TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 12/15 pp360–369
DOI: 10.26599/TST.2022.9010007
Volume 28, Number 2, April 2023

CNN-Based Broad Learning for Cross-Domain Emotion Classification

Rong Zeng, Hongzhan Liu*, Sancheng Peng*, Lihong Cao, Aimin Yang, Chengqing Zong, and Guodong Zhou

Abstract: Cross-domain emotion classification aims to leverage useful information in a source domain to help predict emotion polarity in a target domain in a unsupervised or semi-supervised manner. Due to the domain discrepancy, an emotion classifier trained on source domain may not work well on target domain. Many researchers have focused on traditional cross-domain sentiment classification, which is coarse-grained emotion classification. However, the problem of emotion classification for cross-domain is rarely involved. In this paper, we propose a method, called convolutional neural network (CNN) based broad learning, for cross-domain emotion classification by combining the strength of CNN and broad learning. We first utilized CNN to extract domain-invariant and domain-specific features simultaneously, so as to train two more efficient classifiers by employing broad learning. Then, to take advantage of these two classifiers, we designed a co-training model to boost together for them. Finally, we conducted comparative experiments on four datasets for verifying the effectiveness of our proposed method. The experimental results show that the proposed method can improve the performance of emotion classification more effectively than those baseline methods.

Key words: cross-domain emotion classification; CNN; broad learning; classifier; co-training

1 Introduction

Cross-domain emotion classification (CDEC) aims to utilize useful knowledge in source domain (with sufficient labeled data) to conduct emotion classification in target domain (with few or no labeled data). CDEC is an important task in domain adaptation and emotion classification, which attracts more and more attention from both academia and industry. However, due to the domain discrepancy, an emotion classifier trained on source domain may not works well on target domain. This has motivated some research on domain adaptation method for cross-domain emotion classification.

In recent years, many researchers have focused on traditional cross-domain sentiment classification, and have proposed many methods which divided into two types: The first kind of sentiment classification method is proposed by extracting domain-invariant features[1–7], the other is proposed by extracting domain-invariant and domain-specific features[8–12]. As to the first kind of method, the major limitation of it is that methods can not consider the domain-specific features. As to the other kind of method, although methods consider both domain-invariant and domain-specific features, they may fall into the problem of non-convergence and model collapse, due

• Rong Zeng and Hongzhan Liu are with the Guangdong Provincial Key Laboratory of Nanophotonic Functional Materials and Devices, South China Normal University, Guangzhou 511400, China. E-mail: zengrong980302@163.com; lhzscnu@163.com.
• Sancheng Peng and Lihong Cao are with the Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510006, China. E-mail: psc346@aliyun.com; 201610130@oamail.gdufs.edu.cn.
• Aimin Yang is with the School of Computer Science and Intelligence Education, Lingnan Normal University, Guangzhou 510006, China. E-mail: amyang18@163.com.
• Chengqing Zong is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: cqzong@nlpr.ia.ac.cn.
• Guodong Zhou is with the School of Computer Science and Technology, Soochow University, Suzhou 215031, China. E-mail: gdzhou@suda.edu.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2021-12-22; revised: 2022-03-15; accepted: 2022-03-17

to adoption of generative adversarial networks[13].

Despite their promising results, existing methods for domain adaptation are only exploring the binary sentiment classification[14, 15], which is coarse-grained emotion classification. Fine-grained emotion classification[16] focuses on extracting automatically the emotional states (i.e., happy, moving, angry, sad, fear, disgusted, and surprise) of users from different domains like hotel and movie reviews. However, existing methods for emotion classification only addressing the performance in a single domain, while ignoring cross-domain emotion classification.

To effectively address the above problems, in this study, we proposed a novel method to classify cross-domain emotions by extracting domain-invariant feature (DIF) and domain-specific feature (DSF) simultaneously. Then, we trained a classifier based on DIF with the labeled data from both source and target domain, and train another classifier based on DSF with only a small number of target domain labeled data. Finally, these two classifiers are boosted each other to co-train the target domain unlabeled data.

In summary, the main contributions of this paper are listed as follows: (1) A novel transfer learning method is proposed to address the task of cross-domain emotion classification, by combining deep learning and broad learning models, called convolutional neural network (CNN) based broad learning (CBL). (2) Four real-world datasets have been developed, in which four different domains from the public Chinese E-commerce platforms are involved. (3) The extensive experiments on four real-world datasets have been conducted. The results show that our proposed method can improve the performance of emotion classification more effectively than those baseline methods.

2 Related Work

In this section, we investigate related work in three dimensions. The first dimension is related to cross-domain sentiment classification methods; the second is related to the emotion methods; and the last is related to broad learning models.

(1) Cross-domain sentiment classification

To solve the problem of cross-domain sentiment classification, some methods were proposedby extracting the domain-invariant features. However, these methods do not consider domain-specific features[1–7]. In addition, some methods were also proposed by extracting both the domain-invariant and domain-specific features[8–12]. Despite their promising results, these methods are easy to fall into the problems of non-convergence and model collapse because of using generative adversarial networks.

(2) Emotion classification

In recent years, most of the existing methods on emotion classification are based on deep learning, for example, Kulshreshtha et al.[17] used CNNs, Mohammadi et al.[18] used long short term memory (LSTM), Abdul-Mageed and Ungar[19] and Tafreshi and Diab[20] used gated recurrent neural networks (GRNN), Yu et al.[21] used attention mechanism, and Ahmad et al.[22] used bi-directional long short term memory (Bi-LSTM) for classifying emotions. These methods based on deep learning can effectively overcome the shortcomings of the methods based on dictionary and machine learning to a certain extent. However, they mainly focused on single-domain emotion classification task, few of them involve domain adaptation.

(3) Broad learning

As a flat network, BL[23] has many advantages, such as simple structure, short training time, and good generalization performance, which has become an alternative solution for deep learning. At present, some researchers have used BL for practical applications, such as nonlinear system identification[24], target location[25], image processing[26], automatic control[27], fault diagnosis[28], and emotion classification[29]. However, few researchers have used BL for emotion classification of text. Thus, we attempted to use BL for the task of cross-domain emotion classification.

Different from these methods, we introduced maximum mean discrepancy (MMD)[30], deep learning, and broad learning to concurrently extract DIF and DSF, and train two classifiers based on DIF and DSF.

3 Approach

In this section, we first provide the problem definition and introduce MMD metric which can be utilized to measure discrepancy between the probability distribution of any two random variables. Then, we describe how our proposed method obtains DIF and DSF. At last, we further describe how to fuse these two features by using co-training model. The framework of the proposed method is shown in Fig. 1.

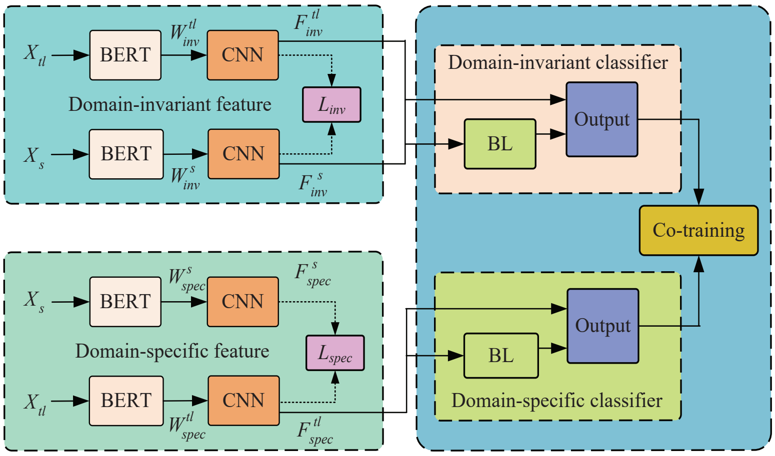


Fig. 1 The framework of the proposed method. BERT indicates bidirectional encoder representation from transformers.

3.1 Problem definition

Two domains D_s and D_t are given to denote a source domain and a target domain, respectively. Suppose that we have a set of labeled data $X_s = \{x^i_s, y^i_s\}^{N_s}_{i=1}$ in source domain, where x^i_s and y^i_s denote the text and label of the i-th sample, respectively, and N_s denotes the total number of labeled data. In addition, suppose that we also have a set of unlabeled data $X_{tu} = \{x^i_{tu}\}^{N_{tu}}_{i=1}$ and some labeled data $X_{tl} = \{x^i_{tl}, y^i_{tl}\}^{N_{tl}}_{i=1}$ in target domain, where x^i_{tu} denotes the text of i-th unlabeled sample in target domain, N_{tu} denotes the total number of unlabeled data, x^i_{tl} and y^i_{tl} denote the text and label of and i-th labeled sample in target domain and N_{tl} denotes the total number of labeled data in target domain. The goal of cross-domain emotion classification is to train a robust classifier using the labeled data in source domain, and then to adopt it for predicting the unlabeled target data.

3.2 Maximum mean discrepancy

MMD was proposed by Borgwardt et al.[30] to measure the probability distribution discrepancy between any two random variables. As an effective non-parametric distribution discrepancy metric, MMD has been widely used to reduce domain shift in domain adaptation.

Given two domains D_s and D_t, MMD is defined as follows:

$$\mathrm{MMD}(X_s, X_t) = \left\| \frac{1}{N_s}\sum_{i=1}^{N_s}\phi(x^i_s) - \frac{1}{N_t}\sum_{i=1}^{N_t}\phi(x^i_t) \right\|^2_{\mathcal{H}} \tag{1}$$

where $\|\cdot\|^2$ denotes 2-norm, $\mathcal{H}$ denotes the reproducing kernel Hilbert space (RKHS), ϕ represents the kernel function that maps the original data to $\mathcal{H}$, and $\|\phi\|_{\mathcal{H}} \leqslant 1$ defines a set of functions in $\mathcal{H}$.

3.3 Feature extraction

To simultaneously extract DIF and DSF, it needs to map data into a domain-invariant space and a domain-specific space using two different mappers.

We first used bidirectional encoder representation from transformers (BERT) to generate word vectors for X_s and X_{tl}, which were described as follows:

$$\begin{aligned}
\boldsymbol{W}^s_{inv} &= \mathrm{BERT}_{inv}(X_s;\ \boldsymbol{\theta}^{\mathrm{BERT}}_{inv}) \in \mathbf{R}^{(N_s l)\times 768},\\
\boldsymbol{W}^{tl}_{inv} &= \mathrm{BERT}_{inv}(X_{tl};\ \boldsymbol{\theta}^{\mathrm{BERT}}_{inv}) \in \mathbf{R}^{(N_{tl} l)\times 768},\\
\boldsymbol{W}^s_{spec} &= \mathrm{BERT}_{spec}(X_s;\ \boldsymbol{\theta}^{\mathrm{BERT}}_{spec}) \in \mathbf{R}^{(N_s l)\times 768},\\
\boldsymbol{W}^{tl}_{spec} &= \mathrm{BERT}_{spec}(X_{tl};\ \boldsymbol{\theta}^{\mathrm{BERT}}_{spec}) \in \mathbf{R}^{(N_{tl} l)\times 768}
\end{aligned} \tag{2}$$

where BERT_{inv} and BERT_{spec} denote the BERT of DIFs and DSFs, $\boldsymbol{\theta}^{\mathrm{BERT}}_{inv}$ and $\boldsymbol{\theta}^{\mathrm{BERT}}_{spec}$ denote their corresponding parameters, l denotes the length of text, $\boldsymbol{W}^s_{inv}$ and $\boldsymbol{W}^{tl}_{inv}$ denote word vector generated by BERT_{inv} for X_s and X_{tl}, respectively, and $\boldsymbol{W}^s_{spec}$ and $\boldsymbol{W}^{tl}_{spec}$ denote word vector generated by BERT_{spec} for X_s and X_{tl}, respectively.

Then, based on these representations, we extracted n-gram feature and salient feature by conducting the convolution and max pooling operations of CNN, respectively. X_s and X_{tl} can be described as follows:

$$\begin{aligned}
\boldsymbol{F}^s_{inv} &= \mathrm{CNN}_{inv}(\boldsymbol{W}^s_{inv};\ \boldsymbol{\theta}^{\mathrm{CNN}}_{inv}) \in \mathbf{R}^{N_s\times q},\\
\boldsymbol{F}^{tl}_{inv} &= \mathrm{CNN}_{inv}(\boldsymbol{W}^{tl}_{inv};\ \boldsymbol{\theta}^{\mathrm{CNN}}_{inv}) \in \mathbf{R}^{N_{tl}\times q},\\
\boldsymbol{F}^s_{spec} &= \mathrm{CNN}_{spec}(\boldsymbol{W}^s_{spec};\ \boldsymbol{\theta}^{\mathrm{CNN}}_{spec}) \in \mathbf{R}^{N_s\times q},\\
\boldsymbol{F}^{tl}_{spec} &= \mathrm{CNN}_{spec}(\boldsymbol{W}^{tl}_{spec};\ \boldsymbol{\theta}^{\mathrm{CNN}}_{spec}) \in \mathbf{R}^{N_{tl}\times q}
\end{aligned} \tag{3}$$

where CNN_{inv} and CNN_{spec} denote the CNN of DIF and DSF, $\boldsymbol{\theta}^{\mathrm{CNN}}_{inv}$ and $\boldsymbol{\theta}^{\mathrm{CNN}}_{spec}$ denote their corresponding parameters, q denotes the dimension of convolution output, $\boldsymbol{F}^s_{inv}$ and $\boldsymbol{F}^{tl}_{inv}$ denote n-gram feature extracted by CNN_{inv} for $\boldsymbol{W}^s_{inv}$ and $\boldsymbol{W}^{tl}_{inv}$, respectively, and $\boldsymbol{F}^s_{spec}$

and $\boldsymbol{F}_{spec}^{tl}$ denote n-gram feature extracted by CNN$_{spec}$ for $\boldsymbol{W}_{spec}^{s}$ and $\boldsymbol{W}_{spec}^{tl}$, respectively.

As to DIF, we hoped it can encode features shared by both source and target domains. From the probability distribution view, we hoped that the distributions of the mapped outputs obtained by DIF from source and target data are similar. Thus, we utilized MMD regularizer onto the features of source data $\boldsymbol{F}_{inv}^{s}$ and target data $\boldsymbol{F}_{inv}^{tl}$. The corresponding loss was defined as follows:

$$L_{sim} = \text{MMD}\left(\boldsymbol{F}_{inv}^{s}, \boldsymbol{F}_{inv}^{tl}\right) \tag{4}$$

The distribution discrepancy between $\boldsymbol{F}_{inv}^{s}$ and $\boldsymbol{F}_{inv}^{tl}$ can be reduced by minimizing the loss L_{sim}, which contributes to obtain DIF.

As to DSF, we hope that it extracts features only from target domain. That is, these features should usually emerge in target domain while hardly emerge in source domain. Thus, we utilized MMD regularizer onto the features of source data $\boldsymbol{F}_{spec}^{s}$ and target data $\boldsymbol{F}_{spec}^{tl}$. The corresponding loss was defined as follows:

$$L_{diff} = -\text{MMD}\left(\boldsymbol{F}_{spec}^{s}, \boldsymbol{F}_{spec}^{tl}\right) \tag{5}$$

The distribution discrepancy between $\boldsymbol{F}_{spec}^{s}$ and $\boldsymbol{F}_{spec}^{tl}$ can be enlarged by minimizing loss L_{diff}, which contributes to obtain DSF.

3.4 BL-Based classifier

To further mine the semantic information of text, we introduced BL to capture the deep semantic features for both source and target domains data, and design two classifiers including domain-invariant classifier (DIC) based on DIF and domain-specific classifier (DSC) based on DSF.

As to DIC, $\boldsymbol{F}_{inv}^{s}$ and $\boldsymbol{F}_{inv}^{tl}$ are concatenated and nonlinearly mapped into n_{inv} groups of enhanced nodes. Thus, the i-th group of enhanced nodes can be represented as follows:

$$\boldsymbol{E}_{inv}^{i} = \varphi\left(\boldsymbol{\theta}_{inv}^{i}\left[\boldsymbol{F}_{inv}^{s}, \boldsymbol{F}_{inv}^{tl}\right] + \boldsymbol{\beta}_{inv}^{i}\right), \quad i = 1, 2, \ldots, n_{inv} \tag{6}$$

where φ denotes nonlinear activation function, and $\boldsymbol{\theta}_{inv}^{i}$ and $\boldsymbol{\beta}_{inv}^{i}$ denote weight matrix and bias matrix generated randomly, respectively.

Let $\boldsymbol{E}_{inv} \triangleq \left[\boldsymbol{E}_{inv}^{1}, \boldsymbol{E}_{inv}^{2}, \ldots, \boldsymbol{E}_{inv}^{n_{inv}}\right]$ denote the matrix of n_{inv} groups of enhancement nodes, which represents all nonlinear features of text data. Thus, the output of DIC can be represented as follows:

$$\hat{\boldsymbol{Y}}_{inv} = \left[\boldsymbol{F}_{inv}^{s}, \boldsymbol{F}_{inv}^{tl}, \boldsymbol{E}_{inv}\right]\boldsymbol{\theta}_{inv}^{\text{BL}} = \boldsymbol{A}_{inv}\boldsymbol{\theta}_{inv}^{\text{BL}} \tag{7}$$

where $\boldsymbol{A}_{inv}$ denotes all the input features of DIC and $\boldsymbol{\theta}_{inv}^{\text{BL}}$ denotes the weight of output layer of DIC.

As DSC only needs to classify target domain data, we conducted nonlinear mapping for $\boldsymbol{F}_{spec}^{tl}$ into n_{spec} groups of enhanced nodes. Thus, the j-th group of enhanced nodes can be represented as follows:

$$\boldsymbol{E}_{spec}^{j} = \varphi\left(\boldsymbol{\theta}_{spec}^{j}\boldsymbol{F}_{spec}^{tl} + \boldsymbol{\beta}_{spec}^{j}\right), \quad j = 1, 2, \ldots, n_{spec} \tag{8}$$

where $\boldsymbol{\theta}_{spec}^{j}$ and $\boldsymbol{\beta}_{spec}^{j}$ denote weight matrix and bias matrix generated randomly, respectively.

Let $\boldsymbol{E}_{spec} \triangleq \left[\boldsymbol{E}_{spec}^{1}, \boldsymbol{E}_{spec}^{2}, \ldots, \boldsymbol{E}_{spec}^{n_{spec}}\right]$ denote the matrix of n_{spec} groups of enhancement nodes, which represents all nonlinear features of text data. Thus, the output of DSC can be represented as follows:

$$\hat{\boldsymbol{Y}}_{spec} = \left[\boldsymbol{F}_{spec}^{tl}, \boldsymbol{E}_{spec}\right]\boldsymbol{\theta}_{spec}^{\text{BL}} = \boldsymbol{A}_{spec}\boldsymbol{\theta}_{spec}^{\text{BL}} \tag{9}$$

where $\boldsymbol{A}_{spec}$ denotes all the input features of DSC and $\boldsymbol{\theta}_{spec}^{\text{BL}}$ denotes the weight of output layer of DSC.

3.5 Co-training

The co-training process of this method was divided into two steps: (1) updating for the corresponding parameters for DIF and DSF; (2) updating for the classifiers DIC and DSC. As to DIF, the purpose of training is to minimize the following loss with respect to parameters $\boldsymbol{\theta}_{inv}^{\text{BERT}}$ and $\boldsymbol{\theta}_{inv}^{\text{CNN}}$:

$$L_{inv} = L_{sim}\left(\boldsymbol{\theta}_{inv}^{\text{BERT}}, \boldsymbol{\theta}_{inv}^{\text{CNN}}\right) + \alpha L_{c}\left(\boldsymbol{\theta}_{inv}^{\text{BERT}}, \boldsymbol{\theta}_{inv}^{\text{CNN}}\right) \tag{10}$$

where α denotes the weight of loss L_c, $L(\boldsymbol{\theta})$ denotes that L_{inv} is optimized on the parameter $\boldsymbol{\theta}$ during training, and L_c denotes the classification loss on domain-invariant feature, which was used to measure the difference between real label and predicted label of source domain and target domain. Thus, it can be represented as follows:

$$L_c = \frac{1}{N_s + N_{tl}}\sum_{i=1}^{N_s} -y_s^i \ln P\left(y_s^i \left|\boldsymbol{F}_{inv}^{si}\right.\right) + \frac{1}{N_s + N_{tl}}\sum_{j=1}^{N_{tl}} -y_{tl}^j \ln P\left(y_{tl}^j \left|\boldsymbol{F}_{inv}^{tlj}\right.\right) \tag{11}$$

where P denotes the probability, and $\boldsymbol{F}_{inv}^{si}$ and $\boldsymbol{F}_{inv}^{tlj}$ denote the one-hot encoding of the class label for the i-th source example $\boldsymbol{F}_{inv}^{s}$ and the j-th target example $\boldsymbol{F}_{inv}^{tl}$, respectively.

At to DSF, the purpose of training was to minimize the following loss with respect to parameters $\boldsymbol{\theta}_{spec}^{\text{BERT}}$ and $\boldsymbol{\theta}_{spec}^{\text{CNN}}$:

$$L_{spec} = L_{diff}\left(\boldsymbol{\theta}_{spec}^{\text{BERT}}, \boldsymbol{\theta}_{spec}^{\text{CNN}}\right) + \gamma L_t\left(\boldsymbol{\theta}_{spec}^{\text{BERT}}, \boldsymbol{\theta}_{spec}^{\text{CNN}}\right) \tag{12}$$

where γ denotes the weight of loss L_t, which is used to measure real label and predicted label of target domain. Thus, it can be represented as follows:

$$L_t = \frac{1}{N_{tl}} \sum_{i=1}^{N_{tl}} -\boldsymbol{Y}_{tl}^{i} \log fc\left(y_{tl}^{i} \middle| \boldsymbol{F}_{spec}^{tli}\right) \tag{13}$$

As to DIC, we needed to solve for an appropriate $\boldsymbol{\theta}_{inv}^{\text{BERT}}$ so that the difference between $\boldsymbol{Y}_{inv}$ and $\hat{\boldsymbol{Y}}_{inv}$ was as small as possible, where $\boldsymbol{Y}_{inv}$ denotes the ground truth label of labeled source data and labeled target data. Therefore, the ridge regression was adopted as objective function, which is expressed as follows:

$$\underset{\boldsymbol{\theta}_{inv}^{\text{BL}}}{\operatorname{argmin}} \left(\left\|\boldsymbol{Y}_{inv} - \hat{\boldsymbol{Y}}_{inv}\right\|_2^2 + \lambda_1 \left\|\boldsymbol{\theta}_{inv}^{\text{BL}}\right\|_2^2\right) \tag{14}$$

where λ_1 denotes regularization parameter.

As to DSC, similarly, the objective function was represented as follows:

$$\underset{\boldsymbol{\theta}_{spec}^{\text{BL}}}{\operatorname{argmin}} \left(\left\|\boldsymbol{Y}_{spec} - \hat{\boldsymbol{Y}}_{spec}\right\|_2^2 + \lambda_2 \left\|\boldsymbol{\theta}_{spec}^{\text{BL}}\right\|_2^2\right) \tag{15}$$

where λ_2 denotes regularization parameter and $\boldsymbol{Y}_{spec}$ denotes the ground truth label of labeled target data.

Thus, ridge regression was adopted as the objective function to obtain optimal solution $\boldsymbol{\theta}_{inv}^{\text{BL}}$, which is represented as follows:

$$\boldsymbol{\theta}_{inv}^{\text{BL}} = \left(\lambda_1 \boldsymbol{I} + \boldsymbol{A}_{inv}\boldsymbol{A}_{inv}^{\text{T}}\right)^{-1} \boldsymbol{A}_{inv}^{\text{T}} \boldsymbol{Y}_{inv} \tag{16}$$

where $\boldsymbol{I}$ denotes identity matrix.

Similarly, to obtain optimal solution $\boldsymbol{\theta}_{spec}^{\text{BL}}$, which is represented as follows:

$$\boldsymbol{\theta}_{spec}^{\text{BL}} = \left(\lambda_2 \boldsymbol{I} + \boldsymbol{A}_{spec}\boldsymbol{A}_{spec}^{\text{T}}\right)^{-1} \boldsymbol{A}_{spec}^{\text{T}} \boldsymbol{Y}_{spec} \tag{17}$$

Here, we regarded the domain-invariant feature and the domain-specific feature as two independent spaces for target domain data. Based on the domain-invariant feature, we trained DIC with respect to parameter $\boldsymbol{\theta}_{inv}^{\text{BL}}$. In addition, based on the domain-specific feature, we train DSC with respect to parameter $\boldsymbol{\theta}_{spec}^{\text{BL}}$. Because the distribution of source data was compatible with that of the target data in domain-invariant space, we employed both enriched source labels and small amount of target labels to train DIC and utilized only the target labels to train DSC. The entire training process is described in Algorithm 1.

4 Experiments

4.1 Datasets

Since there is a lack of public datasets for cross-domain emotion classification, we had collected product reviews as the experimental datasets from the well-known Chinese E-commerce platforms, such as Douban, Ctrip, Jingdong, and Taobao. These datasets contain four domains: Clothing (C), electronics (E), hotel (H), and movie (M). There is a total number of 41 628 reviews for the four datasets, and there are seven emotion classes, including happy, moving, angry, sad, fear, disgusted, and surprise. The detailed statistics are shown in Table 1.

According to these data, we constructed 12 cross-domain seven classification tasks. In each domain adaptation task, there are 1000 labeled source domain examples, 1000 unlabeled target domain examples, and 50 labeled target domain examples for training. In addition, to effectively fine-tune the hyper-parameters, we randomly selected 500 target examples as developing data set, and treated the remainder examples as testing

Algorithm 1 Co-training process for domain adaptation

Input: Labeled source domain data X_s, labeled target domain data X_{tl}, unlabeled target domain data X_{tu};

Output: $\boldsymbol{\theta}_{inv}^{\text{BL}}$ and $\boldsymbol{\theta}_{spec}^{\text{BL}}$;

Repeat
    Obtain DIF based on X_s and X_{tl};
    Train DIC based on $\boldsymbol{F}_{inv}^{s}$ and $\boldsymbol{F}_{inv}^{tl}$;
    Utilize DIC to predict X_{tu};
    Select p samples X_{tu}^{inv} with the highest confidence from X_{tu};
    Obtain DSF based on X_s and X_{tl};
    Train DSC based on $\boldsymbol{F}_{spec}^{tl}$;
    Utilize DSC to predict X_{tu};
    Select p samples X_{tu}^{spec} with the highest confidence from X_{tu};
    Delete samples $X_{tu}^{inv} \cup X_{tu}^{spec}$ from X_{tu}, and add them to X_{tl};
until stopping criterion is met

Table 1 Statistics of different emotion samples on four datasets.

Domain name	Number of classes							Total
	Happy	Moving	Sad	Angry	Fear	Disgusted	Surprise	
Clothing	1699	1035	1819	1483	1014	2014	1237	10 301
Electronics	2316	1453	1727	1602	1405	1405	1410	11 318
Hotel	1426	1498	1661	1537	1175	1456	1256	10 009
Movie	1500	1464	1375	1669	1337	1145	1510	10 000

data set. All of the baseline methods and CBL adopt this setting.

4.2 Baseline methods

We compared our proposed method with the following baselines for cross-domain emotion classification, including Bi-LSTM[31], TextCNN[32], DANN[1], and PBLM[33].

- Bi-LSTM is a non-domain-adaptive method which is trained in source domain and predicts in target domain directly. It is designed based on Bi-LSTM.
- TextCNN is a convolutional neural network for emotion classification without considering contextual information of text.
- DANN is a method which exploits a domain classifier to minimize the discrepancy between two domains via adversarial training manner.
- PBLM is a representation learning model which exploits the structure of the input text.

4.3 Implementation detail

In our experiments, all word embeddings from sentences are initialized as 768-dimension vectors by BERT. CBL was implemented with the Chinese BERT pretrained model proposed by Cui et al.[34], which is composed of 12 transformer blocks and is pre-trained on large number of Chinese corpus (including Chinese Wikipedia, news, Q&A, etc.). The convolution kernel size of CNN is set to two, three, and four, and the dimension of convolution kernel is set to 100. The enhancement nodes of BL consist of 20 groups (i.e., $n_{inv} = n_{spec} = 20$), 50 nodes in each group (i.e., $r_{inv} = r_{spec} = 50$), and tanh activation functions.

Model optimization was implemented using the AdamW update strategy[35] with initial learning rate set to 2×10^{-5} and weights decay set to 0.01. The bandwidth and number of the RBF of MMD are set to 5 and 10, respectively. The hyperparameters corresponding to the best performance on the validation set are obtained by grid search, and the weight of the loss items set $\alpha = \gamma = 0.5$, the factor for each iteration of co-training is set $p = 5$, and the regularization parameter is set $\lambda_1 = \lambda_2 = 0.001$.

4.4 Main results

To demonstrate the effectiveness of CBL, we compared it with other state-of-the-art methods on the task of cross-domain emotion classification. To demonstrate the effects of domain adaptation more intuitively, we also compared CBL with a variant without domain-specific feature, denoted by CBL-s. We utilized classification accuracy to evaluate the models. The experimental results of all methods are shown in Table 2.

Table 2 shows the classification accuracy of different methods on the benchmark datasets. We evaluated our method over 12 transfer pairs, on a total number of 41 628 testing samples. Our proposed CBL consistently achieved the best performance on most of tasks. Bi-LSTM only achieves 56.61% on average because of its poor extraction ability of domain adaptation features. Compared to TextCNN, DANN, and PBLM, CBL outperforms TextCNN by 10.53%, DANN by 7.31%, and PBLM by 5.86% on average, respectively. The possible reasons are that CBL can better exploit automatically domain-invariant and domain-specific representations with BERT, CNN, and BL, which contribute to extract the domain-invariant and domain-specific features.

The average accuracy of CBL-s is 84.56%, which is 1.85% lower than CBL. Through the comparison

Table 2 The accuracy of cross-domain emotion classification by different methods.

Source→target	Accuracy (%)					
	Bi-LSTM	TextCNN	DANN	PBLM	CBL-s	CBL
C→M	45.57	68.11	71.18	73.43	77.14	**80.09**
H→M	63.81	78.29	80.51	80.39	84.11	**85.34**
E→M	51.19	70.30	74.20	77.63	82.54	**83.16**
C→H	58.97	78.80	82.80	85.74	91.22	**92.59**
M→H	61.96	86.92	88.55	87.93	90.61	**92.77**
E→H	55.65	79.31	79.46	81.25	83.85	**86.09**
C→E	60.93	83.12	87.02	88.31	91.28	**92.53**
M→E	51.61	73.02	81.53	82.90	89.34	**92.01**
H→E	60.91	80.56	84.42	83.86	87.21	**90.47**
E→C	59.38	77.31	79.37	80.89	84.50	**84.98**
M→C	53.90	71.76	72.68	72.77	78.90	**80.06**
H→C	55.41	63.01	67.43	71.52	74.02	**76.80**
Average	56.61	75.88	79.10	80.55	84.56	**86.41**

between CBL-s and CBL, we can conclude that CBL can better improve the accuracy of emotion classification.

4.5 Effect of the number of labeled target data

In this subsection, we also provided an analysis under the effect of the number of labeled target data to our proposed CBL. In Figs. 2 and 3, we show the comparison between baseline method DANN and CBL under a setting that some labeled target data were randomly selected, and then mixed them with the training data. Here, we provide the experimental results on two transferring methods, while a similar trend can be obtained in other pairs. The influence on accuracy from the different number of labeled target data on the task H→E and M→C are shown in Figs. 2 and 3, respectively.

From Figs. 2 and 3, it can be seen that with the number of randomly selected labeled target data increasing, the difference between the two methods almost remains unchanged, and CBL can also keep the advantageous position. These trends show that our proposed CBL is more effective with little labeled target data and can further benefit from more labeled target data. We can easily observe that CBL can show continuously better results during the whole training process.

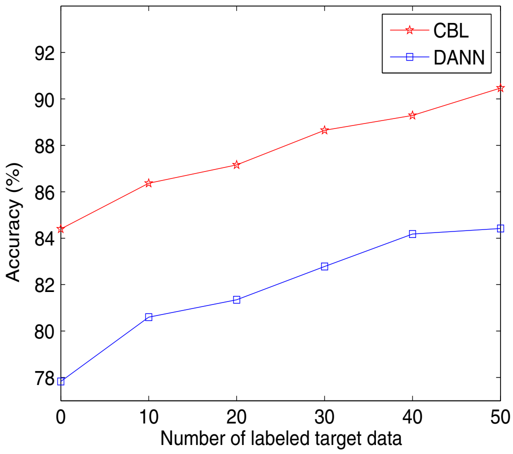


Fig. 2 The influence on accuracy from the different numbers of labeled target data on the task H→E.

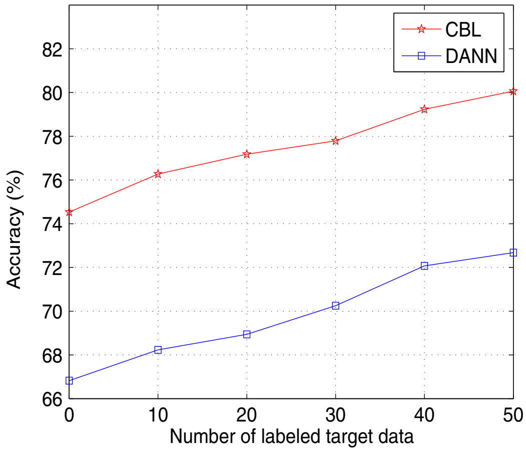


Fig. 3 The influence on accuracy from the different numbers of labeled target data on the task M→C.

5 Conclusion

In this paper, we discussed the significance of CNN and BL for cross-domain emotion classification. Comparing most of the previous methods, which usually extract DIF and DSF by utilizing deep learning models, we showed that the combination of CNN and BL could contribute to be utilized in domain adaptation task. Specifically, we proposed a novel method to simultaneously extract domain-invariant feature and domain-specific feature for target domain data, based on the MMD metric, CNN, and BL. With these two different features, we conducted co-training with labeled data from source domain and a few labeled data from target domain. The extensive experiments demonstrated the superiority of our proposed method CBL. In our future work, we will plan to apply our method for other domain adaptation tasks, and consider to classify multi-source cross-domain emotions.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (No. 61876205), the Natural Science Foundation of Guangdong (No. 2021A1515012652), and the Science and Technology Program of Guangzhou (No. 2019050001).

References

[1] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, Domain-adversarial training of neural networks, *J. Mach. Learn. Res.*, vol. 17, no. 3, pp. 1–35, 2016.

[2] R. D. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, Adaptive semi-supervised learning for cross-domain sentiment classification, in *Proc. 2018 Conf. on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 3467–3476.

[3] J. L. Ji, C. Q. Luo, X. H. Chen, L. X. Yu, and P. Li, Cross-domain sentiment classification via a bifurcated-LSTM, in *Proc. 2nd Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, Melbourne, Australia, 2018, pp. 681–693.

[4] J. Zhou, J. F. Tian, R. Wang, Y. B. Wu, W. M. Xiao, and L. He, SENTIX: A sentiment-aware pre-trained model for cross-domain sentiment analysis, in *Proc. 28th Int. Conf. on Computational Linguistics*, Barcelona, Spain, 2020, pp. 568–579.

[5] Z. Li, X. Li, Y. Wei, L. D. Bing, Y. Zhang, and Q. Yang, Transferable end-to-end aspect-based sentiment analysis

with selective adversarial learning, in *Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing*, Hong Kong, China, 2019, pp. 4590–4600.

[6] M. L. Peng and Q. Zhang, Weighed domain-invariant representation learning for cross-domain sentiment analysis, in *Proc. 28th Int. Conf. on Computational Linguistics*, Barcelona, Spain, 2020, pp. 251–265.

[7] Y. P. Du, M. He, L. L. Wang, and H. T. Zhang, Wasserstein based transfer network for cross-domain sentiment classification, *Know. Based Syst.*, vol. 204, p. 106162, 2020.

[8] C. N. Du, H. F. Sun, J. Y. Wang, Q. Qi, and J. X. Liao, Adversarial and domain-aware BERT for cross-domain sentiment analysis, in *Proc. 58th Ann. Meeting of the Association for Computational Linguistic*, 2020, pp. 4019–4028.

[9] Q. M. Xue, W. Zhang, and H. Y. Zha, Improving domain-adapted sentiment classification by deep adversarial mutual learning, in *Proc. 34th AAAI Conf. on Artificial Intelligence*, New York, NY, USA, 2020, pp. 9362–9369.

[10] R. Sharma, P. Bhattacharyya, S. Dandapat, and H. S. Bhatt, Identifying transferable information across domains for cross-domain sentiment classification, in *Proc. 56th Ann. Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018, pp. 968–978.

[11] X. Y. Qu, Z. K. Zou, Y. Cheng, Y. Yang, and P. Zhou, Adversarial category alignment network for cross-domain sentiment classification, in *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, MN, USA, 2019, pp. 2496–2508.

[12] M. L. Peng, Q. Zhang, Y. G. Jiang, and X. J. Huang, Cross-domain sentiment classification with target domain specific information, in *Proc. 56th Ann. Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 2018, pp. 2505–2513.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Proc. 27th Int. Conf. on Neural Information Processing Systems*, Cambridge, MA, USA, 2014, pp. 2672–2680.

[14] M. Bouazizi and T. Ohtsuki, Multi-class sentiment analysis on twitter: Classification performance and challenges, *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 181–194, 2019.

[15] Y. Bie and Y. Yang, A multitask multiview neural network for end-to-end aspect-based sentiment analysis, *Big Data Mining and Analytics*, vol. 4, no. 3, pp. 195–207, 2021.

[16] S. C. Peng, L. H. Cao, Y. M. Zhou, Z. H. Ouyang, A. M. Yang, X. G. Li, W. J. Jia, and S. Yu, A survey on deep learning for textual emotion analysis in social networks, *Digit. Commun. Netw.*, doi: 10.1016/j.dcan.2021.10.003.

[17] D. Kulshreshtha, P. Goel, and A. K. Singh, How emotional are you? Neural architectures for emotion intensity prediction in microblogs, in *Proc. 27th Int. Conf. on Computational Linguistics*, Santa Fe, NM, USA, 2018, pp. 2914–2926.

[18] E. Mohammadi, H. Amini, and L. Kosseim, Neural feature extraction for contextual emotion detection, in *Proc. Int. Conf. on Recent Advances in Natural Language Processing*, Varna, Bulgaria, 2019, pp. 785–794.

[19] M. Abdul-Mageed and L. Ungar, Emonet: Fine-grained emotion detection with gated recurrent neural networks, in *Proc. 55th Ann. Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 2017, pp. 718–728.

[20] S. Tafreshi and M. Diab, Emotion detection and classification in a multigenre corpus with joint multi-task deep learning, in *Proc. 27th Int. Conf. on Computational Linguistics*, Santa Fe, NM, USA, 2018, pp. 2905–2913.

[21] J. F. Yu, L. Marujo, J. Jiang, P. Karuturi, and W. Brendel, Improving multi-label emotion classification via sentiment classification with dual attention transfer network, in *Proc. 2018 Conf. on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 1097–1102.

[22] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, Borrow from rich cousin: Transfer learning for emotion detection using cross lingual embedding, *Exp. Syst. Appl.*, vol. 139, p. 112851, 2020.

[23] C. L. P. Chen and Z. L. Liu, Broad learning system: An effective and efficient incremental learning system without the need for deep architecture, *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 1, pp. 10–24, 2018.

[24] S. Feng and C. L. P. Chen, Nonlinear system identification using a simplified fuzzy broad learning system: Stability analysis and a comparative study, *Neurocomputing*, vol. 337, pp. 274–286, 2019.

[25] D. X. Zhong and F. G. Liu, RF-OSFBLS: An RFID reader-fault-adaptive localization system based on online sequential fuzzy broad learning system, *Neurocomputing*, vol. 390, pp. 28–39, 2020.

[26] T. Zhang, X. H. Wang, X. M. Xu, and C. L. P. Chen, GCB-Net: Graph convolutional broad network and its application in emotion recognition, *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 379–388, 2022.

[27] J. T. Lin, Z. Liu, C. L. P. Chen, and Y. Zhang, Three-domain fuzzy wavelet broad learning system for tremor estimation, *Know. Based Syst.*, vol. 192, p. 105295, 2019.

[28] W. K. Yu and C. H. Zhao, Broad convolutional neural network based industrial process fault diagnosis with incremental learning capability, *IEEE Trans. Ind. Electron.*, vol. 67, no. 6, pp. 5081–5091, 2020.

[29] S. C. Peng, R. Zeng, H. Z. Liu, G. H. Chen, R. H. Wu, A. M. Yang, and S. Yu, Emotion classification of text based on BERT and broad learning system, in *Proc. 5th The Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Int. Conf. on Web and Big Data*, Guangzhou, China, 2021, pp. 382–396.

[30] K. M. Borgwardt, A. Gretton, M. J. Rasch, H. P. Kriegel, B. Schlkopf, and A. J. Smola, Integrating structured biological data by kernel maximum mean discrepancy, *Bioinformatics*, vol. 22, no. 14, pp. 49–57, 2006.

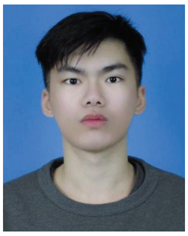
[31] M. Schuster and K. K. Paliwal, Bidirectional recurrent neural networks, *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.

[32] Y. Kim, Convolutional neural networks for sentence classification, in *Proc*. *2014 Conf*. *on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1746–1751.

[33] Y. Ziser and R. Reichart, Pivot based language modeling for improved neural domain adaptation, in *Proc*. *2018 Conf*. *of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, New Orleans, LA, USA, 2018, pp. 1241–1251.

[34] Y. M. Cui, W. X. Che, T. Liu, B. Qin, and Z. Q. Yang, Pre-training with whole word masking for Chinese BERT, *IEEE/ACM Trans*. *Audio Speech Language Process*., vol. 29, pp. 3504–3514, 2021.

[35] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, presented at the 7th Int. Conf. on Learning Representations, New Orleans, LA, USA, 2019.



Rong Zeng received the BE degree in electronic and information engineering from Hefei University of Technology, Hefei, China, in 2020. He is currently pursuing the ME degree at the South China Normal University, Guangzhou, China. His research interests include natural language processing and machine learning.

Aimin Yang received the BE degree from Hunan University of Science and Technology, Xiangtan, China, in 1993; received the ME degree from National University of Defense Technology, Changsha, China, in 2001, and received the PhD degree from Fudan University, Shanghai, China, in 2005. He is currently a professor at Lingnan Normal University, Zhanjiang, China. His research interests include intelligent computing, network traffic classification, nature language processing, and machine learning.

Chengqing Zong received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1998. He is currently a professor with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China. He has authored a book titled *Statistical Natural Language Processing* and coauthored a book titled *Text Data Mining*. He is an ACL Fellow, CAAI Fellow, and CCF Fellow. Also, he is a member of the International Committee on Computational Linguistics (ICCL). From 2019 and 2021, he was the president of the Asian Federation on Natural Language Processing (AFNLP). He has served many top-tier international conferences, such as ACL-IJCNLP 2015 and COLING 2020 as the PC Co-Chair, AAAI 2019, and AAAI 2020 as the area chair. He also serves as associate editor of ACM Transactions on Asian and Low-Resource Language Information Processing and a member of the Editorial Board of *IEEE Intelligent Systems*. His research interests include machine translation, dialog systems, and linguistic cognitive computing as well.

Hongzhan Liu received the PhD degree from Graduate School of Chinese Academy of Sciences, Beijing, China, in 2003. Now he is a professor at the South China Normal University, Guangzhou, China, and he is a member of the Guangdong Provincial Key Laboratory of Nanophotonic Functional Materials and Devices, Guangzhou, China. His research interests include satellite-ground laser communication, precision measurement, sensing, and artificial intelligence technology.

Lihong Cao received the BA degree in English education in 2007 and the MA degree in foreign linguistics and applied linguistics in 2010 from Hunan Normal University, Changsha, China. She is currently a lecturer at Guangdong University of Foreign Studies, Guangzhou, China. Her research interests include applied linguistics, nature language processing, intelligent computing.

Sancheng Peng received the PhD degree in computer science from Central South University, Changsha, China, in 2010. He is a professor at the Guangdong University of Foreign Studies, Guangzhou, China. He has authored or co-authored over 60 technical papers in both journals and conferences, such as *IEEE Communications Surveys and Tutorials*, *IEEE Transactions on Dependable and Secure Computing*, *IEEE Wireless Communications, IEEE Network*, *IEEE Internet of Things Journal*, *Journal of Network and Computer Applications*, *Computer Networks*, *Computer and Security*, *Information Sciences*, *Future Generation Computer Sciences*, *Journal of Computer and System Sciences*, *Journal of Computer Science and Technology*, *IEEE TrustCom*, *IEEE CBD*, *ICA3PP*, *SpaCCS*, and *EUC*. He has served as the guest editor of *Future Generation Computer Systems* and as a PC member for various prestige international conferences. He is a senior member of the CCF and a member of ACM. His current research interests include network and information security, social networks, trusted computing, and mobile computing.

Guodong Zhou received the PhD degree from National University of Singapore, Singapore, in 1999. He is a full professor at the School of Computer Science and Technology and the Director of the Natural Language Processing Laboratory, Soochow University, Suzhou, China. His research interests include information retrieval and natural language processing.