Cross-domain Text Classification with Multiple Domains and Disparate Label Sets

Himanshu S. Bhatt, Manjira Sinha and Shourya Roy
Xerox Research Centre India, Bangalore, INDIA
{Firstname.Lastname}@Xerox.Com

Abstract

Advances in transfer learning have let go the limitations of traditional supervised machine learning algorithms for being dependent on annotated training data for training new models for every new domain. However, several applications encounter scenarios where models need to transfer/adapt across domains when the label sets vary both in terms of count of labels as well as their connotations. This paper presents first-of-its-kind transfer learning algorithm for cross-domain classification with multiple source domains and disparate label sets. It starts with identifying transferable knowledge from across multiple domains that can be useful for learning the target domain task. This knowledge in the form of selective labeled instances from different domains is congregated to form an auxiliary training set which is used for learning the target domain task. Experimental results validate the efficacy of the proposed algorithm against strong baselines on a real world social media and the 20 Newsgroups datasets.

1 Introduction

A fundamental assumption in supervised statistical learning is that training and test data are independently and identically distributed (i.i.d.) samples drawn from a distribution. Otherwise, good performance on test data cannot be guaranteed even if the training error is low. On the other hand, transfer learning techniques allow domains, tasks, and distributions used in training and testing to be different, but related. It works in contrast to traditional supervised techniques on the principle of transferring learned knowledge across domains. Pan and Yang, in their survey paper (2010), described different transfer learning settings depending on if domains and tasks vary as well as labeled data is available in one/more/none of the domains. In this paper, we propose a generic solution for multi-source transfer learning where domains and tasks are different and no labeled data is available in the target domain. This is a relatively less chartered territory and arguably a more generic setting of transfer learning.

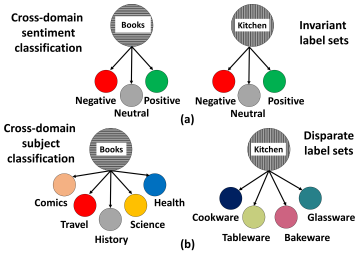

Figure 1: Cross-domain (a) sentiment classification and (b) subject classification. Illustrates (a) invariant and (b) disparate label sets.

Motivating example: Consider a social media consulting company helping brands to monitor their social media channels. Two problems typically of interest are: (i) sentiment classification (*is a post positive/negative/neutral?*) and (ii) subject classification (*what was the subject of a post?*). While sentiment classification attempts to classify a post based on its polarity, subject classification is towards identifying the subject (or topic) of the post, as illustrated in Figure 1. The company has been using standard classification techniques from an off-the-shelf machine learning toolbox. While machine learning toolkit helps them to create and apply statistical models efficiently, the same model can not be applied on a new collection due to variations in data distributions across collections[1]. It requires a few hundreds of manually labeled posts for every task on every collec-

[1]A collection comprises comments/posts pertaining to a particular client/product/services. Domain and collection are used interchangeably.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1641–1650,
Berlin, Germany, August 7-12, 2016.

tion. As social media are extremely high velocity and low retention channels, human labeling efforts act like that proverbial narrow bottleneck. Need of the hour was to reduce, if not eliminate, the human-intensive labeling stage while continue to use machine learning models for new collections.

Several transfer learning techniques exist in the literature which can reduce labeling efforts required for performing tasks in new collections. Tasks such as sentiment classification, named entity recognition (NER), part of speech (POS) tagging that have invariant label sets across domains, have shown to be greatly benefited from these works. On the other hand, tasks like subject classification that have disparate label sets across domains have not been able to gain at pace with the advances in transfer learning. Towards that we formulate the problem of *Cross-domain classification with disparate label sets* as learning an accurate model for the new unlabeled target domain given labeled data from multiple source domains where all domains have (possibly) different label sets.

Our contributions: To the best of our knowledge, this is the first work to explore the problem of cross-domain text classification with multiple source domains and disparate label sets. The other contributions of this work includes a simple yet efficient algorithm which starts with identifying transferable knowledge from across multiple source domains useful for learning the target domain task. Specifically, it identifies relevant class-labels from the source domains such that the instances in those classes can induce class-separability in the target domain. This transferable knowledge is accumulated as an auxiliary training set for an algorithm to learn the target domain classification task followed by suitable transformation of the auxiliary training instances.

Organization of the paper is as follows: Section 2 presents the preliminaries and notation, Section 3 summarizes the related work. Section 4 and 5 present the proposed algorithm and experimental results respectively. Section 6 concludes the paper.

2 Preliminaries and Notations

A domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ is characterized by two components: a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{x_1, x_2, ...x_n\} \in \mathcal{X}$. A task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$ also consists of two components: a label space $\mathcal{Y}$ and an objective predictive function $f(\cdot)$.

In our settings for cross-domain classification with disparate label sets, we assume M source domains, denoted as $\mathcal{D}_{S_i}$, where $i = \{1, 2, ..M\}$. Each source domain has different marginal distribution i.e. $P(X_{S_i}) \neq P(X_{S_j})$ and different label space i.e. $\mathcal{Y}_{S_i} \neq \mathcal{Y}_{S_j}, \forall i, j \in M$. The label space across domains vary both in terms of count of class-labels as well as their connotations; however, a finite set of labeled instances are available from each source domain. The target domain ($\mathcal{D}_T$) consists of a finite set of unlabeled instances, denoted as t_i where $i = \{1, .., N\}$. Let $\mathcal{Y}_T$ be the target domain label space with K class-labels. We assume that the number of classes in the target domain i.e. K is known (analogous to clustering where the number of clusters is given).

3 Related Work

Table 1 summarizes different settings of transfer learning (Pan and Yang, 2010) and how this work differentiates from the existing literature[2]. The first scenario represents the ideal settings of traditional machine learning (Mitchell, 1997) where a model is trained on a fraction of labeled data and performs well for the same task on the future unseen instances from the same domain.

The second scenario where the domains vary while the tasks remain the same is referred to as transductive transfer learning. This is the most extensively studied settings in the transfer learning literature and can be broadly categorized as single and multi-source adaptation. Single source adaptation (Chen et al., 2009; Ando and Zhang, 2005; Daumé III, 2009) primarily aims at minimizing the divergence between the source and target domains either at instance or feature levels. The general idea being identifying a suitable low dimensional space where transformed source and target domains data follow similar distributions and hence, a standard supervised learning algorithm can be trained (Daumé III, 2009; Jiang and Zhai, 2007; Pan et al., 2011; Blitzer et al., 2007; Pan et al., 2010; Dai et al., 2007; Bhatt et al., 2015).

While several existing single source adaptation techniques can be extended to multi-source adaptation, the literature in multi-source adaptation can be broadly categorized as: 1) feature representation approaches (Chattopadhyay et al., 2012; Sun et al., 2011; Duan et al., 2009; Duan et al., 2012;

[2]This is not the complete view of the transfer learning literature; however, covers relevant work that helps motivate/differentiate the novel features of this paper.

Table 1: Summarizing the related work and differentiating the novel features of the proposed algorithm.

Scenario	Settings	Nature of Data	Learning Paradigm	Main Concepts	Our Differentiation
$\mathcal{D}_S = \mathcal{D}_T$, $\mathcal{T}_S = \mathcal{T}_T$	Traditional Machine learning	Labelled data in source domain(s) and unlabeled data in target domain	Source and target domains are exactly the same	Learn models on training set and test on future unseen data	Allows tasks across domains to be different;a more general setting
$\mathcal{D}_S \neq \mathcal{D}_T$, $\mathcal{T}_S = \mathcal{T}_T$	Transductive Transfer Learning	Labelled data in source domain(s) and unlabeled data from the target domain $\mathcal{P}(\mathcal{X}_S) \neq \mathcal{P}(\mathcal{X}_T)$	Single source domain adaptation	Learning common shared representation; instance weighing, parameter transfer	Exploits multiple sources each with disparate label sets.
			Multi-source adaptation	Classifier combination; efficient combination of information from multiple sources; Feature representation	Intelligent selection of transferable knowledge from multiple sources for adaptation.
No conditions on $\mathcal{D}_S$ & $\mathcal{D}_T$, but, $\mathcal{T}_S \neq \mathcal{T}_T$	Inductive Transfer Learning	Unlabeled data in source domain(s) and labeled data in target domain	Self-taught learning	Extracts higher level representations from unlabeled auxiliary data to learn instance-to-label mapping with labeled target instances	Learns instance-to-label mapping in the unlabeled target domain using multiple labeled source domains having different data distributions and label spaces.
		Labeled data is available in all domains	Multi-task learning	Simultaneously learns multiple tasks within (or across) domain(s) by exploiting the common feature subspace shared across the tasks	Learns the optimal class distribution in an unlabeled target domain by minimizing the differences with multiple labeled source domains.
$\mathcal{D}_S \neq \mathcal{D}_T$, $\mathcal{T}_S \neq \mathcal{T}_T$	Kim *et al.* (2015)	Labeled data in source and target domains	Transfer learning with disparate label set	Disparate fine grained label sets across domains, however, same coarse grained labels set can be invoked across domains	No coarse-to-fine label mapping due to heterogeneity of label sets, Assumes no labelled data in target domain.

Bollegala et al., 2013; Crammer et al., 2008; Mansour et al., 2009; Ben-David et al., 2010; Bhatt et al., 2016) and 2) combining pre-trained classifiers (Schweikert and Widmer, 2008; Sun and Shi, 2013; Yang et al., 2007; Xu and Sun, 2012; Sun et al., 2013). Our work differentiates in intelligently exploiting selective transferable knowledge from multiple sources unlike existing approaches where multiple sources contribute in a brute-force manner.

The third scenario where the tasks differ irrespective of the relationship among domains is referred to as inductive transfer learning. Self-taught learning (Raina et al., 2007) and multi-task (Jiang, 2009; Maurer et al., 2012; Xu et al., 2015; Kumar and Daume III, 2012) learning are the two main learning paradigms in this scenario and Table 1 differentiates our work from these.

This work closely relates to the fourth scenario where we allow domains to vary in the marginal probability distributions and the tasks to vary due to different label spaces[3]. The closest prior work by Kim *et al.* (2015) address a sequential labeling problem in NLU where the fine grained label sets across domains differ. However, they assume that there exists a bijective mapping between the coarse and fine-grained label sets across domains. They learn this mapping using labeled instances from the target domain to reduce the problem to a standard domain adaptation problem (Scenario 2).

[3]This work do not consider scenario when domains vary in feature spaces and tasks vary in the objective predictive functions.

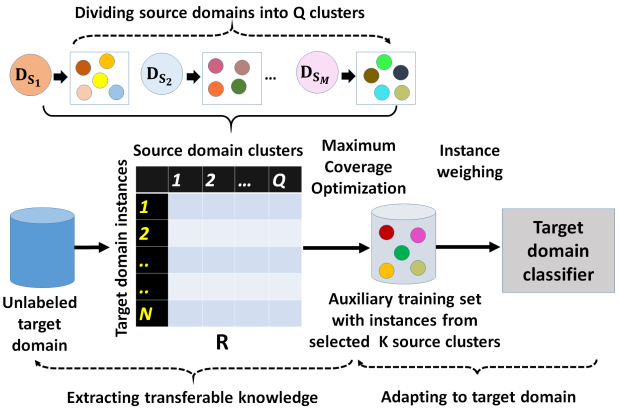

Figure 2: Illustrates different stages of the proposed algorithm.

However, this paper caters to multiple source domains with disparate label sets without assuming availability of any labeled data from the target domain or fine-to-coarse label mappings across domains.

4 Cross-domain Classification with Disparate Label Set

The underlying philosophy of the proposed algorithm is to learn the target domain task by using the available information from multiple source domains. To accomplish this, we have developed an algorithm to identify and extract partial transferable knowledge from multiple sources. This knowledge is then suitably transformed to induce classes in the target domain using the class separation from the source domains. Different stages of the proposed algorithm, as shown in Figure 2, are

elaborated in the next sections.

4.1 Exploiting Multiple Domains

If we had the mappings between the source and target domain label sets, we could have leveraged existing transfer learning approaches. However, heterogeneity of label sets across domains and the unlabeled data from the target domain exacerbate the problem. Our objective is to leverage the knowledge from multiple source domains to induce class-separability in the target domain. Inducing class-separability refers to segregating the target domain into K classes using labeled instances from selective K source domain classes.

Towards this, the proposed algorithm divides each source domain into clusters/groups based on the class-labels such that instances with the same label are grouped in one cluster. All source domains are divided into Q clusters where $Q = \sum_{m=1}^{M} ||Y_m||$ represents the count of class-labels across all sources. $||Y_m||$ being the count of class-labels in the m^{th} source domain. C_q denotes the q^{th} cluster and μ_q denotes its centroid computed as the average of all the members in the cluster. We assert that the target domain instances that have high similarity to a particular source domain cluster can be grouped together. Given N target domain instances and Q source domain clusters, a matrix R (dimension $N \times Q$) is computed based on the similarity of the target domain instances with the source clusters. The i^{th} row of the matrix captures the similarity of the i^{th} target domain instance (t_i) with all the source domain clusters. It captures how different source domain class-labels are associated with the target domain instances and hence, can induce class-separability in the target domain.

4.2 Extracting Transferable Knowledge

The similarity matrix R associates target domain instances to the source domain clusters in proportion to their similarity. However, the objective is to select the optimal K source domain clusters that fit the maximum number of target domain instances. This problem is similar to the well-known combinatorial optimization problem of Maximum Coverage (Vazirani, 2003) where given a collection of P sets, we need to select A sets ($A < P$) such that the size of the union of the selected sets is maximized. In this paper, we are given Q source domain clusters and need to select K clusters such that the corresponding number of associated target domain instances is maximized. As the Maximum Coverage problem is **NP**-hard, we implement a greedy algorithm for selecting the k source domain clusters, as illustrated in Algorithm 1.

Algorithm 1 Selecting K Source Clusters

Input: A matrix R, K = number target domain classes, l= number of selected cluster.
Initialize: $l = 0$, Normalize R such that each row sums up to 1.
repeat:
1: Pick the column in R which has maximum sum of similarity scores for uncovered target domain instances.
2: Mark elements in the chosen column as covered.
3: $l = l + 1$
until: $l = K$
Output: K source domain clusters.

A source domain contributes partially in terms of zero or more class-labels (clusters) identified using the Algorithm 1. Therefore, we refer to the labeled instances from the selected clusters of a source domain as the partial transferable knowledge from that domain. This partial transferable knowledge from across multiple source domains is congregated to form an auxiliary training set, referred to as (AUX).

4.3 Adapting to the Target Domain

The auxiliary training set comprises labeled instances from selected K source domain clusters[4]. Since, the auxiliary set is pulled out from multiple source domains, it follows different data distribution as compared to the target domain. For a classifier, trained on the K-class auxiliary training set, the distributional variations have to be normalized so that it can generalize well on the target domain.

In this research, we proposed to use an instance weighting technique (Jiang and Zhai, 2007) to minimize the distributional variations by deferentially weighting instances in the auxiliary set. Intuitively, the auxiliary training instances similar to the target domain are assigned higher weights while training the classifier and vice versa. The weight for the i^{th} instance in the auxiliary set should be proportional to the ratio $\frac{(P_t(x_i))}{(P_a(x_i))}$. However, since the actual probability distributions

[4]The K classes in auxiliary set induce class-separability in the target domain, however, the actual class-labels across these two may not have any sort of coarse-to-fine mapping.

Algorithm 2 Cross-domain Classification with Disparate Label Sets

Input: M source domains, target domain instances (t_i), $i = (1, ..., N)$, K = number of target domain classes.

Process: Divide M sources into Q clusters *s.t.* $Q = \sum_{q=1}^{M} |Y_m|$. C_q be the q^{th} cluster & μ_q be its centroid computed as shown in Eq 1.

A: Exploiting Multiple Sources:
for $i = 0$: till N **do**
  for $q = 0$: till Q **do**
    $R[i, q] = Sim(\mu_q, t_i)$
  end for
end for

B: Extracting partial knowledge:
1: Pick K columns from R using Algorithm 1.
2: Construct AUX by congregating instances from the selected K source domain class-labels.

C: Adapting to target domain:
1: Minimize distributional variations using instance weighing technique.
2: Train a K-class classifier using AUX.

Output: K-class target domain classifier.

($P_a(x)$ and $P_t(x)$ for the auxiliary set and target domain respectively) are unknown, the instance difference is approximated as $\frac{(P_t(x_i|d=target))}{(P_a(x_i|d=auxilliary))}$, where d is a random variable used to represent whether x_i came from the auxiliary set or the target domain. To calculate this ratio, a binary classifier is trained using the auxiliary set and target domain data with labels {-1} and {+1} respectively. The predicted probabilities from the classifier are used to estimate the ratio as the weight for the i^{th} auxiliary instance x_i. Finally, a K-class classifier is trained on the weighted auxiliary training set to perform classification on the target domain data.

4.4 Algorithm

As shown in Figure 2, the step-by-step flow of the proposed algorithm is summarized below:

1. Divide M source domains into Q clusters, each represented as C_q, $q = \{1, 2, .., Q\}$.

2. Compute centroid of each cluster as the average of the cluster members, as shown in Eq. 1.

$$\mu_q = \frac{1}{||C_q||} \sum_{(i=1;\mathbf{x}_i \in C_q)}^{||C_q||} \mathbf{x}_i \qquad (1)$$

where μ_q is the centroid, $||C_q||$ is the membership count and $\mathbf{x}_i$ is the i^{th} member of C_q.

3. For target instances t_i $\forall i \in N$, compute cosine similarity with all the source domain cluster centroids to form the matrix R (dimensions: $N \times Q$), as shown in Eq. 2

$$R[i, q] = Sim(\mu_q, \mathbf{t}_i) = \frac{\mu_q \cdot \mathbf{t}_i}{||\mu_q||\ ||\mathbf{t}_i||} \qquad (2)$$

4. Run Algorithm 1 on R to select K optimal source clusters (i.e. columns of R).

5. Congregate labeled instances from the selected source domain clusters to form the K-class auxiliary training set.

6. Minimize the divergence between the auxiliary set and target domain using the instance weighing technique, described in Section 4.3.

7. Finally, train a K-class classifier on deferentially weighted auxiliary training instances to perform classification in the target domain.

The K-class classifier trained on the auxiliary training set is an SVM classifier (Chih-Wei Hsu and Lin, 2003) with $L2 - loss$ from the LIBLINEAR library (Fan et al., 2008). The classifier used in the instance weighing technique is again an SVM classifier with RBF kernel. The proposed algorithm uses distributional embedding i.e. Doc2Vec (Le and Mikolov, 2014) to represent instances from the multiple source and target domains. We used an open-source implementation of Doc2Vec (Le and Mikolov, 2014) for learning 400 dimensional vector representation using DBoW.

5 Experimental Evaluation

Comprehensive experiments are performed to evaluate the efficacy of the proposed algorithm for cross-domain classification with disparate label sets across domains on two datasets.

5.1 Datasets

The first dataset is a real-world Online Social Media (OSM) dataset which consists of 74 collections. Each collection comprises comments/tweets that are collected based on user-defined keywords. These keywords are fed to a listening engine which crawls the social media (i.e. Twitter.com) and fetches comments matching the keywords. The task is to classify the comments in a collection

Table 2: Illustrates variability in label sets across some collections from the OSM dataset.

Apple iPhone 6	Apple iOS 8	Apple iPad mini3
Camera	Locking apps & features	Release date & Features
Design	Extensibility features	Apple play & NFC
Review link	General features related marketing	Apple sim card
Apple Play/NFC	Camera features	Touch ID
Comparison to Android	Password with touch integration	iPad mini3 - disappoints
Price	Health & fitness app	
Apple watch	Location & Maps	
	Firmware updates	

Table 3: Table illustrates the collections from the EMPATH database used in this research.

Collection ID	Domain	#Categories
Coll 1	Huwaei	5
Coll 2	Healthcare	9
Coll 3	Whattsapp	8
Coll 4	Apple iOS 8	8
Coll 5	Apple iPhone 6	7

into user-defined categories. These user-defined categories may vary across collections in terms of count as well as their connotations. Table 2 shows an example of the user-defined categories for a few collections related to "Apple" products. In the experiments, one collection is used as unlabeled target collection and the remaining collections are used as the labeled source collections. We randomly selected 5 target collections to report the performance, as described in Table 3.

The second dataset is the 20 Newsgroups (NG) (Lang, 1995) dataset which comprises $20,000$ news articles organized into 6 groups with different sub-groups both in terms of count as well as connotations, as shown in Figure 3(a). Two different experiments are performed on this dataset. In the first experiment ("**Exp-1**"), one group is considered as the target domain and the remaining 5 groups as the source domains. In the second experiment ("**Exp-2**"), one sub-group from each of the first five groups[5] is randomly selected to synthesize a target domain while all the groups (with the remaining sub-groups) are used as source domains. Figure 3(b) shows an example on how to synthesize target domains in "Exp-2". There are 720 possible target domains in this experiment and we report the average performance across all possible target domains, referred to as "Grp 7". The task in both the experiments is to categorize the target domain into its K categories (sub-groups) using labeled data from multiple source domains.

[5]Group-6 has only 1 sub-group, therefore, it is considered for synthesizing target domain in the experiments.

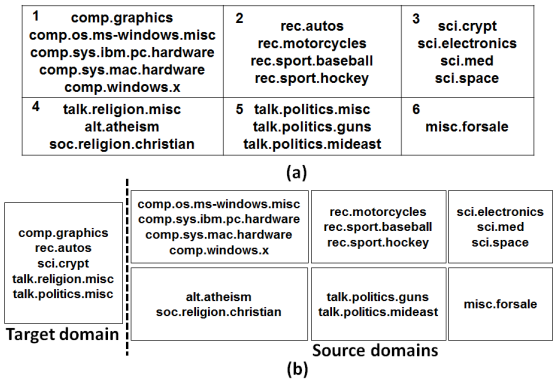

Figure 3: Illustrates (a) different groups (b) target domain synthesis ("EXP 2") on the NG dataset.

5.2 Evaluation Metric

The performance is reported in terms of classification accuracy on the target domain. There is no definite mapping between the actual class-labels in the target domain and the K categories (i.e. induced categories) in the auxiliary training set. Therefore, we sequentially evaluate all possible one-to-one mappings between the K categories in the auxiliary training set and target domain to report results for the best performing mapping.

5.3 Experimental Protocol

The performance of the proposed algorithm is skylined by the in-domain performance (**Gold**), i.e. a classifier trained and tested on the labeled target domain data. We also compared the performance with spherical K-means clustering (Dhillon and Modha, 2001) used to group the target domain data into K categories against the ground truth, referred to **CL**. Spherical K-means clustering is based on cosine similarity and performs better for high-dimensional sparse data such as text.

To compare with a baseline and an existing adaptation algorithm, we selected the most similar source domain[6] with exactly K number of class-labels and report the performance on the best possible mapping, as described in Section 5.2. To compute the baseline (**BL**), a classifier trained on the source domain is used to categorize the target domain. A widely used domain adaptation algorithm, namely structural correspondence learning (**SCL**) (Blitzer et al., 2007) is also applied using the selected source domain.

[6]The most similar source domain is selected using proxy-$\mathcal{A}$ distance (Blitzer et al., 2007) which has good correlation with domain adaptation performance.

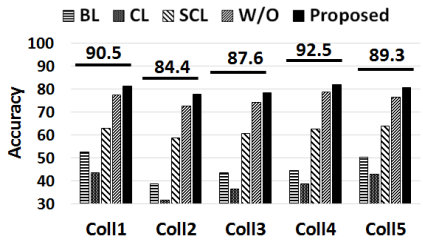

Figure 4: Compares the performance of different techniques on the OSM dataset.

Table 4: Summarizes the performance of the proposed algorithm on the OSM dataset.

Coll ID (#)	BL	CL	SCL	W/O	Proposed	Gold
Coll 1 (5)	52.6	43.7	62.8	77.4	81.4	90.5
Coll 2 (9)	38.6	31.8	58.8	72.5	77.6	84.4
Coll 3 (8)	43.6	36.4	60.7	74.2	78.5	87.6
Coll 4 (8)	44.7	38.8	62.5	78.8	82.1	92.5
Coll 5 (7)	50.5	42.8	64.4	76.6	80.5	89.3

5.4 Results and Analysis

Key observations and analysis from the experimental evaluations are summarized below:

5.4.1 Results on the OSM Dataset

Results in Figure 4 and Table 4 show the efficacy of the proposed algorithm for cross-domain classification with disparate label sets as it outperforms other approaches by at least 15%. Coll ID(#) refers to the target collection and the corresponding count of class-labels. Results in Table 4 also compare the performance of the proposed technique without the distributional normalization of the auxiliary training set, referred to as "W/O". Results suggest that suitably weighing instances from the auxiliary training set mitigates the distributional variations and enhances the cross-domain performance by at least 3.3%.

5.4.2 Results on the 20Newsgroups Dataset

Results in Table 5 show that the proposed algorithm outperforms other techniques for both the experiments by at least 15 % and 18% respectively on the 20 Newsgroups dataset. In Table 5, "-" refers to the cases where a single source domain with the same number of class-labels as in the target domain is not available. In "Exp-1" where the source and target categories vary in terms of counts as well as their connotations, the proposed algorithm efficiently induces the classes in the unlabeled target domain using the partial transferable knowledge from multiple sources. For "Exp-2", it is observed that the performance of the proposed algorithm is better than the performance in "Exp-1" as the target categories have closely related categories (from the same group) in the source domains. Table 5 reports the average performance across all the 720 possible combinations of target domains with a standard deviation of 2.6.

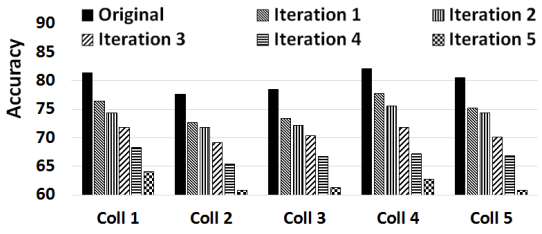

Figure 5: Effects of selected source collections on the OSM dataset.

Table 5: Summarizes the performance of the proposed algorithm on the 20Newsgroups dataset.

Target(#)	BL	CL	SCL	W/O	Proposed	Gold
Grp 1 (5)	-	48.6	-	79.4	80.8	85.6
Grp 2 (4)	62.7	50.2	62.7	78.3	83.6	89.2
Grp 3 (4)	64.3	54.8	64.4	81.6	85.3	90.4
Grp 4 (3)	69.6	55.6	67.3	82.2	86.4	92.5
Grp 5 (3)	69.7	56.4	70.3	83.6	85.3	91.2
Grp 7 (5)	-	52.8	-	84.6	88.4	93.8

5.4.3 Effect of Multiple Source Domains

Table 6 validates our assertion that multiple sources are necessary to induce class-separability in the target domain as a single source is not sufficient to cater to the heterogeneity of class-labels across domains. It also suggests that the proposed algorithm can learn class-separability in the target domain by using arbitrary diverse class-labels from different sources and does not necessarily require class-labels to follow any sort of coarse-to-fine mapping across domains.

To evaluate the effects of using multiple sources, further experiments were performed by varying the number of available source domains. For the OSM dataset, we varied the number of available source collections from 1 to 73 starting with the most similar source collection and repeatedly adding the next most similar collection in the pool of available collections. We observe that even the most similar collection was not independently sufficient to induce classes in the target collection and it was favorable to exploit multiple collections. Moreover, adding collections based on similarity to the target collection had a better likelihood of achieving higher performance as compared to adding random collections.

In another experiment, we first identified the source collections which contributed to learning the target task. We removed these collections and applied the proposed algorithm on the remaining source collections. Figure 5 shows the perfor-

Table 6: Actual target domain class-labels and the corresponding source domain class clusters used to build the auxiliary training set.

Target Collection: Apple iOS 8	Associated Class-labels from multiple source collections
Locking apps & security	Anti Theft Features (**Coll ID:** 776 **on Apple iOS 6 plus**)
Extensibility features	Application update (**Coll ID:**720 **on Apple iOS Features**)
General features related marketing	General press (**Coll ID:** 163 **on XBOX Issues**)
Camera features	Camera (**Coll ID:** 775 **on Apple iPhone 6**)
Password with touch integration	Touch ID (**Coll ID:** 803 **on Apple iPad mini3**)
Health & fitness app	Reproductive health issues (**Coll ID:** 289 **on Healthcare**)
Location & Maps	Events (**Coll ID:** 502 **on L'Oreal**)
Firmware updates	Updates & patches (**Coll ID:** 478 **on Riot Game Support v2**)

mance of the proposed algorithm on 5 such iterations of removing the contributing source collections from the previous iteration. We observed a significant drop in the performance with each iteration which signifies the effectiveness of the proposed algorithm in extracting highly discriminating transferable knowledge from multiple sources.

5.4.4 Comparing with Domain Adaptation

We applied domain adaptation techniques considering the auxiliary training set to be a single source domain with the same number of classes as that in the target domain. We applied two of the widely used domain adaptation techniques, namely SCL (Blitzer et al., 2007) and SFA (Pan et al., 2010) referred to as "AuxSCL" and "AuxSFA" respectively. Results in Table 7 suggest that the proposed algorithm significantly outperforms "AuxSCL" and "AuxSFA" on the two datasets. Generally, existing domain adaptation techniques are built on the co-occurrences of the common features with the domain specific features and hence, capture how domain specific features in one domain behaves w.r.t to the domain specific features in the other domain. They assume homogeneous labels and expect the aligned features across domains to behave similarly for the prediction task. However, these features are misaligned when the label set across domains vary in terms of their connotations.

5.4.5 Effect of Different Representations

The proposed algorithm uses Doc2Vec (Le and Mikolov, 2014) for representing instances from multiple domains. However, the proposed algorithm can build on different representations and hence, we compare its performance with traditional TF-IDF representation (including unigrams and bigrams) and a dense representation using TF-IDF+PCA (reduced to a dimension such that it covers 90% of the variance). We observe that Doc2Vec representation clearly outperforms the other two representations as it addresses the drawbacks of bag-of n-gram models in terms of implicitly inheriting the semantics of the words in a document and offering a more generalizable concise vector representation.

Table 7: Comparing the proposed algorithm with existing domain adaptation algorithms.

Dataset	Target	SCL	SFA	Proposed
OSM	Coll 1	66.2	64.7	81.4
	Coll 2	63.8	62.6	77.6
	Coll 3	64,1	63.4	78.5
	Coll 4	64.2	65.2	82.1
	Coll 5	64.0	63.7	80.5
NG Exp-1	Grp 1	65.2	64.2	80.8
	Grp 2	68.2	65.3	83.6
	Grp 3	69.4	68.4	85.3
	Grp 4	70.3	69.2	86.4
	Grp 5	69.0	68.8	85.3
NG Exp-2	Grp 7	72.6	70.2	88.4

Table 8: Comparing different representations.

Dataset	Target	TF-IDF	TF-IDF +PCA	Doc2Vec
OSM	Coll 1	70.6	76.8	81.4
	Coll 2	69.5	74.2	77.6
	Coll 3	70.2	75.5	78.5
	Coll 4	71.6	77.9	82.1
	Coll 5	70.8	76.8	80.5
NG Exp-1	Grp 1	71.8	75.6	80.8
	Grp 2	73.6	77.5	83.6
	Grp 3	77.4	81.1	85.3
	Grp 4	76.6	82.5	86.4
	Grp 5	75.5	81.4	85.3
NG Exp-2	Grp 7	76.2	83.6	88.4

6 Conclusions

This paper presented the first study on cross-domain text classification in presence of multiple domains with disparate label sets and proposed a novel algorithm for the same. It proposed to extract partial transferable knowledge from across multiple source domains which was beneficial for inducing class-separability in the target domain. The transferable knowledge was assimilated in terms of selective labeled instances from different source domain to form a K-class auxiliary training set. Finally, a classifier was trained using this auxiliary training set, following a distribution normalizing instance weighing technique, to perform the classification task in the target domain. The efficacy of the proposed algorithm for cross-domain classification across disparate label sets will expand the horizon for ML-based algorithms to be more widely applicable in more general and practically observed scenarios.

References

Rie Kubota Ando and Tong Zhang. 2005. A high-performance semi-supervised learning method for text chunking. In *Proceedings of Association for Computational Linguistics*, pages 1–9.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175.

Himanshu Sharad Bhatt, Deepali Semwal, and Shourya Roy. 2015. An iterative similarity based adaptation technique for cross-domain text classification. In *Proceedings of Conference on Natural Language Learning*, pages 52–61.

Himanshu Sharad Bhatt, Arun Rajkumar, and Shourya Roy. 2016. Multi-source iterative adaptation for cross-domain classification. In *In Proceedings of International Joint Conference on Artificial Intelligence*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of Association for Computational Linguistics*, pages 440–447.

Danushka Bollegala, David Weir, and John Carroll. 2013. Cross-domain sentiment classification using a sentiment sensitive thesaurus. *IEEE Transactions on Knowledge and Data Engineering*, 25(8):1719–1731.

Rita Chattopadhyay, Qian Sun, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. 2012. Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data*, 6(4):1–26.

Bo Chen, Wai Lam, Ivor Tsang, and Tak-Lam Wong. 2009. Extracting discriminative concepts for domain adaptation in text mining. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 179–188.

Chih-Chung Chang Chih-Wei Hsu and Chih-Jen Lin. 2003. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.

Koby Crammer, Michael Kearns, and Jennifer Wortman. 2008. Learning from multiple sources. *Journal of Machine Learning Research*, 9(1):1757–1774.

Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2007. Co-clustering based classification for out-of-domain documents. In *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pages 210–219.

Hal Daumé III. 2009. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*.

Inderjit S. Dhillon and Dharmendra S. Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143–175.

Lixin Duan, Ivor W. Tsang, Dong Xu, and Tat-Seng Chua. 2009. Domain adaptation from multiple sources via auxiliary classifiers. In *Proceedings International Conference on Machine Learning*, pages 289–296.

Lixin Duan, Dong Xu, and Ivor Wai-Hung Tsang. 2012. Domain adaptation from multiple sources: a domaindependent regularization approach. *IEEE Transactions on Neural Networks and Learning Systems*, 23(3):504518.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

J. Jiang and C. Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of Association for Computational Linguistics*, volume 7, pages 264–271.

Jing Jiang. 2009. Multi-task transfer learning for weakly-supervised relation extraction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1012–1020. Association for Computational Linguistics.

Young-Bum Kim, Karl Stratos, Ruhi Sarikaya, and Minwoo Jeong. 2015. New transfer learning techniques for disparate label sets. In *Proceedings of Association for Computational Linguistics*.

Abhishek Kumar and Hal Daume III. 2012. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*.

Ken Lang. 1995. NewsWeeder: Learning to filter netnews. In *Proceedings of International Conference on Machine Learning*.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2009. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, pages 1041–1048.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. 2012. Sparse coding for multitask and transfer learning. *arXiv preprint arXiv:1209.0738*.

Thomas M. Mitchell. 1997. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of International Conference on World Wide Web*, pages 751–760.

Sinno Jialin Pan, , Ivor W. Tsang, James T. Kwok, and Qiang Yang. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. 2007. Self-taught learning: Transfer learning from unlabeled data. In *International Conference on Machine Learning*, pages 759–766.

Gabriele Schweikert and Christian Widmer. 2008. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *Advances in Neural Information Processing Systems*, pages 1433–1440.

Shi-Liang Sun and Hong-Lei Shi. 2013. Bayesian multi-source domain adaptations. In *International Conference on Machine Learning and Cybernetics*, pages 24–28.

Qian Sun, Rita Chattopadhyay, Sethuraman Panchanathan, and Jieping Ye. 2011. A two-stage weighting framework for multi-source domain adaptation. In *Advances in Neural Information Processing Systems*, pages 505–513.

Shiliang Sun, Zhijie Xu, and Mo Yang. 2013. Transfer learning with part-based ensembles. In *Multiple Classifier Systems*, volume 7872, pages 271–282.

Vijay V. Vazirani. 2003. *Approximation Algorithms*. Springer-Verlag Berlin Heidelberg.

Zhijie Xu and Shiliang Sun. 2012. Multi-source transfer learning with multi-view adaboost. In *Proceedings of International Conference on Neural Information Processing*, pages 332–339.

Linli Xu, Aiqing Huang, Jianhui Chen, and Enhong Chen. 2015. Exploiting task-feature co-clusters in multi-task learning. In *Proceedings of Association for the Advancement of Artificial Intelligence*, pages 1931–1937.

Jun Yang, Rong Yan, and Alexander G. Hauptmann. 2007. Cross-domain video concept detection using adaptive svms. In *Proceedings of International Conference on Multimedia*, pages 188–197.