MAXIMUM BATCH FROBENIUS NORM FOR MULTI-DOMAIN TEXT CLASSIFICATION

Yuan Wu* *Diana Inkpen*† *Ahmed El-Roby**

* Carleton University, Ottawa, Canada
† University of Ottawa, Ottawa, Canada

ABSTRACT

Multi-domain text classification (MDTC) has obtained remarkable achievements due to the advent of deep learning. Recently, many endeavors are devoted to applying adversarial learning to extract domain-invariant features to yield state-of-the-art results. However, these methods still face one challenge: transforming original features to be domain-invariant distorts the distributions of the original features, degrading the discriminability of the learned features. To address this issue, we first investigate the structure of the batch classification output matrix and theoretically justify that the discriminability of the learned features has a positive correlation with the Frobenius norm of the batch output matrix. Based on this finding, we propose a maximum batch Frobenius norm (MBF) method to boost the feature discriminability for MDTC. Experiments on two MDTC benchmarks show that our MBF approach can effectively advance state-of-the-art performance.

Index Terms— Transfer learning, multi-domain text classification, adversarial learning, Frobenius norm

1. INTRODUCTION

Text classification is a fundamental task in natural language processing (NLP) [1]. In recent years, text classification models have demonstrated significant improvements by employing deep neural networks (DNNs) [2]. These achievements mainly rely on large amounts of labeled data. Unfortunately, in many real-world applications, the availability of labeled data may vary across different domains [3]. For domains like book and movie reviews, abundant labeled data has been made available, while for domains like medical equipment reviews, labeled data is scarce. Therefore, it is of great significance to investigate how to improve the classification accuracy on the target domain by leveraging available resources from related domains.

Multi-domain text classification (MDTC) is proposed to address the above issue. Recent advances in MDTC are reliant on adversarial learning to enhance the transferability of feature representations [4], and shared-private paradigm to benefit the system performance from domain-specific knowledge [5]. Adversarial learning is first proposed by generative adversarial networks (GANs) for image generation [6]. It uses noise data to generate images and enforce the distribution of the generated images to be similar to that of the real images. The success of GANs inspires the adversarial alignment [7, 8, 9]. Adversarial alignment can harness the power of DNNs to learn domain-invariant features that are both discriminative and transferable by deploying a minimax game between a feature extractor and a domain discriminator: the discriminator aims to distinguish features across different domains, while the feature extractor contrives to deceive the discriminator. When these two components reach equilibrium, the learned features are regarded as domain-invariant. As text classification is a well-known domain-specific task [10], the same word in different domains may express different sentiments. For instance, the word *fast* represents negative sentiment in the electronic review ”The battery of the camera runs fast”, while it indicates positive sentiment in the car review ”The car runs fast”. Domain-specific knowledge cannot be ignored in MDTC. Shared-private paradigm is proposed to incorporate domain-specific features to enhance the discriminability of the domain-invariant features [8]. The MDTC models that adopt adversarial learning and shared-private paradigm can yield state-of-the-art performance [8, 11, 12, 13]. However, these models still face one critical issue: as adversarial alignment needs to distort the original feature distributions to generate domain-invariant features, the transferability of the domain-invariant feature is obtained at the expense of sacrificing the discriminability [14].

In this paper, we address the aforementioned issue by proposing a maximum batch Frobenius norm (MBF) approach to boost the feature discriminability. We first investigate the structure of the classification output matrix of a randomly sampled data batch, and then theoretically analyze how the discriminability of the learned features can be measured by the batch classification output matrix. We find that the feature discriminability has a positive correlation with the Frobenius norm of the batch output matrix. Motivated by this finding, our MBF method boosts the feature discriminability by maximizing the Frobenius norm of the batch output matrix. Experiments are conducted on two MDTC benchmarks: the Amazon review dataset and the FDU-MTL dataset. The experimental results show that our MBF method can outperform the state-of-the-art baselines on both datasets.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

2. PROPOSED METHOD

In this paper, we study the MDTC problem, where the text data comes from multiple domains, each with varying amounts of both labeled and unlabeled data. Assume we have M different domains $\{D_i\}_{i=1}^{M}$, the i-th domain D_i contains a limited amount of labeled data $\mathbb{L}_i = \{(\mathbf{x}_j, y_j)\}_{j=1}^{l_i}$ and a large amount of unlabeled data $\mathbb{U}_i = \{\mathbf{x}_j\}_{j=1}^{u_i}$, where l_i and u_i are the numbers of labeled and unlabeled data in D_i. Our goal is to learn a model that maps an instance $\mathbf{x}$ to its corresponding label y and guarantee the model generalizes well over all domains.

2.1. Adversarial Multi-Domain Text Classification

Adversarial MDTC methods, starting from adversarial multi-task learning for text classification (ASP-MTL) [7], have become increasingly influential in MDTC. The main idea is to adversarially align different feature distributions to extract both explanatory and transferable features. Basically, the standard MDTC model has four components: a shared feature extractor $\mathcal{F}_s$, M domain-specific feature extractors $\{\mathcal{F}_d^i\}_{i=1}^{M}$, a domain discriminator $\mathcal{D}$, and a classifier $\mathcal{C}$. The shared feature extractor $\mathcal{F}_s$ learns domain-invariant features, while the i-th domain-specific feature extractor $\mathcal{F}_d^i$ captures domain-specific knowledge of the i-th domain D_i. The two types of feature extractors are supposed to complement each other and maximally capture useful information across domains. The discriminator $\mathcal{D}_i$ is a M-class classifier which takes a domain-invariant feature vector as input and outputs the probability vector for input data; i.e., $\mathcal{D}_i(\mathcal{F}_s(\mathbf{x}))$ denotes the probability of instance $\mathbf{x}$ coming from the i-th domain. In this paper, we use the negative log-likelihood (NLL) loss to encode the multinomial adversarial loss:

$$\mathcal{L}_{Adv} = -\sum_{i=1}^{M} \mathbb{E}_{\mathbf{x}\sim\mathbb{L}_i\cup\mathbb{U}_i} \log(\mathcal{D}_i(\mathcal{F}_s(\mathbf{x}))) \tag{1}$$

By training $\mathcal{F}_s$ adversarially to deceive $\mathcal{D}$, the learned features are made transferable across domains. In addition, the classifier $\mathcal{C}$ are trained together with $\mathcal{F}_s$ and $\{\mathcal{F}_d^i\}_{i=1}^{M}$ to minimize the classification error on the labeled data. This makes the feature representations discriminative across categories. The classifier $\mathcal{C}$ takes the concatenation of a domain-invariant feature and a domain-specific feature as input, and predicts the label probability; i.e., $\mathcal{C}_y([\mathcal{F}_s(\mathbf{x}), \mathcal{F}_d^i(\mathbf{x})])$ denotes the probability of instance $\mathbf{x}$ belonging to label y, where $[\cdot,\cdot]$ represents the concatenation of two vectors. We also use NLL loss to formulate the classification loss:

$$\mathcal{L}_c = -\sum_{i=1}^{M} \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{L}_i} \log(\mathcal{C}_y([\mathcal{F}_s(\mathbf{x}), \mathcal{F}_d^i(\mathbf{x})])) \tag{2}$$

2.2. Frobenius Norm of Batch Output Matrix

DNNs are often trained by using mini-batch stochastic gradient descent (SGD). Here, we set the batch size to B and assume we handle the K-class classification. Then the batch output matrix can be represented as $A \in \mathbb{R}^{B\times K}$. The matrix A should satisfy two conditions: (1) the sum of each entry of A is 1; (2) each entry is non-negative, i.e.:

$$\begin{gathered}\sum_{j=1}^{K} A_{i,j} = 1 \quad \forall i \in 1...B \\ A_{i,j} \geq 0 \quad \forall i \in 1...B, j \in 1...K\end{gathered} \tag{3}$$

In standard supervised machine learning, a well-performed model could learn features with high discriminability by processing abundant labeled data. While in label-scarce scenarios, limited amounts of labeled data can not guarantee the sufficient discriminability of the learned features, and the data without supervision often deteriorate the feature discriminability [15]. Therefore, to learn qualified features in MDTC, we need to optimize the predictions on unlabeled data.

Semi-supervised learning (SSL) is a typical label-scarce machine learning task. In SSL, entropy minimization is often used to minimize the uncertainty of prediction on unlabeled data [16]. In general, lower uncertainty suggests higher discriminability. The entropy of matrix A is formulated as:

$$H(A) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{K} A_{i,j}\log(A_{i,j}) \tag{4}$$

Inspired by entropy minimization [16], stronger discriminability is often achieved by minimizing $H(A)$. When $H(A)$ is optimized to its minimum, for each row of A, only one entry is 1 and the other $K-1$ entries are 0. The minimal $H(A)$ leads to the highest discriminability, where each item A_i is fully determined.

We denote the Frobenius norm of the matrix A as $\|A\|_F$. As $\|A\|_F$ has strict opposite monotonicity with $H(A)$ (See section 2.4 for the theoretical justification), maximizing $\|A\|_F$ is equivalent to minimizing $H(A)$. Therefore, the feature discriminability can be enhanced by maximizing $\|A\|_F$. $\|A\|_F$ is formulated as:

$$\|A\|_F = \sqrt{\sum_{i=1}^{B}\sum_{j=1}^{K} |A_{i,j}|^2} \tag{5}$$

To improve the discriminability of the learned features, we maximize the Frobenius norm of the batch classification output matrix of the unlabeled data. The corresponding loss function is formulated as:

$$\mathcal{L}_{BF} = \sum_{i=1}^{M} \mathbb{E}_{\mathbf{x}\sim\mathbb{U}_i} \frac{1}{B} \|\mathcal{C}^B([\mathcal{F}_s(\mathbf{x}), \mathcal{F}_d^i(\mathbf{x})])\|_F \tag{6}$$

where $\mathcal{C}^B(\cdot)$ represents the batch output matrix with B randomly sampled unlabeled data. Maximizing $\mathcal{L}_{BF}$ could improve the feature discriminability without degrading the transferability, the gradient of the Frobenius norm could be calculated according to [17]. Therefore, $\mathcal{L}_{BF}$ could be applied to optimize the gradient-based DNNs.

2.3. Maximum Batch Frobenius Norm

Based on the above finding, we propose the maximum batch Frobenius norm (MBF) method, whose objective function can be written as:

$$\min_{\mathcal{F}_s, \{\mathcal{F}_d^i\}_{i=1}^M, \mathcal{C}} \max_{\mathcal{D}} \mathcal{L}_c + \alpha\mathcal{L}_{Adv} - \beta\mathcal{L}_{BF} \tag{7}$$

where α and β are hyperparameters that balance different loss functions. The MBF model can be trained with mini-batch SGD and we adopt the alternating fashion [6] to achieve the minimax optimization. The training algorithm is presented in Algorithm 1.

Algorithm 1 SGD training algorithm

```
Input: labeled data $\mathbb{L}_i$ and unlabeled data $\mathbb{U}_i$ in $M$ domains; two hyperparameters $\alpha$ and $\beta$.
for number of training iterations do
    Sample labeled mini-batches from the multiple domains $B^\ell = \{B_1^\ell, \cdots, B_M^\ell\}$.
    Sample unlabeled mini-batches from the multiple domains $B^u = \{B_1^u, \cdots, B_M^u\}$.
    Calculate $loss = \mathcal{L}_\mathcal{C} + \alpha\mathcal{L}_{Adv} - \beta\mathcal{L}_{BF}$ on $B^\ell$ and $B^u$;
    Update $\mathcal{F}_s$, $\{\mathcal{F}_d^i\}_{i=1}^M$, $\mathcal{C}$ by descending along the gradients $\Delta loss$.
    Calculate $l_D = \mathcal{L}_{Adv}$ on $B^\ell$ and $B^u$;
    Update $\mathcal{D}$ by ascending along the gradients $\Delta l_D$.
end for
```

2.4. Theoretical Analysis

As shown in Eq.5, the Frobenius norm of matrix A is the square root sum of all entries in A. The calculating process can be divided into two steps: (1) calculating the quadratic sum of each row in A; (2) calculating the square root of the sum of all rows. As the monotonicity of the square root of the sum of all rows depends on the monotonicity of each row and there is no extra constraint on different rows, we consider the monotonicity of the quadratic sum of each row to analyze the monotonicity of the Frobenius norm. Analogously, we could also draw the monotonicity of the entropy by analyzing the monotonicity of each row for the entropy.

We use $f(A_i)$ to denote the square sum of the i-th row in A, so we have $f(A_i) = \sum_{j=1}^{K} A_{i,j}^2$. To analyze the monotonicity of a function on multiple variables, we could simply analyze the monotonicity of the function on each variable. We make the assumption that the variable $A_{i,K}$ is the only variable dependent on $A_{i,j}$ as the variables are supposed to be independent and $\sum_{j=1}^{K} A_{i,j} = 1$. Then the partial derivative of $f(A_i)$ can be calculated as:

$$\frac{\partial f(A_i)}{\partial A_{i,j}} = 2A_{i,j} - 2A_{i,K} = 4A_{i,j} - 2(1 - \sum_{k=1,k\neq j}^{K-1} A_{i,k}) \tag{8}$$

We can observe that when $A_{i,j} \leq \frac{1}{2} - \frac{1}{2}\sum_{k=1,k\neq j}^{K-1} A_{i,k}$, $f(A_i)$ decreases monotonously, while $A_{i,j} \geq \frac{1}{2} - \frac{1}{2}\sum_{k=1,k\neq j}^{K-1} A_{i,k}$, $f(A_i)$ increases monotonously.

Based on Eq.4, we define the entropy of the i-th row in A as:

$$h(A_i) = -\sum_{j=1}^{K} A_{i,j}\log(A_{i,j}) \tag{9}$$

Furthermore, the partial derivative of $h(A_i)$ can be calculated as:

$$\begin{aligned} \frac{\partial h(A_i)}{\partial A_{i,j}} &= -\log(A_{i,j}) + \log(A_{i,K}) \\ &= \log(\frac{1 - A_{i,j} - \sum_{k=1,k\neq j}^{K-1} A_{i,k}}{A_{i,j}}) \end{aligned} \tag{10}$$

From Eq.10, we can observe that when $A_{i,j} \leq \frac{1}{2} - \frac{1}{2}\sum_{k=1,k\neq j}^{K-1} A_{i,k}$, $h(A_i)$ increases monotonously, while $A_{i,j} \geq \frac{1}{2} - \frac{1}{2}\sum_{k=1,k\neq j}^{K-1} A_{i,k}$, $h(A_i)$ decreases monotonously. In summary, we draw the conclusion that the Frobenius norm and entropy of a matrix have strict opposite monotonicity.

3. EXPERIMENTS

3.1. Dataset

We evaluate our MBF method on two MDTC benchmarks: the Amazon review dataset [18] and the FDU-MTL dataset [7]. The Amazon review dataset contains four domains: books, dvds, electronics, and kitchen. Each domain has 1,000 positive samples and 1,000 negative samples. All data have been pre-processed into a bag of features (unigrams and bigrams), losing word order information. We thus take the 5,000 most frequent features such that each review is encoded as a 5,000-dimensional vector, where feature values are raw counts of the features. The FDU-MTL dataset is a more challenging dataset, containing 16 domains: 14 product review domains (books, electronics, DVDs, kitchen, apparel, camera, health, music, toys, video, baby, magazine, software, and sport) and 2 movie review domains (IMDB and MR). The reviews in this dataset are raw text data only being tokenized by the Stanford tokenizer [19]. Each domain has a development set of 200 samples and a test set of 400 samples. The amounts of training and unlabeled data vary among domains but are roughly 1,400 and 2,000.

3.2. Baselines

We compare our MBF approach with several baselines. The multi-task deep neural network model (MT-DNN) uses a shared layer for different domains and bag-of-words input [20]. The collaborative multi-domain sentiment classification (CMSC) model uses the shared-private paradigm and can be trained by three loss functions: the least square loss (CMSC-L2), the hinge loss (CMSC-SVM), and the log loss (CMSC-Log) [21]. The adversarial multi-task learning for text classification (ASP-MTL) combines adversarial learning, shared-private paradigm, and orthogonality regularizer to train models [7]. The multinomial adversarial network (MAN) tries to encode the domain discrepancy by two f-divergences: the negative log-likelihood loss (MAN-NLL) and the least square loss (MAN-L2) [8]. The dual adversarial Co-learning (DACL) method uses discriminator-based adversarial learning and classifier-based adversarial learning to learn domain-invariant features [11]. The conditional adversarial network (CAN) aligns joint distributions of shared features and label predictions to optimize the adversarial alignment [12].

3.3. Implementation Details

We follow the standard MDTC setting and adopt the same network architectures with the most recent baselines for fair comparisons [8, 11, 12]. For the Amazon review dataset, we employ multi-layer perceptron (MLP) with an input size of 5,000 as feature extractors. Each feature extractor has two hidden layers with sizes 1,000 and 500. The output sizes of $\mathcal{F}_s$ and $\{\mathcal{F}_d^i\}_{i=1}^M$ are 128 and 64, respectively. $\mathcal{C}$ and $\mathcal{D}$ are also MLPs with one hidden layer of the same size as its input ($128 + 64$ for $\mathcal{C}$ and 128 for $\mathcal{D}$). We conduct the 5-fold cross-validation on this dataset and report the 5-fold average results. For the FDU-MTL dataset, a one-layer convolutional neural network (CNN) is used as the feature extractor. The CNN uses different kernel sizes $(3, 4, 5)$, and the number of kernels is 200. The input of the CNN is a 100-dimensional embedding obtained by processing each word of the input sequence through word2vec [22]. For all experiments, we set $\alpha = 0.5$ and $\beta = 1$ and take the convenience to cite the experimental results directly from [11, 12].

Table 1. MDTC results on the Amazon review dataset

Domain	CMSC-LS	CMSC-SVM	CMSC-Log	MAN-L2	MAN-NLL	DACL	CAN	MBF(Proposed)
Books	82.10	82.26	81.81	82.46	82.98	83.45	83.76	**84.58±0.21**
DVDs	82.40	83.48	83.73	83.98	84.03	85.50	84.68	**85.78±0.07**
Electr.	86.12	86.76	86.67	87.22	87.06	87.40	88.34	**89.04±0.11**
Kit.	87.56	88.20	88.23	88.53	88.57	90.00	90.03	**91.45±0.19**
AVG	84.55	85.18	85.11	85.55	85.66	86.59	86.70	**87.71±0.10**

Table 2. MDTC results on the FDU-MTL dataset

Domain	MT-DNN	ASP-MTL	MAN-L2	MAN-NLL	DACL	CAN	MBF(Proposed)
books	82.2	84.0	87.6	86.8	87.5	87.8	**89.1±0.1**
electronics	81.7	86.8	87.4	88.8	90.3	**91.6**	91.0±0.7
dvd	84.2	85.5	88.1	88.6	89.8	89.5	**90.4±0.2**
kitchen	80.7	86.2	89.8	89.9	91.5	90.8	**93.3±0.7**
apparel	85.0	87.0	87.6	87.6	**89.5**	87.0	88.5±0.6
camera	86.2	89.2	91.4	90.7	91.5	**93.5**	92.8±0.8
health	85.7	88.2	89.8	89.4	90.5	90.4	**92.0±0.1**
music	84.7	82.5	85.9	85.5	86.3	**86.9**	85.9±0.4
toys	87.7	88.0	90.0	90.4	91.3	90.0	**92.2±0.5**
video	85.0	84.5	89.5	89.6	88.5	88.8	**90.4±0.7**
baby	88.0	88.2	90.0	90.2	**92.0**	**92.0**	90.8±0.2
magazine	89.5	92.2	92.5	92.9	93.8	**94.5**	93.5±0.2
software	85.7	87.2	90.4	90.9	90.5	90.9	**91.4±0.4**
sports	83.2	85.7	89.0	89.0	89.3	**91.2**	90.3±0.1
IMDB	83.2	85.5	86.6	87.0	87.3	88.5	**89.9±0.1**
MR	75.5	76.7	76.1	76.7	76.0	77.1	**79.2±0.5**
AVG	84.3	86.1	88.2	88.4	89.1	89.4	**90.1±0.2**

3.4. Results

All results are reported on five random trials. The experimental results on the Amazon review dataset are shown in Table 1. We can observe that the MBF method can achieve the best average classification accuracy of 87.71%, outperforming the second-best model CAN by a margin of 1.01%. Moreover, our approach can also beat other baselines on each individual domain. For domain DVDs and kitchen, our MBF method can improve the state-of-the-art by more than 1%. The performance difference between MBF and CAN is more notable than that between CAN and DACL. Table 2 presents the results on the FDU-MTL dataset. It can be noted that the MBF provides the best performance compared with other baselines, outperforming the CAN model by a margin of 0.7%. In particular, our method can obtain the best performance on 9 out of 16 domains. For domain books, kitchen, health, IMDB, and MR domains, we can advance the state-of-the-art performance by more than 1%.

4. CONCLUSION

In this paper, we propose a maximum batch Frobenius norm (MBF) method for MDTC. We find that the discriminability of the learned features can be measured by the Frobenius norm of the batch classification output matrix, and maximizing the Frobenius norm of the batch output matrix can reduce the uncertainty of prediction on unlabeled data, boosting the feature discriminability. Therefore, we use MBF to enhance the discriminability of the learned features without degrading the transferability such that the system performance can be improved. The extensive experiments on two MDTC benchmarks show that our MBF approach effectively advances state-of-the-art performance.

5. REFERENCES

[1] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao, "Deep learning–based text classification: A comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[2] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[3] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[5] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, "Domain separation networks," *Advances in neural information processing systems*, vol. 29, pp. 343–351, 2016.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[7] Pengfei Liu, Xipeng Qiu, and Xuan-Jing Huang, "Adversarial multi-task learning for text classification," in *Proceedings of the 55th Annual Meeting of the ACL (Volume 1: Long Papers)*, 2017, pp. 1–10.

[8] Xilun Chen and Claire Cardie, "Multinomial adversarial networks for multi-domain text classification," in *Proceedings of the 2018 Conference of the NAACL-HLT, Volume 1 (Long Papers)*, 2018, pp. 1226–1240.

[9] Yuan Wu, Diana Inkpen, and Ahmed El-Roby, "Towards category and domain alignment: Category-invariant feature enhancement for adversarial domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 132–141.

[10] Charu C Aggarwal and ChengXiang Zhai, "A survey of text classification algorithms," in *Mining text data*, pp. 163–222. Springer, 2012.

[11] Yuan Wu and Yuhong Guo, "Dual adversarial co-learning for multi-domain text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 6438–6445.

[12] Yuan Wu, Diana Inkpen, and Ahmed El-Roby, "Conditional adversarial networks for multi-domain text classification," in *Proceedings of the Second Workshop on Domain Adaptation for NLP*, 2021, pp. 16–27.

[13] Yuan Wu, Diana Inkpen, and Ahmed El-Roby, "Mixup regularized adversarial networks for multi-domain text classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7733–7737.

[14] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan, "Transferable adversarial training: A general approach to adapting deep classifiers," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4013–4022.

[15] Xiaojin Zhu and Andrew B Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[16] Yves Grandvalet, Yoshua Bengio, et al., "Semi-supervised learning by entropy minimization.," in *CAP*, 2005, pp. 281–296.

[17] Théodore Papadopoulo and Manolis IA Lourakis, "Estimating the jacobian of the singular value decomposition: Theory and applications," in *ECCV*. Springer, 2000, pp. 554–570.

[18] John Blitzer, Mark Dredze, and Fernando Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th annual meeting of the ACL*, 2007, pp. 440–447.

[19] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[20] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," 2015.

[21] Fangzhao Wu and Yongfeng Huang, "Collaborative multi-domain sentiment classification," in *2015 IEEE International Conference on Data Mining*. IEEE, 2015, pp. 459–468.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.