Semi-Supervised Domain Adaptation for Emotion-Related Tasks

Mahshid Hosseini **Cornelia Caragea**
Computer Science
University of Illinois at Chicago
mhosse4@uic.edu cornelia@uic.edu

Abstract

Semi-supervised domain adaptation (SSDA) adopts a model trained from a label-rich source domain to a new but related domain with a few labels of target data. It is shown that, in an SSDA setting, a simple combination of domain adaptation (DA) with semi-supervised learning (SSL) techniques often fails to effectively utilize the target supervision and cannot address distribution shifts across different domains due to the training data bias toward the source-labeled samples. In this paper, inspired by the co-learning of multiple classifiers for the computer vision tasks, we propose to decompose the SSDA framework for emotion-related tasks into two subcomponents of unsupervised domain adaptation (UDA) from the source to the target domain and semi-supervised learning (SSL) in the target domain where the two models iteratively teach each other by interchanging their high confident predictions. We further propose a novel data cartography-based regularization technique for pseudo-label denoising that employs training dynamics to further hone our models' performance. We release our code.[1]

1 Introduction

Large pre-trained language models (Devlin et al., 2019; Liu et al., 2019) have significantly improved many natural language processing (NLP) task performances with the help of large quantities of labeled training data and have become the de facto model for NLP applications. However, obtaining vast troves of annotated data for training in many real-world scenarios is costly and challenging. For example, reliable large annotated emotion or empathy data might not exist in a computer-assisted therapy session (Hosseini and Caragea, 2021a,b, 2023). On top of that, it is shown that a shift in data distribution can substantially affect the performance of such text classification models (Ngo et al., 2022; Blitzer et al., 2006); a model trained on a source domain does not perform well on a dataset from another domain. This deficiency is due to the domain shift across the datasets (Tzeng et al., 2017), which is a problem that is commonly encountered in NLP. To relieve the unsupervised domain adaptation (UDA) bottleneck for textual tasks, recent works attempt to align distributions between source and target domains by extracting the domain-invariant representations (Ganin et al., 2016). However, despite the recent progress, the UDA methods are still impractical as they may yield different new domain-sensitive particularities for large-scale language models.

Recent works showed that the presence of few labeled samples from the target domain in a semi-supervised domain adaption (SSDA) setup can positively impact and significantly boost the performance of the neural models (Qin et al., 2020; Kim and Kim, 2020; Saito et al., 2019). There exists prior work on supervised domain adaptation (Daumé III, 2007) and multi-task learning (Sun et al., 2011) for natural language processing (NLP) tasks. However, despite the importance of semi-supervised domain adaption, in NLP, only a few studies have focused on this problem (Daumé III et al., 2010; Cheng and Pan, 2014). Daumé III et al. (2010) expanded an existing fully supervised domain adaptation technique (Daumé III, 2007) to semi-supervised domain adaptation settings using co-regularization, which originated in the context of multi-view learning (Rosenberg and Bartlett, 2007; Sindhwani and Rosenberg, 2008). Cheng and Pan (2014) also framed the semi-supervised domain adaptation problem as learning with a transformation function and a prediction model under manifold constraints.

Given a large number of labeled data provided in the source domain and only a few target labeled data with the inherent distributional difference, it is shown that in an SSDA setting, a single classifier may likely be dominated by the source domain

[1] https://github.com/Mahhos/CotrainingTrainingDynamics

Findings of the Association for Computational Linguistics: ACL 2023, pages 5402–5410
July 9-14, 2023

(Yang et al., 2021). In this paper, we extend the co-training strategy (Blum and Mitchell, 1998), a semi-supervised learning approach for multi-view data, to a single-view setting for NLP tasks to effectively use unlabeled target data. Co-training trains two classifiers from each view and employs the most confident predictions of the unlabeled data for the two classifiers to teach each other. Inspired by co-training, we propose to decompose the SSDA framework and learn two distinct classifiers to teach each other so that both classifiers can excel in the target domain. Particularly, we employ an unsupervised domain adaptation setup where we leverage the labeled source data and the unlabeled target data to learn one classifier. Furthermore, we employ a semi-supervised learning setup where we learn another classifier using the labeled target data together with the unlabeled target data. We further propose a novel data cartography-based regularization technique for pseudo-label denoising that employs training dynamics (Swayamdipta et al., 2020) to further hone our models' performance.

Our preliminary results on three emotion-related NLP tasks show that transferring knowledge between the two classifiers and incorporating training dynamics to help denoising the generated pseudo-labels can effectively improve the performance on the target domain.

2 Approach

2.1 Co-training with Task Decomposition

Co-training (Blum and Mitchell, 1998) is a well-known semi-supervised learning (SSL) technique that employs two different views of an example (e.g., audio and video) and learns two predictive models that are trained separately on each view. Assuming that each view is sufficient for correct classification, co-training confers the models the capability to teach each other by adding the highly confident predictions of one model (on new unlabeled examples) to the training set of the other. In this way, co-training helps boost a learning algorithm's performance when only a small set of labeled examples is available. Recently, Chen et al. (2011) and Qiao et al. (2018) proposed techniques to perform co-training using single-view data, but they still need additional tasks or objective functions to apply co-training. Along these lines, Yang et al. (2021) proposed to use only single-view data (i.e., only images) to perform co-training in a semi-supervised domain adaptation setup for computer vision tasks, where they leverage the supervision of the labeled data from the source and target domains and combine them with the unlabeled samples from the target domain.

Here we propose our co-training with task decomposition and cartography-based mixup approach, which greatly benefits the text-based emotion-related classification tasks in the few-shot setting. Task decomposition in a co-training paradigm has been initially introduced for computer vision tasks by Yang et al. (2021) to enhance the performance of image classification models. With that, our work is greatly inspired by Yang et al. (2021).

Task Setup. Given the labeled data from the source $\mathcal{D}^S = \{(s_i, y_i)\}_{i=1}^{N_S}$ and the target $\mathcal{D}^T = \{(t_i, y_i)\}_{i=1}^{N_T}$ domains such that $|\mathcal{D}^S| \gg |\mathcal{D}^T|$ and $|\mathcal{D}^T| = \mathbb{K} \times |y|$[2] in the few-shot setting, and the unlabeled data from the target domain $\mathcal{D}^U = \{(u_i)\}_{i=1}^{N_U}$, we first construct two sub-tasks in the semi-supervised domain adaptation setup: one using our $\mathcal{D}^S$ and $\mathcal{D}^U$ to train an *unsupervised domain adaptation* (**UDA**) model θ^{uda}, and one using our $\mathcal{D}^T$ and $\mathcal{D}^U$ to train a *semi-supervised learning* (**SSL**) model θ^{ssl}. We conduct our tasks using mini-batch SGD to update the models' weights in our experiments. In each iteration, we make predictions on $\mathcal{U} = \{u_b\}_{b=1}^{\mathcal{B}}$ (which is sampled from our unlabeled set $\mathcal{D}^U$ with mini-batch size $\mathcal{B}$) using our two models θ^{uda} and θ^{ssl}, and generate pseudo-label sets $\mathcal{U}^{ssl}$ and $\mathcal{U}^{uda}$ that will be filtered based on a threshold τ to update θ^{ssl} and θ^{uda}:

$$\begin{aligned}\mathcal{U}^{ssl} &= \{(u_b, y'_b = \underset{c}{argmax}\, p(c|u_b; \theta^{uda}));\\ &\qquad \text{if } \underset{c}{max}\, p(c|u_b; \theta^{uda})) > \tau\}\\ \mathcal{U}^{uda} &= \{(u_b, y'_b = \underset{c}{argmax}\, p(c|u_b; \theta^{ssl}));\\ &\qquad \text{if } \underset{c}{max}\, p(c|u_b; \theta^{ssl})) > \tau\}\end{aligned}$$

where $p(c|u_b; .)$ is the predicted probability of the unlabeled sample $u_b \in \mathcal{U}$ for a class c. In essence, if the prediction confidence of one model (e.g., θ^{ssl}) on u_b is greater than our pseudo-label selection threshold τ,[3] it would be added to the $\mathcal{U}^{uda}$ to train the other model θ^{uda}. In other words, a model imparts the other model with confident pseudo-labels to learn from and uses the other model's confident pseudo-labels to learn from.

[2] $|y|$ represents the number of classes in y.

[3] We experiment with a range (i.e., $[0.5, 0.6, 0.7]$) for the pseudo-label selection threshold and select $\tau = 0.7$ as our final confidence threshold.

2.2 Data Cartography-based Regularization for Pseudo-label Denoising

It is inevitable to obtain noisy pseudo-labels from each model. First due to the domain shift (in our UDA component), and second given that only a few labeled samples (i.e., $\mathbb{K} = [4, 8, 16]$ per class) are available from the target domain (in our SSL component), which in turn impacts the supervision process. To mitigate noise in the generated pseudo-labels and hone our models' performance, we propose a cartography-based mixup strategy that helps to effectively denoise an incorrect pseudo-label.

Mixup Training. Mixup augments the training data by linearly interpolating training samples and their corresponding labels, based on a simple rule proposed by Zhang et al. (2018):

$$(\widetilde{x}_{ij}, \widetilde{y}_{ij}) := (\lambda x_i + (1-\lambda)x_j, \lambda y_i + (1-\lambda)y_j) \quad (1)$$

where λ is a mixing ratio sampled from a Beta(α, α) distribution with a hyper-parameter α and (x_i, y_i) and (x_j, y_j) are two input examples that are randomly drawn from the training set.

Proposed Approach. Few-shot learning methods generally assume that the training sets *always* include accurately labeled samples. However, this assumption can sometimes be unrealistic. No matter how small, training sets can still contain mislabeled samples (Liang et al., 2022). In other words, it could not be guaranteed that the few-shot training sets were carefully selected to represent their class. In fact, even carefully annotated and selected datasets often hold mislabeled samples (Northcutt et al., 2021; Yang et al., 2020) due to several reasons like ambiguity, automated weakly supervised annotation, or human error. Here, we propose to use a novel Mixup data augmentation technique on the target training data and the pseudo-labels generated by the source domain that is informed by training dynamics to further surpass the noisy data bottleneck and improve the target domain performance in few-shot setting. Our proposed mixup creates vicinal distribution steered by the data maps (Swayamdipta et al., 2020) as described below.

We first characterize each training sample of our few-shot target domain training set $\mathcal{D}^T$ into three groups of easy-to-learn, ambiguous, and hard-to-learn, based on how they contribute to the model learning (i.e., training dynamics). We then sample examples with specific characteristics (emanated from the previous step) to interpolate with the generated pseudo-labels by the source domain $\mathcal{U}^{ssl}$ during our cartography-based mixup process. In our experiments, we measure the statistics using a `RoBERTa-base` model. Sample (x_i, y_i) training dynamics are measured as statistics called confidence and variability computed across the E epochs (Swayamdipta et al., 2020). Confidence is computed as the mean model probability of the true label y_i across epochs, $\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta_e}(y_i|x_i)$; where θ indicates the model parameters and $p_{\theta e}$ denotes the model's probability at the end of the e_{th} epoch. Variability is measured as the standard deviation of the ground-truth probabilities $p_{\theta_e}(y_i|x_i)$ across different epochs, $\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}(p_{\theta_e}(y_i|x_i)-\hat{\mu}_i)}{E}}$.

Intuitively, samples to which the model confidently (i.e., high confidence) and constantly (i.e., low variability) assigns the true, and the same label corresponds to easy-to-learn examples (for the model). On the other hand, samples with low confidence and low variability resemble hard-to-learn examples (for the model), which usually are referred to as mislabeled samples, and examples with high variability that the model is uncertain about during training are ambiguous (to the model). Using these statistics, we select the easy-to-learn samples (i.e., samples that the model *consistently* predicts *correctly* across epochs) to interpolate with the pseudo-labels generated by the source domain. By employing such particularities, our goal is to ensure that we effectively denoise an incorrect pseudo-label by mixing it with the most informative data samples (Swayamdipta et al., 2020), samples with high confidence and low variability which are detected to be *actually* correct. Our mixup approach combines samples at the level of the hidden state representations generated by the task-specific layer on top of the pre-trained language model.

3 Experiments

3.1 Datasets

We perform evaluations on three text classification tasks of emotion detection, sentiment analysis, and empathy detection. We analyze tasks with challenging domain shifts where out-of-domain performance is considerably lower. Furthermore, it is shown that detecting empathy or emotions from text without visual or acoustic information is challenging due to the subjective nature of the annotations (Hosseini and Caragea, 2021b), which makes it difficult to accurately label and interpret the emotions or empathy expressed. Additionally,

Model	News to Health			Reddit to TV Series			Yelp to IMDB		
	K = 4	K = 8	K = 16	K = 4	K = 8	K = 16	K = 4	K = 8	K = 16
Source-Only **		66.63**			24.51**			52.12**	
Target-Only	42.39	44.66	46.00	16.69	18.09	19.96	48.98	50.69	51.55
UDA **		45.26**			21.15**			51.69**	
SSL	44.59	47.85	49.22	18.90	21.08	21.75	50.16	53.24	54.50
SSL + MixText (Chen et al., 2020)	46.67	50.19	53.36	19.82	22.38	23.16	51.10	55.38	55.78
SSL + FliText (Liu et al., 2021)	44.61	46.37	50.23	19.16	19.87	20.67	50.26	55.10	54.92
Unsupervised Data Augmentation (Xie et al., 2020)	47.05	50.10	55.64	20.92	23.75	23.67	52.19	56.07	55.89
Co-training with TD	64.85	67.02	71.00	22.10	24.45	25.87	52.26	57.47	59.53
Co-training with TD + Mixup (Yang et al., 2021)	67.65	70.56	73.63	21.46	22.71	24.53	53.06	57.76	58.10
Co-training with TD + Ours	69.33	72.66	77.33	25.83	27.54	28.85	55.67	61.44	63.66
Supervised Learning-Source **		68.27**			68.55**			95.78**	
Supervised Learning-Target (full)**		84.33**			63.02**			92.12**	

Table 1: Accuracy on empathy (i.e., News to Health), emotion (i.e., Reddit to TV Series) and sentiment (i.e., Yelp to IMDB) (%) for $\mathbb{K} = [4, 8, 16]$ few-shot samples per class, using BERT; TD refers to task decomposition; UDA and SSL refer to the unsupervised data augmentation, and semi-supervised learning, respectively; $**$ means that the result is the same for all our three settings.

most datasets, particularly in empathy detection, are limited in size, with only a few exceptions. However, given the significance of these tasks and the profound impact emotions have on our behavior and daily lives, our objective is to enhance the performance of such tasks and achieve improved detection of emotion-related information from text. We explain our source and target domains datasets below.

Empathy Detection. NewsEmp is a dataset of empathic reactions to news stories, including empathy binary labels released by Buechel et al. (2018) which we use as our source domain dataset. TwitTEmp Dataset (Hosseini and Caragea, 2021a) contains perceived empathy annotated by empathy direction (seeking vs. providing) in the health domain, which we use as our target domain dataset.

Emotion Detection. GoEmotions is an emotion detection dataset from Reddit comments where we use the six basic emotions (joy, anger, fear, sadness, disgust, and surprise) and neutral as our source domain dataset. Meld (Poria et al., 2019) contains dialogues from the popular Friends TV series annotated with the same set of emotion labels, which we use as the target domain dataset.

Sentiment Analysis. Yelp (Zhang et al., 2015) is a dataset for binary sentiment classification, consisting of reviews from Yelp, which is used as our source domain. Our target domain dataset is IMDB movie reviews (Maas et al., 2011), containing sentences of movie reviews and their sentiment.

3.2 Baseline Methods

The details of the experiments are as follows. We use the BERT-base model and $\mathbb{K} = [4, 8, 16]$ in all the experiments.[4] We contrast our proposed approach on emotion, empathy, and sentiment classification tasks with the following baselines: (1) Source-Only, which uses the source domain for fine-tuning BERT (the training portion) and the target domain for the evaluation (the test portion), which is the same for all our three settings; (2) Target-Only, uses the target domain for both training and evaluation of BERT with few-shot data; (3) UDA unsupervised domain adaptation where source domain training set is used to train a model and make predictions on unlabeled data from the target domain. Then, the generated pseudo-labels are added to the source domain training set iteratively based on the selection threshold; (4) SSL semi-supervised learning, where the target domain training set is used to train a model and make predictions on unlabeled data from the target domain. Then, the generated pseudo-labels are added to the target domain training set iteratively based on the selection threshold; (5) MixText (Chen et al., 2020) which guesses low-entropy labels for unlabeled target data and uses Mixup to interpolate labeled and unlabeled samples; (6) FliText leverages convolution networks to achieve faster and lighter semi-supervised text classification; (7) Unsupervised data augmentation enhances training by augmenting unlabeled data and promoting consistency between augmented versions; (8) Co-training with task decomposition where the SSDA is decomposed to two components of UDA and SSL; (9) Co-training with task decomposition and mixup (Yang et al., 2021) where the generated pseudo-labels with the source domain classifier are interpolated

[4] BERT-base yields the best results in our experiments, so we only report the results using this model.

Domain	Sample	Label
Health	Yes. I lost my first wife to cancer at 31 and was wrecked with guilt that I didn't do enough to help her. After a while, I finally realized that we all have a time and when it's up, no one or nothing can change that.	not empathetic
TV Series	Will you marry me?	Fear
IMDB	This solid little horror film is actually one of Renny Harlin's best. The story is pretty routine stuff, but the atmosphere is what really makes it come alive; in fact, the ghost story is almost an afterthought. The real horror comes from the prison setting itself, and Renny H. spares no detail in showing us how bad the conditions are inside that crumbling, leaking, rat-infested old hellhole (with a sadistic warden, too!) Viggo Mortensen is excellent as usual in the lead role, supported by some very authentic-looking prisoners (there are no pretty boys in this cast.) Horror fans should check this one out.	Negative

Table 2: Samples with label errors.

with the target training data; and (10) Standard supervised learning with the full training and test sets of the domains separately (as an upper bound of performance).

3.3 Results

Table 1 compares our proposed approach and baseline methods on our classification tasks for different few-shot settings. We report the average performance on 3 distinct randomly sampled training and development splits with three random seeds to provide a robust measure of our few-shot performance. We make a few remarks below.

As we can see from Table 1, our proposed method achieves higher accuracy on all the few-shot settings than any baseline using task decomposition and cartography-based mixup. The results suggest that incorporating cartography-based mixup to effectively denoise the generated pseudo-labels (see Ours in the tables) results in constant improvement over all the few-shot settings. For example, on empathy (i.e., News to Health) with $K = 4$, our proposed approach increased the performance by a factor of 1.63% compared to the Target-Only and by 1.55% compared to the SSL. Interestingly, we also observe that our proposed approach outperforms standard mixup (i.e., Co-training with TD + Mixup), which signifies the importance and effectiveness of our proposed strategy in using training dynamics to characterize data and identify the correctly-labeled samples (easy-to-learn samples) for denoising the generated pseudo-labels through the mixup process. In the standard mixup, the interpolation process occurs between the generated pseudo-labels and all examples in the target labeled set, where it is possible for some of these examples to have incorrect labels. Table 2 shows examples with erroneous labels from all of our few-shot target labeled sets. Erroneous labels can occur due to human errors in annotations, even with small datasets when using crowdsourcing techniques or relying on human annotators. It is apparent from the table that the standard supervised learning using the full training and test sets from the respective domains results in an increase in performance.

4 Conclusion

In this work, we extend the co-training strategy as a semi-supervised learning approach for multi-view data to a single-view setting for NLP tasks and propose to decompose the SSDA framework and learn two distinct classifiers (one in an semi-supervised setup and another one in a domain adaptation setup) to teach each other so that both classifiers can excel in the target domain. We further propose a novel data cartography-based regularization technique for pseudo-label denoising that employs training dynamics to further hone our models' performance. Our preliminary results show that denoising the pseudo-labels of unlabeled target data using high-quality labeled target data within a co-training framework yields improvements in performance over multiple baselines.

5 Limitations

One potential limitation of our method is that it induces an extra cost of estimating training dynamic statistics of the data samples to characterize them (e.g., easy-to-learn or ambiguous) based on how they incorporate into the model's learning. This may be more expensive for tasks and datasets with a large number of classes. In the future, we will focus on approaches to characterize the training examples on the fly.

Acknowledgments

We acknowledge NSF for support from grants IIS-1912887, IIS-2107487, and ITE-2137846. We also thank our reviewers for their insightful feedback and comments.

References

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.

Avrim Blum and Tom M. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, USA, July 24-26, 1998*, pages 92–100. ACM.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle H. Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *CoRR*, abs/1808.10399.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2147–2157. Association for Computational Linguistics.

Minmin Chen, Kilian Q. Weinberger, and Yixin Chen. 2011. Automatic feature decomposition for single view co-training. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 953–960. Omnipress.

Li Cheng and Sinno Jialin Pan. 2014. Semi-supervised domain adaptation on manifolds. *IEEE Trans. Neural Networks Learn. Syst.*, 25(12):2240–2249.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Co-regularization based semi-supervised domain adaptation. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 478–486. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35.

Mahshid Hosseini and Cornelia Caragea. 2021a. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3713–3724. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2021b. It takes two to empathize: One to seek and one to provide. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13018–13026. AAAI Press.

Mahshid Hosseini and Cornelia Caragea. 2023. Feature normalization and cartography-based demonstrations for prompt-based fine-tuning on emotion-related tasks. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington DC, February 7-14*. AAAI Press.

Taekyung Kim and Changick Kim. 2020. Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation. In *European conference on computer vision*, pages 591–607. Springer.

Kevin J Liang, Samrudhdhi B Rangrej, Vladan Petrovic, and Tal Hassner. 2022. Few-shot learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9089–9098.

Chen Liu, Mengchao Zhang, Zhibing Fu, Panpan Hou, and Yu Li. 2021. Flitext: A faster and lighter semi-supervised text classification with convolution networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2481–2491. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. Unsupervised domain adaptation for text classification via meta self-paced learning. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 4741–4752. International Committee on Computational Linguistics.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics.

Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan L. Yuille. 2018. Deep co-training for semi-supervised image recognition. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, volume 11219 of *Lecture Notes in Computer Science*, pages 142–159. Springer.

Can Qin, Lichen Wang, Qianqian Ma, Yu Yin, Huan Wang, and Yun Fu. 2020. Opposite structure learning for semi-supervised domain adaptation. *CoRR*, abs/2002.02545.

David S. Rosenberg and Peter L. Bartlett. 2007. The rademacher complexity of co-regularized kernel classes. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, volume 2 of *JMLR Proceedings*, pages 396–403. JMLR.org.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. 2019. Semi-supervised domain adaptation via minimax entropy. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8049–8057. IEEE.

Vikas Sindhwani and David S. Rosenberg. 2008. An RKHS for multi-view learning and manifold co-regularization. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 976–983. ACM.

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 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 505–513.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of EMNLP 2020, Online, November 16-20, 2020*, pages 9275–9293. Association for Computational Linguistics.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2962–2971. IEEE Computer Society.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Luyu Yang, Yan Wang, Mingfei Gao, Abhinav Shrivastava, Kilian Q. Weinberger, Wei-Lun Chao, and Ser-Nam Lim. 2021. Deep co-training with task decomposition for semi-supervised domain adaptation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8886–8896. IEEE.

Yuewei Yang, Kevin J. Liang, and Lawrence Carin. 2020. Object detection as a positive-unlabeled problem. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
5

☒ A2. Did you discuss any potential risks of your work?
Our work does not have any potential risks

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1 for introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3.3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3.2

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.