Domain-Aware Contrastive Knowledge Transfer for Multi-domain Imbalanced Data

Zixuan Ke[*]
UIC Computer Science
Chicago, IL
zke4@uic.edu

Mohammad Kachuee
Amazon Alexa AI
Seattle, WA

Sungjin Lee
Amazon Alexa AI
Seattle, WA

{kachum,sungjinl}@amazon.com

Abstract

In many real-world machine learning applications, samples belong to a set of domains e.g., for product reviews each review belongs to a product category. In this paper, we study multi-domain imbalanced learning (MIL), the scenario that there is imbalance not only in classes but also in domains. In the MIL setting, different domains exhibit different patterns and there is a varying degree of similarity and divergence among domains posing opportunities and challenges for transfer learning especially when faced with limited or insufficient training data. We propose a novel domain-aware contrastive knowledge transfer method called DCMI to (1) identify the shared domain knowledge to encourage positive transfer among similar domains (in particular from head domains to tail domains); (2) isolate the domain-specific knowledge to minimize the negative transfer from dissimilar domains. We evaluated the performance of DCMI on three different datasets showing significant improvements in different MIL scenarios.

1 Introduction

The majority of existing works in imbalanced learning focus on the *class imbalance setting* where classes are presented in a long-tailed distribution: a subset of classes (head classes) have sufficient samples, while other uncommon or rare classes (tail classes) are underrepresented by limited samples. This setting is challenging because the model naturally focuses largely on the majority classes and there may be no sufficient data for tail classes to recover their underlying distribution (Liu et al., 2019).

Even though extensive work has been done on the class imbalance problem, the consideration of *domains*[1] is often missed. In many real-world scenarios, data naturally belongs to a set of domains e.g., for an online store, a potential domain assignment for each customer review can be defined based on the corresponding store departments.

A simplistic solution is to ignore domain assignments and train a classifier for all domains, which we refer to as *domain agnostic learning (D-AL)*. D-AL entirely ignores domains and assumes that the model can "automatically" discover the data distribution for domains and learn them equally well. The drawbacks of such an approach are obvious: if the training data is sourced from many domains, updating all parameters may lead the model to focus on the subsets of the data in proportion to their ease of access or frequency. Moreover, if the data from different domains are dissimilar, agnostic learning may cause undesirable convergence dynamics i.e., negative transfer. We, therefore, argue that in the *multi-domain imbalanced learning (MIL)* scenarios, a learning algorithm should consider domain information and leverage them to achieve effective knowledge transfer.

The MIL is a challenging problem. First, different domains may have very different number of samples and show a long-tailed distribution. For example, an intelligent assistant (e.g. Amazon Alexa) may provide a wide variety of skills and different skills may vary largely in number of examples. It is possible that some internal developed skills (e.g. music or whether) have hundreds of thousands of samples while many third-party developed skills may have only less than 10 samples in the same dataset (Kachuee et al., 2021). Second, domains may exhibit different semantic similarities and disparities with each other. For instance, a feature may show positive correlation with a label for cer-

---

[*]Work done as an intern at Amazon Alexa AI.

[1]In this paper, the term *domain* is used to refer to a segmentation of samples, and it should not be confused with the same term also used in the domain adaptation literature studying the distribution shift problem.

tain domains while it is negatively correlated for others. Third, the data-provided domain annotation may not be completely accurate or sufficiently fine-grained. For example, a sentence *"Due to software or hardware issues, my computer cannot open my favorite text book, One hundred Years of Solitude"* may belong to both *computer* and *books* domains while it may have only one domain assignment in the dataset.

Perhaps the most intuitive approach for MIL is *multi-task learning (MTL)*, where separate heads are used for different domains. While MTL considers domains, we will show it performs poorly in our experiments due to the lack of knowledge transfer between the classifiers. We believe that the key to successful MIL is to not only enable but encourage positive transfer learning across domains.

In this paper, we propose Domain-aware Contrastive Knowledge Transfer for Muti-domain Imbalance learning (DCMI). DCMI introduces a novel *domain-aware representation layer* based on domain embeddings which enables fine-grained and scalable representation sharing or separation. Complementary to the data provided domain assignments, we use an auxiliary domain classification task to help determine the relevance of a *sample* to each domain i.e., *soft domain assignments*. DCMI uses a novel contrastive knowledge transfer objective to move the representation from similar domains closer and representation from dissimilar domains further apart. We conduct extensive experiments on three different multi-domain imbalanced datasets to demonstrate the effectiveness of DCMI.

2 Related Work

The recent imbalance learning literature can be organized into the following categories:

Data Resampling. This is one of the most widely used practices to artificially balance the distribution. Two popular options are under-sampling (Buda et al., 2018; More, 2016) and over-sampling (Buda et al., 2018; Sarafianos et al., 2018; Shen et al., 2016). Under-sampling removes data from the head (dominant classes) while over-sampling repeats the data from the tail (minority classes). These approaches can be problematic as discarding tends to remove important samples and duplicating tends to introduce bias or overfitting.

Data Augmentation. Data augmentation has been used to enrich the tail classes. A popular approach is to leverage the Mixup (Zhang et al., 2018)

technique to augment the minority classes. Remix (Chou et al., 2020) assigns the label in favor of minority classes to the mixup samples, Liu et al. (2020) prepares a "feature cloud" for mixing up that has a larger distribution range for tail classes. Kim et al. (2020) adds noise to head classes to generate tail classes. Chu et al. (2020) decomposes the feature spaces and generate tail classes samples by combining class-shared feature from head classes and class-specific features from tail classes. However, this is usually a non-trivial work to generate meaningful samples that can help tail classes.

Loss Reweighting. The basic idea of reweighting is to allocate larger weight for loss terms corresponding to tail classes while less weight for head classes. In class-sensitive cross-entropy loss (Japkowicz and Stephen, 2002), the weight for each class is inversely proportional to the number of samples. Ren et al. (2018) leverages a hold-out evaluation set to minimize the balanced loss.

Regularization. This approach adds an additional regularization term to improve the training for the tail samples. Lin et al. (2017) adds a factor to the standard cross-entropy loss to put more focus on hard, misclassified samples (usually attributed to the minority classes). Cao et al. (2019) proposed to regularize the minority classes strongly so that the generalization error of minority classes can be improved. While regularization is simple and effective, the soft penalty can be insufficient to make the model focus on the tail classes and a large penalty may negatively affect the learning itself.

Parameter isolation. It has been shown that decoupling the learning into representation learning and classifier learning can be quite effective. BBN Zhou et al. (2020) proposed a two-branch approach where the representation learning branch is trained as there is no class imbalance (input random sampling data) while the classifier learning branch applies the reverse sampling technique. The two branches are then combined by a curriculum learning strategy. Wang et al. (2021) further improves BBN by replacing the cross-entropy loss in representation learning branch into a prototypical supervised contrastive loss. This approach offers the opportunity to optimize each part separately but also make it hard to transfer knowledge from head to tail classes

Domain Imbalanced Learning. The above approaches mostly consider the class imbalance but ignore the imbalance across domains. Cheng et al.

(2020) proposed a doubly balancing technique for both class imbalance and cross-domain imbalance, which only limited to two domains, without any explicit mechanism to encourage the positive transfer and avoid the negative transfer.

3 Problem Definition

In this paper, we assume access to a set of samples $(\boldsymbol{x}_i, y_i, j)$ for $i = \{1 \ldots N\}$, $y_i \in \{1 \ldots C\}$, and $j \in \{1 \ldots M\}$. Here, N is the number of samples, C is the number of classes, and M is number of domains, i.e., shared feature space and label set across domains. We assume the scenario where exists (a) *class imbalance*: classes are not evenly distributed in each domain; (b) *domain imbalance*: domains are not evenly distributed, i.e., some domains may have much more or less number of examples than other domains; and (c) *domain divergence*: while some domains are naturally similar to others and thus positively correlated, some domains are naturally dissimilar to others and negatively correlated. Given these assumptions, in *multi-domain imbalanced learning* (MIL) we seek a model to minimize the expected loss for all domains (i.e., macro average).

4 Proposed Method

Fig. 1 presents an overview of the proposed method. In the MIL problem, it is crucial to identify the *shared knowledge* that can be transferred across similar domains to improve the tail domain performance and the *domain-specific knowledge* that needs to be handled carefully to avoid a negative transfer. To obtain domain-aware representations, we leverage domain embeddings to adaptively select the useful representation for each specific domain (Sec. 4.1). Additionally, regardless of the dataset provided domain assignment, in reality, a sample can belong to multiple domains to different degrees. To address this, we propose a *domain classification* task to obtain the relevance of a sample to each domain and transfer the related domain knowledge using a *contrastive* method (Sec. 4.2).

4.1 Domain-Aware Representation

We suggest a domain-aware representation layer to adaptively select the appropriate representation (neurons) for each domain. For a domain j, the corresponding embedding $\boldsymbol{v}^j$ consists of differentiable parameters that can be learned in an end-to-end fashion. Based on this, the sigmoid function is
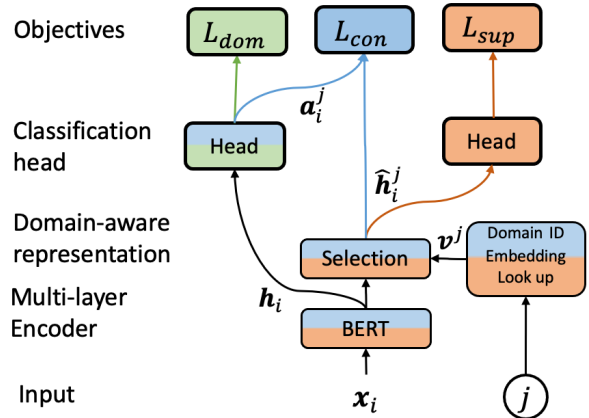

Figure 1: An overview of the DCMI training process. (i) DCMI takes as input a sample $x^{(i)}$ from domain j. (ii) The encoded feature vector $\boldsymbol{h}_i$ is computed using a shared body network (e.g., BERT). (iii) The domain index is used to get the corresponding domain embedding used to compute the domain mask $\boldsymbol{m}_j$ and domain-aware representation $\hat{\boldsymbol{h}}_i^j$. (iv) The supervised classification ($\mathcal{L}_{\text{sup}}$), contrastive ($\mathcal{L}_{\text{con}}$), and domain classification ($\mathcal{L}_{\text{dom}}$) loss terms are computed (see Section 4.2). (v) The flow of gradients from each loss term is controlled such that each term is only used to optimize a subset of trainable parameters as indicated by green, blue, and orange colors in the drawing.

used to find the corresponding domain mask $\boldsymbol{m}^j$:

$$\boldsymbol{m}^j = \sigma(\boldsymbol{v}^j/\tau) \ . \tag{1}$$

Where τ is a temperature variable, linearly annealed from 1 to $\tau_{\min}$ (a small positive value).

To obtain the *domain-aware representation*, we use element-wise multiplication of the output of the body network (i.e., BERT in this paper) $\boldsymbol{h}$ and the mask $\boldsymbol{m}^j$:

$$\hat{\boldsymbol{h}}_i^j = \boldsymbol{h}_i \odot \boldsymbol{m}^j \ . \tag{2}$$

Note that the neurons in $\boldsymbol{m}^j$ may overlap with those in other domain masks to enable knowledge sharing.

To make sure the $\boldsymbol{v}^j$ to have a wide range and its gradient to have a large magnitude, a gradient compensation technique is employed to the original gradient $\boldsymbol{g}$ (Serrà et al., 2018). Specifically,

$$\boldsymbol{g}' = \frac{\tau[\cosh(\boldsymbol{v}^j/\tau) + 1]}{\tau_{\min}[\cosh(\boldsymbol{v}^j) + 1]} \odot \boldsymbol{g} \ . \tag{3}$$

The embedding matrix is trained jointly with the supervised classification objective using a typical cross-entropy loss, denoted by $\mathcal{L}_{\text{sup}}$.

27

4.2 Contrastive Knowledge Transfer

Even though we obtain the domain-aware representation using the suggested domain embedding, there are two limitations: (a) apart from supporting shared features, there is no explicit mechanism to actively encourage knowledge transfer; (b) the dataset provided domains are not necessarily accurate and fine-grained in the real world. Certain examples can be attributed to multiple domains with different degrees of relevance. For example, a review written on a product is usually considered in the general domain of that product (e.g., computers); however, semantically, it may involve discussion of other domains (e.g., the music playback quality of a laptop).

To address the above issues, we employ a domain classification task to estimate the relevance of each sample to different domains. We leverage these relevance/confidence scores as soft labels to conduct contrastive learning, allowing knowledge transfer from similar domains at the instance level.

Domain Classification. To estimate the relevance of different domains for a given sample, we leverage a sigmoid classification head with M output neurons. For training, we employ binary cross-entropy (BCE) loss $\mathcal{L}_{\text{dom}}$ using the dataset provided domain assignments as labels. Using the trained domain classifier, assuming it can generalize and capture domain similarities, we estimate the relevance of sample i to each domain using its sigmoid output score for domain j, denoted by a_i^j.

Note that the domain classification task is only an auxiliary task to be used in the contrastive learning objective explained next. Therefore, we block gradients from this objective to flow outside the domain classifier head.

Contrastive Learning. Fig. 2 shows an illustration of the proposed contrastive objective. Here, for a certain sample, regardless of the dataset provided domain, we compute its domain-aware representations for all domains: $\hat{h}_i^1 \ldots \hat{h}_i^M$. Then, we compute an augmented view of the sample by simply computing a weighted average of domain-aware representations and their normalized relevance:

$$\overline{h}_i = \sum_{j=1}^{M} \frac{a_i^j}{\sum_{j=1}^{M} a_i^j} \hat{h}^j \ . \qquad (4)$$

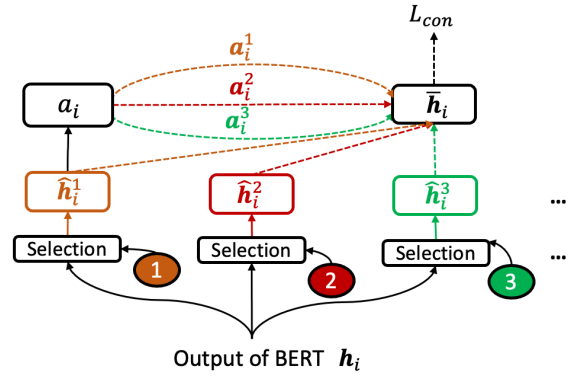Based on this, we define the contrastive objective

Figure 2: An illustration of the contrastive learning objective: (i) Domain-aware representations $\hat{h}_i^j$ are computed for sample i and all domains indexed by j. (ii) Sigmoid outputs of the domain classifier head a_i^j are used to compute a weighted average of domain-aware representations resulting in an augmented view $\overline{h}_i$. (iii) A soft cross-entropy loss based on the augmented view and domain certainties is used as the contrastive objective function.

as:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} a_i^j \log(\sigma(\overline{h}_i \cdot \hat{h}_i^j)) +$$
$$(1 - a_i^j) \log(1 - \sigma(\overline{h}_i \cdot \hat{h}_i^j))) \ , \quad (5)$$

which is essentially a soft cross-entropy loss. Intuitively, the contrastive objective of (5) encourages learning representations that capture the attribution of the augmented view to each domain. Through this objective, similar domains are represented with closer representations and dissimilar domains are moved further apart such that they are easily distinguishable from the augmented view. Note that $\mathcal{L}_{\text{con}}$ is different from the typical contrastive objectives usually used in the literature as it relies on soft domain assignments for the augmented view rather than distinguishing augmented and real data.

As an example, assume that the domain-aware representation $\hat{h}_i^j$ is not a good representation for sample i and lacks knowledge that is potentially transferable from other domains (indicating by a single color in their representation boxes), we can see how $\mathcal{L}_{\text{con}}$ helps (see Fig. 3):

- *Sample i semantically relevant to multiple domains* (domain 1 and domain 3). In this case, a_i^1 and a_i^3 have a large value while a_i^2 has a smaller value. Consequently, $\overline{h}_i$ is mostly the average of $\hat{h}_i^1$ and $\hat{h}_i^3$ (half orange and half
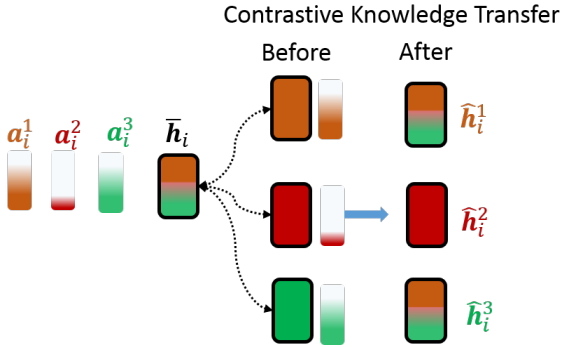
28

Figure 3: A simple example to show the effectiveness of the contrastive knowledge transfer. Orange, red, and green bars show the degrees of relevance to domains 1,2, and 3, respectively. Here, the contrastive objective encourages similar domains (domain 1 and 3) to have similar representations, while the sample belonging to a dissimilar domain (domain 2) is pushed apart in the representation space.

green). Here, updating based on $\mathcal{L}_{con}$ moves $\hat{h}_i^1$ and $\hat{h}_i^3$ closer to $\overline{h}$. In other words, the knowledge transfer is encouraged between the first and third representations for that sample.

- *Sample i is not semantically relevant to a domain* (domain 2). Updating based on $\mathcal{L}_{con}$, $\hat{h}_i^2$ moves further from $\overline{h}_i$ to reflect the difference between them. Consequently, $\hat{h}_i^2$ is discouraged from a negative knowledge transfer. This is expected as $\hat{h}_i^2$ is not relevant to sample i.

4.3 Implementation Details

Final Objective. The final joint training objective is a combination of the supervised classification, domain classification and sample level contrastive loss terms:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_1 \mathcal{L}_{dom} + \lambda_2 \mathcal{L}_{con}, \quad (6)$$

where, λ_1 and λ_2 are hyperparameters to adjust the impact of each term. Note that gradients computed from each objective update different parts of the network as shown in Fig. 1 via different colors. For example, $\mathcal{L}_{dom}$ only updates the domain classifier head, and $\mathcal{L}_{con}$ updates all parameters except those in the supervised classification head.

Architecture. A fully connected layer with softmax output is used as the classification head in the last layer of BERT. We use the embedding of [CLS] as the output of BERT. The training of BERT, follows that of (Xu et al., 2019). We adopt BERT$_{\textbf{BASE}}$ (uncased).

Hyperparameters. Unless otherwise stated, the domain id embeddings have 768 dimensions. We use 0.0025 for τ_{min} in Eq. 3. A dropout layer with the rate of 0.5 is placed between fully connected layers. To find the λ_1 and λ_2 hyperparameters in Eq. 6, we conducted a grid search in the $[0, 5000]$ range using about 200 logarithmic increments. We provide the selected λ_1 and λ_2 for each dataset in Section 5.1.3. For the contrastive objective, an l_2 normalization is applied before computing the contrastive loss. The max length of the number of input tokens is set to 128. We use Adam optimizer and set the learning rate to 3×10^{-5}. For all experiments, we train for 5 epochs using a mini-batch size of 64.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets

We conduct experiments using three datasets: *Document Sentiment Classification (DSC)* (Ni et al., 2019), *Aspect Sentiment Classification (ASC)* (Ke et al., 2021) and *Rumour and Fake News Detection (RFD)* (Zubiaga et al., 2016; Wang, 2017). These datasets have natural class and domain imbalance. For all datasets, we use a random data split of 10% for test, 10% for validation, and the rest for training. To better evaluate the performance of each method in efficient knowledge transfer, we down-sample the training and validation sets of the DSC, ASC, and RFD with a factor of 1000, 10, and 10, respectively. We provide the exact domain and class statistics in the appendix. In addition to these datasets, we conduct additional experiments using an altered version of the ASC dataset with artificially dissimilar domains (Sec. 5.2.2).

DSC. For this dataset, the task is to classify each full product review into one of the two opinion classes (*positive* and *negative*). The training data provides the particular type of product being reviewed as domain information. We adopt the text classification formulation in (Devlin et al., 2019), where the [CLS] token is used to predict the opinion polarity.

To build the DSC dataset, we use 29 domains from the Amazon Review Datasets (Ni et al., 2019) [2], then binarize the ratings by converting 1-2 stars to negative and 4-5 stars to positive.

---

[2] https://nijianmo.github.io/amazon/index.html

ASC. This dataset provides a classification of review sentences on their aspect-level sentiment (one of *positive* and *negative*). For example, the sentence "*The picture is great but the sound is lousy*" about a TV expresses a *positive* opinion about the aspect "picture" and a *negative* opinion about the aspect "sound." We adopt the ASC implementation by Xu et al. (2019), where the aspect term and sentence are concatenated via [SEP] in BERT. The opinion is predicted using the [CLS] token.

The ASC dataset (Ke et al., 2021) consists of 19 domains from 4 sources: (a) *HL5Domains* (Hu and Liu, 2004) with reviews of 5 products; (b) *Liu3Domains* (Liu et al., 2015) with reviews of 3 products; (c) *Ding9Domains* (Ding et al., 2008) with reviews of 9 products; and (d) *SemEval14* with reviews of 2 products - SemEval 2014 Task 4 for laptop and restaurant.

RFD. This dataset is compose of PHEME rumor detection (Zubiaga et al., 2016) and LIAR fake news detection (Wang, 2017) datasets. For rumor detection, the task is to identify whether a piece of given news is a rumor or not, while for the fake news detection, it is to identify fake or real news pieces. We follow Devlin et al. (2019) where the [CLS] token is used for the classification.

The RFD dataset consists of 6 domains from the PHEME dataset (5 domains) of rumor tweets (Zubiaga et al., 2016)[3] and the fake news detection LIAR (Wang, 2017) (1 domain). Note that domains in PHEME defined by different news events (e.g. a specific shooting incident), while the domain in LIAR is defined by news genres (e.g. politics). We intentionally selected this dataset to evaluate the performance of different methods when domains are merely a segmentation of samples rather than following a consistent definition.

5.1.2 Metrics

For each experiment, we report Area Under the ROC Curve (AUC) as the performance measure. Two types of results are reported: *macro* and *micro*. Macro is computed by macro averaging results computed for individual domains. Micro is computed from averaging the performance of all test samples regardless of their domain assignments. Note that there is an imbalance in the frequency of class labels (positive and negative in ASC, DSC; fake and real in RFD) in addition to the imbalance

in the domains for each dataset. To ensure the statistical significance of the results, each experiment is repeated 5 times using random seed and random initialization, reporting the mean and standard deviation of each result.

5.1.3 Comparison Baselines

As the main focus of this study is the domain imbalance, to address class imbalance existing in our benchmarks, we adopt the existing DRS method (Cao et al., 2019) for all experiments. In our comparisons, we use multi-task learning (MTL) and domain-agnostic learning (D-AL) as intuitive and straightforward baselines. Additionally, since little work has been done in MIL, we adapt the recent class imbalance systems to MIL by re-sampling or re-weighting based on the domain statistics. For each case, we follow similar architectures as DCMI to ensure fair comparisons. The compared methods cover various approaches including: loss re-weighting (D-DRW (Cao et al., 2019)), regularization (D-Focal (Lin et al., 2017)), re-sampling (D-DRS (Cao et al., 2019)), parameter isolation (D-BBN (Zhou et al., 2020) and D-HybridSC (Wang et al., 2021)), and mixture-of-experts (D-MDFEND (Nan et al., 2021)). Note that the prefix "D-" in the model name is to indicate that we adapt them to the domain imbalance model.

Among these approaches, D-DRW and D-DRS are re-sampling and re-weighting methods with a deferred training scheduler. As suggested by Cao et al. (2019) the re-sampling or re-weighting are only effective after 80% of epochs have been trained. D-focal is a regularization-based method that uses an carefully designed loss function tailored for imbalanced data. D-BBN and D-HybridSC are two recent parameter isolation approaches that have shown state-of-the-art performance. D-MDFEND is used for multi-domain fake news detection which applies mixture-of-experts to deal with multi-domain transfer and isolation.

Regarding the DCMI hyperparameters i.e. (λ_1, λ_2), we used $(50, 6)$, $(30, 15)$, and $(4, 3)$ for the ASC, DSC, and RFD datasets, respectively. Refer to Section 4.3 for the hyperparameter search space and other implementation details.

5.2 Quantitative Results

5.2.1 Comparison with Other Work

Table 1 presents a comparison of DCMI with other baselines. From this table, DCMI consistently outperforms other competitors for both metrics.

---

[3]https://figshare.com/articles/
dataset/PHEME_dataset_of_rumours_and_
non-rumours/4010619

Model	DSC		ASC		RFD		Altered ASC	
	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
MTL (multitask learning)	74.1±3.1	77.3±3.8	80.0±1.8	84.1±0.7	57.4*	59.1*	76.3±2.9	84.9±2.4
D-AL (domain agnostic)	80.6±3.0	81.3±3.0	82.5±2.3	84.8±1.7	68.8±2.9	70.2±2.6	51.9±1.0	61.1*
D-DRS (Cao et al., 2019)	76.3*	76.6*	84.3±2.7	86.0±2.3	71.4±1.2	72.6±0.9	51.4±0.9	58.3*
D-DRW (Cao et al., 2019)	80.6±3.4	80.9±3.2	76.7*	78.0*	72.6±0.8	74.0±0.6	51.6±1.2	59.1*
D-Focal (Lin et al., 2017)	74.84*	74.97*	75.2*	77.1*	71.4±3.2	72.0±3.4	50.8±0.5	56.7*
D-BBN (Zhou et al., 2020)	79.2±3.7	79.8±3.8	75.6*	77.6*	64.3*	66.1*	49.9±1.4	54.5±3.9
D-HybridSC (Wang et al., 2021)	82.4*	82.4±3.9	83.5±2.2	84.9±2.2	71.2±1.4	72.3±1.2	50.7±1.0	56.7*
D-MDFEND (Nan et al., 2021)	80.5±3.5	80.8*	81.0±3.6	82.8±3.4	69.5±2.0	72.0±2.5	73.8*	83.4*
DCMI (this work)	**83.7**±1.3	**83.8**±1.3	**85.0**±0.7	**87.2**±0.4	**74.2**±1.2	**74.1**±1.0	**77.8**±1.9	**85.2**±1.4

* indicates that we only report the average results and there is a convergence issue due to the small training set or extreme imbalance

Table 1: Comparison of macro and micro averaged AUC results for DCMI (this work) and other baselines.

Specifically, DCMI is much more data-efficient compared to other baselines, as it effectively encourages positive knowledge transfer across domains. Among the three datasets, DCMI has the largest improvement margin for RFD. This can be attributed to the fact that domains in RFD are more diverse than those in ASC and DSC. The sentiment classification domains as in ASC and DSC have similarities as in these tasks positive or negative sentiments are usually expressed with similar words/phrases. For example, wonderful and terrible have similar interpretation for different tasks/domains to express positive or negative sentiment. However, expressions in fake news or rumors are far more diversified, follow more complex semantics, and even contradicting at times. For example, "guns" and "shooting" appear many times in "Charlie Hebdo" domain while they almost never appear in other domains like "Germanwings Flight". Even more interestingly, "Trump" appears frequently in both the fake news of "COVID-19" domains and the real news of "government", therefore it is a significant keyword with different domain interpretation. Under such domain disparities, selectively transferring common knowledge while preventing negative transfer becomes crucial which we believe is addressed by this work.

For the most recent state-of-the-art methods presented in Table 1, we can observe mixed MIL performance results for different datasets indicating less adaptability compared to DCMI. This is perhaps because they do not employ any viable mechanism to explicitly encourage positive transfer.

5.2.2 Extremely Dissimilar Data

We claim that DCMI is capable of adaptively selecting the useful knowledge (neurons) for a given domain and thus robust to extremely dissimilar do-

Model	DSC		ASC		RFD	
	Macro	Micro	Macro	Micro	Macro	Micro
DCMI	**83.7**±1.3	**83.8**±1.3	**85.0**±0.7	**87.2**±0.4	**74.2**±1.2	74.1±1.0
$-\mathcal{L}_{\mathrm{dom}}$	81.9±3.0	82.3±2.7	84.5±1.3	86.7±0.9	73.1±1.7	**74.3**±0.8
$-\mathcal{L}_{\mathrm{dom}}, \mathcal{L}_{\mathrm{con}}$	80.2±3.4	81.0±3.2	82.8±1.6	85.3±1.4	69.5±1.3	69.2±0.9

Table 2: Ablation study of DCMI. "$-\mathcal{L}_{\mathrm{dom}}$" and "$-\mathcal{L}_{\mathrm{con}}$" indicate omitting the domain classification and contrastive loss terms, respectively.

mains. To demonstrate this, we create an artificial case where domains are extremely dissimilar in the dataset by design. Specifically, we divide the ASC dataset into two parts. The first part contains first 10 domains and the second part contains the other 9 domains. We keep the first part as is, while inverting the labels for the second part (i.e., flipping positive to negative and vice versa). Note that in a sentiment classification task such as ASC, domains are highly correlated so inverting labels for half of domains creates a drastic domain disparity.

Table 1 shows the results of using the altered ASC data. We can see all baselines except MTL and D-MDFEND reach on only around 50% AUC. This is because the extremely high domain divergence is causing a severe negative transfer and making it difficult for the majority of baselines to learn a good predictor. However, MTL and D-MDFEND perform better than other baselines, perhaps since negative transfer is reduced due to the use of separate heads for different domains in MTL and mixture-of-experts in D-MDFEND. Nevertheless, DCMI still outperforms MTL and D-MDFEND, confirming that DCMI is not only capable of isolating domain-specific knowledge but also is able to encourage positive transfer among similar domains, which is here for domains within each data part of the altered dataset.

Domains	Review	Label	D-AL	DCMI $-\mathcal{L}_{\text{dom}}, \mathcal{L}_{\text{con}}$	DCMI
Laptop	The nicest part is the low heat output and ultra quiet *operation*.	P.	N.	P.	P.
MicroMP3	The flaw is inside the *Zen*.	N.	P.	N.	N.
Laptop	It feels cheap, the *keyboard* is not very sensitive.	N.	P.	P.	N.
Restaurant	The downstairs bar scene is very cool and chill...	P.	N.	N.	P.
Restaurant	The *sushi* is cut in blocks bigger than my cell phone.	N.	P.	P.	N.

Table 3: Qualitative comparison of predictions for different methods on a set of selected test samples from the ASC dataset (Ke et al., 2021). *Italic* text indicates the aspect in the review. "P." indicates positive and "N." indicates negative assignments.

5.2.3 Ablation Study

We conduct an ablation study to analyze the impact of each objective term. The results of this experiment are presented in Table 2. Here, "-$\mathcal{L}_{\text{dom}}$" indicates DCMI without the domain classification. "-$\mathcal{L}_{\text{dom}}, \mathcal{L}_{\text{con}}$" indicates DCMI without the domain classification and contrastive loss. Note that if we remove the domain-aware representation layer in addition to $\mathcal{L}_{\text{dom}}$ and $\mathcal{L}_{\text{con}}$, DCMI becomes D-AL. Based on the results provided in Table 2 the full DCMI system gives the best results, showing that every suggested component is crucial to the final model performance.

5.3 Qualitative Results

Table 3 shows several examples from ASC test set. For each example, we show the ground truth label (the third column), predictions of D-AL, DCMI and DCMI-[$\mathcal{L}_{\text{dom}}, \mathcal{L}_{\text{con}}$]. By comparing D-AL and DCMI-[$\mathcal{L}_{\text{dom}}, \mathcal{L}_{\text{con}}$], we can see the effectiveness of the domain-aware representation layer. By comparing DCMI and DCMI-[$\mathcal{L}_{\text{dom}}, \mathcal{L}_{\text{con}}$], we can see whether the contrastive knowledge transfer is successful.

In the first row, "quiet" is a positive sentiment word in the "laptop" domain. However, "quite" can indicate negative in other domains (e.g., a "quite" earbud in "MP3" domain indicates negative sentiment). We can see DCMI and DCMI -[$\mathcal{L}_{\text{dom}}, \mathcal{L}_{\text{con}}$] are able to separate the different polarity of the same sentiment word from different domains, while D-AL fails, suggesting that the knowledge selection in DCMI is capable of learning discriminative domain-aware representation.

In the second row, we can see D-AL mistakenly takes the review as positive due to the small amount of training data in the "MP3" domain. DCMI and DCMI -[$\mathcal{L}_{\text{dom}}, \mathcal{L}_{\text{con}}$] can make the correct prediction because of their ability to transfer knowledge from similar domains.

The last three rows of Table 3 showcase where only DCMI is correct. In the "laptop" domain (the third row), "cheap" conveys a negative sentiment in the example. However, "cheap" can indicate positive sentiment in the "laptop" domain if it is talking about the software domain. Therefore, an MIL model that only considers the annotated domain (e.g., DCMI-[$\mathcal{L}_{\text{dom}}, \mathcal{L}_{\text{con}}$]) fails. Similarly, the polarities of "cool" and "chill" depend not only on the dataset provided domain but also on the degrees of domain relevance for a given sample. The last case is an ironic expression, indicating DCMI provides a deeper understanding of the review.

In addition to the presented results, we provide a visual analysis of the domain-aware representation layer using t-SNE in the appendix.

6 Conclusion

In this work, we studied the problem of learning from multi-domain imbalanced data, where not only there is class imbalance but also there is an imbalance among domains with varying degrees of similarity. We proposed a novel technique called DCMI that is capable of identifying the *shared knowledge* that can be transferred to improve the tail domain performance and the *domain-specific knowledge* that needs to be handled carefully to avoid negative transfer. DCMI employs a domain-aware representation layer to adaptively select the relevant knowledge for each domain and leverages a novel contrastive learning objective to encourage knowledge transfer for relevant domains. Based on the experiments using three challenging multi-domain imbalanced datasets, DCMI shows improvements over the current state-of-the-art and demonstrates applicability to different scenarios.

References

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the

class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, pages 1567–1578.

Lu Cheng, Ruocheng Guo, K Selçuk Candan, and Huan Liu. 2020. Representation learning for imbalanced cross-domain classification. In *Proceedings of the 2020 SIAM international conference on data mining*, pages 478–486. SIAM.

Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. 2020. Remix: Rebalanced mixup. In *European Conference on Computer Vision*, pages 95–110. Springer.

Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. 2020. Feature space augmentation for long-tailed data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 694–710. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of ACM SIGKDD*, pages 168–177.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Mohammad Kachuee, Hao Yuan, Young-Bum Kim, and Sungjin Lee. 2021. Self-supervised contrastive learning for efficient user satisfaction prediction in conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4053–4064.

Zixuan Ke, Hu Xu, and Bing Liu. 2021. Adapting BERT for continual learning of a sequence of aspect sentiment classification tasks. In *NAACL-HLT*, pages 4746–4755. Association for Computational Linguistics.

Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. 2020. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13896–13905.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. 2020. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*, pages 2970–2979.

Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for aspect extraction in opinion mining. In *IJCAI*.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546.

Ajinkya More. 2016. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP*, pages 188–197.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR.

Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. 2018. Deep imbalanced attribute classification using visual attention aggregation. In *ECCV*, pages 680–697.

Joan Serrà, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *ICML*, pages 4555–4564.

Li Shen, Zhouchen Lin, and Qingming Huang. 2016. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer.

Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. 2021. Contrastive learning based hybrid networks for long-tailed image classification. In *CVPR*, pages 943–952.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*, pages 422–426. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *NAACL-HLT*, pages 2324–2335. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 9719–9728.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2016. Learning reporting dynamics during breaking news for rumour detection in social media. *CoRR*, abs/1610.07363.

A Detailed Datasets Statistics

In Table 4, 6, and 5, we provide the frequency of samples corresponding to each domain for the ASC, DSC, and RFD datasets.

Domains	Train		Validation		Test	
	N.	P.	N.	P.	N.	P.
Luxury Beauty	1	2	1	1	260	2780
Electronics	61	436	7	54	773	5459
CDs Vinyl	7	99	1	12	89	1243
Appliances	1	1	1	1	3	184
Digital Music	1	12	1	1	401	15883
AMAZON FASHION	1	1	1	1	21	262
Office Products	4	55	1	6	55	693
Books	146	1835	18	229	1834	22946
Gift Cards	1	1	1	1	4	290
Grocery Gourmet Food	7	77	1	9	91	970
Cell Phones Accessories	11	71	1	8	138	890
Prime Pantry	1	9	1	1	692	12160
Home Kitchen	53	457	6	57	663	5719
Magazine Subscriptions	1	1	1	1	22	192
Pet Supplies	19	133	2	16	244	1673
Software	1	1	1	1	222	899
Sports Outdoors	17	193	2	24	212	2415
All Beauty	1	1	1	1	18	498
Automotive	10	118	1	14	132	1475
Musical Instruments	1	16	1	2	1475	20058
Movies TV	29	215	3	26	365	2693
Video Games	4	31	1	3	5502	39327
Tools Home Improvement	13	140	1	17	170	1760
Toys Games	9	126	1	15	121	1575
Patio Lawn Garden	6	52	1	6	83	656
Arts Crafts Sewing	2	35	1	4	2714	43844
Clothing Shoes Jewelry	89	726	11	90	1120	9075
Kindle Store	9	152	1	19	115	1909
Industrial Scientific	1	5	1	1	442	6821

Table 4: The number of samples in each domain and data split for the DSC dataset. "N." indicates negative labels and "P." indicates positive labels.

Dataset	Domains	Train		Validation		Test	
		Fake/Rumour	Real	Fake/Rumour	Real	Fake/Rumour	Real
PHEME	Ferguson	51	17	8	2	258	86
	Charlie Hebdo	97	27	16	4	487	138
	Germanwings Crash	13	14	2	2	70	72
	Sydney Siege	41	31	7	5	210	157
	Ottawa Shooting	25	28	4	4	126	141
LIAR	Politic	199	168	26	16	250	211

Table 5: The number of samples in each domain and data split for the RFD dataset. RFD is composed of PHEME and LIAR data. "N." indicates negative labels and "P." indicates positive labels.

Dataset	Domains	Train		Validation		Test	
		N.	P.	N.	P.	N.	P.
SemEval14	laptop	80	93	66	57	128	341
	restaurant	77	209	26	70	196	728
Ding9Domains	HitachiRouter	3	9	9	18	32	74
	CanonS100	4	6	2	20	11	77
	ipod	3	6	6	13	20	57
	Nokia6600	8	14	15	30	48	134
	DiaperChamp	2	9	5	19	26	70
	CanonD500	1	5	1	13	8	52
	Norton	2	9	8	16	38	60
	MicroMP3	21	9	45	15	170	73
	LinksysRouter	7	3	20	2	59	30
HL5Domains	Creative40G	14	28	35	50	155	184
	ApexAD2600	12	9	26	17	87	85
	Nokia6610	11	5	28	6	114	22
	Nikon4300	8	1	16	4	74	8
	CanonG3	11	2	21	7	89	26
Liu3Domains	Computer	13	4	34	1	101	41
	Router	9	6	19	12	73	50
	Speaker	19	2	31	13	140	36

Table 6: The number of samples in each domain and data split for the ASC dataset. ASC is composed of four datasets. "N." indicates negative labels and "P." indicates positive labels.

B Visual Analysis of the Domain-aware Representation Layer

We visualize sample representations before and after the domain-aware representation layer using for ASC dataset. See Figure 4 for t-SNE visualizations. Here, we color the samples according to their domain assignments.

Before the domain-aware representation layer, we can see the points related to different domains are mixed and hard to differentiate. However, after the domain-aware representation layer, samples with similar colors form clusters, indicating a higher embedding distance for different domains. From this visualization, we can infer that the suggested method is able to learn discriminative domain-aware representations.

(a) Before domain-aware representation layer
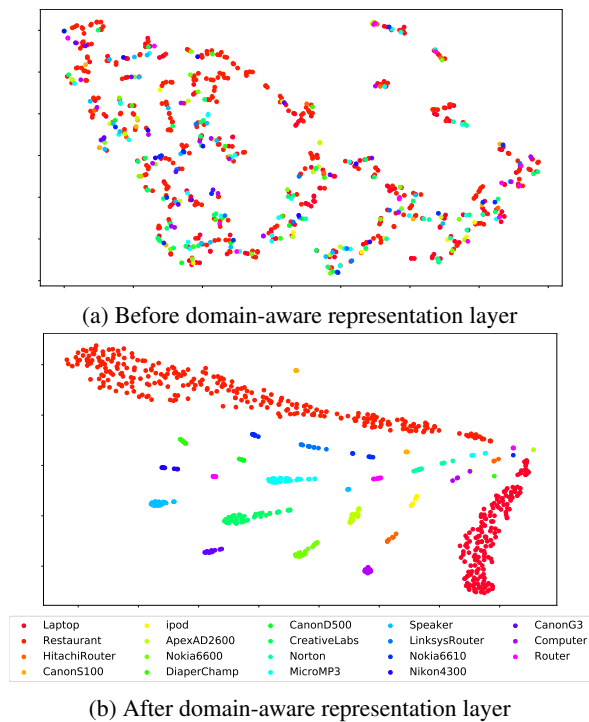

(b) After domain-aware representation layer

Figure 4: t-SNE visualization of sample representation for different domains, (a) before and (b) after the domain-aware representation layer for the ASC dataset (Ke et al., 2021). Figure best viewed in color.