ASTES

*Advances in Science, Technology and Engineering Systems Journal*
*Vol. 6, No. 4, 01-12 (2021)*
www.astesj.com
*Special Issue on Innovation in Computing, Engineering Science & Technology*

**ASTES Journal**
**ISSN: 2415-6698**

# Exploiting Domain-Aware Aspect Similarity for Multi-Source Cross-Domain Sentiment Classification

Kwun-Ping Lai*, Jackie Chun-Sing Ho, Wai Lam

*Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, 999077, China*

ARTICLE INFO

*Article history:*
*Received: 01 May, 2021*
*Accepted: 15 June, 2021*
*Online: 10 July, 2021*

*Keywords:*
*Domain-aware topic model*
*Topic-attention network*
*Adversarial training*
*Artificial neural networks*

ABSTRACT

*We propose a novel framework exploiting domain-aware aspect similarity for solving the multi-source cross-domain sentiment classification problem under the constraint of little labeled data. Existing works mainly focus on identifying the common sentiment features from all domains with weighting based on the coarse-grained domain similarity. We argue that it might not provide an accurate similarity measure due to the negative effect of domain-specific aspects. In addition, existing models usually involve training sub-models using a small portion of the labeled data which might not be appropriate under the constraint of little labeled data. To tickle the above limitations, we propose a domain-aware topic model to exploit the fine-grained domain-aware aspect similarity. We utilize the novel domain-aware linear layer to control the exposure of various domains to latent aspect topics. The model discovers latent aspect topics and also captures the proportion of latent aspect topics of the input. Next, we utilize the proposed topic-attention network for training aspect models capturing the transferable sentiment knowledge regarding particular aspect topics. The framework finally makes predictions according to the aspect proportion of the testing data for adjusting the contribution of various aspect models. Experimental results show that our proposed framework achieves the state-of-the-art performance under the constraint of little labeled data. The framework has 71% classification accuracy when there are only 40 labeled data. The performance increases to around 82% with 200 labeled data. This proves the effectiveness of the fine-grained domain-aware aspect similarity measure.*

## 1 Introduction

Online shopping becomes more and more popular during the pandemic. Product reviews serve as an important information source for product sellers to understand customers, and for potential buyers to make decisions. Automatically analyzing product reviews therefore attracts people's attention. Sentiment classification is one of the important tasks. Given sufficient annotation resources, supervised learning method could generate promising result for sentiment classification. However, it would be very expensive or even impractical to obtain sufficient amount of labeled data for unpopular domains. Large pre-trained model, such as the Bidirectional Encoder Representations from Transformers model (BERT) [1], could be an universal way to solve many kinds of problems without exploiting the structure of the problem. In [2], the author apply large pre-trained model to handle this problem task, which has sufficient labeled data only in source domain but has no labeled data in target domain, with fine tuning on source domain and predicting on target domain. In [3], the author train the large pre-trained model using various sentiment related tasks and show that the model could directly apply to the target domain even without the fine-tuning stage. However, these large pre-trained models do not consider the structure of the problem and they have certain hardware requirement that might not be suitable in some situations. We focus on smaller models, which have a few layers, in this work in order to handle the constraint of little labeled data[1]. Besides using the gigantic pre-trained model, domain adaptation (or cross-domain) [4, 5] attempts to solve this problem by utilizing the knowledge from the source domain(s) with abundant annotation resources and transfers the knowledge to the target domain. This requires the model to learn transferable sentiment knowledge by eliminating the domain discrepancy problem. Domain adversarial training [6, 7] is an ef-

*Corresponding Author: Kwun-Ping Lai, Email: kplai@se.cuhk.edu.hk

[1]To give a brief comparison of our proposed framework and the large pre-trained model, we present the performance of the standard BERT-Large model in the experiment section. We ignore other variants of the large pre-trained models as they are not the major focus of this work.

fective method to capture common sentiment features which are useful in the target domain. Various works using domain adversarial training [8]–[11] achieve good performance for single-source cross-domain sentiment classification. It could be also applied to the large pre-trained model to further boost the performance [12]. Moreover, it is quite typical that multiple source domains are available, the model might be exposed to a wider variety of sentiment information and the amount of annotation requirement for every single domain would be smaller. A simple approach is to combine the data from multiple sources and form a new combined source domain. Existing models tackling single-source cross-domain sentiment classification mentioned above could be directly applied to this new problem setting after merging all source domains. However, the method of combining multiple sources does not guarantee a better performance than using only the best individual source domain [13, 14]. Recent works measure the global domain similarity [15]–[17], i.e. domain similarity between the whole source and target domain, or instance-based domain similarity [18]–[21], i.e. domain similarity between the whole source domain and every single test data point. We observe that these approaches are coarse-grained and ignore the fine-grained aspect relationship buried in every single domain. Domain-specific aspects from the source domain might have negative effect in measuring the similarity between the source domain and the target domain, or the single data point. For instance, we would like to predict the sentiment polarity of some reviews from the Kitchen domain and we have available data from the Book, and the DVD domain. Intuitively, the global domain similarity might not have much difference as both of them are not similar to the target. However, reviews related to the cookbook aspect from the Book domain, or reviews talking about cookery show from the DVD domain might contribute more to the prediction of Kitchen domain. Discovering domain-aware latent aspects and measuring the aspect similarity could be a possible way to address the problem. Based on this idea, we introduce the domain-aware aspect similarity measure based on various discovered domain-shared latent aspect topics using the proposed domain-aware topic model. The negative effect of domain-specific aspects could be reduced.

Existing models measuring domain similarity have another drawback. They usually train a set of expert models with each using a single source domain paired with the target domain. Then, the domain similarity is measured to decide the weighting of each expert model. Another way is to select a subset of data from all source domains which are similar to the target data. We argue that these approaches are not suitable under the constraint of little labeled data as each single sub-model is trained using a small portion of the limited labeled data which might obtain a heavily biased observation. The performance under limited amount of labeled data is underexplored for most of existing methods as they require considerable amount of labeled data for training. In [22], the author study the problem setting applying the constraint. However, they assume equal contribution for every source domain. We study the situation under the constraint of little labeled data and at the same time handling the contribution of source domains using fine-grained domain-aware aspect similarity.

To address the negative effect of domain-specific aspects during the domain similarity measure, and also the limitation of the constraint of little labeled data, we propose a novel framework exploiting domain-aware aspect similarity for measuring the contribution of each aspect model representing the captured knowledge of particular aspects. It is capable of working under the constraint of little labeled data. Specifically, the framework consists of the domain-aware topic model for discovering latent aspect topics and inferring the aspect proportion utilizing a novel aspect topic control mechanism, and the topic-attention network for training multiple aspect models capturing the transferable sentiment knowledge regarding particular aspects. The framework makes predictions using the measured aspect proportion of the testing data, which is a more fine-grained measure than the domain similarity, to decide the contribution of various aspect models. Experimental results show that the proposed domain-aware aspect similarity measure leads to a better performance.

### *1.1 Contributions*

The contributions of this work are as follows:

- We propose a novel framework exploiting the domain-aware aspect similarity to measure the contribution of various aspect models for predicting the sentiment polarity. The proposed domain-aware aspect similarity is a fine-grained measure which is designed to address the negative effect of domain-specific aspects existing in the coarse-grained domain similarity measure.
- We present a novel domain-aware topic model which is capable of discovering domain-specific and domain-shared aspect topics, together with the aspect distribution of the data in an unsupervised way. It is achieved by utilizing the proposed domain-aware linear layer controlling the exposure of different domains to latent aspect topics.
- Experimental results show that our proposed framework achieves the state-of-the-art performance for the multi-source cross-domain sentiment classification under the constraint of little labeled data.

### *1.2 Organization*

The rest of this paper is organized as follows. We present related works regarding cross-domain sentiment classification in Section 2. We describe the problem setting and our proposed framework in Section 3. We conduct extensive experiments and present results in Section 4. Finally, we talk about limitations and furture works in Section 5, and summarize our work in Section 6.

## 2 Related Works

Sentiment analysis [23]–[25] is the computational study of people's opinions, sentiments, emotions, appraisals, and attitudes towards entities [26]. In this work, we focus on textual sentiment data which is based on review of products, and the classification of the sentiment polarity of reviews. We first present the related works of single-source cross-domain sentiment classification. Next, we further extend to multiple-source case.

### 2.1 Single-Source Cross-Domain Sentiment Classification

Early works involve the manual selection of pivots based on predefined measures, such as frequency [27], mutual information [5, 28] and pointwise mutual information [29], which might have limited accuracy.

Recently, the rapid development of deep learning provides an alternative for solving the problem. Domain adversarial training is a promising technique for handling the domain adaptation. In [8], the author make use of memory networks to identify and visualize pivots. Besides pivots, [9] also consider non-pivot features by using the NP-Net network. In [10], the author combine external aspect information for predicting the sentiment.

Large pre-trained models attract people's attention since the BERT model [1] obtains the state-of-the-art performance across various machine learning tasks. Researchers also apply it on the sentiment classification task. Transformer-based models [2, 12, 3] utilize the amazing learning capability of the deep transformer structure to learn a better representation for text data during the pre-training stage and adapt themselves to downstream tasks (sentiment classification in our case) using fine tuning. However, we argue that the deep transformer structure has been encoded with semantic or syntactic knowledge during the pre-training process which makes the direct comparison against shallow models unfair. It also has certain hardware requirement which hinders its application in some situations.

Methods mentioned above focus on individual source only and they do not exploit the structure among domains. Although we can still directly apply these models to solve the problem by either training multiple sub-models and averaging predictions, or merging all source domains into a single domain, having a performance better than using only the single best source is not guaranteed. Therefore, exploring the structure or relationship among various domains is essential.

### 2.2 Multi-Source Cross-Domain Sentiment Classification

Early works assuming equal contribution for every source domain [30]–[32] could be a possible approach to handle the relationship between source domains and the target. Other solutions try to align features from various domains globally [33]–[22]. However, the source domain with higher degree of similarity to the target domain contributing more during the prediction process is a reasonable intuition. These methods fail to capture the domain relation. Recent works try to measure domain contribution in order to further improve the performance.

Researchers propose methods to measure the global domain similarity [15]–[17], i.e. the domain similarity between the whole source and target domain, or the instance-based domain similarity [18]–[21], i.e. the domain similarity between the whole source domain and every single test data point. In [15], the author measure the domain similarity using the proposed sentiment graph. In [17], the author employ a multi-armed bandit controller to handle the dynamic domain selection. In [18], the author compute the attention weight to decide the contribution of various already trained expert models. [20] also utilize the attention mechanism to assign importance weights. They incorporate a Granger-causal objective in their mixture of experts training. The total loss measuring distances of attention weights from desired attributions based on how much the inclusion of each expert reduces the prediction error. Maximum Cluster Difference is used in [19] as the metric to decide how much confidence to put in each source expert for a given example. In [21], the author utilize the output from the domain classifier to determine the weighting of a domain-specific extractor.

These methods measure the coarse-grained domain relation and ignore the fine-grained aspect relationship buried in every single domain. In addition, these methods do not consider the constraint of limited labeled data, which is the main focus of this work.

## 3 Model Descriptions

### 3.1 Problem Setting

The problem setting consists of the source domain group $D_s$ and the target domain $D_t$. The source domain group has $m$ domains $\{D_{s_k}\}|_{k=1}^{m}$ while there is only one target domain. For each source domain, we have two sets of data: i) the labeled data $L = \{x_i^l, y_i^l\}|_{i=1}^{n_L}$ and ii) the unlabeled data $U = \{x_j^u, d_j\}|_{j=1}^{n_U}$ where $n_L$ and $n_U$ are the number of data of labeled and unlabeled data respectively, and $d_j$ is the augmented domain membership indicator. Note that $y_i$ is the sentiment label for the whole review $x_i$ and we do not have any fine-grained aspect-level information. The $k$th source domain can be written as $D_{s_k} = \left\{L_{s_k} = \{x_i^{l,s_k}, y_i^{s_k}\}|_{i=1}^{n_{L_{s_k}}}, U_s = \{x_j^{u,s_k}, d_j^{s_k}\}|_{j=1}^{n_{U_{s_k}}}\right\}$. The data of the target domain has similar structure except that we do not have the sentiment label, i.e. $D_t = \left\{\{x_i^t\}|_{i=1}^{n_t}, U_t = \{x_j^{u,t}, d_j^t\}|_{j=1}^{n_{U_t}}\right\}$ respectively. $n_{L_{s_k}}$ is the number of labeled data and they are the same for all $k$. We set all $d_*^{s_k}$ to $k$ and all $d_*^t$ to $m+1$. The objective of the multi-source cross-domain sentiment classification is to find out a best mapping function $f$ so that given the training data $T = \{D_{s_1}, D_{s_2}, ..., D_{s_m}, D_t\}$, the aim is to predict the label of the target domain labeled data $\overline{\mathbf{y}}^{\mathbf{t}} = f(\mathbf{x}^{\mathbf{t}})$.

### 3.2 Overview of Our Framework

We describe our proposed framework exploiting domain-aware aspect similarity. Specifically, there are two components: i) the domain-aware topic model discovering domain-aware latent aspect topics, ii) the topic-attention network identifying sentiment topic capturing the transferable aspect-based sentiment knowledge. The first component captures both domain-specific and domain-shared latent aspect topics, and infers the aspect distribution of each review. It is an unsupervised model that utilizes only the unlabeled data. It is analogous to the standard topic model which discovers latent topics as well as topic distributions. However, the standard topic model is not capable of controlling discovered latent topics. Our proposed domain-aware topic model is capable of separating discovered latent topics into two groups: we name them as domain-specific aspect topics and domain-shared aspect topics. The topic control is achieved by using the domain-aware linear layer described in the latter subsection. Specifically, the model discover $n_{\texttt{spec}}$ domain-specific aspect topics for every domain, and $n_{\texttt{share}}$ domain-shared

aspect topics which are shared among all domains. Each review has a $n_{\text{spec}} + n_{\text{share}}$ dimensional aspect distribution with the first $n_{\text{spec}}$ dimension corresponding to domain-specific aspect topics and the last $n_{\text{share}}$ dimension corresponding to domain-share aspect topics. Discovered aspect topics and inferred aspect distributions have three important functions:

- By considering only domain-shared aspect topics, the negative effect of domain-specific aspect topics could be minimized for measuring the contribution during the inference process.
- The overall aspect distribution of the testing data reveals the importance of each discovered aspect topic following the assumption that the topic appearing more frequent is more important for the target domain.
- The aspect distribution of the unlabeled data could be used for picking reviews with a high coverage of a particular set of aspect topics.

Based on the domain-shared aspect distribution of the target domain, we divide discovered domain-shared aspect topics into groups with each group having unlabeled reviews from all domains with high aspect proportion forming the training dataset for the second component. Specifically, we divide domain-shared aspect topics into groups based on the overall aspect distribution of the target domain. We aim at separating aspect topics and train an expert model for each group of aspects. Each aspect model focuses on a particular set of aspects so as to boost the learning capability of that set of fine-grained aspect topics. Therefore, we need to construct the dataset carrying the information related to selected aspect topics. We select the unlabeled data from all domains with high aspect proportion of a particular set of aspect topics to form the aspect-based training dataset.

Each of the aspect-based training dataset guides the next component to focus on the corresponding aspect group and identify the related transferable sentiment knowledge. The obtained training dataset is jointly trained with the limited labeled data using the topic-attention network to generate an aspect model for each aspect-based training dataset. The topic-attention network is a compact model which is designed to work effectively under limited training data. The topic-attention network captures two topics simultaneously: i) the sentiment topic and ii) the domain topic. The sentiment topic captures the transferable sentiment knowledge which could be applied to the target domain. The domain topic serves as an auxiliary training task for constructing a strong domain classifier which helps the sentiment topic to identify domain-independent features by using domain adversarial training. These two topics are captured by the corresponding topical query built in the topic-attention layer. These topical queries are learnt automatically during the training process. The limited labeled data works with the sentiment classifier to control the knowledge discovery related to sentiment (sentiment topic captures sentiment knowledge while domain topic does not), while the unlabeled data works with the domain classifier to control the knowledge discovery related to domain. Finally, the framework makes predictions using various aspect models with contribution defined by the aspect distribution of the testing data. For example, if the testing data has a higher coverage regarding aspect group 1, then naturally the prediction made by the aspect model of group 1 should contribute more to the finally prediction as intuitively that aspect model would have more related sentiment knowledge to make judgement. We believe this fine-grained latent aspect similarity would provide a more accurate sentiment prediction than the traditional coarse-grained domain similarity due to the fact that we eliminate the negative effect of domain-specific aspects when measuring the similarity between the testing data and the expert models.

We first describe the architecture of the two components. Then, we describe the procedure of inferring the sentiment polarity of reviews of the target domain.

### 3.3 Domain-Aware Topic Model

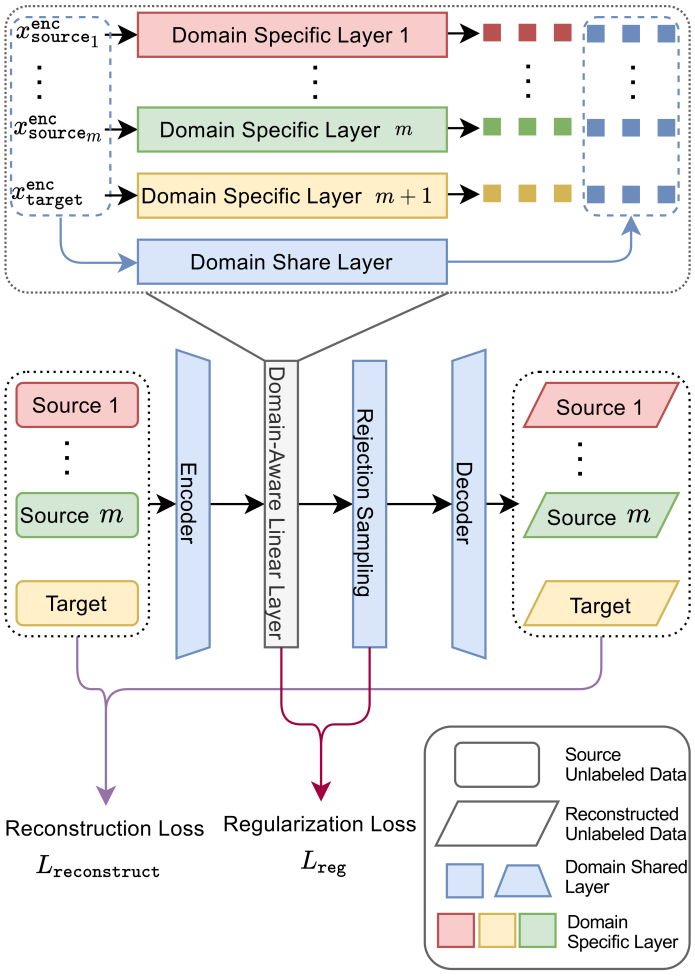


Figure 1: Diagram depicting the proposed domain-aware topic model. The middle part provides a high-level overview of the proposed domain-aware topic model. The model aims at inferring the dense representation of the unlabeled data from all domains in terms of aspect topic proportion. The model discovers domain-specific aspect topics and domain-shared aspect topics utilizing the domain-aware linear layer which is illustrated in the upper part of the figure. The model is trained by minimizing the reconstruction loss calculated by using the input data and the reconstructed data, and the regularization loss based on the inferred $\alpha$ and the predefined Dirichlet prior.

The domain-aware topic model follows the mechanism of the variational autoencoder framework (VAE) [35] which utilizes the encoder for inferring the latent variable (the Dirichlet prior $\alpha$ in our case representing the expected aspect distribution) and the de-

coder for reconstructing the input. Researchers try to apply the VAE network for achieving functionalities of standard topic model in a neural network way, such as inferring the topic proportion of the input and the word distribution of each topic. This provides some advantages such as reducing the difficulty of designing the inference process, leveraging the scalability of neural network, and the easiness of integrating with other neural networks [36]. However, the standard VAE using Gaussian distribution to model the latent variable might not be suitable for text data due to the sparseness of the text data. The Dirichlet distribution used in the topic model [37] has a problem of breaking the back-propagation. Calculating the gradient for the sampling process from the Dirichlet distribution is difficult. Researchers propose approximation methods [38, 39, 40, 41] in order to apply Dirichlet distribution to the neural topic model. We follow the rejection sampling method [42] in this work. Although discovered topics might carry extra information which might be helpful for identifying the hidden structure of the text data, it is not intuitive for applying this information to help the sentiment classification task. We introduce the domain-aware linear layer for controlling the formation of domain-specific and domain-shared aspect topics. To the best of our knowledge, we do not find any similar aspect topic control layer applied for multiple-source cross-domain sentiment classification in related works. The domain-aware linear layer identifies both domain-specific aspect topics and domain-shared aspect topics. We utilize domain-shared aspect topics only which could provide a more accurate measure for calculating the similarity. In addition, the inferred aspect topic proportion is used for constructing the aspect-based training dataset, and determining the level of contribution of each aspect model. Details of the architecture of the model are described below.

#### 3.3.1 Encoder

The input of the encoder is the bag of words of the review. Specifically, we count the occurrence of each vocabulary in each review and we use a vector of dimension $V$ to store the value. This serve as the input representing the review. The encoder is used to infer the Dirichlet prior of the aspect distribution of the input. The bag-of-words input is first transformed using a fully connected layer with RELU activation followed by a dropout layer.

$$\texttt{Layer}^{\texttt{enc}}(x) = \texttt{Dropout}\Big(\texttt{RELU}(W^{\texttt{enc}}x + b^{\texttt{enc}})\Big) \tag{1}$$

#### 3.3.2 Domain-Aware Linear Layer

Next, the output is fed into the domain-aware linear layer for obtaining domain-specific and domain-shared features. The domain-aware linear layer has $m + 1$ sub-layers including $m$ domain-specific sub-layers handling the feature extraction of the corresponding domain and 1 domain-shared sub-layer handling all domains as follows:

$$\texttt{Layer}^{\texttt{DL}}_{d_x}(x) = [W^{\texttt{DL}}_{d_x}x + b^{\texttt{DL}}_{d_x}; W^{\texttt{DL}}_{\texttt{shared}}x + b^{\texttt{DL}}_{\texttt{shared}}] \tag{2}$$

where $d_x$ is the domain ID of the input $x$, and [; ] represents the operation of vector concatenation. The output $x^{\texttt{DL}}$ is batch normalized and passed to the `SoftPlus` function to infer the Dirichlet prior $\alpha$ of the aspect distribution. To make sure each value in $\alpha$ is greater than zero, we set all values smaller than $\alpha_{\texttt{min}}$ to $\alpha_{\texttt{min}}$.

$$\alpha = \texttt{max}\Big(\texttt{SoftPlus}\Big(\texttt{BatchNorm}(x^{\texttt{DL}})\Big), \alpha_{\texttt{min}}\Big) \tag{3}$$

We use the rejection sampling method proposed in [42] to sample the aspect distribution $z$ and at the same time it allows the gradient to back-propagate to $\alpha$.

#### 3.3.3 Decoder

The decoder layer is used for reconstructing the bag-of-word input. The sampled aspect distribution $z$ is transformed by the domain-aware linear layer as follows:

$$\texttt{Layer}^{\texttt{dec}}(x) = [W^{\texttt{dec}}_{d_x}x; W^{\texttt{dec}}_{\texttt{shared}}x] \tag{4}$$

The output $x^{\texttt{dec}}$ is batch normalized and passed to the log-softmax function representing the log probability of generating the word.

$$y = \texttt{ln}\Big(\texttt{Softmax}\Big(\texttt{BatchNorm}(x^{\texttt{dec}})\Big)\Big) \tag{5}$$

#### 3.3.4 Loss Function

The loss function includes the regularization loss and the reconstruction loss. The regularization loss measures the difference of the log probability of generating the aspect distribution $z$ between two prior, $\alpha$ and $\overline{\alpha}$ as follows:

$$L_{\texttt{reg}} = lnP(z|\alpha) - lnP(z|\overline{\alpha}), \quad P(y|x) \sim Dir(x) \tag{6}$$

where $\alpha$ is inferred by the model and $\overline{\alpha}$ is the predefined Dirichlet prior. The reconstruction loss is the log probability of generating the bag-of-word input calculated as follows:

$$L_{\texttt{reconstruct}} = -\sum_{i=1}^{V} y_i x_i \tag{7}$$

where $V$ is the vocabulary size, $y_i$ is the log probability of the $i$th word generated by the model, and $x_i$ is the count of the $i$th word in the input.

### 3.4 Topic-Attention Network

The topic-attention network aims at capturing the transferable sentiment knowledge from the limited labeled data of various source domains. To achieve this goal, the network is designed to capture two topics simultaneously: i) the sentiment topic, and ii) the domain topic. The sentiment topic identifies the transferable sentiment knowledge from the input data while the domain topic helps to train a strong domain classifier. We use the technique of domain adversarial training [6, 7, 43] to maintain the domain independence of the sentiment topic. However, instead of using the standard gradient reversal layer, we use the adversarial loss function [22] to achieve the same purpose with a more stable gradient and a faster convergence. The model has two training tasks: i) the sentiment task for identifying the sentiment knowledge, and ii) the auxiliary domain task for training a strong domain classifier. The adversarial loss function is applied to the domain classifier output of the sentiment topic and the sentiment classifier output of the domain topic to hold the indistinguishability property of these two topics. Details of the architecture of the model is described below.

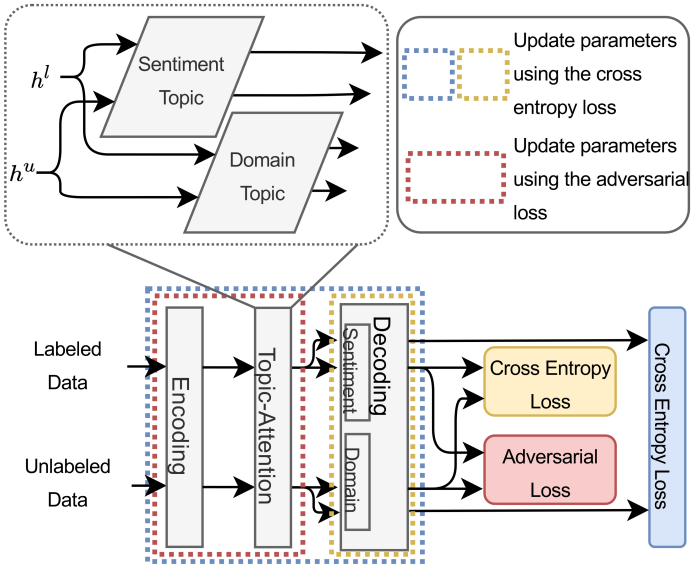


Figure 2: Diagram depicting the proposed topic-attention network. The bottom part provides a high-level overview of the proposed topic-attention network. The network captures two topics, i.e. the sentiment topic and the domain topic, from the review data and classifies the sentiment polarity and the domain membership. The adversarial loss maintains the indistinguishability of topics (domain indistinguishability of the sentiment topic and sentiment indistinguishability for the domain topic). Therefore, the sentiment knowledge captured by the sentiment topic could be transferred to the target domain. The colored dashed boxes show the scope of updating parameters for the corresponding loss.

#### 3.4.1 Encoding Layer

Each word is mapped to the corresponding embedding vector and then transformed by a feed-forward layer with tanh activation for obtaining the feature vector $h$.

$$h = \texttt{tanh}\left(W^{\text{enc}}\,\texttt{Embedding}(x) + b^{\text{enc}}\right) \tag{8}$$

#### 3.4.2 Topic-Attention Layer

The feature vector $h_i$ of the $i$th word is re-weighted by the topical attention weight $\beta_i^k$ calculated as follows:

$$\beta_i^k = \frac{m_i e^{q_k^\top h_i}}{\sum_{i'=1}^{k_w} m_{i'} e^{q_k^\top h_{i'}}} \times n_m \tag{9}$$

where $k$ indicates the topic (either sentiment or domain topic), $m_i$ is the word-level indicator indicating whether the $i$th position is a word or a padding, $n_m$ is the number of non-padding words, and $q_k$ is the topical query vector for topic $k$ learnt by the model. Note that we have two topical query vectors representing two topics. The topical feature vector $t_i^k$ of the topic $k$ and the review $i$ is obtained by summing feature vectors weighted by the corresponding topical attention weight $\beta_*^k$ as follows:

$$t_i^k = \sum_{j=1}^{W_i} \beta_j^k h_j \tag{10}$$

where $W_i$ is the number of words in review $i$. $t_i^k$ represents extracted features of the review by topic $k$.

#### 3.4.3 Decoding Layer

This layer consists of two decoders with each handling one training task, namely the sentiment decoder and the domain decoder for classifying the sentiment polarity and the domain membership respectively. Note that the review feature vector of labeled data is passed to the sentiment decoder while the unlabeled data of the aspect groups is passed to the domain decoder. Although we use the same $t^k$ to represent the input feature vector in the following two equations, they are actually representing the review features captured from the labeled data, and unlabeled data respectively. Specifically, the review feature vector is linearly transformed and passed to the `Softmax` function for obtaining a valid probability distribution.

$$p^{\text{sen},k} = \texttt{Softmax}(W^{\text{sen}} t^k + b^{\text{sen}}) \tag{11}$$

$$p^{\text{dom},k} = \texttt{Softmax}(W^{\text{dom}} t^k + b^{\text{dom}}) \tag{12}$$

Note that there are four outputs generated by the decoding layer, including two outputs generated by the captured features of two topics passing to the sentiment decoder, and similarly the remaining two generated by the domain decoder. The two topics are sentiment and domain topic, i.e. $k = \{\texttt{sen}, \texttt{dom}\}$. Therefore, the four outputs are: $p^{\text{sen,sen}}$ and $p^{\text{sen,dom}}$ coming from the labeled data passing to the sentiment decoder (the first superscript) having specific features captured by the sentiment and domain topic (the second superscript) respectively, and $p^{\text{dom,sen}}$ and $p^{\text{dom,dom}}$ coming from the unlabeled data passing to the domain decoder having specific features captured by the corresponding topic.

#### 3.4.4 Loss Function

We use the standard cross entropy loss to measure the classification performance:

$$L^{\text{sen},k} = -\frac{1}{n_L} \sum_{i=1}^{n_L} \ln p_{i,s_i}^{\text{sen},k} \tag{13}$$

$$L^{\text{dom},k} = -\frac{1}{n_U} \sum_{i=1}^{n_U} \ln p_{i,d_i}^{\text{dom},k} \tag{14}$$

where $s_i$ and $d_i$ are the class indicator specifying the sentiment polarity or the domain membership of the $i$th training data, and $p_{i,c}^*$ is the predicted probability regarding the $c$th class. Therefore, we have four cross entropy losses. The loss generated by the sentiment decoder from the sentiment topic and the loss generated by the domain decoder from the domain topic are used to update all parameters of the model using back-propagation. The remaining two are used to update the parameters of the decoding layer only. We introduce the adversarial loss function for doing adversarial training for both tasks as follows:

$$f_{adv}(p) = \sum_{i=1}^{c} (p_i - \frac{1}{c})^2 \tag{15}$$

where $c$ is the number of classes and $p_i$ is the predicted probability for the class $i$. Note that $c$ for sentiment task is 2 while it is $m + 1$ for the domain task. We use the probability distributions generated by the sentiment decoder from the domain topic $p^{\texttt{sen,dom}}$, and by the domain decoder from the sentiment topic $p^{\texttt{dom,sen}}$, to calculate the adversarial losses, which are used to update the parameters of the encoding layer and the topic-attention layer.

### 3.5 Training Strategy

We first train the domain-aware topic model using the unlabeled data $X^u$ from all domains. The model is then used for predicting the aspect proportion of the unlabeled data $X^u$ and testing data $X^t$ to obtain $\alpha^u$ and $\alpha^t$. Note that the domain-aware topic model is an unsupervised model that does not utilize any labeled data from source domains nor target domain. The aspect score $\theta^t$ of the target domain is calculated using the mean value of the domain-shared aspect part of $\alpha^t$ over all testing data:

$$\theta^t = \frac{1}{n_t} \sum_{i=1}^{n_t} \alpha_i^t[-n_{\texttt{share}} :] \tag{16}$$

where $\alpha_i^t[-n_{\texttt{share}} :]$ represents the last $n_{\texttt{share}}$ dimensions of the vector $\alpha_i^t$. Therefore, $\theta^t$ is a $n_{\texttt{share}}$ dimensional vector with each value representing the importance score of the corresponding aspect topic for the target domain. We divide the domain-shared aspect topics into $k$ groups based on their importance score using $\theta_t$ in descending order. The set $\texttt{topic}_{g_{k'}}$ contains the topic indices of the $k'$th aspect group. For each group $g_{k'}$, we select top $n$ unlabeled data from all domains based on the aspect topic score of the $k'$th group $\omega_{k'}^u$, which is the sum of the corresponding domain-shared aspect proportion of the $k'$th group for the $u$th review using its discovered aspect proportion:

$$\omega_{k'}^u = \sum_{i \in \texttt{topic}_{g_{k'}}} \alpha^u[i] \tag{17}$$

where $\alpha^u[i]$ represents the value in the $i$th dimension of $\alpha^u$.

Next, we train $k$ aspect models using the topic-attention network. For each aspect model, the limited labeled data $X^l, Y^l$ is used for training the sentiment task while the group of selected unlabeled data $g_{k'}$ is used for training the auxiliary domain task. The last step is to utilize the obtained models for predicting the sentiment polarity of all testing data $x^t$. Let $AM_{k'}$ be the aspect model trained by using the dataset $\{X^l, Y^l, g_{k'}\}$, we denote the sentiment prediction of the sentiment topic generated by the model as $p_{k'}^{t'}$ for the target review $x^{t'}$ as follows:

$$p_{k'}^{t'} = AM_{k'}(x^{t'}) \tag{18}$$

Finally, we combine the sentiment predictions of the sentiment topic generated by all aspect models having each contributes according to the aspect proportion of the testing data to obtain the final prediction:

$$p^{t'} = \sum_{i=1}^{k} \omega_i^{t'} p_i^{t'} \tag{19}$$

where $\omega_i^{t'}$ is the contribution of the $i$th aspect model to the final prediction.

## 4 Experiment

### 4.1 Experiment Settings

We use the Amazon review dataset [5] for the evaluation of our proposed framework. The Amazon review dataset is a common benchmark for sentiment classification. We use 5 most common domains, namely Book, DVD, Electronics, Kitchen and Video. For each experiment cross, we reserve one domain as the target domain and use others as source domains. There are 5 combinations in total and we conduct experiments on these 5 crosses. For each domain, we follow the dataset setting in [9] collecting 6000 labeled data, with half positive and half negative polarity. We do further sampling to select a subset of the labeled data to fulfill the constraint of little labeled data. We first construct two lists with each having 3000 elements representing the index of the labeled data of positive and negative class respectively. We randomly shuffle the lists and pick first $n$ indices. Next, we select the labeled data based on these indices. In order to have a comparable result for different size of labeled data, we fix the seed number of the random function so that the runs with different size of labeled data would obtain a same shuffle result. Therefore, the run with 20 labeled data contains the 10 labeled data from the run with 10 labeled data, and also another 10 new labeled data. Similarly, the run with 30 labeled data contains the 20 labeled data from the run with 20 labeled data. With this setting, we can directly estimate the effect of adding additional labeled data and compare the performance directly. We continue the process for other source domains. Finally, we construct 5 datasets having 10 to 50 labeled data for each target domain (there are 40 to 200 labeled data in total as there are 4 source domains). The unlabeled dataset includes all unlabeled data from all domains (including the target domain). All labeled data from the target domain is served as the testing data. We run every single run for 10 times and present the average accuracy with standard deviation in order to obtain a reliable result for model comparison.

### 4.2 Implementation Details

#### 4.2.1 Domain-Aware Topic Model

The Dirichlet prior is set to 0.01. The minimum of inferred prior is set to 0.00001. We set the number of domain-specific and domain-shared topics to 20 and 40 respectively. We divide the domain-shared aspect topics into 5 groups. The domain-aware topic model is trained for 100 warm-up epochs, and stopped after 10 epochs of no improvement.

#### 4.2.2 Topic-Attention Network

We use word2vec[2] embedding [44] to represent each word. We do not further train them to prevent overfitting. The batch size is

[2]It is a distributed representations of words in vector space. It helps various natural language processing task by putting similar words in a closer location.

[3]It is an optimization algorithm with adaptive learning rate. It considers the momentum of the gradient by using the moving average of the gradient. It also uses the moving average of the squared gradient to scale the learning rate of each individual parameter.

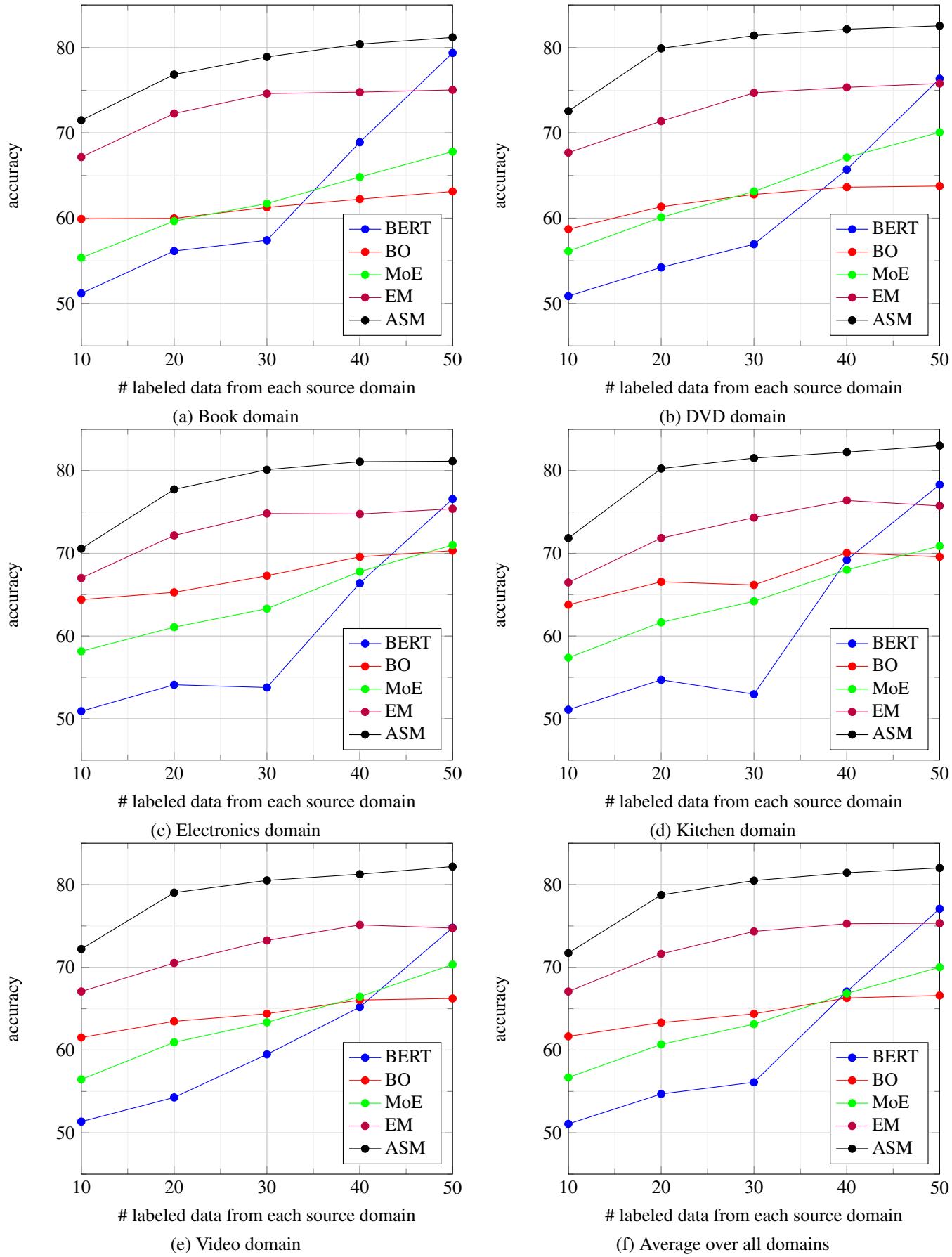


Figure 3: Figure showing the experimental results. Graph (a) to (e) shows the performance of Book, DVD, Electronics, Kitchen and Video domain as target domain respectively. Graph (f) shows the average accuracy of all domains.

set to the number of available labeled data. The topic-attention network is trained for 20 epochs. We use Adam[3] optimizer [45] for back-propagation for both models.

### 4.3 Evaluation Metric

We use accuracy to measure the evaluate the performance of various models. The target is a binary class. Therefore, correct cases involve the true positive (TP) and true negative (TN). Incorrect cases involve the false positive (FP) and the false negative (FN). The accuracy is calculated as follows:

$$\texttt{accuracy} = \frac{\texttt{TP} + \texttt{TN}}{\texttt{TP} + \texttt{TN} + \texttt{FP} + \texttt{FN}} \tag{20}$$

The average accuracy is calculated by taking the average of accuracy scores of multiple runs.

### 4.4 Main Results

Models used for performance comparison are as follows:

- BERT [1]: This is the **B**idirectional **E**ncoder **R**epresentations from **T**ransformers model, which is the popular pre-trained model designed to handle various text mining tasks. We use the BERT-Large model with fine tuning using the labeled data to obtain the prediction.
- BO [46]: This model employs **B**ayesian **o**ptimization for selecting data from source domains and transfer the learnt knowledge to conduct prediction on the target domain.
- MoE [19]: This is the **m**ixture **o**f **e**xpert model. It measures the similarity between single test data to every source domain for deciding the contribution of the expert models.
- EM [22]: This is the **e**nsemble **m**odel. It uses various base learners with different focuses on the training data to capture a diverse sentiment knowledge.
- ASM: This is the proposed framework exploiting the domain-aware **a**spect **s**imilarity **m**easure for obtaining a more accurate measure to adjust the contribution of various aspect models focusing on different aspect sentiment knowledge.

Results are presented in Figure 3 and Table 1. We use the classification accuracy as the metric to measure the performance. The proposed framework achieves the best average accuracy among all crosses. Its average performance is 71.73%, 78.75%, 80.49%, 81.43%, and 82.02% for 10, 20, 30, 40 and 50 labeled data cases respectively, or 40, 80, 120, 160 and 200 labeled data cases in total respectively.

### 4.5 Discussions

Our proposed framework performs substantially better than the comparison models. The proposed framework has an average of 4%, 7%, 6%, 6% and 6% absolute improvement over the second best result for 10, 20, 30, 40 and 50 labeled data cases respectively, or 40, 80, 120, 160 and 200 labeled data cases in total respectively. The variance of the proposed model is comparable to or better than the second best models. The result proves that our proposed framework is very effective for conducting multi-source cross-domain sentiment classification under the constraint of little labeled data. The model can capture transferable sentiment knowledge for predicting the sentiment polarity of the target reviews.

We also do comparative analysis to test the effectiveness of the proposed fine-grained domain-aware aspect similarity measure. It is based on the discovered aspect topics and also the aspect topic proportion for adjusting the contribution of various aspect models. We try to remove these two components to test the performance of the variants. The results are presented in Table 2. The first variant is *rand. select data + avg. pred.*, which means using the unlabeled data selected in a random way instead of using the aspect-based training dataset constructed by the domain-aware topic model, and combining the predictions of various aspect models by averaging them. In other words, the first variant removes both components. The second variant is *avg. pred.*. It keeps the first component (train the aspect models using the aspect-based training dataset) and only removes the second component. Therefore, it assumes equal contribution from various aspect models, just like the first variant. The last one is the proposed framework equipped with both components. Results show that the proposed fine-grained domain-aware aspect similarity measure improves the performance in general except the case having very few labeled data. We think the reason is that the aspect model could not locate the correct aspect sentiment knowledge from the limited data. Thus, the simply averaging the prediction of these biased aspect models would be better than relying on some models. Although the second variant (*avg. pred.*) has a better performance than the full framework in 10 labeled data case, the difference is very small (around 0.18%). Therefore, this comparative analysis could show that the proposed fine-grained domain-aware aspect similarity measure is effective for adjusting the contribution from different discovered aspects.

When comparing with the EM model [22] with similar network architecture but having an equal contribution for the source domains, the result shows that varying the contribution based on the domain-aware aspect similarity leads to a better performance.

We observe that our proposed framework has a small performance gain when giving more labeled training data, besides the case from 10 to 20. The EM model also has similar problem as mentioned in [22]. However, the BERT model [1] has an opposite behavior, which has a steady performance gain. We believe that the reason is due to the compact architecture of the topic-attention network which prevents overfitting the limited labeled data in order to have a better domain adaptation. Increasing the learning capability of the model and at the same time handling domain adaptation could be a future research direction.

## 5 Limitations and Future Works

The proposed framework involves two separate models handling their own jobs. These models do not share any learning parameters. Many works report that the single model handling various tasks would have a better generalization and thus leads to a better performance. One possible future work might consider integrating both

Table 1: Sentiment classification accuracy of different models

| # Labeled Data | Model | Book | DVD | Electronics | Kitchen | Video | Average |
|---|---|---|---|---|---|---|---|
| 10 (40) | BERT | 51.17 ± 1.25 | 50.86 ± 1.06 | 50.91 ± 1.52 | 51.09 ± 2.54 | 51.35 ± 2.21 | 51.07 |
| | BO | 59.90 ± 1.88 | 58.70 ± 3.66 | 64.40 ± 2.30 | 63.77 ± 2.14 | 61.52 ± 3.28 | 61.66 |
| | MoE | 55.35 ± 3.65 | 56.12 ± 3.94 | 58.15 ± 4.87 | 57.37 ± 4.32 | 56.45 ± 3.82 | 56.69 |
| | EM | 67.16 ± 5.03 | 67.68 ± 4.55 | 67.01 ± 5.29 | 66.47 ± 5.40 | 67.08 ± 3.67 | 67.08 |
| | ASM | **71.48 ± 4.70** | **72.56 ± 5.97** | **70.56 ± 4.44** | **71.83 ± 3.97** | **72.21 ± 4.27** | **71.73** |
| 20 (80) | BERT | 56.14 ± 6.14 | 54.22 ± 5.73 | 54.10 ± 4.70 | 54.70 ± 5.27 | 54.27 ± 5.60 | 54.69 |
| | BO | 59.97 ± 1.85 | 61.34 ± 2.65 | 65.28 ± 3.40 | 66.55 ± 2.54 | 63.47 ± 3.03 | 63.32 |
| | MoE | 59.65 ± 4.99 | 60.09 ± 5.66 | 61.07 ± 5.38 | 61.65 ± 5.09 | 60.94 ± 5.24 | 60.68 |
| | EM | 72.27 ± 2.67 | 71.37 ± 3.97 | 72.16 ± 2.74 | 71.84 ± 2.60 | 70.52 ± 1.38 | 71.63 |
| | ASM | **76.85 ± 2.41** | **79.91 ± 1.85** | **77.73 ± 3.51** | **80.24 ± 1.58** | **79.03 ± 1.85** | **78.75** |
| 30 (120) | BERT | 57.40 ± 5.87 | 56.94 ± 6.88 | 53.77 ± 5.40 | 52.96 ± 2.05 | 59.48 ± 7.99 | 56.11 |
| | BO | 61.26 ± 2.03 | 62.78 ± 2.32 | 67.29 ± 2.89 | 66.17 ± 3.11 | 64.39 ± 2.51 | 64.38 |
| | MoE | 61.71 ± 5.47 | 63.13 ± 5.68 | 63.30 ± 6.43 | 64.20 ± 6.06 | 63.36 ± 5.92 | 63.14 |
| | EM | 74.61 ± 2.54 | 74.70 ± 1.35 | 74.81 ± 2.03 | 74.31 ± 1.10 | 73.25 ± 1.97 | 74.34 |
| | ASM | **78.91 ± 2.13** | **81.42 ± 1.07** | **80.11 ± 1.58** | **81.51 ± 1.06** | **80.51 ± 1.75** | **80.49** |
| 40 (160) | BERT | 68.90 ± 8.55 | 65.70 ± 9.25 | 66.38 ± 8.45 | 69.19 ± 8.69 | 65.19 ± 9.77 | 67.07 |
| | BO | 62.23 ± 1.25 | 63.63 ± 2.45 | 69.57 ± 2.07 | 70.04 ± 2.22 | 66.05 ± 1.75 | 66.30 |
| | MoE | 64.82 ± 4.47 | 67.12 ± 5.22 | 67.78 ± 6.05 | 68.00 ± 5.89 | 66.46 ± 5.36 | 66.84 |
| | EM | 74.78 ± 1.06 | 75.34 ± 2.24 | 74.75 ± 1.89 | 76.38 ± 1.16 | 75.13 ± 1.77 | 75.27 |
| | ASM | **80.41 ± 1.13** | **82.16 ± 0.86** | **81.07 ± 1.38** | **82.23 ± 1.02** | **81.26 ± 1.71** | **81.43** |
| 50 (200) | BERT | 79.38 ± 4.79 | 76.36 ± 8.25 | 76.56 ± 7.01 | 78.30 ± 8.08 | 74.79 ± 9.68 | 77.08 |
| | BO | 63.13 ± 2.63 | 63.76 ± 2.14 | 70.32 ± 1.93 | 69.57 ± 2.56 | 66.24 ± 1.67 | 66.60 |
| | MoE | 67.80 ± 2.24 | 70.06 ± 2.59 | 70.99 ± 2.59 | 70.87 ± 2.68 | 70.34 ± 2.76 | 70.01 |
| | EM | 75.04 ± 1.85 | 75.79 ± 1.96 | 75.38 ± 2.34 | 75.73 ± 2.27 | 74.74 ± 1.42 | 75.33 |
| | ASM | **81.20 ± 0.86** | **82.56 ± 0.88** | **81.13 ± 1.30** | **83.02 ± 0.87** | **82.18 ± 1.01** | **82.02** |

Table 2: Comparative analysis of the proposed framework.

| # Labeled Data | Model | Avg. Accuracy |
|---|---|---|
| 10 (40) | rand. select data + avg. pred. | 70.98 |
| | avg. | **71.91** |
| | ASM | 71.73 |
| 20 (80) | rand. select data + avg. pred. | 76.47 |
| | avg. | 78.18 |
| | ASM | **78.75** |
| 30 (120) | rand. select data + avg. pred. | 78.03 |
| | avg. | 79.75 |
| | ASM | **80.49** |
| 40 (160) | rand. select data + avg. pred. | 79.25 |
| | avg. | 80.72 |
| | ASM | **81.43** |
| 50 (200) | rand. select data + avg. pred. | 79.63 |
| | avg. | 81.08 |
| | ASM | **82.02** |

models together forming an unified model to take the advantage of multi-task learning. This might further improve the performance for the sentiment classification task.

# 6 Conclusion

We study the task of multi-source cross-domain sentiment classification under the constraint of little labeled data. We propose a novel framework exploiting domain-aware aspect similarity to identify the contribution of discovered fine-grained aspect topics. This fine-grained similarity measure aims at addressing the negative effect of domain-specific aspects appearing in the existing coarse-grained domain similarity measure, and also the limitation caused by the constraint of little labeled data. Aspect topics are extracted by the proposed domain-aware topic model in an unsupervised way. The topic-attention network then learns the transferable sentiment knowledge based on the selected data related to discovered aspects. The framework finally makes predictions according to the aspect proportion of the testing data for adjusting the contribution of various aspect models. Extensive experiments show that our proposed framework achieves the state-of-the-art performance. The framework achieves a good performance, i.e. around 71%, even though there are only 40 labeled data. The performance reaches around 82% when there are 200 labeled data. This shows that our proposed fine-grained domain-aware aspect similarity measure is very effective under the constraint of little labeled data.

# References


[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018, doi:10.18653/v1/N19-1423.

[2] B. Myagmar, J. Li, S. Kimura, "Cross-Domain Sentiment Classification With Bidirectional Contextualized Transformer Language Models," IEEE Access, 163219–163230, 2019, doi:10.1109/ACCESS.2019.2952360.

[3] J. Zhou, J. Tian, R. Wang, Y. Wu, W. Xiao, L. He, "SentiX: A Sentiment-Aware Pre-Trained Model for Cross-Domain Sentiment Analysis," in Proceedings of the 28th International Conference on Computational Linguistics, 568–579, 2020, doi:10.18653/V1/2020.COLING-MAIN.49.

[4] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, "Analysis of Representa-tions for Domain Adaptation," in Advances in Neural Information Processing Systems, **19**, 2007, doi:10.7551/mitpress/7503.003.0022.

[5] 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, 440–447, 2007.

[6] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, "Domain-adversarial neural networks," NIPS 2014 Workshop on Transfer and Multi-task learning: Theory Meets Practice, 2014, doi:10.1007/978-3-319-58347-1 10.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, "Domain-adversarial training of neural networks," The Journal of Machine Learning Research, **17**(1), 2096–2030, 2016, doi:10.1007/978-3-319-58347-1 10.

[8] Z. Li, Y. Zhang, Y. Wei, Y. Wu, Q. Yang, "End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification," in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2237–2243, 2017, doi:10.24963/IJCAI.2017/311.

[9] Z. Li, Y. Wei, Y. Zhang, Q. Yang, "Hierarchical attention transfer network for cross-domain sentiment classification," in Thirty-Second AAAI Conference on Artificial Intelligence, 5852–5859, 2018, doi:10.1609/AAAI.V33I01.33015773.

[10] 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, 5773–5780, 2019, doi:10.1609/aaai.v33i01.33015773.

[11] Q. Xue, W. Zhang, H. Zha, "Improving domain-adapted sentiment classification by deep adversarial mutual learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 9362–9369, 2020, doi:10.1609/AAAI.V34I05.6477.

[12] 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, 4019–4028, 2020, doi:10.18653/v1/2020.acl-main.370.

[13] S. Zhao, B. Li, C. Reed, P. Xu, K. Keutzer, "Multi-source Domain Adaptation in the Deep Learning Era: A Systematic Survey," arXiv preprint arXiv:2002.12169, 2020.

[14] S. Zhao, Y. Xiao, J. Guo, X. Yue, J. Yang, R. Krishna, P. Xu, K. Keutzer, "Curriculum CycleGAN for Textual Sentiment Domain Adaptation with Multiple Sources," in The Web Conference (WWW), 2021, doi:10.1145/3442381.3449981.

[15] F. Wu, Y. Huang, "Sentiment Domain Adaptation with Multiple Sources," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics ( 1: Long Papers), 301–310, 2016, doi:10.18653/v1/P16-1029.

[16] M. Yu, X. Guo, J. Yi, S. Chang, S. Potdar, Y. Cheng, G. Tesauro, H. Wang, B. Zhou, "Diverse Few-Shot Text Classification with Multiple Metrics," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, **1** (Long Papers), 1206–1215, 2018, doi:10.18653/v1/N18-1109.

[17] H. Guo, R. Pasunuru, M. Bansal, "Multi-Source Domain Adaptation for Text Classification via DistanceNet-Bandits," in The Thirty-Fourth AAAI Conference on Artificial Intelligence, 7830–7838, 2020, doi:10.1609/AAAI.V34I05.6288.

[18] Y.-B. Kim, K. Stratos, D. Kim, "Domain Attention with an Ensemble of Experts," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics ( 1: Long Papers), 643–653, 2017, doi:10.18653/v1/P17-1060.

[19] J. Guo, D. Shah, R. Barzilay, "Multi-Source Domain Adaptation with Mixture of Experts," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 4694–4703, 2018, doi:10.18653/v1/D18-1498.

[20] 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, 1913–1916, 2020, doi:10.1145/3397271.3401314.

[21] Y. Dai, J. Liu, X. Ren, Z. Xu, "Adversarial Training Based Multi-Source Unsupervised Domain Adaptation for Sentiment Analysis," in The Thirty-Fourth AAAI Conference on Artificial Intelligence, 7618–7625, 2020, doi:10.1609/AAAI.V34I05.6262.

[22] K. Lai, J. C. Ho, W. Lam, "Ensemble Model for Multi-Source Cross-Domain Sentiment Classification with Little Labeled Data," in 2020 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, 2020, doi:10.1109/WIIAT50758.2020.00038.

[23] K. Shaukat, T. M. Alam, M. Ahmed, S. Luo, I. A. Hameed, M. S. Iqbal, J. Li, M. A. Iqbal, "A Model to Enhance Governance Issues through Opinion Extraction," in 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 0511–0516, 2020, doi:10.1109/IEMCON51383.2020.9284876.

[24] H. Gupta, S. Pande, A. Khamparia, V. Bhagat, N. Karale, et al., "Twitter Sentiment Analysis Using Deep Learning," in IOP Conference Series: Materials Science and Engineering, 012114, 2021, doi:10.1088/1757-899X/1022/1/012114.

[25] A. Badgaiyya, P. Shankarpale, R. Wankhade, U. Shetye, K. Gholap, S. Pande, "An Application of Sentiment Analysis Based on Hybrid Database of Movie Ratings," International Research Journal of Engineering and Technology (IRJET), **8**, 655–665, 2021.

[26] L. Zhang, S. Wang, B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, **8**(4), e1253, doi:10.1002/widm.1253.

[27] J. Blitzer, R. McDonald, F. Pereira, "Domain adaptation with structural correspondence learning," in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 120–128, 2006, doi:10.3115/1610075.1610094.

[28] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in Proceedings of the 19th International Conference on World Wide Web, 751–760, 2010, doi:10.1145/1772690.1772767.

[29] D. Bollegala, T. Mu, J. Y. Goulermas, "Cross-domain sentiment classification using sentiment sensitive embeddings," IEEE Transactions on Knowledge and Data Engineering, **28**(2), 398–410, 2015, doi:10.1109/TKDE.2015.2475761.

[30] K. Crammer, M. Kearns, J. Wortman, "Learning from Multiple Sources," Journal of Machine Learning Research, **9**(57), 1757–1774, 2008.

[31] S. Li, C. Zong, "Multi-domain Sentiment Classification," in Proceedings of ACL-08: HLT, Short Papers, 257–260, 2008.

[32] P. Luo, F. Zhuang, H. Xiong, Y. Xiong, Q. He, "Transfer learning from multiple source domains via consensus regularization," in Proceedings of the 17th ACM conference on Information and knowledge management, 103–112, 2008, doi:10.1145/1458082.1458099.

[33] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, G. J. Gordon, "Adversarial multiple source domain adaptation," in Advances in Neural Information Processing Systems, 8568–8579, 2018.

[34] X. Chen, C. Cardie, "Multinomial Adversarial Networks for Multi-Domain Text Classification," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, **1** (Long Papers), 1226–1240, 2018, doi:10.18653/v1/N18-1111.

[35] D. P. Kingma, M. Welling, "Auto-Encoding Variational Bayes," in 2nd International Conference on Learning Representations, ICLR, 2014.

[36] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du, W. Buntine, "Topic Modelling Meets Deep Neural Networks: A Survey," arXiv preprint arXiv:2103.00498, 2021.

[37] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, **3**, 993–1022, 2003, doi:10.1016/B978-0-12-411519-4.00006-9.

[38] M. Figurnov, S. Mohamed, A. Mnih, "Implicit Reparameterization Gradients," in Advances in Neural Information Processing Systems, 439–450, Curran Associates, Inc., 2018.

[39] W. Joo, W. Lee, S. Park, I.-C. Moon, "Dirichlet Variational Autoencoder," CoRR, **abs/1901.02739**, 2019, doi:10.1016/j.patcog.2020.107514.

[40] H. Zhang, B. Chen, D. Guo, M. Zhou, "WHAI: Weibull Hybrid Autoencoding Inference for Deep Topic Modeling," in International Conference on Learning Representations, 2018.

[41] C. Naesseth, F. Ruiz, S. Linderman, D. Blei, "Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 489–498, 2017.

[42] S. Burkhardt, S. Kramer, "Decoupling Sparsity and Smoothness in the Dirichlet Variational Autoencoder Topic Model," Journal of Machine Learning Research, **20**(131), 1–27, 2019.

[43] Y. Ganin, V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in Proceedings of the 32nd International Conference on International Conference on Machine Learning, 1180–1189, 2015.

[44] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems, 3111–3119, 2013.

[45] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[46] S. Ruder, B. Plank, "Learning to select data for transfer learning with Bayesian Optimization," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 372–382, 2017, doi:10.18653/v1/D17-1038.