Representation Learning for Imbalanced Cross-Domain Classification

Lu Cheng* Ruocheng Guo* K. Selçuk Candan* Huan Liu*

Abstract

Deep architectures are trained on massive amounts of labeled data to guarantee the performance of classification. In the absence of labeled data, domain adaptation often provides an attractive option given that labeled data of a similar nature but from a different domain is available. Previous work has chiefly focused on learning domain invariant representations but overlooked the issues of label imbalance in a single domain or across domains, which are common in many machine learning applications such as fake news detection. In this paper, we study a new cross-domain classification problem where data in each domain can be imbalanced (*data imbalance*), i.e., the classes are not evenly distributed, and the ratio of the number of positive over negative samples varies across domains (*domain imbalance*). This cross-domain problem is challenging as it entails *covariate bias* in the input feature space and *representation bias* in the latent space where domain invariant representations are learned. To address the challenge, in this paper, we propose an effective approach that leverages a *doubly balancing* strategy to simultaneously control these two types of bias and learn domain invariant representations. To this end, the proposed method aims to learn representations that are (i) robust to data and domain imbalance, (ii) discriminative between classes, and (iii) invariant across domains. Extensive evaluations of two important real-world applications corroborate the effectiveness of the proposed framework.

Keywords:Unsupervised Domain Adaptation, Data Imbalance, Domain Imbalance, Representation Learning

1 Introduction

Learning a transferable classifier in the presence of a *covariate shift* between training and test data is known as *domain adaptation* (DA). When applied to cross-domain classification tasks where both data and domain are imbalanced, the performance of this classifier can be further aggravated by *covariate bias* and *representation bias*. These biases result in distorted estimations of association between features and labels as well as association between representations and labels.

Imbalanced data in classification typically refers to the problems where class distributions are not even.

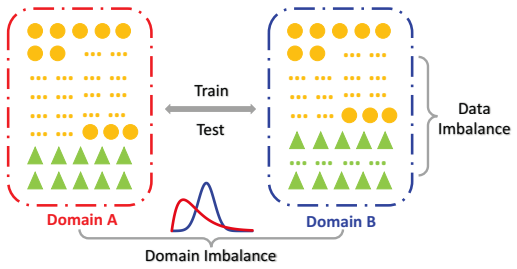


Figure 1: Problem Illustration. The orange dots and the green triangles represent different classes. During training, the cross-domain classifier can access to the imbalanced and labeled data in Domain A and unlabeled data in Domain B. In the test phase, the imbalanced labels in Domain B are available. The *covariate shift* can be aggravated by the imbalanced domains and data.

Data Imbalance can make cross-domain classification tasks more challenging. Take the cross-domain cyberbullying detection task as an example, it is often impossible to train standard classifiers using data from one social media platform (e.g., Twitter[1]) to accurately predict bullying instances on another platform, say Formspring[2], given the number of positive samples (i.e., bullying instances) is limited [6, 5]. As data labeling is often time-consuming, expensive and sometimes even impossible due to data privacy issues, we may have to rely on other domains that have sufficiently large labeled data for training. In addition, cross-domain classification also commonly confronts a *domain imbalance* issue. For instance, the ratio of number of bullying over number of non-bullying samples can differ across various social media platforms. Fig. 1 illustrates a cross-domain classification problem where data in both domains are not balanced (data imbalance) and the ratio of positive and negative classes in the two domains are also different (domain imbalance).

Among the many DA methods, feature mapping has shown tremendous success with the superior advantages of deep neural architectures [24]. The underlying principle is to map features from different domains to a common space and enforce the emergence of domain-

*Computer Science and Engineering, Arizona State University, Tempe. {lcheng35, rguo12, candan, huanliu}@asu.edu

[1]https://twitter.com/
[2]https://spring.me/

invariant representations that reduce domain discrepancy [2]. Previous work on DA shows satisfactory performance in tasks such as cross-domain image classification [10, 17], however, these models are not robust to data imbalance and domain imbalance that often occur between different domains with extremely skewed data. Consequently, domain invariant representations learned using standard DA methods can suffer from covariate bias and representation bias, resulting in low-quality representation that is neither aligned well between domains nor discriminated effectively between classes.

To address these challenges, we introduce a simple doubly balancing framework that consists of a *covariate balancing* (CB) [14] component and a *representation balancing* (RB) component. On one hand, CB is commonly used in observational studies to produce balanced data via sample re-weighing [14]. Hence, we use CB to deal with covariate bias in the input feature space. On the other hand, as DA aims to learn domain invariant representations, it is also necessary to account for the representation bias in the latent space via RB. Standard CB methods may not be ablel to handle the representation bias due to the noise induced by transforming features from the input space to latent space. The doubly balancing strategy, therefore, seeks to jointly learn an optimal re-weighing strategy via CB and RB to account for both covariate and representation bias.

In this paper, we propose a novel Robust Domain Invariant Representation Larning (RIRL) model that leverages the doubly balancing strategy to learn representations for the imbalanced cross-domain classification task. In essence, RIRL seeks to find a trade-off between the predictive accuracy within the source domain and the imbalance error induced by CB and RB across different domains. In this work, we use text classification for illustration because textual data is imbalanced in nature [18], but our model is generic and can be used in other applications. The main contributions of this paper are as follows:

- **Problem Formulation**: We formally define the problem of representation learning for imbalanced cross-domain classification. To the best of our knowledge, this is the first work in DA that considers both *covariate bias* and *representation bias* induced by *domain imbalance* and *data imbalance*.
- **Algorithms**: We propose a new domain invariant representation learning framework, RIRL, that leverages a new doubly balancing strategy. This framework mainly consists of three components: (i) a main component that minimizes the prediction error in the source domain; (ii) a CB component that aims to re-weigh and balance the skewed data distributions; and (iii) an RB component that makes the representation distribution in the target domain mimic the distribution in the source domain.
- **Evaluation**: We perform empirical evaluation on two real-world applications, where data is typically imbalanced. Experimental results show that the proposed framework often outperforms the state-of-the-art methods.

2 Related Work

Most methods for learning domain invariant representations are feature-based. This includes asymmetric [15, 9] and symmetric feature-based methods [10, 24]. One such method seeks to minimize maximum mean discrepancy (MMD) [12] metric between distributions of the source and target domains in the shared space. For example, the deep domain confusion (DDC) method [29] tries to integrate MMD in the last fully connected layer. Correlation alignment (CORAL) method [26, 27] is proposed to align the second order statistics of the source and target domains with a linear transformation. It is then extended to a deep CORAL which learns a nonlinear transformation that aligns the correlations of layer activation in deep neural network [28]. Domain adversarial neural network (DANN) [1, 11] learns domain invariant features by a minimax game between the domain classifier and the feature extractor. Different from the previous asymmetric feature-based methods, the Wasserstein Distance Guided Representation Learning (WDGRL) model [24] is a symmetric approach that projects both domains to a common latent space to learn domain invariant representations. Besides learning shared representations, domain separation network (DSN) [3] attempts to explicitly separate private representations for each domain and shared ones between the source and target domains.

Causal feature selection targets at a similar problem regarding feature transportability/generalizability [20, 4]. Causal feature selection differs from standard feature selection in that it attempts to learn causal associations between features and labels rather than correlated relationship. Causal feature selection gets closer to the causal mechanisms and reveal a more refined notion of relevance, i.e. *causes* [13]. In [4], the author proposes a propensity score matching technique for learning causal associations between word features and class labels in document classification task. The proposed causal text classifier seeks to identify generalizable features that make valid predictions when applied to out-of-domain data [4].

Data imbalance and domain imbalance often occur in many machine learning applications, nevertheless, existing DA approaches fail to address this problem. Covariate and representation bias brought by data and domain imbalance can lead to systematic differences between the distributions over domains, thus can severely affect the quality of the learned representations.

3 Problem Definition

We use a binary text classification task (e.g., cyberbullying and fake news detection) as an example for problem illustration. Suppose that a corpus of text is from the input space X, each $x \in X$ is represented as a d-dimensional vector, i.e., $x \in \mathbb{R}^d$. The associated labels (output) $y \in \{0, 1\}$ are from the label space Y with number of instances of one category substantially larger than that of the other. We further assume that there exist two distributions $\mathcal{S}(x, y)$ and $\mathcal{T}(x, y)$ on $X \otimes Y$, which are referred as the source and target distributions. $\mathcal{S}$ is assumed to be "shifted" from $\mathcal{T}$ by some *covariate shift* [10] and both distributions are assumed to be unknown, imbalanced. In addition, the ratio of classes in these two domains can be different.

Let $N_{\mathcal{S}}$, $N_{\mathcal{T}}$ be the number of instances in $\mathcal{S}$, $\mathcal{T}$. Consequently, we can access to $N_{\mathcal{S}} + N_{\mathcal{T}}$ training samples $\{x_{\mathcal{S}1}, x_{\mathcal{S}2}, ..., x_{\mathcal{S}N_{\mathcal{S}}}, x_{\mathcal{T}1}, x_{\mathcal{T}2}, ..., x_{\mathcal{T}N_{\mathcal{T}}}\}$ from both source and target domains distributed according to the marginal distributions $\mathcal{S}(x)$ and $\mathcal{T}(x)$. Labels of samples from the source distribution $\{y_{\mathcal{S}1}, y_{\mathcal{S}2}, ..., y_{\mathcal{S}N_{\mathcal{S}}}\}$ are known while those of the target distribution $\{y_{\mathcal{T}1}, y_{\mathcal{T}2}, ..., y_{\mathcal{T}N_{\mathcal{T}}}\}$ are not available during training. Suppose we have $P_{\mathcal{S}}(Q_{\mathcal{S}})$ positive(negative) samples in $\mathcal{S}$, and $P_{\mathcal{T}}(Q_{\mathcal{T}})$ positive(negative) samples in $\mathcal{T}$. Let D be the dimension of learned representations, the proposed imbalanced cross-domain classification problem can then be defined as

Definition 1 (Imbalanced Cross-Domain Classification)**.** Given labeled samples $(x_{\mathcal{S}i}, y_{\mathcal{S}i})$ from the source domain and unlabeled samples $x_{\mathcal{T}j}$ from the target domain, $\frac{P_{\mathcal{S}}}{Q_{\mathcal{S}}} \neq 1$ and $\frac{P_{\mathcal{T}}}{Q_{\mathcal{T}}} \neq 1$. In addition, $\frac{P_{\mathcal{S}}}{Q_{\mathcal{S}}} \neq \frac{P_{\mathcal{T}}}{Q_{\mathcal{T}}}$; we seek a representation learning model $\Phi : \mathcal{X} \to \mathbb{R}^D$ and a hypothesis $h : \mathbb{R}^D \to \{0, 1\}$ to minimize the expected classification error $\mathbb{E}_{j \in \mathcal{T}}[L(\hat{y}_{\mathcal{T}j}, y_{\mathcal{T}j})]$ over the instance j in the target domain.

4 RIRL: The Framework

In this section, we describe the details of the proposed RIRL framework, that for each input x from the target domain, predicts its label $y \in Y$ using the classifier trained on the labeled data in the source domain. The proposed deep feed-forward framework consists of three components: (1) a label predictor that minimizes the prediction error in the source domain; (2) a CB component that aims to resolve covariate bias in the input space; and (3) an RB component that is designed to reduce representation bias in the latent space. The RIRL workflow can be viewed in Fig. 2.

4.1 A Naïve Classifier for Imbalanced Data

RIRL builds upon a naïve classifier that accounts for data imbalance by re-weighing training samples in the source domain. First, we project the input $x \in \mathbb{R}^d$ in the source domain to a D-dimensional latent space by the *representation functions* of the form $\Phi : \mathcal{X} \to \mathbb{R}^D$, where $\mathbb{R}^D$ denotes the representation space. The feature mapping may include several hidden layers and we denote the vector of parameters of all layers as θ_Φ. Then to perform the classification task, we seek a hypothesis h with parameters θ_h which transfers the latent representations to the label space Y. Now, an imbalanced classifier can be learned through the following objective function:

$$\min_{\theta_\Phi, \theta_h} \frac{1}{N_{\mathcal{S}}} \sum_{i=1}^{N_{\mathcal{S}}} v_i \cdot L(h(\Phi(x_{\mathcal{S}i}, \theta_\Phi), \theta_h), y_{\mathcal{S}i}), \tag{4.1}$$
$$\text{with } v_i = \begin{cases} \frac{P_{\mathcal{S}}+Q_{\mathcal{S}}}{P_{\mathcal{S}}}, & y_{\mathcal{S}i} = 1 \\ \frac{P_{\mathcal{S}}+Q_{\mathcal{S}}}{Q_{\mathcal{S}}}, & y_{\mathcal{S}i} = 0. \end{cases}$$

Here, $L(\cdot, \cdot)$ is the loss function for binary label prediction (e.g. logistic loss, cross-entropy loss) and the weight v_i compensates for the imbalanced count of positive and negative samples in the source domain. It is computed by the inverse proportion of class sample i belongs to.

To generalize the classifier to make better predictions for unseen data in the target domain, we further add a squared ℓ_2-norm for model parameters to prevent the model from overfitting.

$$\min_{\theta_\Phi, \theta_h} \frac{1}{N_s} \sum_{i=1}^{N_s} v_i \cdot L(h(\Phi(x_{si}, \theta_\Phi), \theta_h), y_{\mathcal{S}i}) + \alpha \mathcal{C}(h), \tag{4.2}$$

where α is the parameter that controls the model complexity $\mathcal{C}(h)$.

4.2 Covariate Balancing between Domains

CB has been widely used to reduce covariate bias in observational studies [14]. A common approach is to re-weigh the instances in one of two groups with sample weights $W \in \mathbb{R}^{N_{\mathcal{T}} \times 1}$ so that the data distributions of these two groups are closer [17]. Formally, the objective to learn the sample weights W is defined as:

$$\min_W \|\bar{x}_t - \Sigma_{j:T_j=0} w_j \cdot x_j\|_2^2, \tag{4.3}$$

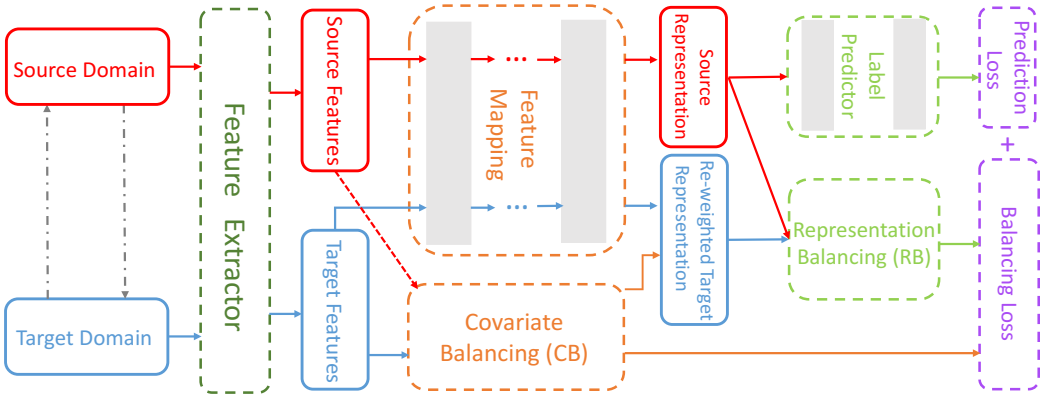


Figure 2: The proposed RIRL framework includes a feature extractor (dark green), a deep feature mapping component (orange) and a deep label predictor (green), which together form a standard deep feed-forward framework for the classification task. Robust cross-domain classification is achieved by leveraging CB for the input data and RB for the learned representations over the source and target domains. Best viewed in colors.

where $T_j \in \{0, 1\}$ represents the group sample j belongs to. $\bar{x}_t$ denotes the average value of variables in group with $T_j = 1$. Solutions to Eq. 4.3 can be used to produce more balanced samples for both groups under the following assumption:

Assumption 1 (Neglect Assumption)**.** *For all x in the support of X_i,*

$$P(R_i = r|X_i = x) > 0, \quad P_r(x), Q_r(x) > 0, \quad r \in \{\mathcal{S}, \mathcal{T}\},$$

where R_i denotes the domain to which sample i belongs. This assumption states that number of positive and negative samples in both domains should be larger than zero. Here, we adapt Eq. 4.3 to a more flexible model that later can be used to reweigh the non-linear relationship in RB. Given a mapping function l that projects data in the target domain to the sample weights space, we introduce the objective function for CB below:

$$\begin{aligned} &\min_{\theta_w} \|\bar{x}_s - \Sigma_{j\in\mathcal{T}} w_j \cdot x_j\|_2^2, \\ &\text{s.t.} \qquad \mathbf{1}^T W = 1, \quad W \geq 0, \end{aligned} \tag{4.4}$$

where $w_j = l(\theta_w, x_j)$, $\theta_w \in \mathbb{R}^d$ denotes the mapping parameters of l and $\bar{x}_s \in \mathbb{R}^d$ is a vector with entries equal to the mean of each variable in the source domain. The term $\mathbf{1}^T W = 1$ normalizes the sample weights in the target domain to add up to one. $W \geq 0$ ensures the sample weights are non-negative. We then incorporate Eq. 4.4 to the classifier in Sec. 4.1 and the objective function for the cross-domain classification is

$$\begin{aligned} &\min_{\theta_\Phi, \theta_h, \theta_w} \frac{1}{N_\mathcal{S}} \sum_{i=1}^{N_\mathcal{S}} v_i \cdot L(h(\Phi(x_{\mathcal{S}i}, \theta_\Phi), \theta_h), y_{\mathcal{S}i}) \\ &\qquad +\alpha\mathcal{C}(h) + \beta\|\bar{x}_s - \Sigma_{j\in\mathcal{T}} w_j \cdot x_{\mathcal{T}j}\|_2^2, \\ &\text{s.t.} \qquad \mathbf{1}^T W = 1, \quad W \geq 0, \end{aligned} \tag{4.5}$$

where β controls the contribution of CB. To resolve the representation bias and enforce the similarities between the representations across domains, we next introduce the RB component.

4.3 Robust Domain Invariant Representation via Doubly Balancing Standard DA approaches may fail to learn high-quality representations when applied to scenarios with data and domain imbalance due to the representation bias. In this section, we propose to incorporate an RB component that seeks to reduce representation bias and the domain discrepancies via the enforcement of similarity between source and target representations. Together with CB, RIRL jointly learns an optimal re-weighing strategy and the representation function Φ to simultaneously account for covariate and representation biases.

The goal of Φ is to push the distributions of source and target domains into a new common space $\mathcal{R}$. This is traditionally achieved by minimizing the distance between the representation distributions of source and target domains. However, this method may not be effective when the learned representations are polluted by representation bias along with covairate bias. Consequently, here, we introduce a doubly balancing technique that leverages both CB and RB. Similar to CB, the core idea of RB is to re-weigh the representation distribution in

the target domain with sample weights W and make the re-weighted representation distribution in the target domain mimic the distribution in the source domain.

Let $p_\Phi^{\mathcal{S}}$, $p_\Phi^{\mathcal{T}}$ be the representation distributions of source and target domains induced over $\mathcal{R}$ such that they satisfy Assumption 1. We measure the distance between $p_\Phi^{\mathcal{S}}$ and $p_\Phi^{\mathcal{T}}$ using *Integral Probability Metrics* (IPMs) [19]. Following previous work [24, 8, 23], we use family of norm-1 reproducing kernel Hilbet space (RKHS) functions, particularly, the *maximum mean discrepancy* (MMD) [12]. RIRL adopts MMD with a linear kernel and a mixture of RBF kernels because they are sufficiently rich to uniquely identify the distance between $p_\Phi^{\mathcal{S}}$, $p_\Phi^{\mathcal{T}}$ and also easy to implement. The squared linear MMD for $p_\Phi^{\mathcal{S}}, p_\Phi^{\mathcal{T}}$ on the samples $X_{\mathcal{S}}$, $X_{\mathcal{T}}$ can then be calculated by the following equation:

$$\text{IPM}_{Linear}(p_\Phi^{\mathcal{S}}, p_\Phi^{\mathcal{T}}) = \sum_{i=1}^{D} (\bar{x}_{\mathcal{S}i} - \bar{x}_{\mathcal{T}i})^2. \tag{4.6}$$

MMD with RBF kernels is computed by

$$\text{IPM}_{RBF}(p_\Phi^{\mathcal{S}}, p_\Phi^{\mathcal{T}}) = \frac{1}{N_{\mathcal{S}}(N_{\mathcal{S}}-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{N_{\mathcal{S}}} k(x_{\mathcal{S}i}, x_{\mathcal{S}j}) + \frac{1}{N_{\mathcal{T}}(N_{\mathcal{T}}-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{N_{\mathcal{T}}} k(x_{\mathcal{T}i}, x_{\mathcal{T}j}) - \frac{2}{N_{\mathcal{S}} N_{\mathcal{T}}} \sum_{i,j=1}^{N_{\mathcal{S}}, N_{\mathcal{T}}} k(x_{\mathcal{S}i}, x_{\mathcal{T}j}) \tag{4.7}$$

where $k(\cdot,\cdot)$ is the RBF kernels of the form

$$k(x,y) = \exp\left(- \frac{\|x-y\|^2}{2\sigma^2} \right), x \in X_{\mathcal{S}}, y \in X_{\mathcal{T}}. \tag{4.8}$$

We then formulate RB as

$$\text{IPM}_{\mathcal{F}}\big(\{\Phi(x_{\mathcal{S}i})\}_{i\in\mathcal{S}}, \{w_j \Phi(x_{\mathcal{T}j})\}_{j\in\mathcal{T}}\big). \tag{4.9}$$

This leads to the following optimization problem:

$$\begin{aligned} \min_{\theta_\Phi, \theta_h, \theta_w} & \frac{1}{N_{\mathcal{S}}} \sum_{i=1}^{N_{\mathcal{S}}} v_i \cdot L(h(\Phi(x_{\mathcal{S}i}, \theta_\Phi), \theta_h), y_{\mathcal{S}i}) \\ & + \alpha \mathcal{C}(h) + \beta \|\bar{x}_{\mathcal{S}} - \Sigma_{j\in\mathcal{T}} w_j \cdot x_{\mathcal{T}j}\|_2^2 \\ & + \gamma \text{IPM}_{\mathcal{F}}\big(\{\Phi(x_{\mathcal{S}i})\}_{i\in\mathcal{S}}, \{w_j \Phi(x_{\mathcal{T}j})\}_{j\in\mathcal{T}}\big), \\ \text{s.t.} \quad & \mathbf{1}^T W = 1, \quad W \geq 0, \end{aligned} \tag{4.10}$$

where γ, together with α and β, controls the trade-off between predictive accuracy and imbalance error from CB and RB. Essentially, the sample weight W is jointly optimized through CB and RB, therefore, accounts for both covariate and representation bias. We normalize Φ through either projection or batch-normalization with fixed scale. Prediction of samples in the target domain is then computed by $\hat{y}_t = h(\Phi(x_t, \theta_\Phi), \theta_h), t \in \mathcal{T}$.

We train the models by minimizing Eq. 4.10 using the Adam optimization algorithm [16], where the error is backpropagated through the hypothesis, CB and representation networks with one mini-batch at a time. After each iteration, we further normalize the updated W and Φ, and keep W non-negative to satisfy the constraints. As we see in the next section, the learned representations are domain invariant, discriminative, and robust to both data and domain imbalance.

Table 1: Dataset Statistics

Dataset	#Neg	#Pos	#Total	Ratio
Formspring	12,036	1,126	13,162	10.7:1
Twitter	11,335	3,645	14,980	3.1:1
Gossip	3,728	934	4,662	4.0:1
Fake	2,118	869	2,987	2.4:1

5 Experiments

For evaluations, we seek to answer the following research questions: (1) How effective is the proposed framework RIRL compared to the following three types of methods: (i) existing DA algorithms, (ii) text classification models which aim at learning more generalizable features and (iii) common classification models that account for data imbalance? (2) How does each module, i.e., CB and RB, affects the performance of RIRL in cross-domain imbalanced text classification tasks? (3) How does the performance of RIRL vary as the classes ratio ($Q_{\mathcal{S}}/P_{\mathcal{S}}$) changes in the source domain? (4) How robust is RIRL to different settings of hyperparamters? To answer these research questions, we test RIRL on cross-domain cyberbullying detection and fake news detection, and compare our model with various baseline models. We also exam the robustness of RIRL via parameters sensitivity analysis.

5.1 Dataset Descriptions. We first provide details of the four real-world datasets used in fake news and cyberbullying detection, respectively.

Fake News Datasets. The first dataset for fake news detection is the *Gossip* dataset [25] crawled from GossipCop[3], a website for fact-checking entertainment stories aggregated from various media outlets. The second dataset is a kaggle[4] dataset which includes the fake and real news crawled from multiple news organizations. The original kaggle dataset has approximate 1:1 ratio of

[3] https://www.gossipcop.com/

[4] https://www.kaggle.com/jruvika/fake-news-detection

Table 2: Performance comparisons of various methods w.r.t the cyberbullying detection task.

Models		S-LR	S-NN	CTC	CORAL	DANN	WDGRL	RIL	RIRL-C	RIRL-L	RIRL-R
Twitter →	F1	0.227	0.213	0.255	0.212	0.258	0.243	0.259	0.267	0.279	**0.283**
Formspring	AUC	0.632	0.620	0.615	0.587	0.637	0.613	0.625	0.699	0.703	**0.712**
Twitter ←	F1	0.496	0.500	0.462	0.383	0.362	0.125	0.430	0.496	0.457	**0.508**
Formspring	AUC	0.669	0.673	0.648	0.604	0.602	0.513	0.631	0.713	0.757	**0.778**

Table 3: Performance comparisons of various methods w.r.t the fake news detection task.

Models		S-LR	S-NN	CTC	CORAL	DANN	WDGRL	RIL	RIRL-C	RIRL-L	RIRL-R
Fake →	F1	0.128	0.124	**0.259**	0.173	0.206	0.183	0.182	0.191	0.218	0.244
Gossip	AUC	0.394	0.425	0.468	0.487	0.477	0.496	0.430	0.451	**0.504**	0.497
Fake ←	F1	0.264	0.264	0.173	0.259	0.240	0.040	0.217	0.264	0.275	**0.299**
Gossip	AUC	0.414	0.415	0.475	0.479	0.499	0.505	0.387	0.447	**0.511**	0.454

#real news over #fake news, which is not common in practice [25]. Therefore, we randomly sample half of the positive samples and together with all the negative samples to form a new dataset *Fake*.

Cyberbullying Datasets. For cyberbullying detection task, we use two real-world datasets crawled from Twitter and Formspring. These two social media platforms are often reported with the most occurrences of cyberbullying instances. *Formspring* and *Twitter* datasets have been collected and used in [21] and [7], respectively. Basic statistics of the four datasets used for experiments can be seen in Table 1.

5.2 Experimental Settings To test the effectiveness of the proposed model, we compare RIRL with the following baselines:

SMOTE This is a common approach to constructing classifiers from imbalanced datasets. It combines the method of oversampling the minority class and undersampling the majority class to generate a balanced dataset. We apply SMOTE to two common classification models: Logistic Regression (**S-LR**) and Multi-layer Perceptron classifier (**S-NN**).

CORAL This is the correlation alignment that minimizes domain discrepancy by aligning the second-order statistics of the source and target distributions and can be applied to the layer activation in neural networks.

DANN This is an adversarial representation learning approach with a domain classifier aimed at distinguishing the learned source/target features while a feature extractor attempts to confuse the domain classifier.

WDGRL This model utilizes a neural network to estimate empirical Wasserstein distance between the source and target samples and optimizes the feature network to minimize the estimated Wasserstein distance in an adversarial manner.

CTC A text classifier that accounts for covariate bias and seeks to learn causal relationships between word features and document labels. This model applies a propensity score matching method [22] to identify features that are more generalizable to different domains.

RIL The variant of RIRL without CB term.

RIRL-C The variant of RIRL without RB terms, i.e. the naïve text classifier (CB does not have effects on the prediction without RB).

RIRL-L RIRL using MMD with linear kernel.

RIRL-R RIRL using MMD with RBF kernel.

Note that we decided not to perform sampling methods on the input data to make it balanced for the domain adaptation models among baselines because preliminary experimental results showed that it can, in fact, harm the F1 scores of most of these models. For data preprocessing, we perform standard text preprocessing procedures such as stop-words removal, stemming for all datasets and extract the Bag of Words features as the model input. RIRL is implemented as a feed-forward neural network with two hidden layers for the feature mapping and one hidden layer for the label prediction. The mapping function l is a linear function followed by ReLU activation. The batch size is set to 128 and the layer size is set to 100 for all the models. We perform sensitivity analyses for hyperparamters α, β, γ (Sec. 5.5) and selected the parameters that give the best performance. Specifically, we set $\alpha = 0.01, 0.1$; $\beta = 1\text{e}-3, 0.1$ and $\gamma = 1\text{e}-7, 1\text{e}-3$ in the cyberbullying and fake news detection tasks, respectively. We use the recommended parameters setting for the baseline models. The evaluation methods include two widely used metrics - AUC and F1 score. Different from accuracy, AUC provides an aggregate measure of performance across all possible classification thresholds, therefore, is a more appropriate metric to evaluate models with imbalanced data [7].

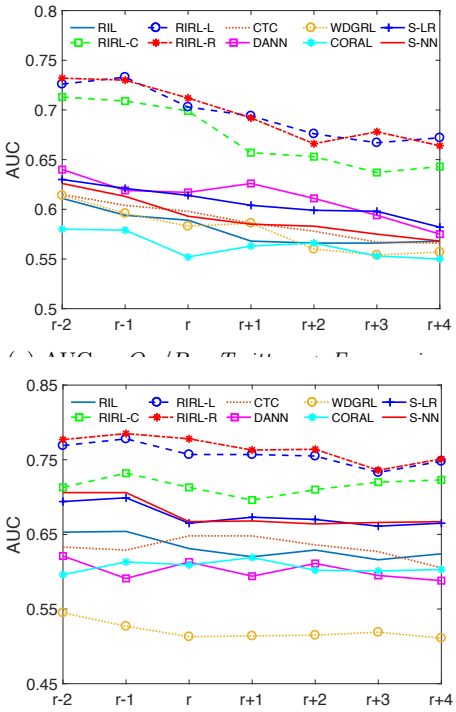
5.3 Quantitative Evaluation Results To answer the first two research questions, we run experiments on the two cross-domain imbalanced classification tasks and present the results in Table 2-3. We highlight the best performance with bold text and underline the second best results. We list the following findings:

- The proposed models mostly achieve the best performance regarding all evaluation metrics. Specifically, for the cyberbullying detection datasets, RIRL-L/RIRL-R presents the highest or second highest F1 and AUC scores. The improvement is significant, especially for AUC. For instance, in *Twitter* → *Formspring* task, RIRL-R improves AUC score by 11.8% compared to the best baseline DANN.
- Compared to its variants RIRL-C and RIL, RIRL achieves the best AUC and F1 scores in both tasks. This result sheds light on the effectiveness of the proposed doubly balancing strategy. Surprisingly, the naïve classifier RIRL-C often shows better performance than RIL as well as other domain invariant representation learning methods. It manifests the importance of re-weighing samples for imbalanced classification.
- Comparing the results of RIRL-L with that of RIRL-R, we can observe that the two IPM distance metrics, i.e., MMD with linear and with RBF kernels, have similar influence on learning domain invariant representations. We can choose either one when applied in real-world applications.

To answer the third question, we further perform experiments on the cyberbullying datasets[5] to investigate the influence of various class ratios $r = Q_S/P_S$ in the source data on AUC score. We randomly sample data from source domain to vary r among the range $\{r-2, r-1, ..., r+4\}$ for training and keep the ratio in test data fixed. We present the results in Fig. 3. A larger Q_S/P_S implies a more imbalanced dataset in the source domain. We can observe from the results that the proposed models consistently achieve the best AUC scores when varying the class ratio from close to 1 to a relatively large value. It corroborates the effectiveness of the doubly balancing strategy in dealing with extremely imbalanced as well as more balanced datasets.

5.4 Qualitative Examination via Visualization To further investigate the quality of these domain invariant representations learned by various models, we

[5]Similar results from fake news detection are omitted here for space limitation.

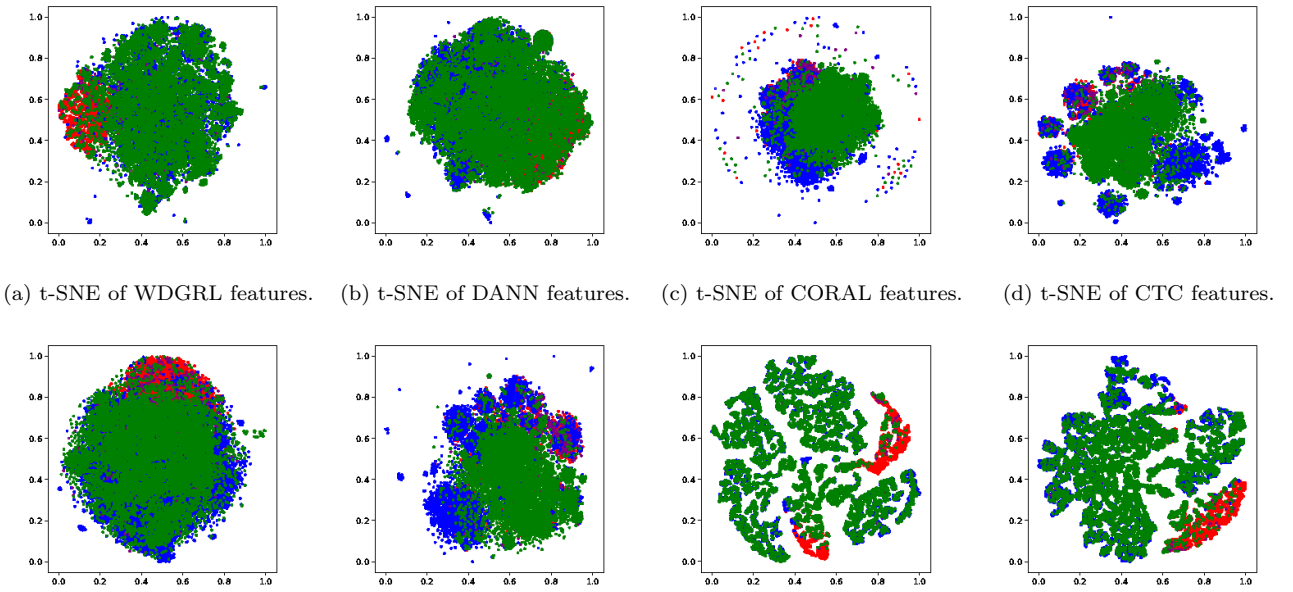


(b) AUC vs Q_S/P_S: *Formspring* → *Twitter*.

Figure 3: Performance evaluation w.r.t. different ratios Q_S/P_S in the source domain. Best viewed in colors.

randomly choose the task *Formspring* → *Twitter* and plot in Fig. 4 the t-SNE visualization following [24]. In these figures, red and blue dots represent positive (bullying) and negative (normal) samples of the source domain, purple and green dots denote positive and negative samples of the target domain. An effective transferable feature mapping should cluster together the red (blue) and purple (green) dots, meanwhile classification can be effectively conducted between purple and green points. As we can observe from the results of t-SNE embeddings, RIRL-L and RIRL-R can discriminate better between the positive (purple) and negative (green) classes in the target domain meanwhile the categories between source and target domains are aligned much better (red → purple, blue → green) compared to baseline models. These observations further validate the advantages of the doubly balancing strategy in learning transferable and robust representations.

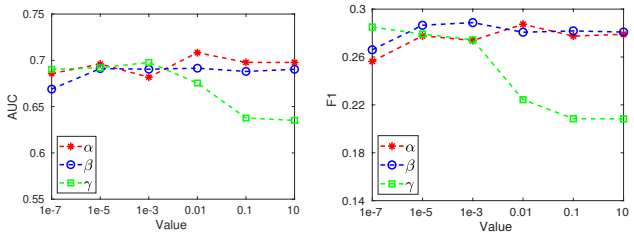
5.5 Parameter Analysis The RIRL framework has mainly three parameters, α, β, and γ that control the trade-off between the prediction accuracy and imbalance error. To answer the last research question, we randomly select the task *Twitter* → *Formspring* and run a set of experiments regarding different values of



(a) t-SNE of WDGRL features. (b) t-SNE of DANN features. (c) t-SNE of CORAL features. (d) t-SNE of CTC features.

(e) t-SNE of RIL features. (f) t-SNE of RIRL-C features. (g) t-SNE of RIRL-L features. (h) t-SNE of RIRL-R features.

Figure 4: Feature visualization of the *Twitter* → *Formspring* task. Best viewed in colors.



(a) Effect on AUC score (b) Effect on F1 score

Figure 5: Parameter analyses for α, β, γ. The x-axis denotes different parameter values. Best viewed in colors.

each parameter. In particular, α, β, γ are set to {1e-7, 1e-5, 1e-3, 0.01, 1, 10} and we vary one parameter at a time and fix the other two. Results presented in Fig. 5 indicate that RIRL is robust to α and β. Specifically, performance of RIRL presents an increasing trend as β becomes larger, i.e., when the model pays more attention to the CB term. RIRL is robust to large β because it is also balanced by the RB term. For model complexity, the best performance is achieved when α is set around 0.01 and as it gets larger, classification performance of RIRL on the target domain tends to degrade due to the over-emphasis on the sparsity of the model parameters. RIRL is robust to γ in a certain range, an extremely large value of γ can considerably aggravate the prediction performance.

6 Conclusions & Future Work

In this paper, we study a novel problem of cross-domain classification where data and domains are imbalanced. We propose an effective doubly balancing strategy (CB and RB) that accounts for both covariate and representation biases induced by the imbalanced cross-domain classification task. RIRL learns high-quality domain invariant representations because the feature alignment is conducted in both input data space and latent representation space. Extensive experimental results on two text classification tasks show that RIRL can learn invariant representations that are robust to both domain and data imbalance.

Our work opens several future directions. First, given that imbalanced data are ubiquitous in real-world applications, it is expected to apply our model to other machine learning tasks such as spam detection and image classification. Second, RIRL is currently designed for binary classification tasks, additional constraints and potential issues of multi-class cross-domain classification tasks need to be further investigated. Lastly, it is also interesting to explore other CB methods such as entropy balancing [14] and Inverse Probability Weighting [22], and exam the influence of different CB and RB methods for addressing data and domain imbalance.

Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) Grants #1610282, #1633381, and #1909555.

References

[1] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NIPS*, pages 343–351. NIPS, 2016.

[4] Minmin Chen, Yixin Chen, and Kilian Q Weinberger. Feature selection as causal inference:experiments with text classification. In *CoNLL 2017*, pages 163–172, 2017.

[5] Lu Cheng, Ruocheng Guo, and Huan Liu. Robust cyberbullying detection with causal interpretation. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 169–175. ACM, 2019.

[6] Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. Hierarchical attention networks for cyberbullying detection on the instagram social network. In *SDM*. SIAM, 2019.

[7] Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. Pi-bully: Personalized cyberbullying detection with peer influence. In *IJCAI*. IJCAI, 2019.

[8] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *ECML PKDD*, pages 274–289. Springer, 2014.

[9] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016.

[12] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *NIPS*, pages 513–520, 2007.

[13] Isabelle Guyon, Constantin Aliferis, et al. Causal feature selection. In *Computational methods of feature selection*, pages 75–97. Chapman and Hall/CRC, 2007.

[14] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[15] Melih Kandemir. Asymmetric transfer learning with deep gaussian processes. In *ICML*, pages 730–738. ICML, 2015.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD*, pages 265–274. ACM, 2017.

[18] Ying Liu, Han Tong Loh, and Aixin Sun. Imbalanced text classification: A term weighting approach. *Expert systems with Applications*, 36(1):690–701, 2009.

[19] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[20] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *ICDMW*, pages 540–547. IEEE, 2011.

[21] Kelly Reynolds, April Kontostathis, and Lynne Edwards. Using machine learning to detect cyberbullying. In *Machine learning and applications and workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE, 2011.

[22] Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[23] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *arXiv preprint arXiv:1606.03976*, 2016.

[24] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *AAAI*, 2018.

[25] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[26] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.

[27] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*, pages 153–171. Springer, 2017.

[28] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450. Springer, 2016.

[29] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.