CIRRELT-2023-06

Sentiment Polarity of Transportation Users' Opinions: Domain Adaptation from Multiple Sources

Banafsheh Mehri
Martin Trepanier
Yves Goussard

January 2023

Bureau de Montréal
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
Tél : 1-514-343-7575
Télécopie : 1-514-343-7121

Bureau de Québec
Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
Tél : 1-418-656-2073
Télécopie : 1-418-656-2624

Sentiment Polarity of Transportation Users' Opinions: Domain Adaptation from Multiple Sources

Banafsheh Mehri[1,2,*], Martin Trépanier[1,3], Yves Goussard[2]

[1] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)

[2] Department of Electrical Engineering, Polytechnique Montréal

[3] Department of Mathematics and Industrial Enginnering, Polytechnique Montréal

Abstract. Social media platforms allow transportation agencies to engage personally with transportation users and can provide a new approach for collecting information regarding mobility and transportation services. Nevertheless, the huge amount of data on social media platforms makes it difficult to extract meaningful information to decide which messages are more important and to prioritize which messages to respond to first. Sentiment analysis on the social media data can be used to support intelligent transportation systems (ITSs) with automatically discovering people's feelings regarding transportation services through users' textual posts in social media platforms such as Twitter. Nevertheless, the difficulty in generating annotated datasets for sentiment analysis in transportation domain raises the need for developing new approaches such as domain adaptation where available annotated datasets from other domains can be employed. In this paper we show that domain adaptation is theoretically and practically useful for sentiment classification of transportation users' opinions. We particularly introduce a domain adaptation method using multi-source domains where the training data is sampled from multiple sources where source domains might be different not only from the target domain but also from each other. In this paper we propose a deep learning approach for knowledge transfer from multiple labeled source domains to an unlabeled target domain (sentiment analysis in transportation domain) by aligning their feature distributions. We report the results of a series of experiments that we conduct comparing against previous approaches to demonstrate the performance of our proposed method.

Keywords: Intelligent Transportation Systems (ITSs), Twitter, sentiment analysis, domain adaptation, multi-source domain adaptation.

Acknowledgements. The authors gratefully acknowledge the financial support provided by the Natural Sciences and Engineering Council of Canada (NSERC), and the Société de transport de Montréal (STM) for data provisioning.

---

* Corresponding author: banafsheh.mehri@polymtl.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2023

I. Introduction

THE task of categorizing texts according to their polarity or Sentiment Analysis can be useful to support intelligent transportation systems (ITSs) in helping to assess the passengers' feelings or opinions through posted messages on social media platforms. To be able to automatically perform the sentiment analysis task, we need to train models on annotated datasets. However, most of machine learning algorithms require massive amount of annotated data in order to be trained in their full capacity. In conventional machine learning models, both the training dataset and the test dataset come from the same distribution which equates to a single domain. A domain consists of feature space and a marginal probability distribution i.e., the features of the data and the distribution of those features in the dataset. As a result, the probability of a classifier trained within this distribution delivering accurate class predictions is very high. However, in many practical applications such as transportation, it is often time consuming, labor-intensive and expensive if not impossible to obtain enough annotated data to train a model for specific tasks. In domain adaptation, the goal is to train a model on one dataset (source) for which label or annotation is available and secure good performance on another dataset (target) whose label or annotation is not available. Thus, transferring the learned knowledge from a different annotated source domain to a different target domain with sparsely annotated data or without any annotated data is the goal we try to achieve with domain adaptation. Nevertheless, due to the presence of the domain shift [1]–[3] between training and test samples coming from two different distributions, the joint probability distributions are different in the two domains. Hence, the direct transfer of models across domains can cause the trained model to suffer from high generalization error; which leads to a poor model performance. Domain adaptation focuses on minimizing the impact of the domain shift when establishing knowledge transfer from a labeled source domain to an unlabeled target domain. Many researchers focused on aligning the features across the source domain and the target domain by learning a domain-invariant latent space to minimize the domain-shift in the latent space; and improves the generalization power of the model. This alignment of features across source and target domains in image processing field is mostly done by using discrepancy minimization [4]–[6], adversarial objective [7]–[9], or domain-specific transformations [10], [11]. Accordingly, there are several domain adaptation methods that have been adapted to sentiment classification [12]–[18]. Nonetheless, the majority of attempts in this field are focused on single source domain adaptation where the source samples are collected from a single domain. This neglects the more realistic and complex extended scenario where training data can be collected from multiple sources with different distributions. In particular, useful and unified applications of sentiment detection specially in the field of transportation may require different types of documents, due to lack of annotated data. In addition, since the sentiment of a piece of text depends on general properties of language and to a large extent stable across different domains, we intuitively expect to successfully adapt methods for transportation sentiment detection to multiple sources domain adaptation approaches. However, combining all sources into one domain and simply applying single source domain adaptation methods on them might give off poor performance results [19]. The reason resides in the existence of the domain shift; not only between each source-target pair but also between sources themselves. In this study, we propose a method to address domain adaptation from multiple sources for sentiment classification of transportation-related tweets that aligns the source domains with the target domain while aligning the source domains with each other simultaneously. The following are the main contributions of this work:

- We introduce the concept of Domain Adaptation to the public transportation fields, where performing direct sentiment analysis is an issue due to the lack of annotated data in this field.
- We provide theoretical insights for moment-based approaches in multiple source domain adaptation settings.
- We transfer the knowledge learned from multiple labeled source domains to an unlabeled target domain by dynamically aligning moments of their feature distributions by proposing a deep learning approach, for Multi-Source Domain Adaptation.
- We conduct series of experiments to showcase the performance of our proposed method compared to some of the state-of-the-art approaches for domain adaptation.

The rest of this paper is organized as follows: we begin by reviewing the related works in section II. Next, we describe the problem statement, theoretical analysis and our proposed approach in section III. The experiment results are discussed in section IV. Finally, we present the conclusion and discussion in section V.

II. BACKGROUND

In this section, we first quickly point out to sentiment analysis and Twitter data for transportation studies and we review the existing research on sentiment analysis in transportation systems. Then we focus on the single- and multi-source domain adaptation.

A. *Sentiment Analysis in Transportation Systems*

Sentiment analysis aims to determine the judgment of a writer based on a given piece of writing with respect to a given topic. People use social media platforms such as Twitter to share their opinions, regarding various topics including transportation systems and services. Hence, processing and extracting actionable information about transportation systems and services from Twitter have become an attractive research domain in intelligent transportation systems [20]–[24]. The abundantly accessible online reviews and recommendations made sentiment analysis an interesting research area in both academia and industry. In the last two decades, several methods have been introduced to analyze emotions and opinions from social media [25]–[30]. In particular, regarding the use of sentiment analysis of social media data in transportation, Chen et al. [31] developed a system to discover transportation safety-related topics from tweets and identify the sentiment polarity of those tweets to support safety enhancements. Their prototype system was able to retrieve the sentiment of safety-related tweets and geocode them to street locations to identify potential safety bottlenecks. Collins et al. [32] used a sentiment analysis method to analyse transit riders' satisfaction in the city of Chicago. They then used transit rider satisfaction evaluation as a metric to evaluate timeliness, safety, cleanliness, and ridership and showed the results in the form of positive, negative, and neutral opinion. Schweitzer [33] studied the capacity of social media communication analysis for transportation planning and used tweets to evaluate users' opinions about public transit. Das et al. [34] used sentiment analysis on Twitter data to evaluate the popularity of a bike sharing program of Capital Bike share in Washington DC. Their results revealed higher positive sentiments towards the current system. Luong et al. [35] provide an interactive online interface to display and monitor real-time feedback and sentiment along different lines in the area's light rail system in Los Angeles. Their system was capable of providing new insights into the way that transit users present their opinions, engage with government agencies, react to events/policies, and share information with others. Windasari et al. [36] proposed a system that detect public sentiments on Twitter post about online transportation services. They performed sentiment analysis using Support Vector Machines (SVM), and presented results into positive and negative sentiment.

These aforementioned studies provided valuable insight into the applications of sentiment analysis via social media data in public transit. However, all these studies relied on manual annotated data for training purposes which might not be reachable or it is prohibitively expensive and time-consuming to obtain the labeled dataset in the presence of huge amount of data in most real-world scenarios.

B. Single Source Domain Adaptation

Generally in machine learning methods and applications we assume that models for each domain are trained and tested using data with the same features of the data drawn from the same distribution. Under this assumption, the uniform convergence theory then guarantees that a model's empirical training error is closed to the true error. Hence, the learned model will make accurate predictions for new samples. A domain consists of feature space and a marginal probability distribution i.e., the features of the data and the distribution of those features in the dataset. Nevertheless, in many real-word applications, it is not possible for many domains to obtain large quantities of labeled data to train and test models from the same distribution. To deal with the lack of large labeled dataset for building a predictive model for a domain, first solution might be applying models trained on a large-scale labeled domain directly to another unlabeled domain. However, standard models cannot cope with changes in data distributions between training and test phases, hence, will not perform well. Domain adaptation deals with the setting in which the training and testing data are sampled from different distributions. Single-source domain adaptation then focuses on transferring source distribution information to the target distribution where there are one labeled source domain and one unlabeled target domain and it has been approached in a variety of manners with variety of theoretical underpinnings analysis [37]–[39]. One of these approaches to single source domain adaptation is to initiate shift invariant representations via domain Invariant feature learning by aligning the source and target domains. It focuses on creating a domain-invariant feature representation i.e features are encouraged to follow the same distribution regardless of whether the input data is from the source or the target domain. These representations aim to be invariant to the shift in distribution between source and target domain. There are several representation learning approaches, such as domain adversarial networks [7], [40]–[42] and the marginalized stacked denoising autoencoders [43] and the denoising autoencoder [14]. Each one of these methods show some degree of success to tackle the problem of single source domain adaptation. Nevertheless, multi-source domain adaptation presents a more complex situation.

C. Multiple Source Domain adaptation

When the training data is collected from multiple labeled source domains with different distributions, combining multiple sources into a single source oftentimes results in a poorer performance rather than simply using one of the sources and ignoring the others. Many researchers treated all source domains as of equal importance for the target domain prediction [44]–[46]. Here, the primary focus of unsupervised multiple source domain adaptation is on several source domains with fully labeled data and one target domain. The one target domain data is unlabeled but available during the training process, and the label sets of all source domains and the target domain are the same. Here, the challenging part is to learn domains' relationship. There is a variety of works in multiple source domain adaptation area. In general, following the same approach as in single domain adaptation, many methods learn domain shared invariant components for the target domain prediction [47]. Some methods try to learn a global domain similarity metric using only labeled data from sources [48], [49]. Many approaches are based on finding a way to determine the weight when focusing on the source domain marginal distribution and using a weighted

combination of numbers of hypotheses from source domains [48], [50]–[52]. Many of these algorithms employed unlabeled samples from target data as well to find a distributed weighted combination of multiple source domains for prediction of target samples, to obtain a supplementary training set that mimics the target domain samples or to align source-target pairs via an alignment loss [50], [53], [54]. There also exist several adversarial methods which globally align the source domains to the target domain [55], [56]. In addition, the focus of mixture of experts approach is to combine the prediction of a learned domain specific representation and global shared representations [57]–[59].
In this study, we introduce a model of multi-source domain adaptation for binary classification in the aforementioned setting where we classify positive and negative user reactions from transportation tweets. We propose to address domain adaptation from multiple sources for sentiment classification of transportation-related tweets by minimizing the moment distance between the target domain and each source domain. We use per-domain classifier which enhances this alignment by maximizing the disagreement between two classifiers while minimizing it with respect to the feature extraction.

III. Methodology

In this section, we explain the detail of our proposed method. We present the problem settings, definitions, theoretical insights and training steps associated with our proposed approach to multi-source domain adaptation by moment-based approach for sentiment analysis towards transportation performance.

A. Definitions and theoretical foundations

We use domain to represent a distribution $D = (\mu, f)$ which is defined by a probability measure μ, on input space X and a labeling function $f : X \longrightarrow$ [0, 1].

Let D_S be the labeled source domain and D_T the unlabeled target domain. Then D_{S_1}, D_{S_2}, $\cdots$, D_{S_N} are multiple sources, where N is the number of sources and all domains are defined by bounded rational measures on input space X.

Let the source data X_i and the corresponding label Y_i in the ith source domain D_{S_i} drawn from the distribution $p_i(x, y)$ be X_i = $\{X_i^j\}_{j=1}^{N_i}$ and Y_i = $\{Y_i^j\}_{j=1}^{N_i}$, where N_i is the number of source samples. Suppose the target data X_T and the corresponding label Y_T drawn from distribution $p_T(x, y)$ be X_T = $\{X_T^j\}_{j=1}^{N_T}$ and Y_T = $\{Y_T^j\}_{j=1}^{N_T}$, where N_T is the number of target samples. Let the N_{TL} be the number of labeled target samples.

A hypothesis is a function $h : X \longrightarrow$ [0, 1]. The aim of unsupervised multi-source domain adaptation problem is to find a hypothesis in the given hypothesis space H, which minimizes the testing target error on D_T where N_{TL} = 0.

Remind that a hypothesis is a function $h : X \longrightarrow$ [0, 1] and a labeling function $f : X \longrightarrow$ [0, 1].
Then, the error of a hypothesis h with regard to f under distribution D_S is defined as:
$\epsilon_S(h, f) = E_{X \sim D_S}[|h(X) - f(X)|]$

When f and h are binary classification functions, this definition reduces to the probability that h disagrees with f under D_S as follows: $E_{X \sim D_S}[|h(X) - f(X)|]$ = $P_{X \sim D_S}(h(X) \neq f(X))$

We use the risk of hypothesis h as the error of h with regard to a ground truth labeling function under source distribution D_S, i.e., $\epsilon_S(h, f_S)$ = $\epsilon_S(h)$ and $\epsilon_T(h)$ is the true risk on the target domain. Then, $\hat{\epsilon}_S(h)$ denotes the empirical risk of h on the source domain D_S and $\hat{\epsilon}_T(h)$ is the empirical risk of h on the target domain.

Now, the weighted multi-source error of a hypothesis h for some weight vector α = (α_1, ... , α_N) is defined as: $\epsilon_\alpha(h)$ = $\sum_{j=1}^{N} \alpha_j \epsilon_j(h)$, where: $\sum_{j=1}^{N} \alpha_j = 1$. The $\hat{\epsilon}_\alpha(h)$ denotes the empirical α-weighted source error.
Following [60] we define the moment distance between source domain D_S and target domain D_T as:

$$d_M(D_S, D_T) = \sum_{k=1}^{2} \Big(\frac{1}{N} \sum_{i=1}^{N} \|E(X_i^k) - E(X_T^k)\|_2 + \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|E(X_i^k) - E(X_j^k)\|_2 \Big) \quad (1)$$

where :
X_1 , X_2 , ..., X_N , X_T are collections of i.i.d. samples from D_1 , D_2 , ..., D_N , D_T, respectively.

Given a compact domain $\chi \subset R^n$ and two distributions $D = (\mu, f)$, $D' = (\mu', f')$ on χ, the k-th order moment of each probability distribution is given by the Riemann–Stieltjes integral and based on [60], the order-wise moment divergence between probability measures μ, μ' is :

$$d_{m^k}(\mu, \mu') = d_{m^k}(D, D') = \sum_{i \in \Delta_k} \left| \int_\chi \prod_{j=1}^{n} (x_j)^{i_j} d\mu(x) - \int_\chi \prod_{j=1}^{n} (x_j)^{i_j} d\mu'(x) \right| \quad (2)$$

where $\{\Delta_k = (i_1, i_2, ..., i_n) \in N_0^n | \sum_{j=1}^{n} i_j = k\}$

We now provide insights for error bound on target domain data for moment-based approach built on theories presented in [61]. Let $\mathcal{H}$ be a hypothesis space of VC dimension[1] d. For each $j \in \{1, ..., N\}$, let S_j be a labeled sample of $\beta_j m$ generated by drawing the $\beta_j m$ points from the source domain D_j and labeled according to the ground truth labeling function f_j where $\sum_j \beta_j = 1$. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_{\alpha(h)}$ for a fixed weight vector α on these samples and $\overset{*}{h}_T = min_{h \in \mathcal{H}} \epsilon_T(h)$ is the target error minimizer, then there exist N integers $\{n_\epsilon^j\}_{j=1}^N$ and N constants $\{a_{n_\epsilon^j}\}_{j=1}^N$, for any $\epsilon > 0$ and any $\delta \in (0, 1)$, such that with probability at least $1 - \delta$,

$$\epsilon_T(\hat{h}) \leq \epsilon_T(\overset{*}{h}_T) + \sum_{j=1}^{N} \alpha_j \left(2\lambda_j + a_{n_\epsilon^j} \sum_{k=1}^{n_\epsilon^j} d_{m^k}(D_j, D_T) \right) + \epsilon + 4\sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}\right)\left(\frac{2d(log(\frac{2m}{d}) + 1) + 2log(\frac{4}{\delta})}{m}\right)} \quad (3)$$

where $\lambda_j = min_{h \in \mathcal{H}} \{\epsilon_T(h) + \epsilon_j(h)\}$.

The complete proof of the above-mentioned theorem is provided and discussed in [60].

According to this theorem; the upper bound on the target error of the learned hypothesis motivates our multi-source domain adaptation approach which aligns the moments between each source-target pair. In addition, according to the triangle inequality, the target error of the learned hypothesis is lower bounded by the pairwise moment divergence between source domains. This motivates the algorithms to align the moments between source domain pairs. As an example, for two source domains D_1, D_2 and one target domain D_T, according to triangle inequality the lower bound is as:
$d_{m^k}(D_1, D_T) + d_{m^k}(D_2, D_T) \geq d_{m^k}(D_1, D_2)$

B. Multi-source domain adaptation by moment-based approach

Motivated by above-mentioned theoretical insights, we propose a moment-base approach to tackle the multi-source domain adaptation problem using deep neural networks. In the multi-source setting our model receives data from N distinct sources. Each source S_j is associated with an unknown distribution D_j over input points and unknown labeling function f_j. The model receives a lot of m labeled samples, with $m_j = \beta_j m$ from each source S_j. The objective is to use these samples to train a model to perform well on a target domain $< D_T, f_T >$, which may not be one of the sources. Given a vector α = (α_1, ... , α_N) of domain weights with $\sum_{j=1}^{N} \alpha_j = 1$, we examine algorithms that minimize convex combinations of source error with regard to the labeled examples from each source domain D_j.
The first phase in our proposed method is the feature extraction phase where we train the feature extractor G to take inputs X_S and X_T and maps the source domains and the target domain into a common feature space. Hence, the feature extractor produces the embedding representations of the input data. In the second phase, the moment-based constituent minimizes the moment distance between source domains and target domain as well as the moment distance between source domains themselves. The last phase is the classification phase where we train N pairs of classifiers on labeled source domains with cross entropy loss to output the sentiment of tweets. We drive the weight vector W for the combination of the classifiers' output based on sources accuracy:

$$w_i = \frac{acc_i}{\sum_{j=1}^{N-1} acc_j}$$

[1]Vapnik–Chervonenkis dimension

Where, the N-th domain is the target domain and $W = (w_1, w_2, ..., w_{N-1})$; $\sum_{i=1}^{N-1} w_i = 1$.

To train our model, considering that we wish to align $P(x|y)$ and $P(x)$ in the meantime, we follow the training steps recommended by [62], where we use two classifiers per domain.
Step I: We consider N pairs of classifiers $C' = \{(C_1, C_1'), (C_2, , C_2')..., (C_N, , C_N')\}$ which consists of two classifiers per domain. We train C' and the feature extractor G to perform the classification task for classifying the multi-source samples. The objective is:

$$\min_{G,C'} \sum_{i=1}^{N} \mathcal{L}_{D_i} + \lambda \min_{G} d_M(D_S, D_T) \tag{4}$$

where $\mathcal{L}_{D_i}$ denotes the softmax cross entropy loss for the classifier C_i' on domain D_i and λ denotes the trade off parameter.
Step II: We train the classifier pairs for a fixed feature extractor G. This way we train the classifiers to increase the disagreement, so they can detect the target samples excluded by the support of the source. The objective is:

$$\min_{C'} \sum_{i=1}^{N} \mathcal{L}_{D_i} - \sum_{i}^{N} |P_{C_I}(D_T) - P_{C_I}'(D_T)| \tag{5}$$

where the disagreement of classifiers is defined as the Euclidean distance between the outputs of the classifiers. The $P_{C_I}(D_T)$ and $P_{C_I}'(D_T)$ are the outputs of C_i and C_i' on the target domain D_T.
Step III: We train the feature extractor G to minimize the disagreement for fixed classifiers. The objective is:

$$\min_{G'} \sum_{i}^{N} |P_{C_I}(D_T) - P_{C_I}'(D_T)| \tag{6}$$

All the above-mentioned steps are performed until the convergence of the whole network.

C. Neural Network Architecture

In our model, the feature extractor is composed of two biLSTM (bidirectional Long-Short Term Memory) layers; each layer containing two LSTMs and 1024 nodes. The embeddings are initialized with GloVe [2] [63]. Then mappings of inputs into vector representations are produced by concatenating the last states of each bi-LSTMs in the second layer. The fc layer is a fully connected neural network that produces the probability distribution of labels. Training loss is defined as the cross-entropy. We use mini-batch stochastic gradient descent (SGD) with an Adagrad optimizer, batch size of 128 and learning rate of 0.2. The classifier is a single fc layer with 2048 nodes.

IV. Experiments

In this section we begin with describing the data sources that we used for our experiments. Then we proceed to report results on different scenarios using our datasets and compare the results of our proposed method with those of baseline models.

A. Data Sources

To conduct our experiments and to put in test the performance of our proposed approach we use several datasets that we describe here. We mainly use Amazon, Yelp and Airline datasets in several scenarios for training and we use Canada metro Twitter dataset for the test phase.

Domain	Positive tweet	Negative tweet
Canada metro	Just experienced the best Metro ride of my life @stm_Orange. Thanks for the tip! #climatemarch #MTL	@stm_Orange Terrible service STM! No reliability. How are you planing on coping with all the users that will be forced to take the metro due to the Deux Montagnes train???

Fig. 1. Example of positive and negative tweets in the target domain (Canada metro)

1) Canada metro: The Canadian metro data contains tweets about subways in Canadian cities such as Toronto, Montreal, Calgary, Edmonton, Ottawa, Vancouver and Waterloo scraped from Twitter API in the timeline from 2015 to 2020. To conduct our experiments for this study we use a part of this dataset that contains 44 000 tweets. An example of positive and negative tweets is shown in Figure 1.

[2]Global Vectors for Word Representation

2) Amazon: The Amazon dataset consists of reviews from the Amazon website for different types of products. After removing duplicate items, it contains 82.83 million reviews for 24 different domains. For this study, we chose reviews from the domains such as Automotive, Books, DVDs, Electronics, kitchen and Mobile. Each review consists of a title, text body of the review and a rating from 1 to 5 scale. We label reviews with ranks 5 4 as "positive", reviews with ranks 2 and 1 as "negative". We discard reviews with rank 3.
Figure 1 shows a positive and a negative review from each of these four domains. We then keep the text body of the review and the sentiment labels for each category of products. In total we kept 100 000 reviews for each of the product categories such as Automotive, Books, DVDs, Electronics and kitchen of which 50 000 reviews are labeled as "positive" and 50 000 reviews of each category are labeled as "negative". For the category Mobile, we were able to keep 48 000 reviews consists of equal amounts of reviews with "positive" and "negative" labels.

3) Yelp: The Yelp dataset includes reviews for different categories, such as restaurants, shopping, hotels and travel, etc. from Phoenix, Las Vegas, Madison, Waterloo and Edinburgh. This dataset contains information about 42,153 businesses, 320,002 business attributes, 31,617 check-in sets, 403,210 tips and 1,125,458 text reviews. We use only restaurant reviews from the dataset provided by Yelp as part of their dataset Challenge 2014 for training and testing prediction models. We were able to preserve data balance for 38 000 records of which 19 000 reviews are labeled as "positive" and the same amounts of reviews are labeled as "negative".

4) Airline: Airline dataset originally from "Crowdflower's Data for Everyone library" [3] is the Twitter data about the problems of major American Airlines from February of 2015. For this study we use a version of this data provided for Kaggle competition which is slightly different from the original source and it contains sentiment tags. We use 3 600 records of the data from six of U.S airlines that consists of 1 800 tweets with "positive" tags and 1 800 tweets with "negative" tags.
Figure 2 shows an example of positive and negative reviews for source domains.

Domain	Positive review	Negative review
Airline	VirginAmerica you have the absolute best team and customer service ever. Every time I fly with you I'm delighted.Thank you!	Yeah sorry but there's always a problem with United. And you have an international reputation for having problems.
Automative	Wow.. the picture supplied looks great BUT the actual grille looks even better! I purchased this for our dealership for a used Nissan to help move the car and it did! An accessory, at a small price that enhanced the car to move it! Well worth the prices spent!	I couldn't use the filter. The threads on the bottom of the filter were no good. It wouldn't hold the fitting that the wires plug into.
Books	Excellent writing. The character development nabbed me from page 1, tho' physical descriptions of protagonists is usually not recommended. It works here. Ms Coplin is an author I'll be watching out for. Truly gifted and disciplined. A modern classic.	The editing in this book is so poor that it's pretty much unreadable. Also, the book is greatly overpriced for the amount of information it contains, which is unfortunately quite limited. Definitely not worth purchasing.
DVDs	We could not find this movie anywhere in stores! I am absolutely thrilled to have found this! The movie came in excellent condition and my kids and I are very excited to have it. I would highly recommend this seller and this movie!	As others have warned, this is a horrible reproduction of a truly fantastic film. The dvd I received had many bad video spots and then stopped just before the charge scene. I wish I had my old vcr copy still. I am surprised Amazon hasn't pulled this from the shelf.
Electronics	Product arrived in timely manner, everything fit and quality appears to be good. Will recommend it to friends.	Overall disappointed with product. First time I went to mount this to my radio the mounting screw broke and now the attachment piece comes off the radio on occasion. When this needs replacing, which I think will be soon, I will look for other brands for replacement.
Kitchen	This is just so simple to use. It never gets old, it never gets damaged and it's always there to serve me! From what I see and experience, I know that this made my life easier and more enjoyable when cooking	It smells like smoke or burnt rubber or something any time I use it, and it's only one speed. It could be better.
Mobile	This works like a charm! I've had some problems with off brand cables with my Apple products before but these work like a charm. The length allows you to use your device while it is charging without being tied closely to the outlet. It is a very useful product. Thanks!	After a bit of water fell on it, it stopped working. This product is a complete sham. SAVE YOUR MONEY!

Fig. 2. Example of positive and negative reviews for source domains

B. Baseline models

To evaluate and to demonstrate the performance of our proposed method, we use three of state-of the-art algorithms in the domain as our baseline models that we describe below. All the experiments are implemented using PyTorch from Torch library[4].

Domain-Adversarial Training of Neural Networks [7] (DANN): Based on Unsupervised Domain Adaptation by Backpropagation [64] it is an adversarial learning based deep net approach for domain adaptation problem. The DANN architecture consists of three parts: label predictor, domain classifier and feature extractor. The architecture takes advantage of a label predictor module that predicts the class labels and is used both during training and at test time; and a domain classifier that discriminates between the source and the target domains during training. This way it promotes the emergence of features that are discriminative for

[3]https://appen.com/open-source-datasets/
[4]http://pytorch.org

Sources	MDAN	DANN	MSDA	proposed method
Auto + Book	0.7807	0.7796	0.6090	0.8668
Auto + Elec	0.7784	0.7720	0.6251	**0.8723**
Auto + kit	0.7885	0.7807	0.6022	**0.8869**
Auto + Mob	0.7892	0.7816	0.6236	0.8699
Auto + DVD	0.7803	0.7818	0.6038	0.8596
Air + Yel	0.7793	0.7893	0.6328	0.8638
Book + Elec	0.7939	0.7910	0.6381	0.8604
Book + kit	0.7911	0.7893	0.6261	0.8603
Book + Mob	0.7797	0.7807	0.6166	0.8684
Book + DVD	0.7836	0.7320	0.6083	0.8586
Elec + kit	0.7886	0.7614	0.6888	0.8559
Elec + Mob	0.7930	0.7129	0.6769	0.8610
Elec + DVD	0.7996	0.7940	0.6284	0.8571
kit + Mob	0.7265	0.7390	0.6129	0.8369
kit + DVD	0.7480	0.7747	0.6897	0.8206
Mob + DVD	0.7951	0.7133	0.6606	0.8303

TABLE I

SENTIMENT CLASSIFICATION ACCURACY WITH TWO SOURCE DOMAINS AND CANADA METRO AS TARGET DOMAIN

the main learning task on the source domain and indiscriminate with respect to the shift between the domains. In order to extend the DANN method to multi-source domain adaptation, we combine all the source domains into a single one and train it using DANN.

Marginalized Stacked Denoising Autoencoders [43] (MSDA): It is an unsupervised representation learning based algorithm for domain adaptation. In particular, *denoising autoencoders* learns data representations by reconstruction, recovering original features from data that are artificially corrupted with noise. The architecture consists of two stages of representation learning and domain adaptation. First, the model learns features in an unsupervised fashion on the union of the source and target datasets. Once a MSDA is trained, the output of all layers, combined with the original features are concatenated and form the new representation. Then, all inputs are transformed into the new feature space. The second phase is applying MSDA to domain adaptation by a linear Support Vector Machine (SVM). The SVM is then trained on the transformed source inputs and tested on the target domain.

Multisource Domain Adversarial Networks [55] (MDAN): It approaches multi-source domain adaptation by optimizing task-adaptive generalization bounds. The MDAN Network architecture consists of a feature extractor, a domain classifier, and a task learning that are combined in one training process. In the smooth version of MDAN all the domain classification risks over all source domains are combined and backpropagated adaptive with gradient reversal.

C. Experimental Results

In this section we report the results of our experiments with different scenarios. Altogether, we conducted over 400 experiments of which we report the results of 60 major experiments here.

In the results below, the abbreviations: *Auto* stands for Automotive, *Book* stands for Books, *DVD* stands for DVDs, *Elec* stands for Electronics, *kit* stands for kitchen, *Mob* stands for Mobile, *Yel* stands for Yelp and *Air* stands for Airline.

In table I and table II we report the experimental results of multi source domains with two sources and three sources & more respectively for our proposed method as well as the baseline methods. As depicted in Table I in all the reported scenarios our proposed method outperforms all the baseline methods in terms of model accuracy with two source domains. Nonetheless, the very best results being *0.8723* & *0.8869* are among the scenarios where one of the sources is from *Automotive* domain.
Table II demonstrates anew the performance of our proposed method that exceeds the best accuracy of baselines in all reported scenarios for three and more source domains. Nevertheless, once again we see *Automotive* domain as one of the source domains where our model achieves best accuracy. This behaviour can be because of the closeness of the *Automotive* domain to the *Metro* domain, however, the inquiry of this phenomenon is beyond the scope of this study.
Overall, we see the accuracy of *0.9196* as the best result which our proposed method achieved by using all the source domains that we have. Figure 3 also demonstrates the effect of the number of source domains on the accuracy of the proposed method. As we see more source domains yields better accuracy to the extent that the best result is achieved with using all eight source domains.

In Table III and IV, we report the results of some of experiments regarding the study of the performance of the proposed method through procedures in which we remove its essential parts such as feature extraction part and moment-based module. With regards to the results shown in Table III, the feature extraction part contributes to the performance of our proposed

Sources	MDAN	DANN	MSDA	proposed method
Auto + Book + Elec	0.7877	0.7815	0.6701	0.8734
Auto + Book + kit	0.7972	0.7981	0.6945	0.8796
Auto + Book + Mob	0.7961	0.7931	0.6462	0.8845
Auto + Book + DVD	0.7800	0.7904	0.6617	0.8686
Auto + Elec + kit	0.7843	0.7889	0.6469	0.8885
Auto + Elec + Mob	0.7940	0.7948	0.6440	0.8827
Auto + Elec + DVD	0.7930	0.7881	0.6454	0.8725
Auto + kit + Mob	0.7888	0.7924	0.6344	0.8781
Auto + kit + DVD	0.7974	0.7992	0.6850	0.8619
Auto + Mob + DVD	0.7986	0.7872	0.6839	0.8609
Auto + Book + Elec + kit	0.7895	0.7808	0.6568	0.8899
Auto + Book + Elec + Mob	0.7851	0.7929	0.6748	0.8914
Auto + Book + Elec + DVD	0.7992	0.7989	0.6376	0.8900
Auto + Book + kit + Mob	0.7949	0.7835	0.6572	0.8908
Auto + Book + kit + DVD	0.7920	0.7852	0.6816	0.8892
Auto + Book + Mob + DVD	0.7806	0.7900	0.6691	0.8912
Auto + Elec + Mob + kit + DVD	0.7822	0.7823	0.6665	0.8952
Auto + Book + Elec + kit + Mob	0.7829	0.7918	0.6893	0.8937
Auto + Book + Elec + kit + DVD	0.7866	0.7838	0.6830	0.8907
Auto + Book + Elec + kit + Mob + DVD + Yel + Air	0.7957	0.7920	0.6973	**0.9196**

TABLE II
SENTIMENT CLASSIFICATION ACCURACY WITH THREE AND MORE SOURCE DOMAINS AND CANADA METRO AS TARGET DOMAIN

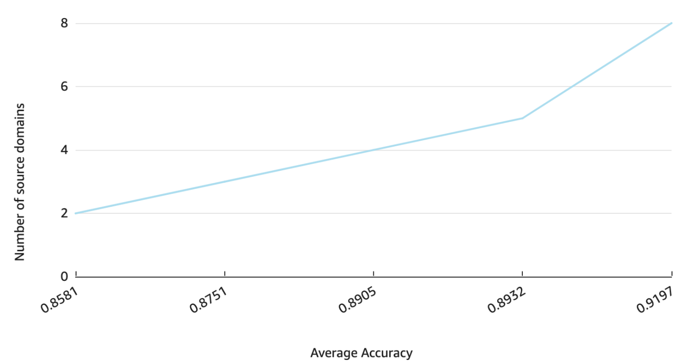


Fig. 3. The Effect of Number of Source Domains on Accuracy of the Method

method by improving the results by *0.079* up to *0.137* in different scenarios. In addition, Table IV depicts the contribution of the moment-based module to the performance of the proposed method by up to *0.184* increase in the accuracy.

V. CONCLUSION

Traditional machine learning methodologies assume that the training and test sets come from the same distributions. Hence, a model learned from the labeled training data is expected to perform well on the test data. Despite what has just been mentioned, in real-world applications, when the training and test data come from distinct distributions, this assumption may not always hold true. There would be a discrepancy amongst domain distributions in this situation. Therefore, in this case, applying the trained model to the new dataset directly could result in performance loss. In this article, to identify the sentiment polarity of transportation users' opinions we approached the problem of lack of annotated data for sentiment analysis in transportation field. We studied the possibility of tackling this problem by means of transfer learning using multi source domains. We theoretically justified our approach. We theoretically justified our approach. We focused on developing a unsupervised domain adaptation, where the labels are only available in the source domains. This is challenging due to the presence of notable domain shift between target domain and source domains as well as between source domains themselves. To mitigate the domain gap we proposed a functional solution to align multiple source domains with the target domain as well as with each others. To study the performance of the proposed model and the effectiveness of its components, we conduct a comprehensive set of experiments that confirms the proposed method is capable of outperforming baseline models in terms of accuracy. A future research direction is to address transportation sentiment analysis when the target domain language is different from source languages.

VI. APPENDIX

In the Section I we report the results of baseline models and our proposed methods with single source domain and Section II shows feature visualization with t-SNE plot.

Sources	without feature extraction part	with feature extraction part
Auto + Elec	0.7353	0.8723
kit + DVD	0.74063	0.8206
Auto + Elec + kit	0.7621	0.8885
Auto + Mob + DVD	0.7536	0.8609
Auto + Book + Elec + M	0.7847	0.8914
Auto + Book + K + DVD	0.7811	0.8892
Auto + Elec + Mob + kit + DVD	0.7901	0.8952
Auto + Elec + Mob + kit + DVD + Book + Yel + Air	0.7858	0.9196

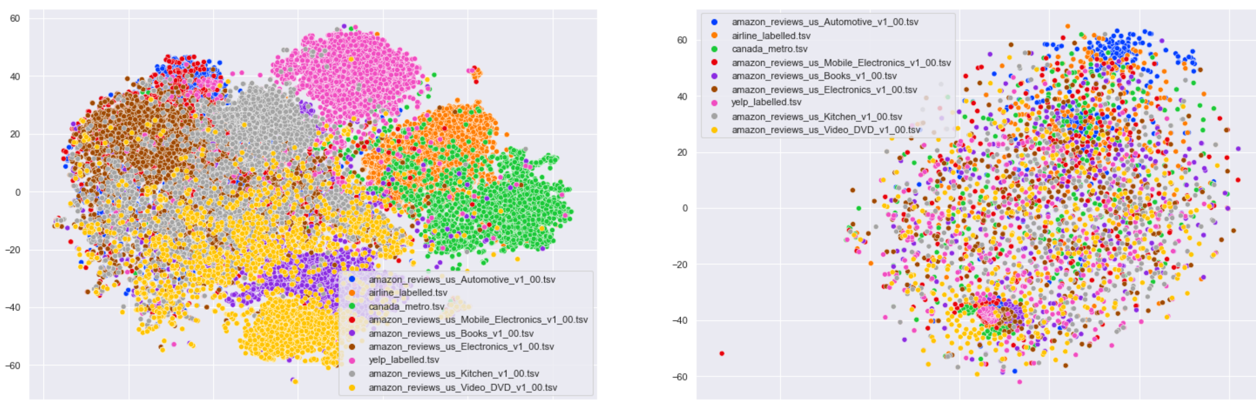
TABLE III
RESULTS OF PROPOSED METHOD WITH/WITH OUT REPRESENTATION PART

Sources	without moment-based module	with moment-based module
Auto + Elec	0.6878	0.8723
kit + DVD	0.6971	0.8206
Auto + Elec + kit	0.7389	0.8885
Auto + Mob + DVD	0.7289	0.8609
Auto + Book + Elec + M	0.7632	0.8914
Auto + Book + kit + DVD	0.7755	0.8892
Auto + Elec + Mob + kit + DVD	0.7731	0.8952
Auto + Elec + Mob + kit + DVD + Book + Yel + Air	0.7987	0.9196

TABLE IV
RESULTS OF PROPOSED METHOD WITH/WITH OUT MOMENT-BASED MODULE

Section I

As shown in Table V for all single source domain scenarios the proposed methods shows better performance compare to baseline models which is expected. Considering the formula (3), single source can be a special case of our proposed method when $N = 1$.



(a) Data Representation Before Feature Extraction

(b) Data Representation after Feature Extraction

Fig. 4. Data Representation (Best viewed in colour)

Section II

In Figure 4 points are sampled from source domains & target domain. Part (a) & part (b) show the data representation for different domains before and after feature extraction part of our proposed method respectively. Different colours denote different domains. The feature extraction part maps the source domains into a common latent feature space.

Sources	MDAN	DANN	MSDA	proposed method
Air	0.6991	0.7052	0.5986	0.7718
Auto	0.7188	0.7184	0.5843	0.7543
Book	0.7305	0.7234	0.6230	0.7694
Elec	0.7215	0.7248	0.6142	0.7799
kit	0.7208	0.7230	0.6030	0.7400
Mob	0.6936	0.7003	0.5908	0.7655
DVD	0.7084	0.7042	0.6047	0.7558
Yel	0.7127	0.7191	0.6199	0.7431

TABLE V
SENTIMENT CLASSIFICATION ACCURACY FOR SINGLE SOURCE DOMAIN AND CANADA METRO AS TARGET DOMAIN

REFERENCES

[1] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Dataset shift in machine learning. The MIT Press, 2009.
[2] A. Storkey, "When training and test sets are different: characterizing learning transfer," Dataset shift in machine learning, vol. 30, pp. 3–28, 2009.
[3] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in CVPR 2011. IEEE, 2011, pp. 1521–1528.
[4] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012, pp. 2066–2073.
[5] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," IEEE Transactions on Neural Networks, vol. 22, no. 2, pp. 199–210, 2010.
[6] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in European conference on computer vision. Springer, 2016, pp. 443–450.
[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," The journal of machine learning research, vol. 17, no. 1, pp. 2096–2030, 2016.
[8] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," arXiv preprint arXiv:1705.10667, 2017.
[9] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7167–7176.
[10] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han, "Domain-specific batch normalization for unsupervised domain adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 7354–7362.
[11] S. Roy, A. Siarohin, E. Sangineto, S. R. Bulo, N. Sebe, and E. Ricci, "Unsupervised domain adaptation using feature-whitening and consensus loss," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9471–9480.
[12] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in Proceedings of the 45th annual meeting of the association of computational linguistics, 2007, pp. 440–447.
[13] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in Proceedings of the 19th international conference on World wide web, 2010, pp. 751–760.
[14] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in ICML, 2011.
[15] D. Bollegala, D. Weir, and J. Carroll, "Cross-domain sentiment classification using a sentiment sensitive thesaurus," IEEE transactions on knowledge and data engineering, vol. 25, no. 8, pp. 1719–1731, 2012.
[16] R. Xia, X. Hu, J. Lu, J. Yang, and C. Zong, "Instance selection and instance weighting for cross-domain sentiment classification via pu learning," in Twenty-Third International Joint Conference on Artificial Intelligence, 2013.
[17] J. Yu and J. Jiang, "Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification." Association for Computational Linguistics, 2016.
[18] Z. Li, Y. Wei, Y. Zhang, and Q. Yang, "Hierarchical attention transfer network for cross-domain sentiment classification," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.
[19] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," arXiv preprint arXiv:1810.11910, 2018.
[20] S. M. Grant-Muller, A. Gal-Tzur, E. Minkov, S. Nocera, T. Kuflik, and I. Shoor, "Enhancing transport data collection through social media sources: methods, challenges and opportunities for textual data," IET Intelligent Transport Systems, vol. 9, no. 4, pp. 407–417, 2014.
[21] S. Das, X. Sun, and A. Dutta, "Text mining and topic modeling of compendiums of papers from transportation research board annual meetings," Transportation Research Record, vol. 2552, no. 1, pp. 48–56, 2016.
[22] V. Effendy, A. Novantirani, and M. K. Sabariah, "Sentiment analysis on twitter about the use of city public transportation using support vector machine method," Intl. J. ICT, vol. 2, no. 1, pp. 57–66, 2016.
[23] F. Ali, D. Kwak, P. Khan, S. R. Islam, K. H. Kim, and K. S. Kwak, "Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling," Transportation Research Part C: Emerging Technologies, vol. 77, pp. 33–48, 2017.
[24] J. F. F. Pereira, "Social media text processing and semantic analysis for smart cities," arXiv preprint arXiv:1709.03406, 2017.
[25] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources," in Proceedings of the first SIGMM workshop on Social media, 2009, pp. 3–10.
[26] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in International conference on discovery science. Springer, 2010, pp. 1–15.
[27] N. F. Da Silva, E. R. Hruschka, and E. R. Hruschka Jr, "Tweet sentiment analysis with classifier ensembles," Decision Support Systems, vol. 66, pp. 170–179, 2014.
[28] C. Clavel and Z. Callejas, "Sentiment analysis: from opinion mining to human-agent interaction," IEEE Transactions on affective computing, vol. 7, no. 1, pp. 74–93, 2015.
[29] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep convolution neural networks for twitter sentiment analysis," IEEE Access, vol. 6, pp. 23 253–23 260, 2018.
[30] U. Naseem, I. Razzak, K. Musial, and M. Imran, "Transformer based deep intelligent contextual embedding for twitter sentiment analysis," Future Generation Computer Systems, vol. 113, pp. 58–69, 2020.
[31] F. Chen, R. Krishnan et al., "Transportation sentiment analysis for safety enhancement," 2013.
[32] C. Collins, S. Hasan, and S. V. Ukkusuri, "A novel transit rider satisfaction metric: Rider sentiments measured from online social media data," Journal of Public Transportation, vol. 16, no. 2, p. 2, 2013.
[33] L. Schweitzer, "Planning and social media: a case study of public transit and stigma on twitter," Journal of the American Planning Association, vol. 80, no. 3, pp. 218–238, 2014.
[34] S. Das, X. Sun, A. Dutta et al., "Investigating user ridership sentiments for bike sharing programs," Journal of Transportation Technologies, vol. 5, no. 02, p. 69, 2015.

[35] T. T. Luong and D. Houston, "Public opinions of light rail service in los angeles, an analysis using twitter data," IConference 2015 Proceedings, 2015.
[36] I. P. Windasari, F. N. Uzzi, and K. I. Satoto, "Sentiment analysis on twitter posts: An analysis of positive or negative opinion on gojek," in 2017 4th international conference on information technology, computer, and electrical engineering (ICITACEE). IEEE, 2017, pp. 266–269.
[37]
[38] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.
[39] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," IEEE transactions on pattern analysis and machine intelligence, 2019.
[40] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.
[41] Y. Zhang, R. Barzilay, and T. Jaakkola, "Aspect-augmented adversarial networks for domain adaptation," Transactions of the Association for Computational Linguistics, vol. 5, pp. 515–528, 2017.
[42] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," arXiv preprint arXiv:1412.3474, 2014.
[43] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," arXiv preprint arXiv:1206.4683, 2012.
[44] S. Li and C. Zong, "Multi-domain sentiment classification," in Proceedings of ACL-08: HLT, Short Papers, 2008, pp. 257–260.
[45] P. Luo, F. Zhuang, H. Xiong, Y. Xiong, and Q. He, "Transfer learning from multiple source domains via consensus regularization," in Proceedings of the 17th ACM conference on Information and knowledge management, 2008, pp. 103–112.
[46] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources." Journal of Machine Learning Research, vol. 9, no. 8, 2008.
[47] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in International Conference on Machine Learning. PMLR, 2013, pp. 10–18.
[48] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua, "Domain adaptation from multiple sources via auxiliary classifiers," in Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 289–296.
[49] M. Yu, X. Guo, J. Yi, S. Chang, S. Potdar, Y. Cheng, G. Tesauro, H. Wang, and B. Zhou, "Diverse few-shot text classification with multiple metrics," arXiv preprint arXiv:1805.07513, 2018.
[50] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation with multiple sources," Advances in neural information processing systems, vol. 21, pp. 1041–1048, 2008.
[51] J. Gao, W. Fan, J. Jiang, and J. Han, "Knowledge transfer via multiple model local structure mapping," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 283–291.
[52] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Multisource domain adaptation and its application to early detection of fatigue," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 6, no. 4, pp. 1–26, 2012.
[53] H. S. Bhatt, M. Sinha, and S. Roy, "Cross-domain text classification with multiple domains and disparate label sets," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1641–1650.
[54] Y. Zhu, F. Zhuang, and D. Wang, "Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, 2019, pp. 5989–5996.
[55] H. Zhao, S. Zhang, G. Wu, G. J. Gordon et al., "Multiple source domain adaptation with adversarial learning," 2018.
[56] X. Chen and C. Cardie, "Multinomial adversarial networks for multi-domain text classification," arXiv preprint arXiv:1802.05694, 2018.
[57] Y.-B. Kim, K. Stratos, and D. Kim, "Domain attention with an ensemble of experts," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 643–653.
[58] J. Guo, D. J. Shah, and R. Barzilay, "Multi-source domain adaptation with mixture of experts," arXiv preprint arXiv:1809.02256, 2018.
[59] Y. Li, T. Baldwin, and T. Cohn, "What's in a domain? learning domain-robust text representations using adversarial training," arXiv preprint arXiv:1805.06088, 2018.
[60] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1406–1415.
[61] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," Machine learning, vol. 79, no. 1, pp. 151–175, 2010.
[62] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3723–3732.
[63] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[64] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in International conference on machine learning. PMLR, 2015, pp. 1180–1189.