# TACIT: A Target-Agnostic Feature Disentanglement Framework for Cross-Domain Text Classification

**Rui Song**[1], **Fausto Giunchiglia**[1, 2, 3], **Yingji Li**[3], **Mingjie Tian**[1], **Hao Xu**[*1, 3, 4]
[1]School of Artificial Intelligence, Jilin University, Changchun 130012, China
[2]Department of Information Engineering and Computer Science, University of Trento, Via Sommarive 938123, Trento, Italy
[3]College of Computer Science and Technology, Jilin University, Changchun 130012, China
[4]Chongqing Research Institute, Jilin University, Chongqing 401123, China
{songrui20, yingji21, mjtian19}@mails.jlu.edu.cn, fausto@disi.unitn.it, xuhao@jlu.edu.cn

## Abstract

Cross-domain text classification aims to transfer models from label-rich source domains to label-poor target domains, giving it a wide range of practical applications. Many approaches promote cross-domain generalization by capturing domain-invariant features. However, these methods rely on unlabeled samples provided by the target domains, which renders the model ineffective when the target domain is agnostic. Furthermore, the models are easily disturbed by shortcut learning in the source domain, which also hinders the improvement of domain generalization ability. To solve the aforementioned issues, this paper proposes TACIT, a target domain agnostic feature disentanglement framework which adaptively decouples robust and unrobust features by Variational Auto-Encoders. Additionally, to encourage the separation of unrobust features from robust features, we design a feature distillation task that compels unrobust features to approximate the output of the teacher. The teacher model is trained with a few easy samples that are easy to carry potential unknown shortcuts. Experimental results verify that our framework achieves comparable results to state-of-the-art baselines while utilizing only source domain data.

## Introduction

In recent years, natural language processing (NLP) models based on deep networks have made significant progress and have even surpassed human-level performance. But these methods often rely on manually labeled data, and the inconsistency between the distribution of labeled training domains and the unlabeled target domains poses a challenge for deploying these methods in practical applications (Ben-David, Rabinovitz, and Reichart 2020). To address this challenge, Unsupervised Domain Adaptation (UDA) has emerged as a solution. UDA aims to generalize a model trained on labeled data from source domains to perform well on a target domain without labeled data. By employing UDA, models can overcome their dependency on labeled data from the target domain, which has attracted considerable attention from researchers.

In UDA, cross-domain text classification is a basic but challenging task because of the differences in text expressions among the source and target domains. To enhance the performance of cross-domain text classification, numerous researches have focused on improving the generalization ability by extracting domain invariant features, including pivot-based methods (Ziser and Reichart 2018; Peng et al. 2018; Ben-David, Rabinovitz, and Reichart 2020), task-specific knowledge-based methods (Zhou et al. 2020), domain adversarial training methods (Wu and Shi 2022), and class-aware methods (Ye et al. 2020; Luo et al. 2022). Besides, there are approaches that use language models to perform self-supervised tasks to capture task-agnostic features in the target domain (Karouzos, Paraskevopoulos, and Potamianos 2021). These methods take full advantage of the commonality between the source and target domains to encourage model generalization.

However, these approaches still face **two main challenges**. **First**, capturing domain-invariant features tends to depend heavily on the target domain, which makes the model ineffective when the target domain is agnostic. Besides, the training of the generalized model requires consideration of additional target domain samples, which adds training and deployment costs. **Second**, the models are susceptible to shortcut learning[1] in the source domain, which also hinders the improvement of domain generalization ability (Geirhos et al. 2020; Feng et al. 2023).

To overcome the above challenges, we propose a **T**arget-**A**gnostic framework for **C**ross-domain text class**I**fica**T**ion (TACIT). It is inspired by the work of feature disentanglement (Huang et al. 2021) as well as Variational Auto-Encoder (VAE) for text generation (Bao et al. 2019). The aim of TACIT is to separate robust and unrobust features from the potential latent feature space of the source domain, and use robust features to promote cross-domain generalization performance. Moreover, we design a feature distillation task to encourage further separation of the unrobust features from the robust features. The teacher model in the distillation task learns from easy samples in the training set

*Corresponding author


[1]Shorcuts are defined as simple decision rules that can not be applied to more challenging scenarios, such as cross-domain generalization. Shortcut learning occurs when the model relies excessively on superficial correlations in the source domain, disregarding domain-specific features crucial for accurate classification in the target domain. Therefore, mitigating shorcuts can improve cross-domain generalization (Moon et al. 2021). In our study, shortcuts are not predefined, but included in easy samples.

to ensure that it itself carries unrobust features. As a result, TACIT can use only source domain samples for cross-domain text classification without any target domain data and additional target domain training. Experiments on four publicly available datasets confirm that TACIT is capable of going beyond state-of-the-art approaches. Overall, our contributions are as follows:

- We propose a feature disentanglement framework for separating robust and unrobust features and facilitating the model's ability to generalise across domains in target domain agnostic scenario.
- We train an unrobust teacher model with easy samples, and design a feature distillation task to encourage further decoupling of unrobust features.
- We experimentally confirm that the proposed TACIT can be compared with some of the most advanced methods in the absence of target domain data.

## Related Work

In this section, we list some of the work related to TACIT, including cross-domain text classification and entanglement methods in NLP.

### Cross-Domain Text Classification

The high cost of acquiring large amounts of labeled data for each domain has prompted research into cross-domain text classification with the help of Unsupervised Domain Adaptation techniques (Blitzer, Dredze, and Pereira 2007; Yu and Jiang 2016; Ramponi and Plank 2020). Most of the previous works facilitate generalization by capturing pivots common to source and target domains (Li et al. 2018; Ziser and Reichart 2018; Peng et al. 2018; Ben-David, Rabinovitz, and Reichart 2020), where pivots are key features or attributes that act as a bridge, enabling the model to transfer knowledge learned from labeled source data to the unlabeled target data. Another common approaches are domain adversarial training, which enhance the generalization ability of the model by allowing it to distinguish between source domain and target domain data (Ganin et al. 2016; Qu et al. 2019; Wu and Shi 2022). Task-specific knowledge-based methods introduce additional task-related knowledge to facilitate generalization. For example, SENTIX uses existing lexicons and annotations at both token and sentence levels to retrain the language model (Zhou et al. 2020). (Li et al. 2022) helps cross-domain generalization by extracting sentiment-driven semantic graphs from Abstract Meaning Representation. Class-aware methods extracts better category-invariant features by learning more discriminative source domain labels (Ye et al. 2020; Luo et al. 2022). Besides, there are approaches that use language models to perform self-supervised tasks to capture task-agnostic features in the target domain (Du et al. 2020; Karouzos, Paraskevopoulos, and Potamianos 2021). They re-train the language model by performing cloze tasks in the target domain, so that the features of the target domain can be captured without any labels.

In contrast to the previous research methods, our approach adopts a more stringent criteria where the target domain is completely agnostic, and even unlabeled texts are not provided. This means there is no need to retrain the model specifically for the target domain when performing a new task in that domain. This flexibility allows for seamless application of our approach across diverse target domains without any target-specific training requirements.

### Textual Feature Disentanglement

The disentanglement of latent space is first explored in the field of computer vision, and features of images (such as rotation and color) have been successfully disentangled (Chen et al. 2016). In NLP tasks, it is used to address the decoupling of latent representations of text, such as text style and content (John et al. 2019), syntax and semantics (Bao et al. 2019), opinions and plots in user reviews (Pergola, Gui, and He 2021), fairness representation and bias against sensitive attributes (Colombo et al. 2022). They rely on Variational Auto-Encoders or some variations (Kingma and Welling 2014), to restore the original feature from the space of disentanglement. In addition, there are methods to facilitate the separation of specific feature spaces by imposing regularization constraints on different tasks (John et al. 2019; Huang et al. 2021). In this paper, inspired by the above disentangled methods, we promote the effect of cross-domain text classification by separating robust and unrobust features.

## Proposed Framework

This section elaborates on the proposed framework TACIT. First, to facilitate the narration, we first give the problem formulation and some symbolic definitions. Subsequently, we gradually describe the composition of TACIT as shown in Figure 1. TACIT mainly contains a student model based on VAE and an easy teacher model with unrobust features. In the process of feature disentanglement of the student, the separated unrobust features are encouraged to learn from the teacher for better decoupling effect.

### Problem Formulation

Similiar to (Wu and Shi 2022), we consider two different scenarios: a source domain relative to a target domain and multiple source domains relative to a target domain. For any number of source domains $\mathcal{S}^l = \{x_i^l, y_i^l\}_{i=1}^{N_s^l}$ with labeled datas, our goal is to get a fully trained language model $\mathcal{M}$ with a classification head $\mathcal{F}(.)$, which has good generalization ability in the target domain $\mathcal{T} = \{x_i^t\}_{i=1}^{N_t}$ without any label. Here, $N_s^l$ and $N_t$ represents the number of samples from different source domains and the target domain, where $l \geq 1$ denotes the minimum number of source domain is 1. For any text $x_i$, it contains $m + 2$ tokens $\{[CLS], t_1, ...t_m, [SEP]\}$ where $[CLS]$ is used to obtain the representation $h_i$ of the text output by $\mathcal{M}$. Then, $\mathcal{F}(h_i)$ maps the representation to the appropriate label $y_i$. In some methods, despite the absence of any labeling, data from the target domain $\mathcal{T}$ can still help train $\mathcal{M}$. In our approach, only the source domain $\mathcal{S}^l$ can be used.

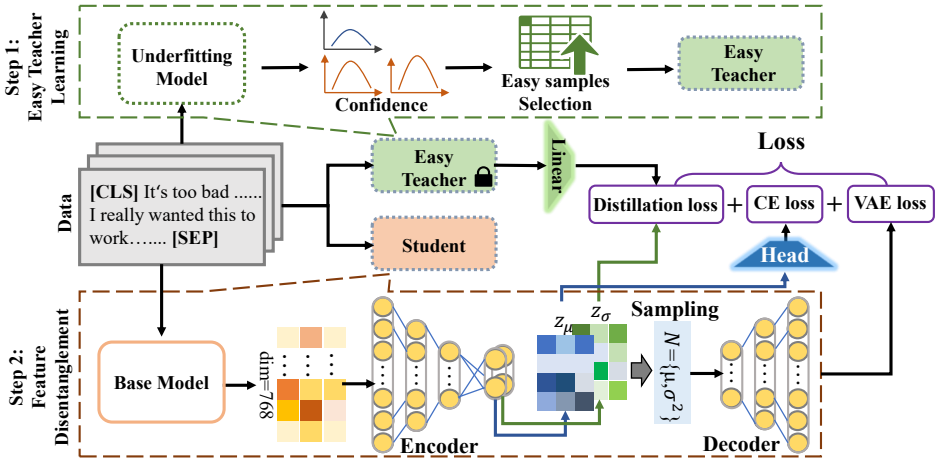


Figure 1: TACIT's overall architecture and processing flow. It consists of two main steps and three tasks. In Step 1, an underfitting model selects a subset of easy samples from the source domain based on the confidence. Subsequently, such samples are used to train a teacher model. In Step 2, the output features of the base model are fed into VAE for disentanglement. The robust feature $z_\mu$ is used to predict the sample labels. Then, the unrobust feature $z_\sigma$ is scheduled to be learned from the teacher's output $\hat{z}$ through feature distillation. Finally, cross-entropy loss, VAE loss and distillation loss are used to co-optimize the model. 🔒 indicates that model parameters are not updated during training.

## Student: Feature Disentanglement Based on VAE

Under the premise that the target domain is agnostic, we expect that the model can disentangle robust and unrobust features in the continuous latent feature space, and the former is used for effective cross-domain generalization, while the latter is discarded as task-irrelevant features. Inspired by some related work on textual feature disentanglement (Bao et al. 2019; John et al. 2019), we adopt VAE to separate robust and unrobust features from sample feature space (Kingma and Welling 2014).

Specifically, we use a probabilistic latent variable $z$ to encode the representation $h$, and then decode $h$ from $z$:

$$p(h) = \int p(z)p(h|z)\,\mathrm{d}z, \tag{1}$$

where $p(z)$ denotes the prior which is the standard normal $\mathcal{N}(0, \mathrm{I})$. To optimize VAE, the following loss according to the evidence lower bound (ELBO) is defined:

$$\mathcal{L}_{vae} = -\mathbb{E}_{q(z|h)}[logp(h|z)] + KL(q(z|h)||p(z)), \tag{2}$$

where $q(z|h)$ is the posterior given by the decoder, which is formed by $\mathcal{N}(\mu, \mathrm{diag}\ \sigma^2)$. $KL$ is Kullback-Leibler divergence. Here, $\mu$ and $\sigma^2$ can be regarded as independent of each other under the premise of the standard normal (Kawata and Sakamoto 1949; Fotopoulos 2007), we present the relevant proof in the Appendix. Therefore, we use their corresponding representations to represent robust and unrobust features, instead of a simple feature split of $z$ (John et al. 2019). In practice, they can be modeled by two independent linear transformations and represented as $z_\mu$, $z_\sigma$.

Next, to ensure the robustness of $z_\mu$, it should be able to help the model make correct predictions. Therefore, a classification head is used to predict the label of the current sample from $z_\mu$ by cross entropy (CE):

$$\mathcal{L}_{ce} = CE(softmax(Head(z_\mu))), \tag{3}$$

where $Head(.)$ is modeled using a linear transformation, which maps the input representations to the latent label space. By optimizing the above loss, it is possible to ensure the effectiveness of robust features for the classification task.

## Teacher: Easy Samples Learning

Now, we have two premises, $z_\mu$ and $z_\sigma$ are disentangled and $z_\mu$ is used for robust label prediction. Several studies have shown that additional tasks targeting different features can help to further disentangle the features (John et al. 2019; Huang et al. 2021). Therefore, a natural idea is to add an extra task for $z_\sigma$ making it produce unrobust predictions. With difficulty, when the target domain is agnostic, producing unrobust predictions that are not conducive to cross-domain generalization is unavailable. Therefore, we train an easy teacher model for generating unrobust features and guiding $z_\sigma$ indirectly. The acquisition of teacher model is inspired by some unknown biases mitigation approaches (Utama, Moosavi, and Gurevych 2020a,b), where a shallow model is easily affected by easy samples. We expect to extract easy samples from the training set and train the teacher model to learn the unrobust features contained in the easy samples.

**Easy Samples Extraction**. Previous studies have proved that easy samples can be easily fitted by models with fewer parameters (Lai et al. 2021). Besides, the model is also more likely to make overconfident predictions for the easy samples (Du et al. 2021). Therefore, we obtain an underfitting shallow model to determine whether the sample is an easy sample. Specifically, we train a DistilBERT[2] or Dis-

[2]https://huggingface.co/distilbert-base-uncased

tilRoBERTa[3] for 2 epochs on all the training samples and rank the samples based on confidence. Confidence denotes the largest value in the predicted probability distribution. So if a sample can obtain a large confidence in the case of underfitting, it could be an easy samples. Top 35%[4] of the samples are considered as easy samples.

**Teacher training**. Subsequently, the easy samples are fed into a new distillation model for teacher learning. Unlike the underfitting models described above, we expect the teacher model to capture as much knowledge as possible from the easy samples, so the teacher model is trained until convergence. The training process for the teacher model is the same as for the student model, with details in Section .

## Distillation: Unrobust Features Distillation

Different from most previous distillation methods for distilling logits, the unrobust features in TACIT do not perform the label prediction task. Therefore, our approach works on features as the distillation target. For each sample in the training set, the teacher model produces an unrobust feature $\tilde{h}$, even if the sample is not an easy sample. To align with $z_\sigma$, $\tilde{h}$ is fed to a simple linear transformation to get $\tilde{z} \in \mathbb{R}^{64}$. To make two different features comparable, we normalize them with a whitening operation, which is implemented by a non-parametric layer normalization operator without scaling and bias (Wei et al. 2022). Then, a smooth $l_1$ loss is used as the loss function for feature distillation:

$$\mathcal{L}_{distill} = \begin{cases} \frac{1}{2}(\zeta(z_\sigma) - \zeta(\tilde{z}))^2/\beta, & |\zeta(z_\sigma) - \zeta(\tilde{z})| \leq \beta \\ |\zeta(z_\sigma) - \zeta(\tilde{z})| - \frac{1}{2}\beta, & otherwise, \end{cases} \tag{4}$$

where $\zeta(.)$ indicates the whitening operation, $\beta$ is a fixed parameter set to 2.0. By optimizing the above loss, the entangled feature $z_\sigma$ can be approached to the features of the unrobust teacher, thus further achieving separation from the robust features.

## Training and Inference

Finally, the main body of training is the student model, so the overall loss function is the joint loss of three different loss functions:

$$\mathcal{L} = (1 - \lambda_1 - \lambda_2) * \mathcal{L}_{ce} + \lambda_1 * \mathcal{L}_{vae} + \lambda_2 * \mathcal{L}_{distill}, \tag{5}$$

where $\lambda_1$ and $\lambda_2$ are the weighted coefficient. Throughout the training process, all parameters of the teacher are frozen as it only provides prior knowledge of unrobust features.

In the process of inference, the encoder part of the student model is used to predict the label of a new sample by the robust feature $z_\mu$, regardless of whether the new sample comes from the source or target domain. After the above process, TACIT does not require any unlabeled data from the target domain for domain adversarial training, but only uses the source domain data to obtain robust model.

[3]https://www.huggingface.co/distilroberta-base

[4]In practice, the top 30% samples are sometimes difficult to guarantee that the teacher model can give intelligent predictions, so we choose a slightly higher proportion.

# Experiments

In this section, we present the datasets required for the experiments, the baselines for comparisons, the results in the single-source and multi-source domains, and the corresponding experimental results with the framework analysis.

## Datasets

We evaluate the proposed TACIT on the most widely used Amazon reviews dataset (Blitzer, Dredze, and Pereira 2007), which contains binary sentiment classification tasks from four different domains: Books (B), DVDs (D), Electronics (E), and Kitchen (K). Each domain contains 1000 positive samples and 1000 negative samples. For each domain, we use a five-fold cross-validation protocol, where 20% of the samples are randomly selected as the development set, and the optimal model on the development set is saved for the target domain generalization test. Publicly available data divisions are used to make fair comparisons (Ben-David, Rabinovitz, and Reichart 2020). Then, compliance with the previous work (Wu and Shi 2022), we give different configurations for the single-source and multiple-source cases. For single-source domains, we train on one dataset and test on the other three. Thus a total of 4*3 = 12 tasks are constructed (Ziser and Reichart 2018). For multi-source domains, we train the model on any three datasets and test it on the remaining one. Thus a total of 4*1=4 tasks are constructed. In addition to the widely used Amazon reviews dataset, we have also compared the proposed approach on a variety of tasks and models. See Appendix for the details and results.

## Baselines

We compare TACIT with the following state-of-the-art approaches to validate the competitiveness of the proposed method:

- **DAAT** (Du et al. 2020). It encourages BERT to capture domain-invariant features through domain-adversarial training, thus improving generalization capabilities.
- **R-PERL** (Ben-David, Rabinovitz, and Reichart 2020). It extends BERT with a pivot-based variant of the Masked Language Modeling (MLM) objective.
- **CFd** (Ye et al. 2020). It introduces class-aware feature self-distillation to self-distill PLM's features into a feature adaptation module, which makes the features from the same class are more tightly clustered.
- **UDALM** (Karouzos, Paraskevopoulos, and Potamianos 2021). It continues the pretraining of BERT on unlabeled target domain data using the MLM task, then trains a task classifier with source domain labeled data.
- **COBE** (Luo et al. 2022). It improves the contrastive learning loss of negative samples within one batch, so that the representations of different classes become further away in the potential space. It is more generalizable on similar tasks by giving more reasonable determinations on categories.
- **AdSPT** (Wu and Shi 2022). It trains vanilla language model with soft prompt tuning and an adversarial training

object, thus alleviating the domain discrepancy of MLM task.

- **Vanilla**. For comparison, we also fine-tune the basic language models BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) with the cross-entropy loss function.

To evaluate the baselines, we use accuracy as the evaluation metric following (Karouzos, Paraskevopoulos, and Potamianos 2021; Wu and Shi 2022). With the exception of AdSPT and CFd, which use RoBERTa and XLM-R, the other approaches use BERT as the backbone language model. We report the optimal results given in the original papers to prevent duplicated code from failing to achieve the results reported in the paper. In addition, we also replicate UDALM using RoBERTa as the backbone based on the official code for a fair comparison, as it is the optimal model on BERT.

## Experimental Details

We initialize our model with $\text{BERT}_{base}$ and $\text{RoBERTa}_{base}$ as the backbone. Accordingly, to ensure that student models can be aligned to the teachers, the teacher models are the corresponding distillation versions, DistilBERT (Sanh et al. 2019) and DistilRoBERTa. All models are trained 10 epochs with batch size 64. The learning rate is set to 1e-5, and the optimizer is AdamW (Loshchilov and Hutter 2019). The weight of the loss function is set to $\lambda_1 = 0.001$ and $\lambda_2 = 0.1$ (See Section  for a detailed discussion). For Encoder and Decoder, we use two symmetric three-layer MLPs where the activation function is ReLU and the hidden layer sizes are 356, 128, and 64, respectively. All experiments are conducted with Pytorch and HuggingFace Transformers on four NVIDIA GeForce RTX 2080 Ti GPUs. Our code is available online[5].

## Results

We report the experimental results of BERT and RoBERTa as the backbones in Table 1. We find that the proposed TACIT is close to the state-of-the-art approaches in both single-source (Table 1) and multi-sources configurations (Table 2), even without using any target domain samples. We also find that different data sources have different impacts on the target domain (Figure 2). The specific descriptions and discussions are as follows.

**Results on single source**. When using BERT as the backbone, UADLM achieves the best results (91.74%) because it performs BERT's MLM training on the target domain, which improves the ability to model the target domain context. But the average result of TACIT is only 0.42% less than that of UDALM without additional tasks for the target domain. This proves that the proposed method can influence the performance of the target domain through reasonable modeling of the source domain. Besides, TACIT works better than CFd (0.69%) and COBE (0.93%), which shows that the disentanglement of robust and unrobust features is more efficient than reasonable in-class modeling.

While using RoBERTa as the backbone, we can observe that TACIT achieves the best average result, which is 0.25% higher than UDALM. This is because RoBERTa is larger and contains more task-related knowledge than BERT. Our feature disentanglement method can assist RoBERTa to make more informed feature choices without the stimulation of the target domain. A more straightforward example can be found in Vanilla, where just fine-tuned RoBERTa yielded better result (91.84%) than $\text{UDALM}_{bert}$, suggesting that RoBERTa is better suited to the task of cross-domain semantic classification. Therefore, in the subsequent parameter exploration and ablation experiment, we used RoBERTa as the backbone.

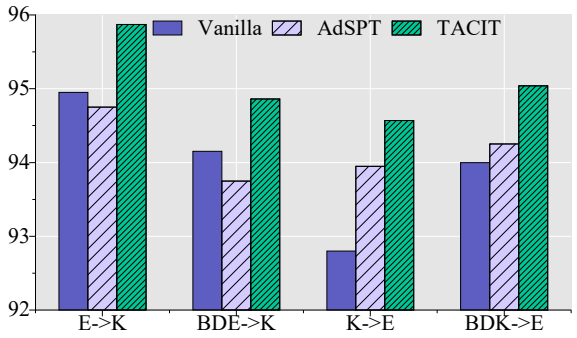


Figure 2: Comparison of single-source and multi-source experimental results on similar data sets K and E.

**Results on multi-sources**. We observe that increasing the source domain generally improves the performance of the target domain on average, due to the commonality among similar tasks. For Vanilla and TACIT, multiple source configurations create 1.28% and 1.08% boosts. But for AdSPT, the improvement is only 0.61%. This suggests that AdSPT is not sensitive to changes in the source domain, which may be due to the fact that it partially relies on data from the target domain, whereas Vanilla and TACIT fully rely on the source domain. It indicates that similar multi-source configurations can better stimulate TACIT's performance improvement. But multi-source configuration is not valid in all cases, as explained in the following paragraph.

**Single-source v.s. Multi-source**. Subsequently, with the same target domain, we further compare the results for single and multiple sources in Figure 2. We select datasets K and E with similar feature distributions. As reported in previous work, generalization between similarly distributed datasets tends to have better experimental results (Wu and Shi 2022), but with the introduction of less similar datasets, a decrease in generalizability may be observed, e.g., the results of BDE→K are inferior to those of E→K. On the contrary, this phenomenon disappears when E is used as the target domain, which suggests that different datasets perform differently when used as source and target domains, indicating the importance of dataset selection in cross-domain text classification tasks.

## Parameter Selection

In the cross-domain generalization task, the grid search of parameters is difficult because of the need to consider multiple target domains. Therefore, we compare the loss values

[5]https://github.com/songruiecho/TACIT

| Source→Target | BERT | | | | | | | RoBERTa | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Vanilla | DAAT | R-PERL | CFd | COBE | UDALM | TACIT | Vanilla | UDALM | AdSPT | TACIT |
| B→D | 88.96 | 89.70 | 87.80 | 87.65 | 90.05 | 90.97 | 91.42 | 91.45 | 92.18 | 92.00 | **92.65** |
| B→E | 86.15 | 89.57 | 87.20 | 91.30 | 90.45 | 91.69 | 91.68 | 93.19 | 93.55 | **93.75** | 93.81 |
| B→K | 89.05 | 90.75 | 90.20 | 92.45 | 92.90 | 93.21 | 92.73 | 93.35 | **95.32** | 93.10 | 95.03 |
| D→B | 89.40 | 90.86 | 85.60 | 91.50 | 90.98 | 91.00 | 91.33 | 91.51 | 93.34 | 92.15 | **93.57** |
| D→E | 86.55 | 89.30 | 89.30 | 91.55 | 90.67 | 92.30 | 91.83 | 90.42 | 93.60 | **94.00** | 93.16 |
| D→K | 87.53 | 87.53 | 90.40 | 92.45 | 92.00 | 93.66 | 91.55 | 92.85 | 93.21 | 93.25 | **94.40** |
| E→B | 86.50 | 88.91 | 90.20 | 88.65 | 87.90 | 90.61 | 89.62 | 91.40 | 91.80 | **92.70** | **92.70** |
| E→D | 87.59 | 90.13 | 84.80 | 88.20 | 87.87 | 88.83 | 89.25 | 89.28 | **93.38** | 93.15 | 92.06 |
| E→K | 91.60 | 93.18 | 91.20 | 93.60 | 93.33 | 94.43 | 94.18 | 94.95 | 94.85 | 94.75 | **95.87** |
| K→B | 87.55 | 87.98 | 83.00 | 89.75 | 88.38 | 90.29 | 89.70 | 91.00 | 92.74 | 92.35 | **93.06** |
| K→D | 87.95 | 88.81 | 85.60 | 87.80 | 87.43 | 89.54 | 89.20 | 89.83 | 92.33 | **92.55** | 91.97 |
| K→E | 90.45 | 91.72 | 91.20 | 92.60 | 92.58 | 94.34 | 93.40 | 92.80 | 93.56 | 93.95 | **94.57** |
| Avg | 88.25 | 90.12 | 87.50 | 90.63 | 90.39 | 91.74 | 91.32 | 91.84 | 93.32 | 93.14 | **93.57** |

Table 1: Single source cross-domain generalization performance for TACIT and baselines. The boldface indicates the optimal results. For each model we report the average results across the five folds. 'Vanilla' denotes fine-tuning on the source domain labeled data. 'Source' denotes training on the source and 'Target' means testing on the target dataset. 'Avg' represents the average of all cross-domain generalization tasks.

| Source→Target | Vanilla | AdSPT | TACIT |
|---|---|---|---|
| DEK→B | 92.70 | 93.50 | **93.64** |
| BEK→D | 91.63 | 93.50 | **95.06** |
| BDK→E | 94.00 | 94.25 | **95.04** |
| BDE→K | 94.15 | 93.75 | **94.86** |
| Avg | 93.12 | 93.75 | **94.65** |

Table 2: Cross-domain generalization results of multiple training sources. The boldface indicates the optimal results. AdSPT is the only method reporting multiple sources in the baselines, so we use RoBERTa$_{base}$ as the backbone for comparison.

for different tasks to determine a rough parameter magnitude, rather than manually adjusting for different datasets. As shown in Figure 3, the cross entropy loss of main task and distillation loss are about the same order, while the VAE loss is much larger. So we set $\lambda_1 = 0.001$ and $\lambda_2 = 0.1$ to ensure that the auxiliary tasks do not unduly affect the optimization of the main task. Although this may lead to non-optimal results, parameter tuning in the face of a new task is economized and is more conducive to task migration easily.

## Ablation Study

To further verify the effectiveness of the proposed method, the following three variants of TACIT are tested:

- TACIT$_{-distill}$. It means that the feature distillation module is not used, and only VAE is used for disentanglement.
- TACIT$_{-random}$. It means randomly selecting 35% of the samples as the training data for the teacher model, rather than selecting the samples with high confidence.
- TACIT$_{-vae}$. It means that VAE is not used, but the output of Encoder is fed directly to two different linear transformations, one whose output is used to predict labels and the other whose output is used for feature distillation.

The results of the ablation studies are shown in Figure 4. All three variants cause TACIT performance degradation, both

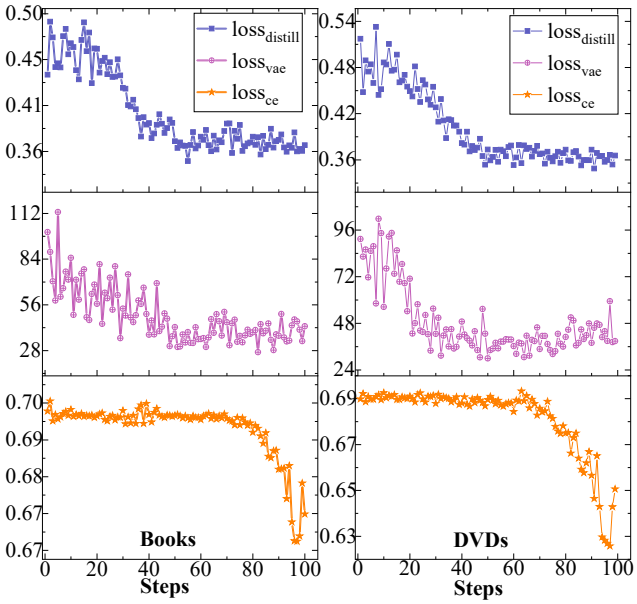


Figure 3: The changes of loss on fold-1 with Books and DVDs as source domains during the model training process. Different styles of lines represent different datasets as well as loss values.

in individual tasks and overall averages. But there are differences between them. Firstly, we observe that TACIT$_{-vae}$ causes the most performance degradation in all but a few cases (K→B). This shows that the biggest factor affecting TACIT is sufficient decoupling of features. If VAE is removed, then the independence between features is abandoned. As a result, the robust features can not be separated. Secondly, we also observe a decline in TACIT$_{-distill}$'s performance, as it is further disentangling features through different tasks. Because it does not destroy the overall architecture of feature disentanglement, the impact is small. Thirdly, TACIT$_{-random}$ also reduces the generalization effect of the model, which shows that easy sample selection based on

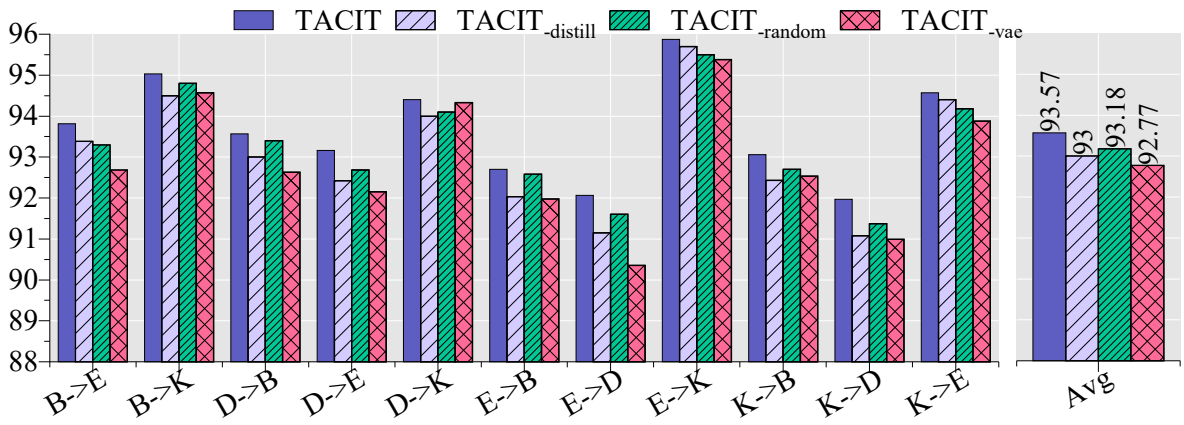


Figure 4: Comparison of ablation results of different cross-domain generalization tasks, where different colors and styles of bars indicate different TACIT variants.

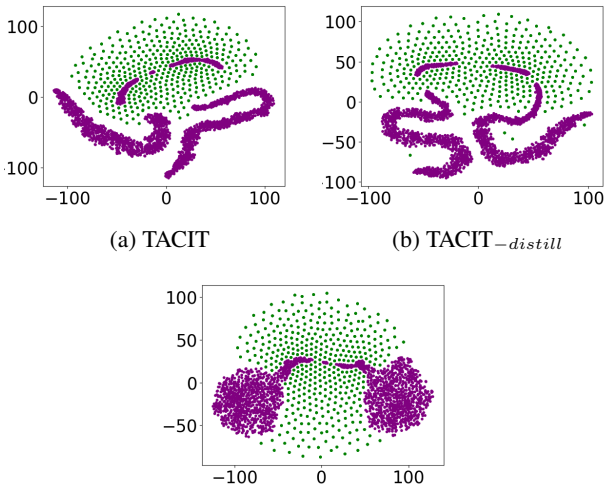


Figure 5: Feature visualisation results of $z_\mu$ and $z_\sigma$ for TACIT and the two corresponding variants TACIT$_{-distill}$ and TACIT$_{-vae}$ on B→D, where the green nodes indicate $z_\sigma$ and the purple nodes indicate $z_\mu$.

confidence is more advantageous than random easy sample selection. In addition, TACIT$_{-random}$ also brings minimal performance degradation, indicating that even with random sample selection, feature distillation still brings some positive effects compared to TACIT$_{-distill}$. Finally, with the exception of B→E, all variants of TACIT achieve better results than Vanilla. Therefore, to sum up, the core components of TACIT all make positive contributions to the improvement of generalization.

## Visualisation

Further, through the visualisation of the representations, we determine the impact of feature disentanglement on cross-domain generalisation. Specifically, tSNE is used to project 64-dimensional features into a two-dimensional space (van der Maaten and Hinton 2008). In Figure 5, we show the visualisation results of B→D. The representation $z_\mu$ is used for classification so that smooth clusters can be obtained with model optimization, which can be observed in three subgraphs. But in Figure 5c, the purple clusters are not as smooth as Figure 5a and Figure 5b, suggesting that VAE can enhance the results of label classification. Besides, the three subgraphs demonstrate differences in $z_\sigma$. Specifically, in Figure 5a, green cluster is more compact and more clearly distinguishable from the purple clusters, suggesting a good separation of the two features. For TACIT$_{-distill}$ in Figure 5b, the green cluster is much looser and some nodes demonstrate a tendency to stray, which suggests that deleting feature distillation has had some negative effects. The most significant impact on the results is the deletion of VAE as shown in Figure 5c, which directly results in green nodes spanning the entire space. The above observations correspond to the results in Section , which further illustrates the effectiveness of the proposed method for feature disentanglement.

## Conclusion

In this paper, facing the challenge of target domain agnostic in cross-domain text classification, we propose a feature disentanglement framework TACIT based only on source domain. TACIT is built on the premise that robust features contribute to classification, while unrobust features are irrelevant. The disentanglement of robust and unrobust features is achieved by variational autoencoders, and this feature separation is exacerbated by additional feature distillation tasks. The experiment of common cross-domain text classification datasets proves that the proposed method can achieve comparable results as the optimal method without using any target domain data.

In the future work, we will explore more judicious methods of easy sample selection to train a more unrobust teacher model. In addition, other language models will be further explored to rate the generalizability of the proposed method.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC), "From Learning Outcome to Proactive Learning: Towards a Humancentered AI Based Approach to Intervention on Learning Motivation" (No. 62077027), and the major project of the National Natural Science Foundation of China (NSFC) "Research on Major Theoretical and Practice Issues in Innovation-Driven Entrepreneurship" (Grant No. 72091310), Project 1 "Developing Theory on Innovation-Driven Entrepreneurship in the Digital Economy" (Grant No. 72091315). The work was also supported by the Education Department of Jilin Province, China (JJKH20200993K) and the Department of Science and Technology of Jilin Province, China (20200801002GH).

## References

Bao, Y.; Zhou, H.; Huang, S.; Li, L.; Mou, L.; Vechtomova, O.; Dai, X.; and Chen, J. 2019. Generating Sentences from Disentangled Syntactic and Semantic Spaces. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, 6008–6019.

Ben-David, E.; Rabinovitz, C.; and Reichart, R. 2020. PERL: Pivot-based Domain Adaptation for Pre-trained Deep Contextualized Embedding Models. *Trans. Assoc. Comput. Linguistics*, 8: 504–521.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, 2172–2180.

Colombo, P.; Staerman, G.; Noiry, N.; and Piantanida, P. 2022. Learning Disentangled Textual Representations via Statistical Measures of Similarity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, 2614–2630.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, 4171–4186.

Du, C.; Sun, H.; Wang, J.; Qi, Q.; and Liao, J. 2020. Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, 4019–4028.

Du, M.; Manjunatha, V.; Jain, R.; Deshpande, R.; Dernoncourt, F.; Gu, J.; Sun, T.; and Hu, X. 2021. Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, 915–929.

Feng, Y.; Li, B.; Qin, L.; Xu, X.; and Che, W. 2023. A Two-Stage Framework with Self-Supervised Distillation For Cross-Domain Text Classification. *CoRR*, abs/2304.09820.

Fotopoulos, S. B. 2007. All of Nonparametric Statistics. *Technometrics*, 49(1): 103.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. S. 2016. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.*, 17: 59:1–59:35.

Geirhos, R.; Jacobsen, J.; Michaelis, C.; Zemel, R. S.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11): 665–673.

Huang, Y.; Zhang, Y.; Chen, J.; Wang, X.; and Yang, D. 2021. Continual Learning for Text Classification with Information Disentanglement Based Regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, 2736–2746.

John, V.; Mou, L.; Bahuleyan, H.; and Vechtomova, O. 2019. Disentangled Representation Learning for Non-Parallel Text Style Transfer. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, 424–434.

Karouzos, C.; Paraskevopoulos, G.; and Potamianos, A. 2021. UDALM: Unsupervised Domain Adaptation through Language Modeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, 2579–2590.

Kawata, T.; and Sakamoto, H. 1949. On the characterisation of the normal population by the independence of the sample mean and the sample variance. *Journal of the Mathematical Society of Japan*, 1(2): 111–115.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*.

Lai, Y.; Zhang, C.; Feng, Y.; Huang, Q.; and Zhao, D. 2021. Why Machine Reading Comprehension Models Learn Shortcuts? In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 989–1002.

Li, S.; Wang, Z.; Jiang, X.; and Zhou, G. 2022. Cross-Domain Sentiment Classification using Semantic Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 289–299.

Li, Z.; Wei, Y.; Zhang, Y.; and Yang, Q. 2018. Hierarchical Attention Transfer Network for Cross-Domain Sentiment Classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, 5852–5859.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V.

2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019*.

Luo, Y.; Guo, F.; Liu, Z.; and Zhang, Y. 2022. Mere Contrastive Learning for Cross-Domain Sentiment Analysis. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022*, 7099–7111.

Moon, S. J.; Mo, S.; Lee, K.; Lee, J.; and Shin, J. 2021. MASKER: Masked Keyword Regularization for Reliable Text Classification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 13578–13586.

Peng, M.; Zhang, Q.; Jiang, Y.; and Huang, X. 2018. Cross-Domain Sentiment Classification with Target Domain Specific Information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, 2505–2513.

Pergola, G.; Gui, L.; and He, Y. 2021. A Disentangled Adversarial Neural Topic Model for Separating Opinions from Plots in User Reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, 2870–2883.

Qu, X.; Zou, Z.; Cheng, Y.; Yang, Y.; and Zhou, P. 2019. Adversarial Category Alignment Network for Cross-domain Sentiment Classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, 2496–2508.

Ramponi, A.; and Plank, B. 2020. Neural Unsupervised Domain Adaptation in NLP - A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*, 6838–6855.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Utama, P. A.; Moosavi, N. S.; and Gurevych, I. 2020a. Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, 8717–8729.

Utama, P. A.; Moosavi, N. S.; and Gurevych, I. 2020b. Towards Debiasing NLU Models from Unknown Biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, 7597–7610.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Wei, Y.; Hu, H.; Xie, Z.; Zhang, Z.; Cao, Y.; Bao, J.; Chen, D.; and Guo, B. 2022. Contrastive Learning Rivals Masked Image Modeling in Fine-tuning via Feature Distillation. *CoRR*, abs/2205.14141.

Wu, H.; and Shi, X. 2022. Adversarial Soft Prompt Tuning for Cross-Domain Sentiment Analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL, May 22-27, 2022*, 2438–2447.

Ye, H.; Tan, Q.; He, R.; Li, J.; Ng, H. T.; and Bing, L. 2020. Feature Adaptation of Pre-Trained Language Models across Languages and Domains with Robust Self-Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, 7386–7399.

Yu, J.; and Jiang, J. 2016. Learning Sentence Embeddings with Auxiliary Tasks for Cross-Domain Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, 236–246.

Zhou, J.; Tian, J.; Wang, R.; Wu, Y.; Xiao, W.; and He, L. 2020. SentiX: A Sentiment-Aware Pre-Trained Model for Cross-Domain Sentiment Analysis. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*, 568–579.

Ziser, Y.; and Reichart, R. 2018. Pivot Based Language Modeling for Improved Neural Domain Adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, 1241–1251.