, and the results demonstrated that our CTDA model significantly outperforms state-of-the-art approaches.



## 1. Introduction

Internet interactions and surveys are common in all aspects of life. Sentiment classification aims to automatically determine sentiment polarity by analyzing the reviews and opinions of network users, which plays a decisive role in product recommendation [1] or public opinion polls [2]. Naturally, opinions differ across on a wide variety of products or services, which can be considered to come from different domains. At present, no labeled data are missing in many domains, and it costs too much to label data manually; therefore, unsupervised cross-domain sentiment classification (UCSC) has attracted a lot of attention [3,4]. UCSC aims to predict the emotional trends of unannotated samples in the target domain by analyzing the annotated samples in the source domain. Most existing studies have focused on single-source cross-domain sentiment analysis. In practice, the sources of annotated data are extensive, giving multi-source cross-domain sentiment classification (MCSC) practical value.

UCSC is generally based on the high similarity of emotional expression, such as "happy" and "sad" corresponding to positive and negative responsed, respectively in many domains. However, due to domain discrepancy, it is supposed that these opinions have different distributions and language characteristics; even, similar words or phrases represent different entities or emotional tendencies in different domains. E.g. "apple" represents a brand in the electronics domain, but a fruit in the kitchen domain, "it runs fast" is positive for a car in the automotive domain, but negative for battery in the electronics domain. Therefore, when a sentiment classifier trained in one domain and applied to another domain, it may not perform well. To achieve UCSC, three existing problems must be solved: (1) domain shift caused by data from different domains, (2) retention of domain-private features when eliminating domain differences, and (3) the negative transfer problem.

In view of these difficulties, scholars have proposed some solutions [5–7]. Generally, there are two approaches, one is to construct a sentiment classifier by generating domain-invariant and domain-private word features [8,9], and the other is to bridge the domain gaps using domain adaptation methods [10,11]. In early work, whether the features were divided into shared or private spaces mainly depended on whether these words were statistically identical. Subsequently, pivots and non-pivots extracted domain-shared and domain-private emotion words by applying

* Corresponding author.
*E-mail addresses:* 17111012@bjtu.edu.cn (Y. Fu), liuyun@bjtu.edu.cn (Y. Liu).

https://doi.org/10.1016/j.knosys.2022.108649
0950-7051/

deep neural networks [12]. However, these methods also have some problems, such as the domain-invariant features including some irrelevant domain-private features and some shared domain features being divided into private spaces, which weakens the discrimination ability of the sentiment classifier in the UCSC task. As a sentiment migration method, domain adaptation [13,14] has been proposed to induce domain alignment, which mainly applies domain adversarial training or minimizes the feature distance between domains. To enhance domain adaptation, some methods use additional emotional information, such as prior information or emotional dictionaries.

These one-to-one cross-domain learning approaches cannot be directly utilized in multi-source scenarios. In addition to sentiment migration, multi-source domains cause data quality problems. Examples include some irrelevant source domains that are very different from the target domain or even some malicious source domains that perform a poisoning attack. Thus, MCSC methods not only need to consider sentiment transfer from multi-source domains but should also focus on selecting multi-source domains. Existing MCSC approaches are few, and they mainly study multi-source sentiment transfer but ignore the choice of source domains. For example, Wu et al. [15] proposed the transfer of sentiment knowledge from multiple sources, including global and specific information, using a multitask learning method. Zhao et al. [16] proposed multi-source domain adaptation to minimize distance constraints for deep domain fusion by joint learning. Dai et al. [17] proposed a shared and private knowledge transfer structure with multitask learning. Currently, the following challenges persist in MCSC tasks: (i) How to choose suitable sources from multiple source domains and how to weight the selected sources for effective sentiment transfer. The traditional strategy involves directly combining the knowledge of all source domains to transfer it to the target domain. However, irrelevant or malicious sources have negative effects on the performance of MCSC tasks. (ii) How can the knowledge of multiple source domains be transferred to the target domain? No suitable scheme currently exists for balancing domain adaptation and domain-private information.

Driven by the above survey, we propose a contrastive transformer-based domain adaptation (CTDA) approach, which aims to provide an effective multi-source selector and a discriminative domain adaptation method for MCSC tasks. Specifically, we first design a mixed selection strategy to weigh all related sources or select the Top-K sources according to the spatial similarity between the source and target domains. We then extract document features by building an adaptor and discriminator: the adaptor learns transferable domain-shared information by minimizing the Wasserstein distance between both domains, and the discriminator preserves the domain-private information through contrastive learning. Finally, we use a weighted classifier to predict the emotional tendency of the target domain according to selected multi-source data.

The following are the main contributions of this study:

- We propose a mixed multi-source domain selection strategy to weight all related sources or pick out Top-K sources, which selects relevant sources and eliminates irrelevant or malicious sources to reduce the impact of negative transfer.
- We propose a domain adaptor and discriminator to automatically capture the features containing domain-shared and domain-private information, which best supports the final prediction to achieve cross-domain adaptation.
- We propose a novel weighting strategy to aggregate multiple source classifiers to build a sentiment predictor, that can emphasize the importance of different source domains.
- We performed extensive experiments on the FDU-MTL and Amazon review datasets, and the results demonstrate that our CTDA framework can achieve significant performance on unsupervised MCSC tasks.

In the remaining paper, Section 2 presents related work; In Section 3, the proposed approach is introduced; where, Section 3.3 details the multi-source selection strategy; Section 3.4 gives the contrastive transformer domain alignment method and Section 3.5 describes the classifier weighting component; In Section 4, experimental details and analysis of results are described; In Section 5, conclusions and future work are presented.

## 2. Related work

In this section, we introduce related work on domain adaptation and cross-domain sentiment classification, and compare these methods with the proposed CTDA model.

### Domain adaptation

Domain adaptation aims to bridge the gap between the source and target domains. It involves transferring domain-shared information in the cross-domain classification task. Many domain adaptation methods have been proposed for the UCSC task to address the domain migration problem, such as feature-based adaption [18], instance-based adaption [19] and model-based adaption [20]. Among them, feature-based adaptive methods have been widely studied; they mainly use adversarial [21], discrepancy [22] and self-supervision loss [23] as the alignment loss.

The adversarial training method usually builds a domain adaptor to encourage domain confusion with an opposing objective and can be trained by utilizing minimizing–maximizing GAN [24] or the gradient reversal layer [25]. Discrepancy-based methods primarily reduce the discrepancy by explicitly minimizing the distance between the source and target domains, such as the maximum mean discrepancy (MMD) [26], correlation alignment (CORAL) [27] and Wasserstein distance [28]. Self-supervised learning methods leverage auxiliary tasks to shorten the feature space between both domains, and generative [29] and contrastive learning [30] are often used. Although these domain adaptation methods have made remarkable progress in UCSC, they suffer from performance decay when directly applied to MCSC tasks because the negative transfer [31] caused by the data quality of multiple sources.

### Cross-domain sentiment classification

Several studies have focused on UCSC to transfer emotion information from labeled source domains to the unlabeled target domain. Depending on the number of available source domains, UCSC tasks can be divided into single-source and multi-source UCSC tasks.

***Single-source UCSC*:**

The effectiveness of cross-domain sentiment classification mainly depends on the correlation between different domains and affective transfer approaches. At present, the mainstream methods for single-source UCSC include feature-based transfer and domain adaptation.

Feature-based transfer methods require the generation of pivot or non-pivot features, which means the extraction of domain-shared and domain-private emotion words. Early extraction methods applied statistical information or required manual selection; For example, Bollegala et al. [32] proposed PMI to select pivot features, which combines the occurrences of individual features. They then improved the previous method and proposed the PPMI, which defines only negative PMI values as zero [33]. The current extraction methods mainly apply deep learning. Yu et al. [34] jointly learned hidden features and a prediction model with a CNN to predict the emotional tendency of the target domain. Li et al. [35] presented a hierarchical attention transfer network

**Table 1**
Notations and definitions.

| Symbols | Definition |
|---|---|
| $D_{S_i}/D_T$ | the $i$th source domain/target domain |
| $n_{S_i}/n_T$ | the number of the $i$th source/target samples |
| $P_S/P_T$ | the source/target marginal distribution |
| $\tilde{D^p_{S_i}}/\tilde{D^q_{S_i}}$ | the augmented sample of the $i$th source domain by back-translation with German-to-English/Chinese-to-English |
| $\tilde{D^p_T}/\tilde{D^q_T}$ | the augmented sample of target domain by back-translation with German-to-English/Chinese-to-English |
| $\tilde{f^p_{S_i}}/\tilde{f^q_{S_i}}$ | the feature of $\tilde{D^p_{S_i}}/\tilde{D^q_{S_i}}$ |
| $f^p_T/f^q_T$ | the feature of $\tilde{D^p_T}/\tilde{D^q_T}$ |
| $L^{S_i}_{con}$ | the contrastive loss function of the $i$th source domain |
| $L^T_{con}$ | the contrastive loss function of the target domain |
| $L_{con}$ | the total contrastive loss function |
| $L_{Wa}$ | the domain adaptor loss function |
| $L_{sent}$ | the sentiment loss function |
| $C_T$ | the classification output of target domain obtained by the multiple source classifiers |
| $f^j_T$ | the feature of target domain generated by the $j$th source feature extractor |
| $\alpha_{S_i}/\alpha'_{S_i}$ | the hypothetical/experimental weight coefficient of the $i$th source classifier |

(HATN), that includes a P-net structure and an NP-net structure to acquire pivots and non-pivots.

Domain adaptation methods of single-source UCSC realize sentiment transfer by domain adversarial training or minimization of the feature distance between domains. Li et al. [36] proposed a classical adversarial memory network (AMN) that automatically captured domain-shared features for UCSC. Louizos et al. [37] presented an MMD [38] regularizer to extract domain-invariant features by reducing the distance between the two domains.

***Multi-source UCSC*:**

When the number of available source domains is large, the multi-source sentiment transfer strategy is no longer limited to UCSC methods, which also needs to consider how multiple sources are integrated. Most multi-source UCSC studies focus on joint migration from numerous source domains to the target domain. Himanshu et al. [39] presented a multi-source iterative domain adaptation algorithm that chooses the best K sources from both the similarity and complementarity properties of the domains, builds shared representations from the source domains, and learns target-specific features iteratively. Wu et al. [15] proposed domain-private sentiment classifiers that combine four types of sentiment information from multi-source domains, which requires additional emotional dictionary information. Zhao et al. [16] proposed an MCSC method for transferring emotional knowledge using a joint learning method. Yang et al. [40] applied a capsule network and granger-causal weights to realize the multi-source domain adaptation. Dai et al. [17] proposed an adversarial training method for a multi-source UCSC task, which includes two frameworks: the first is a novel weighting scheme to acquire pseudo-labels for the target domain and the second trains a private extractor with these pseudo-labels. Dai et al. [41] proposed a two-stage multitask learning methodology to transfer two levels of sentiment information from multiple sources to the unlabeled target domain.

Although these methods utilize sentiment information from multiple source domains by integrating knowledge migration, they rarely consider negative transfer owing to the data quality of multi-source domains. Thus, we proposed a mixed multi-source selection strategy to resist negative transfers. Because sentiment classification rules of the source domain can be moved to the judgment strategy of the target domain rather than the simply transferring shared emotion information, we consider both the domain-shared feature alignment and the retention of domain-private characteristics in the proposed CTDA. In contrast to the methods in existing MCSC studies, we built a domain adaptation architecture with a Wasserstein distance on different domains and a contrastive loss on each domain, thus taking target-specific information into consideration when learning domain-invariant features.

## 3. Methodology

In this section, we introduce the problem formulation and notation. Second, we provide an overview of the CTDA architecture. Subsequently, a Multi-source selection strategy for appropriate source domains is introduced. We then provide a detailed introduction to the proposed contrastive transformer domain alignment structure, which includes the specific compositions of the domain adaptor, domain discriminator and classifier, and as well as their joint training process. Finally, the classifier weighting component is presented.

### 3.1. Problem formulation and notations

To clearly define the problem, we present some formulations for the MCSC task. Let a domain $D = \{X, P(X)\}$, where $X$ is the feature presentation and $P(X)$ is the predicted marginal distribution of data in this domain. We define multiple source domains as $D_S = \{D_{S_1}, \ldots, D_{S_i}, \ldots, D_{S_N}\}$, where $N$ is the number of available source domains. For the $i$th source domain, $D_{S_i} = \{X_{S_i}, Y_{S_i}\}_1^{n_{S_i}}$, where $n_{S_i}$ is the number of samples and $Y_{S_i}$ is the label. The unlabeled target domain is defined as $D_T = \{X_T\}_1^{n_T}$, where $n_T$ is the number of samples in the target domain. Table 1 summarizes the main notations used in this article.

### 3.2. Overview of the architecture

In the UCSC task, negative transfer is the main factor affecting the classification performance and is currently a challenge. Herein, we overcome negative transfer by optimizing the data quality of different domains. We propose a CTDA approach for the MCSC task, and its flowchart is shown in Fig. 1. To optimize the data quality of the source domains, we first propose a mixed multi-source selection strategy to transfer emotional knowledge from appropriate source domains. To further optimize data quality in both domains, we propose a contrastive transformer domain alignment structure for the extraction of features, including domain-shared and domain-private information. Finally, we present a weight composition with a combination

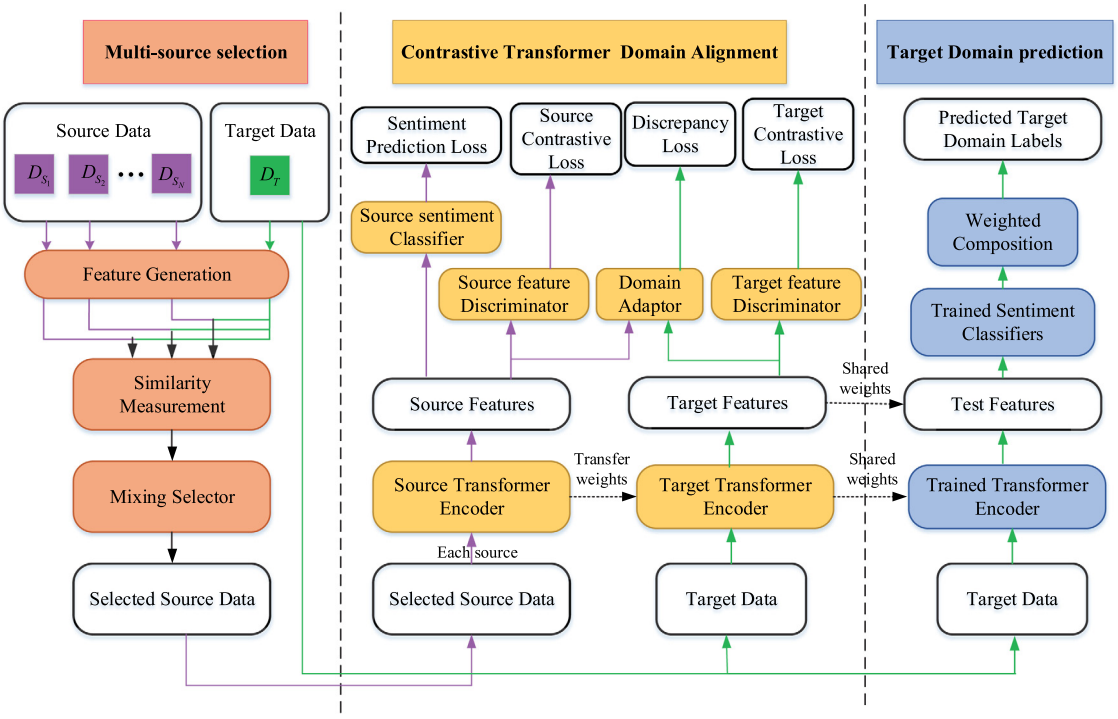



**Fig. 1.** The flowchart overview of our work.

of trained source classifiers for the sentiment prediction of the target domain.

CTDA principally involves multi-source selection, contrastive transformer domain alignment, and the classifier weighting component. (i) During multi-source selection, we use a mixed selector to select the related sources from available source domains, and the selector includes two selection methods: the weighted selection method and Top-K selection method. Specifically, we first generate the feature distribution of different domains via GloVe embedding and using a BERT structure. We then measure the distance between each source and target domain to assess their distribution similarity. Finally, we decide which selection method to use according to the dispersion of similarity. (ii) During contrastive transformer domain alignment, we train different sentiment classifiers for different source and target domain pairs by extracting domain-shared and domain-private features. Specifically, the selected source data and target date are first sent to the feature encoder to obtain the context features. Afterward, we, respectively, construct a domain adaptor with discrepancy loss to generate the domain-shared features and a two feature discriminator with contrastive loss to retain the domain-private features. Next, we apply the features combined the domain-shared and domain-private information to generate the sentiment classifiers with the labeled source data. Finally, four loss functions are jointly trained to obtain the transferable or shared parameters of the model. (iii) In the target domain prediction stage, we propose a classifier weighting component to account for varying degrees of importance for different source classifiers to predict the sentiment polarity of a target domain. In the following section, we introduce the internal compositions of these structures in detail.

### 3.3. Multi-source selection strategy

The structure of the proposed multi-source domain selector is shown in Fig. 2. In this structure, a mixed selection strategy is proposed by assessing the similarity of data distribution between each source and target domains.

We first generate the sample distribution of different domains by mapping the texts to the feature vector space. Because BERT can construct bidirectional contextual representation using a transformer encoder by discriminating surrounding sentences, and a large number of experiments have proved that the pre-training structure enables BERT have strong language understanding ability [42,43]. Thus, in order to combine the semantics of context, we apply a pretrained BERT network to encode the each sample into a feature vector. For the input of the $i$th source domain and target domain, the BERT network encode them to the features, and they are presented as follows:

$$R_{S_i} = BERT(X_{S_i}) \tag{1}$$

$$R_T = BERT(X_T) \tag{2}$$

Generally, Kullback–Leibler (KL) divergence can measure the matching degree of two probability distributions, and the greater the difference between two distributions, the greater the KL divergence. Thus, we consider the KL divergency of features between each source domain and target domain to assess their similarity. In order to make up for the shortcomings of asymmetry of KL divergence, we apply the sum of bi-directional KL divergence to calculate the divergence between each source and target domains. It can be calculated as follows:

$$KL_i = KL(g_{S_i} \| g_T) + KL(g_T \| g_{S_i}) \tag{3}$$

$$g_{S_i} = exp(norm(g'_{S_i})),\ g'_{S_i} = \frac{1}{n_{S_i}} \sum_{k=1}^{n_{S_i}} R_{S_i}(k) \tag{4}$$

$$g_T = exp(norm(g'_T)),\ g'_T = \frac{1}{n_T} \sum_{k=1}^{n_T} R_T(k) \tag{5}$$

where, *norm* denotes the L2 normalization operation and $k$ present the $k$th sample in the $i$th source domain.

Then, we define its similarity score as follows.

$$SC_i = \beta KL_i \tag{6}$$

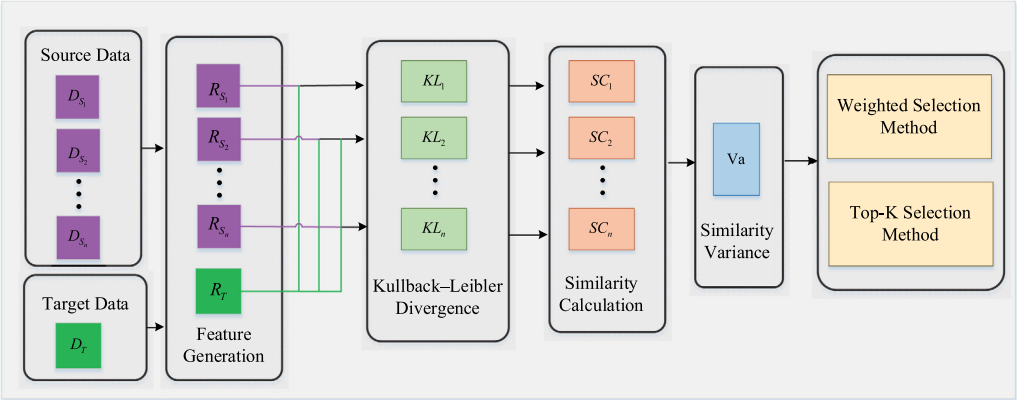



**Fig. 2.** The structure of multi-source domain selector.

where, $\beta$ is a coefficient that gauges the divergence value between 1 and 10. Actually, $\beta$ does not change the distance between domains, and it only maps the value of $KL_i$ to the interval of 0–10, making the similarity score a convenient value to observe, so as to help us intuitively evaluate the difference between domains.

Inspired by Zhang's study [44], the multi-source selector relies on the theory "the nearer distance of feature distributions is, the more related instance is". We present a mixed selection strategy according to the similarity variance, and it is calculated as follows:

$$Va^2 = \frac{\sum_{i=1}^{N}(SC_i - M)^2}{N} \tag{7}$$

where, $M$ is mean value of similarity scores and $N$ is the number of source domains.

Our mixed selection strategy include two selection methods, and which method to choose depends on the similarity variance. These two selection methods is defined as follows:

1.*Weighted Selection Method*

When the similarity variance is relatively small, the distribution difference between different pairs of the source and target domains fluctuates little. It means that all source domains and target domain are located in the same similarity level almost. In this condition, we design a weighted selection strategy that exploits the knowledge of all the sources to predict the target domain. Because the contribution of the knowledge transfer from each source domain is different, the weighted selection method is to sum all the weighted source domains for the final predictor. Let us denote the weight of each source as $\alpha_{S_i}$, the transferred domains are given by

$$P_T = \sum_{i=1}^{n_{S_i}} \alpha_{S_i} P(X_{S_i}) \tag{8}$$

where,

$$\alpha_{S_i} = \frac{SC_i}{\sum_{i=1}^{n_{S_i}} SC_i} \tag{9}$$

In the Eq. (8), $P(X_{S_i})$ indicates the marginal distribution of target domain predicted by the $i$th source predictor, and $P_T$ indicates the final the marginal distribution of target domain.

2.*Top-K Selection Method*

When the similarity variance is relatively large, the distribution difference between different pairs of the source and target domains fluctuates greatly. It indicates that the distribution differences between some domain pairs are small, but the distribution differences between others are large. Thus, under the circumstances, we choose Top-K related source domains as the transferable domains, and eliminate other source domains to reduce the negative transfer. The specific calculation process of Top-K selection method is the same as the weighted selection method. In this paper, we set Top-30% to verify the feasibility of the algorithm.

### 3.4. Contrastive transformer domain alignment

In this section, the proposed contrastive transformer domain alignment framework is described, which transfers the efficient information from selected source domains to the unlabeled target domain. As illustrated in Fig. 3, for each pair of source domain and target domain, our framework consists of a domain adaptor, two domain discriminators, and a sentiment classifier. Specially, the domain adaptor eliminates the domain shift by evaluating Wasserstein distance to obtain domain-shared information; the domain discriminators respectively retain the domain-private information of two domains by the contrastive learning method; the sentiment classifier is obtained by training the labeled data of source domains with the features including domain-shared and domain-private information.

As Fig. 3 shown, in the stage of feature extraction, we apply a transformer structure which includes Glove word embedding mapping, the pretrained BERT network and an MLP layer as our feature extractor. Note that the original data is encoded for the computation of domain adaptor and sentiment classifier; while their augmented data is encoded for the computation of domain discriminators. Here, we first describe the feature generation process of these two types of data.

Following the definition in the previous section, we take the source domain data as an example, and the Glove word embedding mappings of the $k$th sample in the $i$th source domain is expressed as follows:

$$w_{S_i}(k) = Glove(x_{S_i}(k)) \tag{10}$$

Next, we apply BERT to construct bidirectional contextual representation with a transformer encoder by discriminating surrounding sentences. Different from the pre-training structure of BERT in Section 3.1, we fine-tune the parameters of BERT according to subsequent constraints. Thus, the corresponding BERT output is presented as follows:

$$b_{S_i}(k) = BERT(w_{S_i}(k)) \tag{11}$$

CLIM [23] has proved that the MLP projection can learn better discriminative features in terms of BERT model. Thus, we adopt an MLP with one hidden layer to get the projected representations. It is assumed that the features of original data in source domain

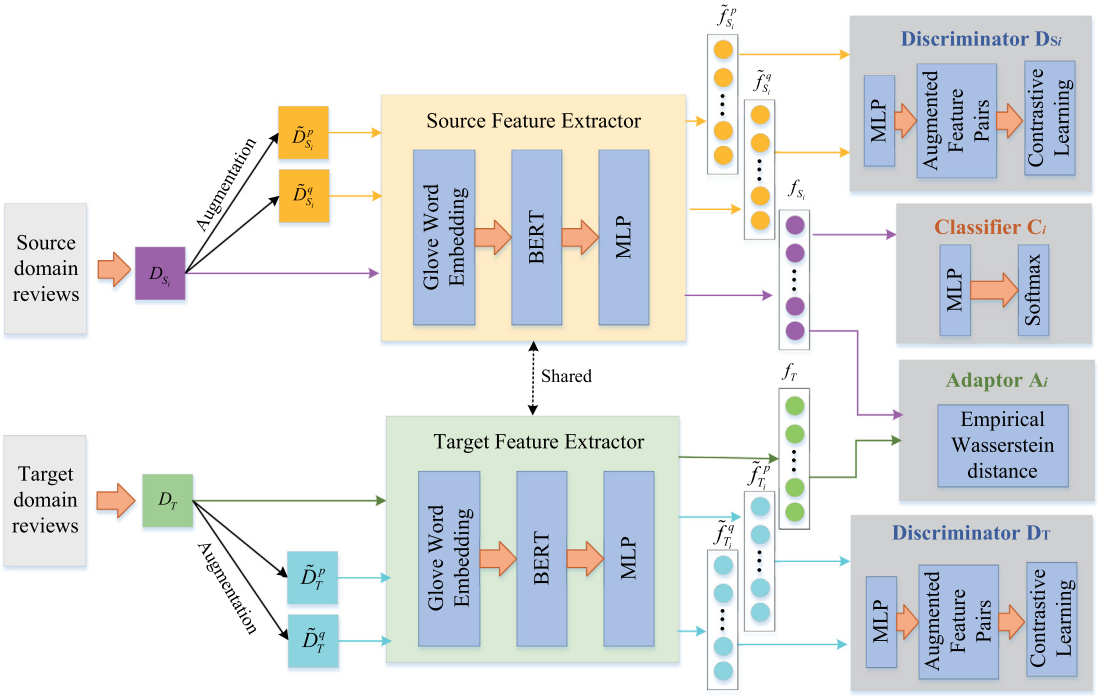



**Fig. 3.** The structure of transformer encoding feature extractor.

and target domain are presented $f_{S_i}$ and $f_T$, and their features of two augmented data respectively are presented as $\tilde{f}_{S_i}^p$ and $\tilde{f}_{S_i}^q$ for the source domain, $\tilde{f}_T^p$ and $\tilde{f}_T^q$ for the target domain. When the weight and bias of MLP layer respectively are $A$ and $c$, the feature output of $b_{S_i}$ is presented as follows:

$$f_{S_i}(k) = Ab_{S_i}(k) + c \tag{12}$$

Because the parameters of the feature extractors between both domains are shared, the calculation process of other features $f_T$, $\tilde{f}_{S_i}^p$, $\tilde{f}_{S_i}^q$, $\tilde{f}_T^p$ and $\tilde{f}_T^q$ are similar to the formula (12).

### 3.4.1. Domain adaptor

For the unsupervised cross-domain classification, the most critical goal is to eliminate domain shift by reducing domain differences. As an effective method, domain adaptation aims at capturing the domain-shared features by minimizing the distances between two domains. Our domain adaptor applies the Wasserstein distance to estimate domain differences and optimize the feature extractor in an adversarial manner, whose theoretical advantages are its gradient property and promising generalization bound.

For the document features of source domain and target domain, such as $f_{S_i}$ and $f_T$, their Wasserstein distances can be presented as follows.

$$W_a(f_{S_i}, f_T) = \sup_{\|f_w\|_L \leqslant 1} E_{f_{S_i}}\left[f_w\left(f_{S_i}\right)\right] - E_{f_T}\left[f_w\left(f_T\right)\right] \tag{13}$$

where, $f_w$ represents the mapping function for document features that is set to satisfy the 1-Lipschitz constraint. And its trainable parameter is $\theta_w$. Thus, the Wasserstein distance can approximately be presented as $L_{wf}$.

To achieve domain confusion, we minimize the distance $L_{wf}$ between two domains.

$$L_{wf}(f_{S_i}, f_T) = \frac{1}{n_{S_i}} \sum_{f_{S_i} \in D_{S_i}} f_w(f_{S_i}) - \frac{1}{n_t} \sum_{f_T \in D_t} f_w(f_t) \tag{14}$$

Because the function $f_w$ needs to meet the Lipschitz constraint, we clip the weights in the range [-c, c] after each gradient updating. Here, to avoid the gradient vanishing or gradient explosion caused by the weight clipping, we propose a gradient penalty function $L_{wg}$. It trains the parameter $\theta_w$ in the process of domain confusion.

$$L_{wg}(f_{S_i}, f_T) = \left\| \nabla_{\hat{d}} f_{wf}\left(\hat{d}\right) \right\| - 1 \tag{15}$$

where, $\hat{d}$ is the random point in the concatenation of $f_{S_i}$ and $f_T$.

The Wasserstein distance is obtained by calculating the following loss function:

$$L_w = max_{\theta_w}\left\{L_{wf}(f_{S_i}, f_T) - \lambda \cdot L_{wg}(f_{S_i}, f_T)\right\} \tag{16}$$

where, $\lambda$ represents the balancing coefficient.

For the calculation of Wasserstein distance, we first train the optimality of $L_w$ by learning feature representations iteratively. After the optimization is achieved, we fix parameters and set $\lambda = 0$ to minimize the Wasserstein distance $L_w$. By iteratively learning features with lower Wasserstein distance, the adversarial objective can finally learn domain-invariant features. Thus, minimizing the loss function of domain adaptor is presented as follows.

$$minL_{W_a}(\theta_e) = min_{\theta_e} L_w \tag{17}$$

where, $\theta_e$ indicate the parameters of the feature extractor.

### 3.4.2. Domain discriminator

During the domain alignment process, domain adaptation has caused the loss of domain-private information while capturing the domain-shared features. To address this issue, the domain discriminator is proposed to preserve the domain-private features with contrastive learning methods. As existing studies [23] have proved, contrastive learning can obtain discriminative sample characteristics through self-supervised method. The core of contrastive loss function is to shorten the distance between original data and positive samples and push away the distance between original data and negative samples. Thus, for the data of each

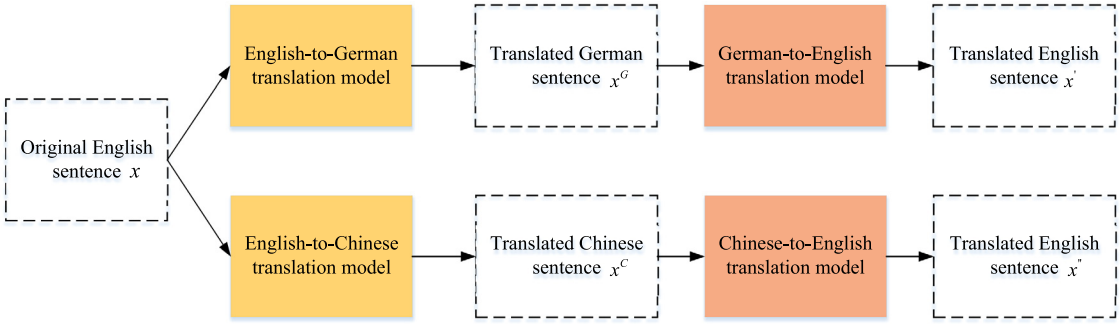



**Fig. 4.** The workflow of data augmentation using back-translation.

domain, we construct the augmented data of original data as positive samples, and other data as negative samples.

We apply back-translation as our augmentation method to generate positive samples, and Fig. 4 shows the workflow of data augmentation using back-translation. We apply two back translation models to augment the document data, which include the back translation between English and German, and between English and Chinese. For the contrastive learning, the positive samples $\tilde{D}^p_{S_i}$ and $\tilde{D}^q_{S_i}$ are augmented data for each document $D_{S_i}$ in each source domain, and the positive samples $\tilde{D_T}^p$ and $\tilde{D}^q_T$ are augmented data for each document $D_T$ in target domain. For each domain discriminator, the inputs include the augmented features generated from two types of augmented data, and their expressions are $\tilde{f}^p_{S_i}$ and $\tilde{f}^q_{S_i}$ for each source domain, $\tilde{f}^p_T$ and $\tilde{f}^p_T$ for the target domain. In this paper, we apply InfoNCE loss [45] to realize the contrastive learning of different domains.

In the contrastive learning process, we use a MLP layer to map representations to the space where contrastive loss is applied. E.g., for a feature of source augmented data $\tilde{f}^p_{S_i}$, the output of MLP layer can be presented as follows:

$$h^p_{S_i}(k) = g(\tilde{f}^p_{S_i}(k)) = W^{(2)}Re\left(W^{(1)}\tilde{f}^p_{S_i}(k) + o^{(1)}\right) + o^{(2)} \tag{18}$$

where, $W^{(1)}$ and $W^{(2)}$ are weights, $o^{(1)}$ and $o^{(2)}$ are biases, and $Re$ is a ReLU non-linearity function.

Taking the source domain discriminator as an example, we define a contrastive loss function for the contrastive prediction task. Given a set $\tilde{f}^j_{S_i}$ including a positive pair of samples $\tilde{f}^p_{S_i}$ and $\tilde{f}^q_{S_i}$, the contrastive prediction task aims to identify $\tilde{f}^q_{S_i}$ in $\tilde{f}^j_{S_i(j\neq p)}$ for a given $\tilde{f}^p_{S_i}$. We define a minibatch of $M$ samples randomly and compute the loss function on pairs of augmented samples from this minibatch, producing $2M$ data points. Thus, given a positive pair $\tilde{f}^p_{S_i}$ and $\tilde{f}^q_{S_i}$, we treat the other $2(M-1)$ augmented data with a minibatch as negative examples. Then, the loss function for a positive pair is defined as:

$$l(p,q) = -log\frac{exp(sim(\tilde{f}^p_{S_i}, \tilde{f}^q_{S_i})/\delta)}{\sum_{j=1}^{2M}\mathbb{I}_{[j\neq p]}exp(sim(\tilde{f}^p_{S_i}, \tilde{f}^j_{S_i})/\delta)} \tag{19}$$

where,

$$sim(\tilde{f}^p_{S_i}, \tilde{f}^q_{S_i}) = \frac{(\tilde{f}^p_{S_i})^T\tilde{f}^q_{S_i}}{\left\|\tilde{f}^p_{S_i}\right\|\left\|\tilde{f}^q_{S_i}\right\|} \tag{20}$$

and $\mathbb{I}_{[j\neq p]} \in \{0, 1\}$ represents an indicator function evaluating to 1 if $j \neq p$, and $\delta$ is a temperature parameter.

The contrastive loss is computed across all positive pairs, both $\tilde{f}^p_{S_i}, \tilde{f}^q_{S_i}$ and $\tilde{f}^q_{S_i}, \tilde{f}^p_{S_i}$, in a mini-batch. Although the positive pairs of $\tilde{f}^p_{S_i}, \tilde{f}^q_{S_i}$ are consistent, their negative pairs are different. Thus, inspired by Chen's work [45], the final contrastive loss function is formulated as:

$$L^{S_i}_{con} = \frac{1}{2M}\sum_{k=1}^{M}[l(p(k), q(k)) + l(q(k), p(k))] \tag{21}$$

The objective of the discriminator is to preserve the domain-private features of the source domains by minimizing the contrast loss function $L^{S_i}_{con}$, thus, we only focus on the parameters of the feature extractor, regardless of the parameters of the discriminator $g(\cdot)$.

We construct parameter-shared domain discriminators for each pair of source and target domains, and simultaneously train them for the augmented samples from different domains. Thus, the final contrastive loss includes two parts, and it is presented as follows:

$$L_{con} = L^{S_i}_{con} + L^{T}_{con} \tag{22}$$

### 3.4.3. Sentiment classifier

The final goal of CTDA is to design a sentiment predictor to determine the sentiment polarity of unlabeled target domain, here, we apply feature-driven method to construct the sentiment classifier, where the domain adaptor and domain discriminator provide domain-shared and domain-private information for the features. We apply a MLP to generate the outputs of sentiment classifier. As Fig. 3 shown, the predicted output of sentiment classifier in a source domain is presented as:

$$y^p_{S_i}(k) = Softmax(W^{(3)}f_{S_i}(k) + o^{(3)}) \tag{23}$$

Then, for the annotated samples of source domain, they are used to compute the cross-entropy loss by minimizing predicted sentiment label and true sentiment label. Thus, sentiment loss function can be described as follows.

$$L_{sent} = -\frac{1}{n_{S_i}}\sum_{k=1}^{n_{S_i}} y^p_{S_i}(k)ln(y^t_{S_i}(k)) + (1 - y^p_{S_i}(k))(1 - ln(y^t_{S_i}(k))) \tag{24}$$

where, $y^t_{S_i}(k)$ presents the ture sentiment label.

### 3.4.4. Joint training

In this work, the adaptation loss, contrastive estimation loss and classification loss are jointly optimized. For the final domain adaptation, the total contrastive transformer domain alignment loss can be summarized as follows:

$$L_{totle} = \sum_{j=1}^{\acute{N}} L^j_{totle} = \sum_{j=1}^{\acute{N}}(\sigma L^j_{W_a} + \tau L^j_{con} + L^j_{sent}) \tag{25}$$

where, $\acute{N}$ presents the number of source domains determined by the multi-source selection strategy.

Here, our objective is to optimize the parameters of feature extractor by minimizing the loss function $L_{total}$. Actually, the parameters of different pairs of source and target domains are not shared, so we need to optimize different model parameters for the final sentiment prediction with the combination of trained sentiment classifiers. Then, the optimal parameter set $\theta_{opt}$ can be represented as follows :

$$\theta_{opt} = \underset{\theta}{argmin}\, L_{totle} = \sum_{j=1}^{\acute{N}} \underset{\theta^j}{argmin}\, L^j_{totle} \tag{26}$$

**Algorithm 1** Framework of ensemble learning for our multi-source selector.

**Input:**
Available source domain $D_S = \{D_{S_1}, \ldots, D_{S_i}, \ldots, D_{S_N}\}$;
Target domain $D_T$;
**Output:**
Selected source domains for the target domain $D_T$;
1: Get the feature representation $R_{S_i}$ for each source domain using Eqs. (1);
2: Get the feature representation $R_T$ for target domain using Eqs. (2);
3: Calculate the similarity score $SC_i$ of domain pairs formed by each source domain and target domain using Eqs. (3)–(6);
4: Calculate the similarity variance $Va^2$ with similarity scores of all domain pairs using Eqs. (7);
5: **return** $Va^2$; According to the value of variance, decide whether to select all related sources or the Top-30% related sources.

### 3.5. Training algorithm for CTDA model

In order to further introduce the training process of the proposed model, we present the algorithms of multi-source selector and multi-source domain adaptation method. Algorithm 1 summarizes the learning process of our multi-source selector. The mixed selection strategy mainly depends on the threshold of similarity variance, which will be given according to the experimental results. Algorithm 2 summarizes the training algorithm of our multi-source domain adaptation method. Iterative methods are used to obtain the optimal set of parameters $\theta_{opt}$, and trained parameters are used to predict the sentiment polarity of target domain.

### 3.6. Classifier weighting component

For the MCSC task, the combination method of different source classifiers has a direct impact on the prediction performance. Therefore, we propose a novel classifier weighting component for the application of selected source domains. Corresponding to the model of each source domain, we extract the feature $f^j_T$ of target domain based on the learned encoder and obtain the sentiment prediction $C^j_T(f^j_T)$ using trained sentiment classifier. Let the number of selected source domains is $\acute{N}$, the different predictions from each source classifier are combined to obtain the final result:

$$C_T = \sum_{j=1}^{\acute{N}} \acute{\alpha_{S_i}} C^j_T(f^j_T) \tag{27}$$

Because the learned document features contain emotional and semantic information and the feature space can better represent the distance relationship between both domains than original data, we use a new weight component $\acute{\alpha_{S_i}}$ instead of the assumed weight $\alpha_{S_i}$ of Eq. (9). The proposed weighting strategy is to emphasize more relevant sources and suppress the less relevant ones. We apply the estimated Wasserstein distance $L_{Wi}$ between the $i$th source and target from the trained model and map the distance to a standard Gaussian Distribution $\mathbb{N}(0, 1)$. Therefore, the weight of each domain $\acute{\alpha_{S_i}}$ can be computed as follows.

$$\acute{\alpha_{S_i}} = \frac{e^{\frac{-L^2_{Wi}}{2}}}{\sum_{i=1}^{n_{S_i}} e^{\frac{-L^2_{Wi}}{2}}} \tag{28}$$

**Algorithm 2** Training algorithm for our multi-source domain adaptation method.

**Input:**
Selected source domain $D_S = \{D_{S_1}, \ldots, D_{S_i}, \ldots, D_{S_{\acute{N}}}\}$;
Target domain $D_T$;
**Output:**
Parameter set $\theta_{opt}$;
1: Initialize parameter set $\theta$;
2: Give the number of training iterations T;
3: **for** t=1:T **do**
4: **for** i=1:$\acute{N}$ **do**
5: **for** number=1:batches **do**
6: Sample minibatch of m source sample from the $i$th source domain;
7: Sample minibatch of m target sample;
8: Extract source feature $f_{S_i}$ using Eqs. (10)–(12);
9: Extract target feature $f_T$;
10: compute augmented source features $\tilde{f}^p_{S_i}$ and $\tilde{f}^q_{S_i}$;
11: compute augmented target features $\tilde{f}^p_T$ and $\tilde{f}^q_T$;
12: Feed $f_{S_i}$ and $f_T$ to Adaptor $A_i$;
13: Compute adversarial loss $L_{W_a}$ using Eqs. (13)–(17);
14: Update $A_i$ by $L_{W_a}$;
15: Feed $\tilde{f}^p_{S_i}$ and $\tilde{f}^q_{S_i}$ to Discriminator $D_{S_i}$;
16: Feed $\tilde{f}^p_T$ and $\tilde{f}^q_T$ to Discriminator $D_T$;
17: Compute contrastive loss $L_{con}$ using Eqs. (18)–(22);
18: Update discriminators $D_{S_i}$ and $D_T$ by $L_{con}$;
19: Feed $f_{S_i}$ to Classifier $C_i$;
20: Compute sentiment loss $L_{sent}$ using Eqs. (23)–(25);
21: Update sentiment classifier $C_i$ by $L_{sent}$;
22: Update all parameters by min $L_{total}$;
23: **end for**
24: **end for**
25: **end for**
26: **return** the optimized parameter set $\theta_{opt}$;

## 4. Experiment

In this section, we describe the design of a large number of experiments to verify the performance of the proposed model based on two publicly published datasets. We empirically evaluate the CTDA, addressing the following points.
(1) Can the mixed selection strategy solve the negative transfer issues?
(2) Can CTDA capture features including domain-shared and domain-private information and obtain better performance than baseline experiments?
(3) Is the suggestion of classifier weighting component effective?

### 4.1. Datasets

To compare the models, we used two widely used datasets to conduct experiments. These are briefly described as follows:

**Table 2**
Statistics of the FDU-MTL review dataset.

| Domain | Training | Val. | Testing | Unlab. | Avg.L | Vocab. |
|---|---|---|---|---|---|---|
| Books | 1600 | 400 | 1600 | 2000 | 159 | 62 K |
| DVD | 1600 | 400 | 1600 | 2000 | 173 | 69 K |
| Electronics | 1598 | 400 | 1500 | 2000 | 101 | 30 K |
| Kitchen | 1600 | 400 | 1600 | 2000 | 89 | 28 K |
| Music | 1600 | 400 | 1600 | 2000 | 136 | 60 K |
| Toys | 1600 | 400 | 1600 | 2000 | 90 | 28 K |
| Vedio | 1600 | 400 | 1600 | 2000 | 156 | 57 K |
| Software | 1515 | 400 | 1500 | 2000 | 129 | 26 K |
| Baby | 1500 | 400 | 1500 | 2000 | 104 | 26 K |
| Apparel | 1600 | 400 | 1600 | 2000 | 57 | 21 K |
| Magazines | 1570 | 400 | 1500 | 2000 | 117 | 30 K |
| Camera | 1597 | 400 | 1500 | 2000 | 130 | 26 K |
| Health | 1600 | 400 | 1600 | 2000 | 81 | 26 K |
| Sports | 1515 | 400 | 1500 | 2000 | 129 | 30 K |
| IMDB | 1600 | 400 | 1600 | 2000 | 269 | 44 K |
| MR | 1600 | 400 | 1600 | 2000 | 21 | 12 K |

**Table 3**
Statistics of the Amazon review dataset.

| Domain | Training | Val. | Testing | Unlab. |
|---|---|---|---|---|
| B | 2000 | 600 | 1400 | 4465 |
| D | 2000 | 600 | 1400 | 5586 |
| E | 2000 | 600 | 1400 | 5681 |
| K | 2000 | 600 | 1400 | 5945 |

**FDU-MTL** [46]: contains a 16-domain dataset. This dataset[1] comes from 14 Amazon domains, which include books, electronics, DVD, kitchen, music, toys, video, software, baby, apparel, magazines, camera, health, and sports, as well as two additional movie reviews from the IMDB[2] and MR[3] dataset. In each domain, the number of training set vary across domains from 1400 to 2000 and the number of validation set is 600. Table 2 presents the relevant statistics, where Val. indicates the validation set, Unlab. indicates the number of unlabeled data, Avg.L indicates the average length of reviews, Vocab. indicates the size of vocabulary.

**Amazon review dataset** [47]: contains a 4-domain dataset. This dataset[4] is the most commonly used dataset in the MCSC task and includes 4 domains: Books (B), Electronics (E), DVD (D), Kitchen Appliances (K). In the experiments, each domain was selected as target domain in turn, and the remaining domains were selected as source domains. Each domain had 2000 samples, of which 1000 were positive reviews and 1000 were negative reviews. Table 3 presents the corresponding statistics.

The FDU-MTL and Amazon review datasets were used in the experiments. As shown in Tables 2 and 3, when a domain is used as the source domain, we use its training set to train the model and its validation set to obtain the optimal parameters from the trained model. When a domain is used as the target domain, we use its testing set to predict the sentiment polarity. All unlabeled data were used to assess domain differences.

### 4.2. Experiment setup

Our experiments applied GloVe word embedding and Bert structure ; one is in the multi-source selection structure, and the other is in the contrast transformer domain alignment structure. GloVe word embedding is applied to map the words from reviews to the vector space, and the BERT structure is used to encode the document features. GloVe word embedding comprises 300 dimension vectors, and BERT comprises 12 layers, 12 self-attention heads, and 768 hidden units. In the multi-source selection structure, we aim to measure the similarities in feature distribution between the source and target domains. Thus, we directly fed the reviews into the two structures to obtain the corresponding feature vectors. However, in the contrast transformer domain alignment structure, it is necessary to obtain the most discriminating features through the training model. Thus, word embedding is initialized by GloVe word embedding and can be fine-tuned during the training process.

When training our model, we set the batch size to 128. An Adam optimizer was applied to minimize joint training loss, and the learning rate was 1.0e−4. Based on the theory presented in [48], the temperature parameter $\lambda$ of the contrastive loss is defined as 0.05. For the Wasserstein distance, we penalized the gradients at random points along the straight line between the source and target pairs, and the parameter $\delta$ was set to 10 as suggested in [49]. To reduce overfitting, the dropout is used in different modules and the dropout ratio is set as 0.5. The number of training epochs ranged from 32 to 128. The weights of the three loss functions $\sigma$ and $\tau$ can vary across different domains, and a related study on its effect on the prediction performance will be demonstrated later through a sensitivity analysis.

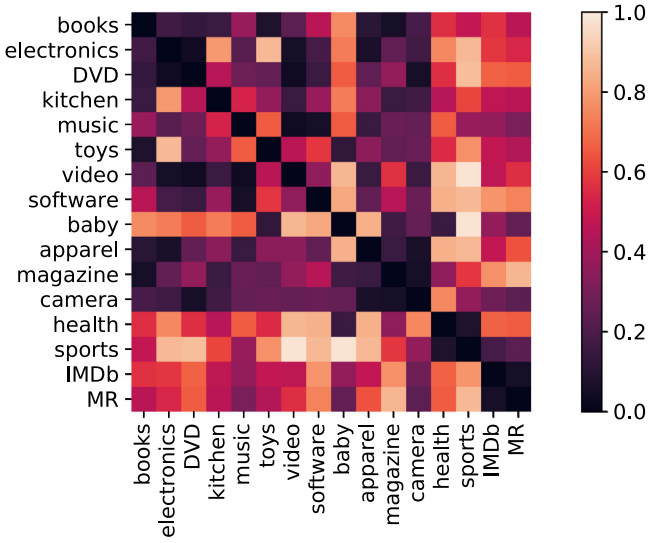



**Fig. 5.** The heatmap of standardized bi-directional KL divergence on FDU-MTL dataset.

### 4.3. Experiment results

In this section, we conduct experiments to verify the effect of the multi-source selection strategy, compare the performances against those in prior work, perform an ablation study to verify the importance of the different components, and analyze the sensitivity of the parameters.

#### 4.3.1. Effect of multi-source selection strategy

The FDU-MTL dataset includes 16 domains, which have sufficient sources to allow the selection of different strategies, and we only conducted experiments on the FDU-MTL review dataset to verify the multi-source selection strategy. Fig. 5 shows a heatmap of the standardized bidirectional KL divergence on the FDU-MTL

[1] Dataset can be found at http://jmcauley.ucsd.edu/data/amazon/index_2014.html.
[2] Dataset can be found at http://ai.stanford.edu/~amaas/data/sentiment/.
[3] Dataset can be found at http://www.cs.cornell.edu/people/pabo/movie-review-data/.
[4] Dataset can be found at http://www.cs.jhu.edu/~mdredze/datasets/sentiment/.

**Table 4**
The variance of the similarity scores of the target domain and its each source domain.

| Target | Books | Electronics | DVD | Kitchen | Music | Toys | Video | Software | Baby | Apparel | Magazines | Camera | Health | Sports | IMDB | MR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variance | 0.05 | 0.08 | 0.05 | 0.04 | 0.06 | 0.07 | 0.08 | 0.08 | 0.13 | 0.10 | 0.05 | 0.02 | 0.11 | 0.12 | 0.13 | 0.14 |

dataset. The smaller the KL divergence between the two domains, the higher their similarity is. Corresponding to the heatmap, lower numbers or darker shades indicate closer proximity between them. Fig. 5 indicates that some domains contain more dark blocks than light blocks, such as books, DVDs, and videos, whereas some domains contain fewer dark blocks than light blocks, such as health, sports, and MR. Thus, in all domains, some domains have more related domains, whereas others have only a few related domains, which promotes the creation of a multi-source domain selection strategy.

To evaluate the degree of dispersion in the similarity between the target domain and other source domains, we computed the variation in similarity scores for each target domain. Table 4 shows the variance of the similarity scores of the domain pairs formed by each source and target domain for each target domain. Table 4 indicates two orders of magnitude, 0.1, and 0.01, in the variance. In probability theory, variance is used to measure the deviation between discrete values and their mean. For the target domain with a variance of 0.1, the similarity scores of the domain pairs formed by the target domain and all source domains are relatively scattered; therefore, in all source domains, some domains are more related to it than others. However, for the target domain with a variance of 0.01, the opposite holds; all the source domains have similar relevance to the target domain. Therefore, for domains with a variance magnitude of 0.1 orders of, such as baby, apparel, health, sports, IMDB, and MR, we apply the Top-K selection method to select sources, and for domains with a variance of 0.01 order of magnitude, such as books, music, software, electronics, toys, DVD, video, magazines, kitchen, and camera, we use the weighted selection method to select sources.

To verify the effect of the proposed multi-source selection strategy, we compared the performance of the proposed algorithm with selection methods from random Top-K sources, average all sources, single best source, referred to as "Random Top-30%", "Average All", "Single Best" respectively. Fig. 6 shows the results of different approaches with different selection strategies on the FDU-MTL dataset. Fig. 6 indicates that the proposed algorithm outperforms "Random Top-30%", "Average All" and "Single Best" for all domains, which proves that the simple selection strategy may introduce too many irrelevant domains as the source domain, resulting in negative transfer. Thus, the overall analysis suggests that the proposed mixed selection strategy mitigates negative transfer issues and leads to significant performance improvements.

The Amazon review dataset contain only four domains; when one domain is as the target domain, only the other three domains can be multi-source domains. For our Top-30% selection method, the number of available source domains is less than one, thus, we only apply our weighted selection method to apply all source domains to transfer the sentiment for each target domain.

*4.3.2. Comparisons with prior works*

In this section, we perform experiments to evaluate the performance of our proposed model by comparing it with a series of baseline methods. For the two datasets, we compared the performance of MSCS with that of existing methods.

***Competitor methods on FDU-MTL are briefly described as follows:***

**ASP**[1], **ASP**[2] [47]: these two models are based on adversarial multi-task learning, which respectively has single-channel model and bi-channel in the task-specific layer. The published results from the original paper is reported in this paper.

**MAN** [3]: it proposed a multinomial adversarial network, which represents a family of theoretically sound adversarial networks to classify multi-source dataset. The published results from Dai et al. [17] is reported in this paper.

**Meta** [9]: this model applies the shared meta-network and multi-task learning to capture the shared and task-specific semantic composition. The published results from the original paper is reported in this paper.

**WS-UDA, 2ST-UDA** [17]: these two models are two end-to-end transfer learning frameworks, which respectively provide a weighting scheme and a two-stage learning process. The published results from the original paper is reported in this paper.

**BERT-base** [42]: it provides a powerful pre-trained model and obtains strong baselines for many downstream tasks, here is for the MSCS task. The published results from Dai et al. [41] is reported in this paper.

**Distil** [43]: it is a distilled version of BERT, which is provide a smaller, faster, and cheaper model than BERT-base. The published results from Dai et al. [41] is reported in this paper.

**SDA, TOE** [41]: these two models are based on the shared-private features to achieve sentiment transfer across domain, and SDA only applies the closest source domain and TOE is through a well-designed ensemble method. The published results from the original paper is reported in this paper.

Table 5 lists the classification accuracies of the proposed CTDA model and other baseline methods on the FDU-MTL reviews dataset. We first used proposed a mixed selection strategy to identify suitable source domains for each target domain, forming 16 multi-source cross-domain tasks for the experiments. Table 5 indicates that the proposed CTDA model almost achieves the best performance on all tasks, and only two tasks were consistent with current best performances. Compared with approaches of the latest multi-source adaptation methods with deep neural networks, the average accuracy of our model was higher than that of WS-UDA by 1.3%, 2ST-UDA by 1.0%, BERT by 4.0%, Distil by 6.6%, SDA by 1.6%, and TOE by 1.4% . For ASP[1], ASP[2], and MAN, applying adversarial multitask learning to complete multi-source domain adaptation reveals that the structure of our model is superior to that of classical methods. For Meta, WS-UDA, 2ST-UDA and TOE, the other methods also consider the domain-shared and domain-private features when performing domain adaptation; however, our model is better because our multi-source selection strategy reduces negative transfer. BERT and Distil also apply similar methods to extract the domain-shared features and perform the SCMC task. Apparently, the contrastive learning of our model can better capture the domain-private characteristics. SDA only applies the closest source domain, thus losing useful transferable information from other source domains. Therefore, its model is less effective than our multi-source models.

***Competitor Methods for Amazon Competitor methods on Amazon review dataset include:***

**mSDA** [10]: it applies marginalized stacked denoising autoencoders to capture the shared-domain feature for the domain transfer. The reported results are from Guo et al. [4].

**DANN** [24]: it uses domain-adversarial neural networks to provide transferable shared information for cross-domain tasks. The reported results are from Dai et al. [41].

**MDAN**[1], **MDAN**[2] [50]: these two models are multi-source adaptation models: the MDAN[1] optimizes directly their boundary of

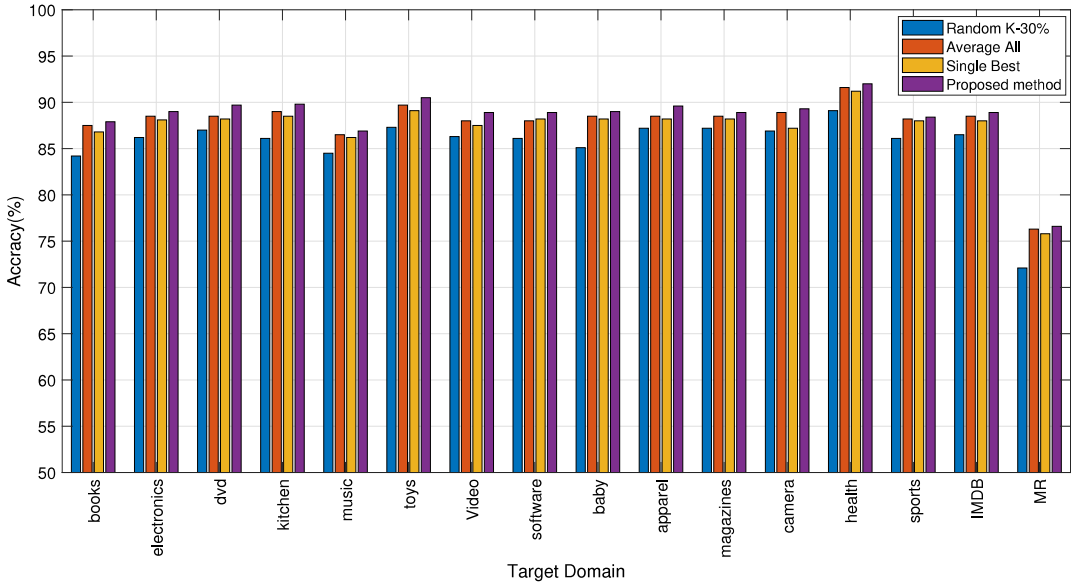



**Fig. 6.** Compares different approaches to combine multiple sources on the FDU-MTL dataset.

**Table 5**
Classification accuracy on the FDU-MTL review dataset.

| Target domain | ASP[1] | ASP[2] | MAN | Meta | WS-UDA | 2ST-UDA | BERT | Distil | SDA | TOE | CTDA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Books | 83.2 | 83.7 | 84.5 | 86.3 | 84.6 | 86.3 | 81.8 | 84.3 | 87.7 | 87.8 | **87.9** |
| DVD | 85.5 | 85.7 | 86.3 | 86.5 | 88.9 | 88.5 | 82.5 | 83.0 | 87.9 | 88.3 | **89.0** |
| Electronics | 82.2 | 83.2 | 86.5 | 86.0 | 87.9 | 89.5 | 88.3 | 83.5 | 89.0 | 88.0 | **89.7** |
| Kitchen | 83.7 | 85.0 | 87.3 | 86.3 | 88.7 | 89.3 | 86.3 | 84.5 | **89.8** | 89.0 | **89.8** |
| Music | 82.5 | 81.7 | 84.0 | 85.7 | 84.3 | 84.1 | 80.0 | 82.8 | 86.2 | 86.8 | **86.9** |
| Toys | 87.0 | 88.2 | 87.3 | 85.3 | 89.1 | 89.0 | 88.5 | 88.0 | 89.3 | 90.3 | **90.5** |
| Video | 85.2 | 85.2 | 87.0 | 85.5 | 88.7 | 87.5 | 81.0 | 78.5 | 87.5 | 87.3 | **88.9** |
| Software | 85.5 | 88.2 | 86.3 | 86.5 | 88.4 | 88.8 | 87.5 | 76.8 | 85.7 | 88.8 | **88.9** |
| Baby | 86.5 | 88.0 | 87.0 | 86.0 | 87.4 | **89.0** | 85.8 | 85.5 | 88.6 | 88.3 | **89.0** |
| Apparel | 87.5 | 86.2 | 86.0 | 86.0 | 90.1 | 89.5 | 84.8 | 81.3 | 87.6 | 86.0 | **89.6** |
| Magazines | 91.2 | 90.5 | 86.8 | 90.3 | 86.6 | 88.8 | 82.5 | 83.5 | 84.1 | 85.8 | **88.9** |
| Camera | 88.2 | 89.7 | 83.8 | 87.0 | 90.2 | 89.1 | 87.3 | 83.0 | 87.4 | 87.3 | **89.3** |
| Health | 87.7 | 86.5 | 88.5 | 88.7 | 87.8 | 88.5 | 91.8 | 85.5 | 89.1 | 88.3 | **92.0** |
| Sports | 86.7 | 86.5 | 88.3 | 85.7 | 88.2 | 88.3 | 86.5 | 85.8 | 87.2 | 87.8 | **88.4** |
| IMDB | 87.5 | 86.7 | 84.3 | 87.3 | 88.7 | 88.3 | 80.8 | 74.3 | 84.8 | 87.8 | **88.9** |
| MR | 75.2 | 76.5 | 73.3 | 75.5 | 74.6 | 73.3 | 74.5 | 68.8 | 76.4 | 74.3 | **76.6** |
| Average | 85.3 | 85.7 | 85.4 | 85.9 | 87.1 | 87.4 | 84.4 | 81.8 | 86.8 | 87.0 | **88.4** |

distributions by adversarial training, whereas the MDAN[2] is to approximates the boundary between both domains smoothly. The reported results are from Dai et al. [41].

**MoE** [4]: it applies mixture of experts to verify the effect of cross domain classification with the combination of different expert networks. The reported results are from the original paper.

**MDAJL** [16]: it proposes a joint learning method and soft parameter sharing technology for the MCSC task. The reported results are from the original paper.

**SDA** [41]: it applies the shared-private architecture to transfer knowledge from the closest source domain. The reported results are from the original paper.

**WS-UDA, 2ST-UDA** [17]: these two models are two end-to-end transfer learning frameworks, which respectively provide a weighting scheme and a two-stage learning method. The reported results are from the original paper.

Table 6 lists the classification accuracies of the proposed CTDA model and other baseline methods on the Amazon reviews dataset. In this experiment, because there are not many source domains to choose from, we directly use the remaining three domains as source domains for each target domain, forming four multi-source cross-domain tasks. Table 6 indicates that compared with other benchmark models, the proposed method obtains the best classification accuracy on all tasks, which demonstrates the effectiveness of our model. The average accuracy of our model is higher than that of mSDA by 3.16%, MDAN[1] by 2.86%, MDAN[2] by 1.9%, DANN by 2.61%, MoE by 0.28%, MDAJL by 11.02%, SDA by 1.54%, WS-UDA by 1.87%, and 2ST-UDA by 0.48%. Unlike the best models, MoE and 2ST-UDA, our model introduces a domain discriminator to enhance private features, which improves the classification performance.

#### 4.3.3. Ablation study

To verify the validity of each component in the proposed model, we performed ablation studies on two datasets, and the results are presented in Tables 7 and 8. In these tables, 'w/o' BERT means MLP replacing BERT, 'w/o' Con. represents model deletes contrastive learning modules, '-W+GRL' represents classical gradient reverse layer (GRL) method replacing Wasserstein-distance, '-a+s' represents the averaging all source classifiers replacing the proposed classifier weighting component. As shown, when a component of the model is replaced or deleted, the experimental results are correspondingly reduced. From the results of the two tables, the following observations can be made:

1. When the input features are extracted using only MLP rather than BERT, the performance of the model is reduced by 2.9% and

**Table 6**
Classification accuracy on the Amazon review dataset (CTDA[1] indicates that the proposed model does not contain multi-source selection strategy).

| Target domain | mSDA | DANN | MDAN[1] | MDAN[2] | MoE | MDAJL | SDA | WS-UDA | 2ST-UDA | CTDA[1] |
|---|---|---|---|---|---|---|---|---|---|---|
| B | 76.98 | 77.89 | 78.45 | 78.63 | 79.42 | 78.80 | 78.68 | 79.39 | 79.92 | **79.97** |
| D | 78.61 | 78.86 | 77.97 | 80.65 | 83.35 | 80.20 | 81.23 | 80.14 | 83.86 | **83.90** |
| E | 81.98 | 84.91 | 84.83 | 85.34 | 86.62 | 81.20 | 85.06 | 83.81 | 85.11 | **86.63** |
| K | 84.26 | 86.39 | 85.80 | 86.26 | 87.96 | 54.30 | 87.33 | 87.66 | 87.68 | **87.97** |
| Average | 80.46 | 82.01 | 81.76 | 82.72 | 84.34 | 73.60 | 83.08 | 82.75 | 84.14 | **84.62** |

**Table 7**
Classification accuracy of CTDA without some components on the FDU-MTL review dataset.

| Target domain | CTDA (w/o BERT) | CTDA (w/o Con.) | CTDA (-W+GRL) | CTDA (-a+s) | CTDA |
|---|---|---|---|---|---|
| Book | 83.1 | 84.8 | 85.2 | 87.5 | **87.9** |
| DVD | 85.1 | 86.2 | 87.5 | 88.5 | **89.0** |
| Electronics | 85.4 | 86.1 | 87.4 | 88.5 | **89.7** |
| Kitchen | 86.1 | 87.5 | 88.9 | 89.0 | **89.8** |
| Music | 84.3 | 86.1 | 87.2 | 86.5 | **86.9** |
| Toys | 87.8 | 87.9 | 88.1 | 89.7 | **90.5** |
| Video | 86.2 | 85.3 | 87.2 | 87.8 | **88.9** |
| Software | 86.0 | 86.7 | 87.6 | 88.0 | **88.9** |
| Baby | 86.9 | 87.1 | 87.3 | 88.5 | **89.0** |
| Apparel | 87.5 | 87.9 | 88.4 | 88.5 | **89.6** |
| Magazines | 86.4 | 87.4 | 88.2 | 88.5 | **88.9** |
| Camera | 87.1 | 87.8 | 88.2 | 88.9 | **89.3** |
| Health | 88.7 | 88.9 | 90.9 | 91.6 | **92.0** |
| Sports | 86.1 | 87.3 | 87.4 | 88.2 | **88.4** |
| IMDB | 86.1 | 86.8 | 87.5 | 88.5 | **88.9** |
| MR | 74.6 | 75.1 | 75.2 | 76.3 | **76.6** |
| Average | 85.5 | 86.2 | 87.0 | 87.8 | **88.4** |

**Table 8**
Classification accuracy of CTDA[1] without some components on the Amazon reviews dataset.

| Target domain | CTDA[1] (w/o BERT) | CTDA[1] (w/o Con.) | CTDA[1] (-W+GRL) | CTDA[1] (-a+s) | CTDA[1] |
|---|---|---|---|---|---|
| B | 76.88 | 77.65 | 78.54 | 79.56 | **79.97** |
| D | 80.00 | 81.44 | 82.65 | 83.77 | **83.90** |
| E | 83.22 | 84.16 | 85.27 | 86.34 | **86.63** |
| K | 85.01 | 85.11 | 86.88 | 87.55 | **87.97** |
| Average | 81.28 | 82.09 | 83.36 | 84.31 | **84.62** |

3.34% for the FDU-MTL and Amazon datasets, respectively, which are the most severe declines in all ablation studies. This indicates that BERT has a powerful ability to encode context semantics and is appropriate for the feature extraction of the MCSC task.

2. When contrastive learning modules were deleted, the performances of the model were reduced by 2.4% and 2.53% for the two datasets, respectively. This is because our model loses the protection of domain-private features, which is critical to emotion classification.

3. To investigate the impact of the Wasserstein distance on the model, we used the general GRL method to replace the Wasserstein distance. The results in Tables 7 and 8 show that the Wasserstein distance achieves a higher accuracy than GRL for all MCSC tasks.

4. To verify the effect of the proposed classifier weighting component, we averaged all the source classifiers to predict the performance of the target domain. Tables 7 and 8 show that the accuracies are lower than those of our model when using the classifier weighting component. This also proves that the classifier weighting component is effective.

*4.3.4. Sensitivity analysis*

In this section, we analyze the sensitivity of the learning parameter for the FDU-MTL and Amazon review datasets. As mentioned in the previous section, parameters $\sigma$ and $\tau$ represent the loss coefficients of domain adaptation and contrastive learning, respectively. For the two datasets, we considered each domain as the target domain to investigate the effect of parameters $\sigma$ and $\tau$ on the performance of the model.

We tested the prediction accuracy for various combinations of $\sigma$=[0.01, 0.1, 0.3, 0.5, 0.7, 1] and $\tau$=[0.01, 0.1, 0.3, 0.5, 0.7, 1], and the results are shown in Figs. 7 and 8. The $z$-axis indicates the accuracy of the test set and the $y$-axis and $x$-axis indicate the values of $\sigma$ and $\tau$, respectively. Noticeably, the accuracy is affected by the coefficients $\sigma$ and $\tau$, and the highest result can be obtained when $\sigma$ and $\tau$ are set to the appropriate values. Therefore, according to the information in Figs. 7 and 8, we selected the parameter pair corresponding to the marked maximum value in the figures as the final values of the parameters $\sigma$ and $\tau$.

## 5. Conclusion and future work

This paper presents a novel approach called CTDA for MCSC tasks. We first propose a mixed selection strategy to select appropriate source domains when there are a large number of available domains and then design a contrastive transformer-based domain adaptation approach to predict the sentiment polarity of the target domain. In particular, domain adaptation based on the Wasserstein distance was applied for shared feature extraction, and a domain discriminator based on contrastive learning was designed for private feature protection. Finally, a classifier weighting component is designed to emphasize the importance of different source classifiers. We performed extensive experiments on the FDU-MTL and Amazon product review datasets. The experimental results show that our approach can effectively select suitable source domains to reduce negative transfer and significantly improve the performance of MCSC.

In the future, we intend to expand the proposed multi-source framework to multimodal cross-domain sentiment classification between texts, speech, and images. Specifically, we intend to develop methods to eliminate domain differences between different signal modes.

## CRediT authorship contribution statement

**Yanping Fu:** Conceptualization, Methodology, Writing – original draft, Software, Visualization, Writing – reviewing and editing. **Yun Liu:** Data curation, Investigation, Validation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported in part by Beijing Nova Program, China (No. Z211100002121120) from Beijing Municipal Science & Technology Commission, China.

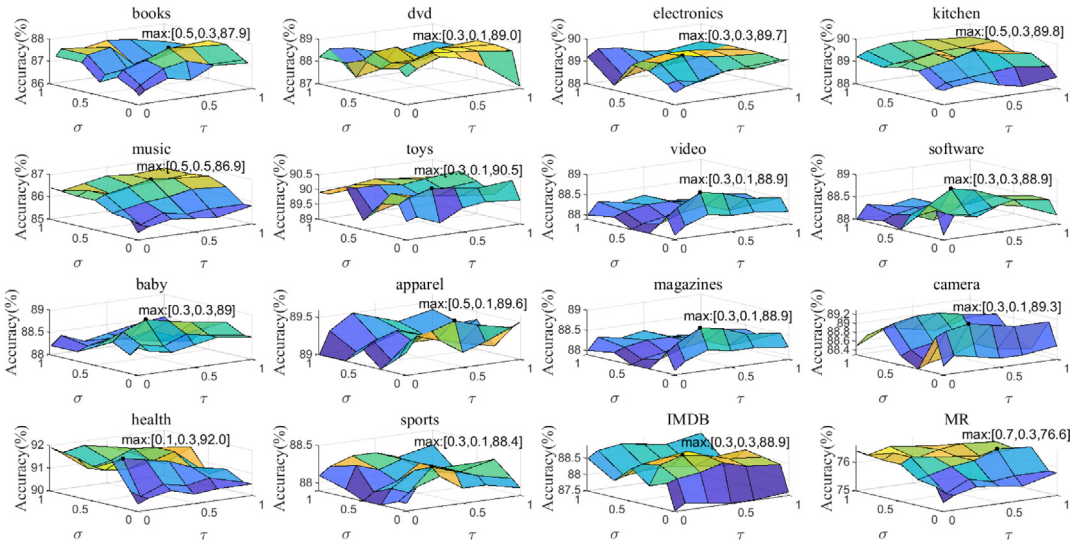



**Fig. 7.** The sensitivity of accuracy value with respect to the parameters $\sigma$ and $\tau$ which change from 0.01 to 1.0 for the FDU-MTL review dataset.

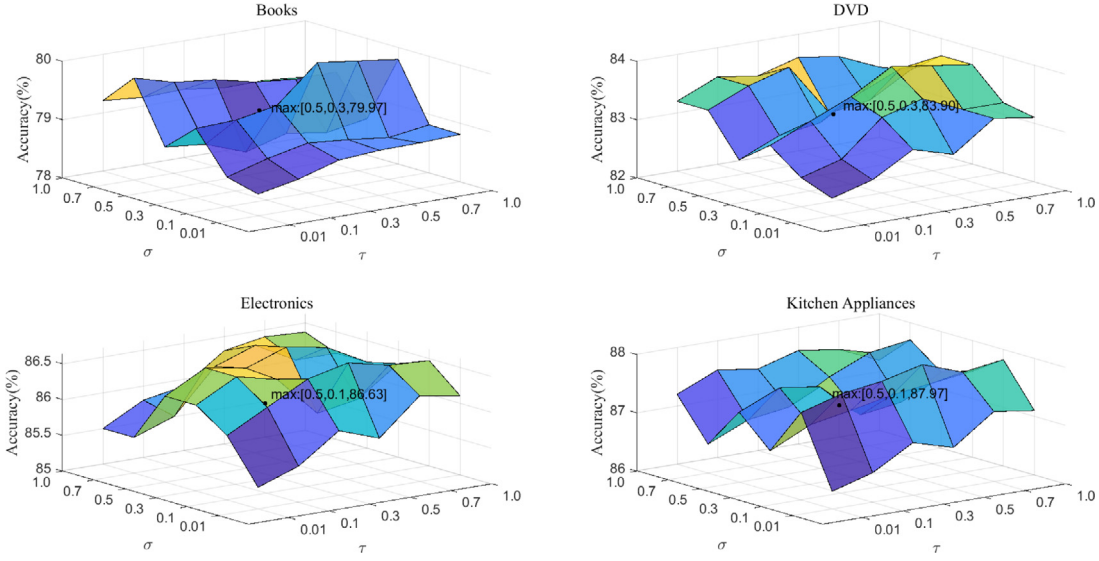



**Fig. 8.** The sensitivity of accuracy value with respect to the parameters $\sigma$ and $\tau$ which change from 0.01 to 1.0 for the Amazon reviews dataset.

## References


[1] V.K. Singh, M. Mukherjee, G.K. Mehta, Combining collaborative filtering and sentiment classification for improved movie recommendations, in: International Workshop on Multi-Disciplinary Trends in Artificial Intelligence, Springer, 2011, pp. 38–50.

[2] Y. Mejova, P. Srinivasan, B. Boynton, GOP primary season on twitter: " popular" political sentiment in social media, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, 2013, pp. 517–526.

[3] X. Chen, C. Cardie, Multinomial adversarial networks for multi-domain text classification, 2018, arXiv preprint arXiv:1802.05694.

[4] J. Guo, D.J. Shah, R. Barzilay, Multi-source domain adaptation with mixture of experts, 2018, arXiv preprint arXiv:1809.02256.

[5] Y. Fu, Y. Liu, Cross-domain sentiment classification based on key pivot and non-pivot extraction, Knowl.-Based Syst. 228 (2021) 107280.

[6] K.-P. Lai, J.C.-S. Ho, W. Lam, Cross-domain sentiment classification using topic attention and dual-task adversarial training, in: International Conference on Artificial Neural Networks, Springer, 2020, pp. 571–583.

[7] K. Zhang, H. Zhang, Q. Liu, H. Zhao, H. Zhu, E. Chen, Interactive attention transfer network for cross-domain sentiment classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, 2019, pp. 5773–5780.

[8] M. Peng, Q. Zhang, Y.-g. Jiang, X.-J. Huang, 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), 2018, pp. 2505–2513.

[9] J. Chen, X. Qiu, P. Liu, X. Huang, Meta multi-task learning for sequence modeling, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018.

[10] M. Chen, Z. Xu, K. Weinberger, F. Sha, Marginalized denoising autoencoders for domain adaptation, 2012, arXiv preprint arXiv:1206.4683.

[11] R. Sharma, P. Bhattacharyya, S. Dandapat, H.S. Bhatt, Identifying transferable information across domains for cross-domain sentiment classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 968–978.

[12] L. Li, W. Ye, M. Long, Y. Tang, J. Xu, J. Wang, Simultaneous learning of pivots and representations for cross-domain sentiment classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 05, 2020, pp. 8220–8227.

[13] S. Gao, H. Li, A cross-domain adaptation method for sentiment classification using probabilistic latent analysis, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 1047–1052.

[14] A. Rietzler, S. Stabinger, P. Opitz, S. Engl, Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification, 2019, arXiv preprint arXiv:1908.11860.

[15] F. Wu, Y. Huang, Z. Yuan, Domain-specific sentiment classification via fusing sentiment knowledge from multiple sources, Inf. Fusion 35 (2017) 26–37.

[16] C. Zhao, S. Wang, D. Li, Multi-source domain adaptation with joint learning for cross-domain sentiment classification, Knowl.-Based Syst. 191 (2020) 105254.
[17] Y. Dai, J. Liu, X. Ren, Z. Xu, Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 05, 2020, pp. 7618–7625.
[18] N. Ponomareva, M. Thelwall, Do neighbours help? an exploration of graph-based algorithms for cross-domain sentiment classification, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 655–665.
[19] C. Gong, J. Yu, R. Xia, Unified feature and instance based domain adaptation for end-to-end aspect-based sentiment analysis, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020, pp. 7035–7045.
[20] K. Zhang, H. Zhang, Q. Liu, H. Zhao, H. Zhu, E. Chen, Interactive attention transfer network for cross-domain sentiment classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, 2019, pp. 5773–5780.
[21] C. Du, H. Sun, J. Wang, Q. Qi, J. Liao, Adversarial and domain-aware BERT for cross-domain sentiment analysis, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4019–4028.
[22] S. Zhang, L. Jiang, H. Peng, Q. Dai, J. Tan, Discriminative representation learning for cross-domain sentiment classification, in: PAKDD (2), Springer, 2021, pp. 54–66.
[23] T. Li, X. Chen, S. Zhang, Z. Dong, K. Keutzer, Cross-domain sentiment classification with contrastive learning and mutual information maximization, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 8203–8207.
[24] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (1) (2016) 2096-2030.
[25] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7167–7176.
[26] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, PMLR, 2015, pp. 97–105.
[27] J. Zhuo, S. Wang, W. Zhang, Q. Huang, Deep unsupervised convolutional domain adaptation, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 261–269.
[28] Y. Du, M. He, L. Wang, H. Zhang, Wasserstein based transfer network for cross-domain sentiment classification, Knowl.-Based Syst. 204 (2020) 106162.
[29] M. Ghifary, W.B. Kleijn, M. Zhang, D. Balduzzi, W. Li, Deep reconstruction-classification networks for unsupervised domain adaptation, in: European Conference on Computer Vision, Springer, 2016, pp. 597–613.
[30] T. Li, X. Chen, S. Zhang, Z. Dong, K. Keutzer, Cross-domain sentiment classification with in-domain contrastive learning, 2020, arXiv preprint arXiv:2012.02943.
[31] T.-T. Vu, D. Phung, G. Haffari, Effective unsupervised domain adaptation with adversarially trained language models, 2020, arXiv preprint arXiv:2010.01739.
[32] D. Bollegala, D. Weir, J. Carroll, Cross-domain sentiment classification using a sentiment sensitive thesaurus, IEEE Trans. Knowl. Data Eng. 25 (8) (2012) 1719–1731.
[33] D. Bollegala, D. Weir, J.A. Carroll, Learning to predict distributions of words across domains, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 613–623.
[34] J. Yu, J. Jiang, Learning Sentence Embeddings with Auxiliary Tasks for Cross-Domain Sentiment Classification, Association for Computational Linguistics, 2016.
[35] Z. Li, Y. Wei, Y. Zhang, Q. Yang, Hierarchical attention transfer network for cross-domain sentiment classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018.
[36] Z. Li, Y. Zhang, Y. Wei, Y. Wu, Q. Yang, End-to-end adversarial memory network for cross-domain sentiment classification, in: IJCAI, 2017, pp. 2237–2243.
[37] C. Louizos, K. Swersky, Y. Li, M. Welling, R. Zemel, The variational fair autoencoder, 2015, arXiv preprint arXiv:1511.00830.
[38] K.M. Borgwardt, A. Gretton, M.J. Rasch, H.-P. Kriegel, B. Schölkopf, A.J. Smola, Integrating structured biological data by kernel maximum mean discrepancy, Bioinformatics 22 (14) (2006) e49–e57.
[39] H.S. Bhatt, A. Rajkumar, S. Roy, Multi-source iterative adaptation for cross-domain classification, in: IJCAI, 2016, pp. 3691–3697.
[40] M. Yang, Y. Shen, X. Chen, C. Li, Multi-source domain adaptation for sentiment classification with granger causal inference, in: Proceedings of the 43rd International Acm Sigir Conference on Research and Development in Information Retrieval, 2020, pp. 1913–1916.
[41] Y. Dai, J. Liu, J. Zhang, H. Fu, Z. Xu, Unsupervised sentiment analysis by transferring multi-source knowledge, 2021, arXiv preprint arXiv:2105.11902.
[42] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[43] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019, arXiv preprint arXiv:1910.01108.
[44] J. Zhang, W. Zhou, X. Chen, W. Yao, L. Cao, Multisource selective transfer framework in multiobjective optimization problems, IEEE Trans. Evol. Comput. 24 (3) (2019) 424–438.
[45] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
[46] J. Blitzer, M. Dredze, F. Pereira, Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 440–447.
[47] P. Liu, X. Qiu, X. Huang, Adversarial multi-task learning for text classification, 2017, arXiv preprint arXiv:1704.05742.
[48] J.M. Giorgi, O. Nitski, G.D. Bader, B. Wang, Declutr: Deep contrastive learning for unsupervised textual representations, 2020, arXiv preprint arXiv:2006.03659.
[49] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of wasserstein gans, 2017, arXiv preprint arXiv:1704.00028.
[50] Z.-H. Zhou, M. Li, Tri-training: Exploiting unlabeled data using three classifiers, IEEE Trans. Knowl. Data Eng. 17 (11) (2005) 1529–1541.