Weakly-supervised Text Classification with Wasserstein Barycenters Regularization

Jihong Ouyang[1,2], **Yiming Wang**[1,2], **Ximing Li**[1,2*], **Changchun Li**[1,2]
[1]College of Computer Science and Technology, Jilin University, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
ouyj@jlu.edu.cn, {yimingw17, liximing86, changchunli93}@gmail.com

Abstract

Weakly-supervised text classification aims to train predictive models with unlabeled texts and a few representative words of classes, referred to as *category words*, rather than labeled texts. These weak supervisions are much more cheaper and easy to collect in real-world scenarios. To resolve this task, we propose a novel deep classification model, namely Weakly-supervised Text Classification with Wasserstein Barycenter Regularization (WTC-WBR). Specifically, we initialize the pseudo-labels of texts by using the category word occurrences, and formulate a weakly self-training framework to iteratively update the weakly-supervised targets by combining the pseudo-labels with the sharpened predictions. Most importantly, we suggest a Wasserstein barycenter regularization with the weakly-supervised targets on the deep feature space. The intuition is that the texts tend to be close to the corresponding Wasserstein barycenter indicated by weakly-supervised targets. Another benefit is that the regularization can capture the geometric information of deep feature space to boost the discriminative power of deep features. Experimental results demonstrate that WTC-WBR outperforms the existing weakly-supervised baselines, and achieves comparable performance to semi-supervised and supervised baselines.

1 Introduction

Text classification is a significant and fundamental task in natural language processing, with many real-world applications, *e.g.,* document tagging, sentiment analysis, and question answering, to name a few. Basically, the task aims to train predictive models with a collection of manually labeled texts, enabling to automatically infer the labels of unseen texts. During the past decades, it has been well investigated by the community, suggesting a number of conventional text classifiers and the emerging deep classification models [Li *et al.*, 2021].

The existing text classifiers, especially the deep classification models, have achieved great success with promising performance [Lan *et al.*, 2020]. However, to guarantee the effectiveness, they often require abundant training texts with accurate labels, which are expensive and difficult to collect. Due to the manual annotation burden, only the training texts with cheaper weak supervision are available in many real-world scenarios. Such cheaper supervision tends to be inaccurate, incomplete, and ambiguous, potentially resulting in performance degradation even by a large margin. Naturally, how to learn strong text classifiers with weak supervision now becomes an urgent demand, and the community has paid more attention to the weakly-supervised methods [Zhou, 2018].

The weakly-supervised scenario we now concern is even more challenging, where we are only given by unlabeled texts with the sets of **category words** for classes as the only available supervised signals [Li *et al.*, 2016; Li and Yang, 2018]. To be specific, the category words are defined as a few representative words of classes, *e.g.,* label names, label descriptions, and hot words, supporting the basis knowledge of classes. For example, in terms of news articles, the label names such as "sports", "politics", and "bussiness" definitely express the corresponding classes [Meng *et al.*, 2020].

Contrary to training texts with accurate labels, the category words are much cheaper to be collected by human annotators [Druck *et al.*, 2008]. Unfortunately, they provide very weak and limited supervision. For example, on the dataset AG News with category words of label names, only about 3.4% texts contain category words. To make matters worse, only 2.1% texts contain the category words from the relevant labels. Training with such weak supervision is intractable. To handle this weakly-supervised task, several methods have been developed by simultaneously expanding and taking full advantage of the limited supervised signals within the category words. For example, the popular existing techniques include: propagating supervised signals among texts with manifold regularization [Li *et al.*, 2018], generating pseudo-texts with category words [Meng *et al.*, 2018], and applying the language model to expand category words [Meng *et al.*, 2020].

In this paper, our motivation is to solve for the aforementioned weakly-supervised task by regularizing the supervised signals with Wasserstein barycenter. Accordingly, we propose a novel deep classification model, namely **W**eakly-supervised **T**ext **C**lassification with **W**asserstein **B**arycenter **R**egularization (**WTC-WBR**). By referring to [Li *et al.*, 2016;

*Corresponding Author

Li *et al.*, 2018], we initialize the pseudo-labels of texts by using the category word occurrences. Because the pseudo-labels may be inaccurate and a number of texts may even contain no category words, we formulate a weakly self-training framework, *i.e.,* iteratively updating the weakly-supervised targets by combining the pseudo-labels with the sharpened predictions. Most importantly, we suggest a Wasserstein barycenter regularization with weakly-supervised targets on the deep feature space. The intuition is that the texts tend to be close to the corresponding Wasserstein barycenter indicated by weakly-supervised targets. Another benefit is that the regularization can capture the geometric information of deep feature space to enhance the discriminative power of deep features. The experiments have been conducted on several prevalent text datasets. Empirical results show that WTC-WBR performs better than existing weakly-supervised baselines and achieves competitive performance comparing with even semi-supervised and supervised baselines.

In summary, the major contributions of this paper are as follows:

- We develop a novel deep classification model named WTC-WBR, which trains the text classifier over unlabeled texts with category words.
- We formulate a weakly self-training framework, and suggest a Wasserstein barycenter regularization to boost the supervision updating.
- Empirical results show that WTC-WBR outperforms existing weakly-supervised baselines, and is even on a par with semi-supervised and supervised baselines.

2 Preliminary

We now briefly introduce the preliminaries of Wasserstein distance [Bogachev and Kolesnikov, 2012] and Wasserstein barycenter [Agueh and Carlier, 2011].

2.1 Wasserstein Distance in Discrete Space

Formally, the Wasserstein distance measures the distances between probability distributions from the perspective of geometry.

Definition 1. Consider a discrete state space $\mathbf{\Omega} = \{\mathbf{w}_1, \ldots, \mathbf{w}_V\}$, where $\mathbf{w}_i$ represents the feature vector of each state i. The p-order Wasserstein distance between two discrete probability distributions $\boldsymbol{\upsilon}_1$ and $\boldsymbol{\upsilon}_2$ on $\mathbf{\Omega}$ is defined as follows:

$$W_p^p(\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2; \mathbf{M}) = \min_{\mathbf{T} \in \mathcal{T}(\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2)} \langle \mathbf{T}, \mathbf{M} \rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius dot-product; $\mathbf{M} \in \mathbb{R}_+^{V \times V}$ is the cost matrix of each state pair, *i.e.,* each element $\mathbf{M}_{ij} = d_\Omega^p(\mathbf{w}_i, \mathbf{w}_j)$; and $\mathcal{T}(\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2)$ is the set of joint distributions of $\boldsymbol{\upsilon}_1$ and $\boldsymbol{\upsilon}_2$.

With an auxiliary entropy regularization, a more efficient regularized version of Wasserstein distance [Cuturi, 2013] is defined below:

$$W_{p,\gamma}^p(\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2; \mathbf{M}) = \min_{\mathbf{T} \in \mathcal{T}(\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2)} \Big\{ \langle \mathbf{T}, \mathbf{M} \rangle - \frac{1}{\gamma} \mathcal{H}(\mathbf{T}) \Big\}, \tag{2}$$

where $\mathcal{H}(\mathbf{T}) = -\langle \mathbf{T}, \ln(\mathbf{T}) \rangle$ and γ is the scaling parameter. Here, we concentrate on the case of $p = 2$, and to make the notations simple, we denote by $W(\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2; \mathbf{M})$ the regularized Wasserstein distance of Eq.2.

By applying the Sinkhorn method [Cuturi, 2013], the optimum $\mathbf{T}^*$ and gradients of $\boldsymbol{\upsilon}_1$ and $\boldsymbol{\upsilon}_2$ can be calculated below:

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leftarrow (\boldsymbol{\upsilon}_1 \oslash \boldsymbol{\kappa}\boldsymbol{\beta}, \boldsymbol{\upsilon}_2 \oslash \boldsymbol{\kappa}^\top \boldsymbol{\alpha}), \tag{3}$$

$$\mathbf{T}^* = \mathrm{diag}(\boldsymbol{\alpha}) \boldsymbol{\kappa} \mathrm{diag}(\boldsymbol{\beta}), \tag{4}$$

$$\frac{\partial W(\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2; \mathbf{M})}{\partial \boldsymbol{\upsilon}_1} = -\frac{\ln(\boldsymbol{\alpha})}{\gamma} + \frac{\ln(\boldsymbol{\alpha})^\top \mathbf{1}}{\gamma V} \mathbf{1},$$

$$\frac{\partial W(\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2; \mathbf{M})}{\partial \boldsymbol{\upsilon}_2} = -\frac{\ln(\boldsymbol{\beta})}{\gamma} + \frac{\ln(\boldsymbol{\beta})^\top \mathbf{1}}{\gamma V} \mathbf{1}, \tag{5}$$

where $\boldsymbol{\kappa} = \exp(-\gamma \mathbf{M})$; $\mathbf{1}$ is the all-one vector; and $\oslash$ denotes the element-wise division.

2.2 Wasserstein Barycenter

The Wasserstein barycenter is a minimizer of a weighted average of squared Wasserstein distances [Agueh and Carlier, 2011], providing an efficient notion for constructing geometric prototypes in the Wasserstein space.

Definition 2. The Wasserstein barycenter of S probability distributions $\mathbf{\Upsilon} = \{\boldsymbol{\upsilon}_1, \ldots, \boldsymbol{\upsilon}_S\}$ with barycentric weights $\mathbf{\Lambda} = \{\lambda_1, \ldots, \lambda_S\}$ is defined below:

$$\boldsymbol{\mu}^* \coloneqq \min_{\boldsymbol{\mu}} \sum_{s=1}^{S} \lambda_s W(\boldsymbol{\upsilon}_s, \boldsymbol{\mu}). \tag{6}$$

Generally speaking, the Wasserstein barycenter can be regarded as a special case of unbalanced optimal transport problem [Chizat *et al.*, 2018], solved with an efficient iterative method derived from the Sinkhorn method. At each iteration m, it is updated by the following equations:

$$a_s^{(m)} = \frac{\boldsymbol{\upsilon}_s}{\boldsymbol{\kappa} b_s^{(m-1)}}, \qquad s = 1, \cdots, S \tag{7}$$

$$\boldsymbol{\mu}^{(m)} = \prod_{s=1}^{S} \Big(\boldsymbol{\kappa}^\top a_s^{(m)} \Big)^{\lambda_s}, \tag{8}$$

$$b_s^{(m)} = \frac{\boldsymbol{\mu}^{(m)}}{\boldsymbol{\kappa}^\top a_s^{(m)}}, \qquad s = 1, \cdots, S \tag{9}$$

The method continues this update loop until the termination condition is reached, and finally outputs the solution of Wasserstein barycenter.

3 The Proposed WTC-WBR Approach

For clarity, we first formulate the weakly-supervised scenario concerned in this paper. The training dataset we faced is composed of a set of unlabeled raw texts $\mathcal{U} = \{\mathbf{x}_d\}_{d=1}^D$ and sets of category words for K classes $\mathcal{C} = \{\mathcal{C}_k\}_{k=1}^K$. For each class k, it is associated with a small set of category words $\mathcal{C}_k$, which is treated as the only available supervision. Because the set $\mathcal{C}_k$ of each class has very few category words, the supervision information must be very weak and limited. Our goal is to induce the predictive model over the training dataset $\{\mathcal{U}, \mathcal{C}\}$,

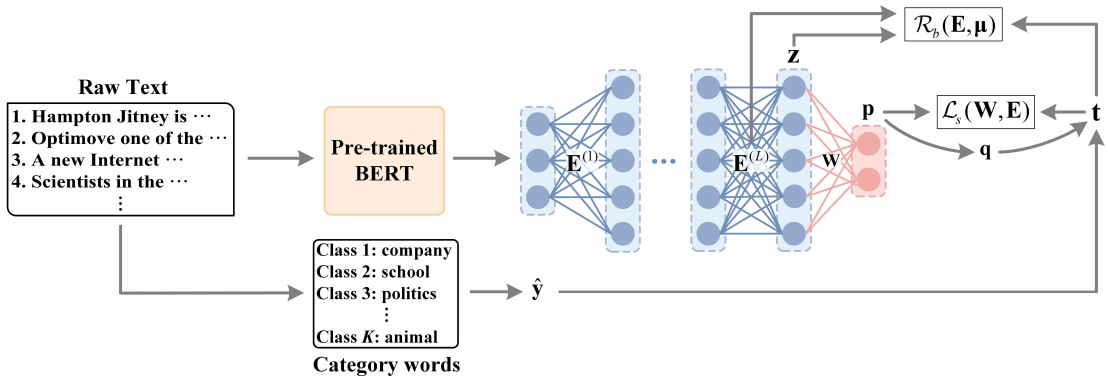

Figure 1: The overall framework of WTC-WBR.

where the model can infer the most relevant labels for unseen texts.

To handle this task, we propose a novel deep classification model named **WTC-WBR**, whose overall framework is illustrated in Fig.1. Formally, the objective of WTC-WBR is composed of two parts, **weakly self-training loss with category words** and **Wasserstein barycenter regularization**. In the following, we introduce them in detail.

3.1 Weakly Self-Training Loss with Category Words

Naturally, one intuitive idea of learning with category words is to form pseudo-labels for unlabeled texts with the category word occurrence information before training text classifiers [Li *et al.*, 2016; Li *et al.*, 2018]. For each unlabeled text d, we can compute its pseudo-label vector $\widehat{\mathbf{y}}_d$ by the following formula:

$$\widehat{\mathbf{y}}_{dk} = \frac{\mathbf{s}_{dk}}{\sum_{k'} \mathbf{s}_{dk'} + \epsilon}, \qquad k = 1, \cdots, K, \tag{10}$$

where $\mathbf{s}_{dk}$ denotes the total number of category words of class k appearing in text d; and ϵ is a smoothing parameter in case of zero-division. Accordingly, we can then train the text classifier with the pseudo-training dataset $\{(\mathbf{x}_d, \widehat{\mathbf{y}}_d)\}_{d=i}^{D}$ by minimizing the following objective:

$$\mathcal{L}(\mathbf{W}, \mathbf{E}, \mathbf{B}) = \frac{1}{D} \sum_{d=1}^{D} \ell\left(f_{\mathbf{W}}\left(f_{\mathbf{E}}\left(f_{\mathbf{B}}\left(\mathbf{x}_d\right)\right)\right), \widehat{\mathbf{y}}_d\right). \tag{11}$$

where $\ell(,\cdot,)$ is the loss function; $f_{\mathbf{B}}(\cdot)$ is the pre-trained BERT model with parameters $\mathbf{B}$; $f_{\mathbf{E}}(\cdot)$ is a L-layer feature encoder parameterized by $\mathbf{E} = \{\mathbf{E}^{(l)}\}_{l=1}^{L}$; and $f_{\mathbf{W}}(\cdot)$ is the classification layer parameterized by $\mathbf{W}$. To make the notations simple, we denote by $\mathbf{z}_{1:D}$ and $\mathbf{p}_{1:D}$ the deep features and predictions, respectively:

$$\mathbf{z}_d = f_{\mathbf{E}}(f_{\mathbf{B}}(\mathbf{x}_d)), \quad \mathbf{p}_d = f_{\mathbf{W}}(\mathbf{z}_d), \qquad d = 1, \cdots, D, \tag{12}$$

Unfortunately, because the category words are scarce, the pseudo-labels tend to be inaccurate, where each text may contain the category words of its irrelevant classes. To make matters worse, many texts may contain no category words, resulting in the useless all-zero pseudo-label vectors [Meng *et al.*, 2020]. To resolve the problems, we propose to train the model in a weakly self-training manner, which can simultaneously refine the inaccurate and all-zero supervision. Specifically, for each text d, we form a weakly-supervised target vector $\mathbf{t}_d$ by combining the pseudo-labels $\widehat{\mathbf{y}}_d$ and sharpened predictions $\mathbf{q}_d$, formulated below:

$$\mathbf{t}_{dk} = \frac{\rho\widehat{\mathbf{y}}_{dk} + (1-\rho)\mathbf{q}_{dk}}{\sum_{k'} \rho\widehat{\mathbf{y}}_{dk'} + (1-\rho)\mathbf{q}_{dk'}},$$

$$\mathbf{q}_{dk} = \frac{\mathbf{g}_{dk}}{\sum_{k'} \mathbf{g}_{dk'}}, \quad \mathbf{g}_{dk} = \frac{\mathbf{p}_{dk}^2}{\sum_{d'} \mathbf{p}_{d'k}}, \qquad k = 1, \cdots, K, \tag{13}$$

where ρ is a scaling parameter used to control the importance between $\widehat{\mathbf{y}}$ and $\mathbf{q}$. Following the assumption that the predictions become more accurate along with the model update [Xie *et al.*, 2019], we adaptively tune ρ by applying a log-schedule decreasing annealing to weaken the importance of the pseudo-labels. At each epoch t, it is computed as follows:

$$\rho = 1 - (\alpha * (\rho_{final} - \rho_{init}) + \rho_{init}), \tag{14}$$

$$\alpha = 1 - \exp(-5 * \frac{t}{T}), \tag{15}$$

where ρ_{init} and ρ_{final} are coefficient parameters; and T is the maximum number of epochs.

We replace $\widehat{\mathbf{y}}_{1:D}$ with $\mathbf{t}_{1:D}$ as the predictive targets of training texts, so as to formulate a weakly self-training loss. Here, we specify it by applying the KL-divergence loss function below:

$$\mathcal{L}_s(\mathbf{W}, \mathbf{E}, \mathbf{B}) = \frac{1}{D} \sum_{d=1}^{D} \sum_{k=1}^{K} \mathbf{t}_{dk} \log \frac{\mathbf{p}_{dk}}{\mathbf{t}_{dk}} \tag{16}$$

3.2 Wasserstein Barycenter Regularization

To further refine the weakly-supervised signals, we regularize them by minimizing the distances between the texts and the barycenters of relevant labels indicated by the weakly-supervised targets. Specifically, we measure the distances between the deep features of texts $\mathbf{z}_{1:D}$ and the trainable label barycenters $\boldsymbol{\mu}_{1:K}$ with the Wasserstein distance. This is actually equivalent to formulating a Wasserstein barycenter objective for each label as follows:

$$\mathcal{R}_b(\mathbf{E}, \mathbf{B}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \sum_{d=1}^{D} \frac{\lambda_{dk}}{D} W(\widehat{\mathbf{z}}_d, \widehat{\boldsymbol{\mu}}_k; \mathcal{M}(\mathbf{E})), \tag{17}$$

Algorithm 1 Model fitting for WTC-WBR

Input: Training dataset $\{\mathcal{U}, \mathcal{C}\}$
Output: A trained text classifier $f_{\mathbf{W}}(f_{\mathbf{E}}(f_{\mathbf{B}}(\cdot)))$
1: Employ the pre-trained BERT model with parameters $\mathbf{B}$;
2: Initialize the parameters $\{\mathbf{W}, \mathbf{E}\}$ randomly;
3: Calculate the pseudo-labels $\widehat{\mathbf{y}}_{1:d}$ by Eq.(10);
4: Initialize $\boldsymbol{\mu}_{1:K}$ with Eqs.(7), (8), (9);
5: **for** $iter = 1$ to T **do**:
6: Calculate each $\mathbf{T}^*_{dk}$ with mini-batches;
7: Calculate $\mathcal{M}(\mathbf{E})$ by Eq.(19);
8: Calculate gradients of $\{\mathbf{W}, \mathbf{E}, \mathbf{B}, \boldsymbol{\mu}\}$ with mini-batches;
9: Update $\{\mathbf{W}, \mathbf{E}, \mathbf{B}, \boldsymbol{\mu}\}$ with Adam;
10: **end for**

where each barycentric weight λ_{dk} is derived from the weakly-supervised target $\mathbf{t}_d$:

$$\lambda_{dk} = \begin{cases} 1, & \text{if } k = \arg\max(\mathbf{t}_d) \\ 0, & \text{otherwise} \end{cases}; \tag{18}$$

$\widehat{\mathbf{z}}_d = softmax(\mathbf{z}_d)$ and $\widehat{\boldsymbol{\mu}}_k = softmax(\boldsymbol{\mu}_k)$ aim to ensure normalized discrete distributions; and specially, each attribute of the deep feature $\mathbf{z}$ is represented by the top layer weight $\mathbf{E}^{(L)}$ of the feature encoder, thus the each element of the cost matrix $\mathcal{M}(\mathbf{E})$ can be calculated by the cosine distance between any two attributes:

$$\mathcal{M}(\mathbf{E})_{ij} = \frac{1 - \cos(\mathbf{E}^{(L)}_i, \mathbf{E}^{(L)}_j)}{2} \tag{19}$$

Accordingly, with the above definition of $\mathcal{M}(\mathbf{E})$, we kindly consider that the regularization of Eq.(17) brings another benefit, where it can capture the geometric information on the deep feature space, enabling to boost the discriminative power.

3.3 Full Objective and Model Fitting

By combining Eqs.(16) and (17), we show the full objective of WTC-WBR with respect to the trainable parameters $\{\mathbf{W}, \mathbf{E}, \mathbf{B}, \boldsymbol{\mu}\}$ as follows:

$$\mathcal{L}(\mathbf{W}, \mathbf{E}, \mathbf{B}, \boldsymbol{\mu}) = \mathcal{L}_s(\mathbf{W}, \mathbf{E}, \mathbf{B}) + \eta \mathcal{R}_b(\mathbf{E}, \boldsymbol{\mu}), \tag{20}$$

where η is the regularization parameter.

We initialize each Wasserstein barycenter $\boldsymbol{\mu}_k$ by performing the loops of Eqs.(7), (8), and (9) until convergence with pseudo-labels calculated by Eq.(10). We then adopt the gradient-based method to update each parameter of interest. For $\{\mathbf{W}, \mathbf{B}\}$, the gradients can be directly calculated by backpropagation. For $\mathbf{E}$, we perform a few number of inner loops with Eqs.(3) and (4) to estimate the optimum $\mathbf{T}^*_{dk}$ for each $W(\widehat{\mathbf{z}}_d, \widehat{\boldsymbol{\mu}}_k; \mathcal{M}(\mathbf{E}))$ by fixing the current parameters. Accordingly, the full objective can be written as follows:

$$\mathcal{L}_s(\mathbf{W}, \mathbf{E}) + \eta \sum_{k=1}^{k} \sum_{d=1}^{D} \frac{\lambda_{dk}}{D} \langle \mathbf{T}^*_{dk}, \mathcal{M}(\mathbf{E}) \rangle \tag{21}$$

Its gradient can be then calculated by backpropagation. For each $\boldsymbol{\mu}_k$, its gradient can be directly calculated by referring to Eq.(5).

Dataset	#Train	#Test	#Class	#AvgLc
IMDB	25,000	25,000	2	1
AG News	120,000	7,600	4	1
DBPedia	560,000	70,000	14	1.4

Table 1: Summary of dataset statistics. "AvgLc" denotes the average number of category words for classes.

To efficiently handle a large number of texts, we form noisy gradients of $\{\mathbf{W}, \mathbf{E}, \mathbf{B}, \boldsymbol{\mu}\}$ by randomly drawing a mini-batch of texts at each iteration with the spirit of stochastic optimization. For clarity, we summarize the full model fitting process in *Algorithm* 1.

4 Experiment

Datasets. We explore the proposed WTC-WBR method on 3 prevalent datasets from various domains: **IMDB** from movie review sentiment, **AG News** from news topic, and **DBPedia** from Wikipedia topic [Meng *et al.*, 2020]. Following the protocol in [Meng *et al.*, 2020], we employ the label names as category words, where each label name contains at most 3 words. The dataset statistics are listed in Table 1.

Baseline methods. To study the effectiveness of WTC-WBR, we compare it with 9 existing text classification methods: 5 weakly-supervised methods, **Dataless** [Chang *et al.*, 2008], **WeSTClass**[1] [Meng *et al.*, 2018], **LOTClass**[2] [Meng *et al.*, 2020], **X-Class**[3] [Wang *et al.*, 2021], **ClassKG**[4] [Zhang *et al.*, 2021]; 2 semi-supervised method, **UDA**[5] [Xie *et al.*, 2019] and **MixText**[6] [Chen *et al.*, 2020]; and 2 supervised methods, **BERT**[7] [Devlin *et al.*, 2019] and **XLNet**[8] [Yang *et al.*, 2019]. **WTC-ST** is the ablative version of WTC-WBR, which trains the classifier with the weakly self-training loss only. Specially, we compare the versions without BERT fine-tuning on WTC-WBR and WTC-ST, and annotate as *static*.

Implementation details. For WTC-WBR and WTC-ST, we feed each text into the pre-trained `BERT-base-uncased` encoder and feed the averaged pooling of token embeddings into the feature encoder. The maximum sequence lengths are set as 512, 200 and 200 for IMDB, AG News and DBPedia, respectively. We apply the Adam optimizer and the learning rates are tuned over $1e-7 \sim 7e-4$. We pre-train 2 epochs with the weakly self-training loss, and further train the full objective 10 epochs. For the *static* WTC-WBR and WTC-ST, we feed the static averaged pooling of token embeddings into feature encoder. The maximum sequence lengths are all set as 512. The batch size is 256. We pre-train 5 epochs and further train the full objective with 130, 20, and 20 epochs for IMDB,

[1]https://github.com/yumeng5/WeSTClass
[2]https://github.com/yumeng5/LOTClass
[3]https://github.com/ZihanWangKi/XClass
[4]https://github.com/zhanglu-cst/ClassKG
[5]https://github.com/google-research/uda
[6]https://github.com/GT-SALT/MixText
[7]https://github.com/huggingface/transformers
[8]https://github.com/zihangdai/xlnet

Supervision Pattern	Methods	IMDB	AG News	DBPedia
Weakly-supervised	**DataLess** [Chang *et al.*, 2008]	0.505	0.696	0.634
	WeSTClass [Meng *et al.*, 2018]	0.774	0.823	0.811
	LOTClass [Meng *et al.*, 2020]	0.865	0.864	0.911
	X-Class [Wang *et al.*, 2021]	0.828⋆	0.846⋆	0.917⋆
	ClassKG [Wang *et al.*, 2021]	0.874	0.888	0.980
	WTC-ST (*static*) (**Ours**)	0.808	0.864	0.956
	WTC-WBR (*static*) (**Ours**)	0.868	0.880	0.974
	WTC-ST (**Ours**)	0.871	0.886	0.978
	WTC-WBR (**Ours**)	**0.884**	**0.897**	**0.984**
Semi-supervised	**UDA** [Xie *et al.*, 2019])	0.908	0.912	0.991
	MixText [Chen *et al.*, 2020])	0.913	0.915	0.992
Supervised	**BERT** [Devlin *et al.*, 2019]	0.945	0.944	0.993
	XLNet [Yang *et al.*, 2019]	0.968	0.956	0.994

Table 2: Experimental results of classification accuracy. The best scores among weakly-supervised methods are indicated in boldface. We apply the results reported by the original papers and mark out our re-production results by "⋆". Results of semi-supervised methods are trained with 2500 labeled documents per class.

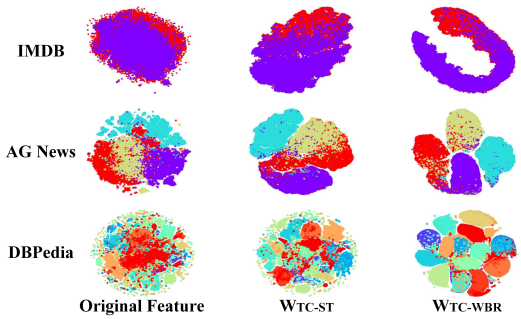

Figure 2: The t-SNE visualization of the original features, deep features learned by the *static* versions of WTC-ST and WTC-WBR on IMDB, AG News, and DBPedia.

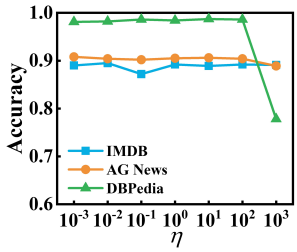

Figure 3: The accuracy performance varying the regularization parameter η on IMDB, AG News and DBPedia.

AG News, and DBPedia, respectively. The learning rates are tuned over $5e-5 \sim 1e-3$. For both two versions, we firstly train the model on the texts which contain category words due to the lack of category-word-covered texts. We adopt a one-layer MLP as the feature encoder and a one-layer MLP as the classification layer. The dimensions are 768-200-K, and we apply $\tanh$ as the activation function. ρ_{init} and ρ_{final} are set as 0.05 and 0.99. We have varied the regularization parameter η from the set $\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$ for examination and empirically set η as 100, 1 and 1 for IMDB, AG News and DBPedia, respectively. We implemented our method by PyTorch and run it on 1 RTX A6000 GPU in a Ubuntu platform of 32G memories.

4.1 Performance Comparison

We compare WTC-WBR with baseline methods by the classification accuracy on the test examples. For each dataset, we independently run WTC-WBR 5 times and show the averaged results in Table 2. First, we can observe that WTC-WBR significantly outperforms all weakly-supervised baselines, including conventional Dataless methods and recent neural competitors in all settings. Besides, it can be surprisingly seen that WTC-WBR is comparable to the semi-supervised and supervised methods, further demonstrating the effectiveness of WTC-WBR.

Ablative Study. We observe that the ablative version WTC-ST is on a par with the baseline methods. This indicates that applying the predictions to self-training can be significant to gain better results in the weakly-supervised tasks. WTC-WBR consistently performs better than WTC-ST without the Wasserstein barycenter regularization. Those results directly indicate the positive impact of the proposed regularization in weakly-supervised tasks. And it is worth mentioning that the two static versions are also on a par with baseline methods, indicating the effectiveness of the proposed regularization.

4.2 Feature Visualization

We investigate the discriminative capabilities of original features, and deep features of WTC-ST and WTC-WBR by using t-SNE. We examined both the static (*i.e.,* without fine-tuning) and fine-tuned versions of models. Early results show all versions show similar trends, and here we only plot the results of static ones due to the space limitation. We apply the TSNE-CUDA tool[9], and plot the results of t-SNE in Fig.2. It can be clearly seen that the deep features of WTC-ST and WTC-WBR are much more discriminative than the original ones on all datasets, so that they can improve the classification performance. More importantly, we can observe that the features of WTC-WBR are also better than those of WTC-ST, further indicating the effectiveness of the Wasserstein barycenter regularization.

[9]https://github.com/CannyLab/tsne-cuda

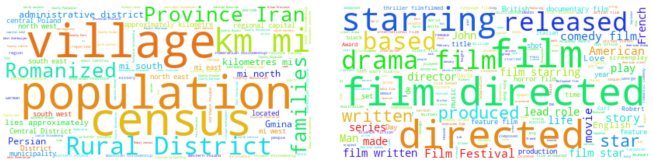

Figure 4: Two examples of word cloud graphs of classes across DBPedia.

4.3 Parameter Evaluation

We examined the regularization parameter η on WTC-WBR varying over $\{10^{-3}, 10^{-2}, \ldots, 10^{2}, 10^{3}\}$ and plot accuracy results in Fig.3. We may find that our WTC-WBR shows insensitivity overall when $\eta \leq 10^2$, and a relatively high parameter leads to performance deficiency when $\eta = 10^3$, especially on DBPedia. In a word, our WTC-WBR shows insensitivity on regularization parameter.

4.4 Case Study

We draw the word cloud graph for each class on the test samples annotated by the predictions.[10] We show two of these graphs on DBPedia with category words "villiage" and "film" in Fig.4. It can be clearly seen that the words with highest weights are highly correlated with their category words. This indirectly indicates WTC-WBR can take full advantage of the only available category words.

5 Related Work

Here, we briefly review related studies on weakly-supervised text classification and Wasserstein barycenter.

5.1 Weakly-supervised Text Classification

We concentrate on the weakly-supervised paradigm of learning with unlabeled texts and category words of classes. Generally, the existing methods can be divided into two groups: *shallow weakly-supervised models* and *deep weakly-supervised neural models*.

The shallow weakly-supervised models generate pseudo-labels with category words before applying traditional text classifiers, or jointly update the pseudo-labels and the classifier [Li *et al.*, 2016; Li and Yang, 2018; Li *et al.*, 2018]. For example, the method of [Li and Yang, 2018] iteratively updates the pseudo-labels, whose confidences are measured by a mixture of category word occurrences and label predictions. A Naïve Bayes classifier is jointly trained with the pseudo-labels with high confidences. Besides, there are some topic modeling-based methods, which define the category-topic priors with category word occurrences and classify texts with the estimated category-topic distributions [Li *et al.*, 2016; Li *et al.*, 2018]. Various techniques are proposed to further enhance the model training, such as background-topics used to filter out the less discriminative background knowledge [Li *et al.*, 2016] and manifold regularization used to spread supervised signals among neighboring texts [Li *et al.*, 2018].

[10] https://github.com/amueller/word_cloud

The recent deep weakly-supervised neural models [Meng *et al.*, 2020; Mekala and Shang, 2020; Wang *et al.*, 2021] are mainly built on the pre-trained language models such as BERT. The LOTClass model [Meng *et al.*, 2020] expands the given category words to more category indicative words by applying the word prediction task of BERT, and then fine-tunes the BERT by a novel masked category prediction task with category indicative words. The X-Class model [Wang *et al.*, 2021] feeds the raw texts into BERT to get contextualized word embeddings, and forms the label prototypes with the embeddings of category words. The pseudo-labels can be then generated with label prototypes and embeddings of texts to fine-tune BERT. Comparing with those arts, our WTC-WBR applies a novel Wasserstein barycenter regularization to refine the weakly-supervised targets.

5.2 Wasserstein Barycenter

The Wasserstein distance [Bogachev and Kolesnikov, 2012] provides a strong way to measure the distances between probability distributions from the perspective of geometry. Many techniques have been developed to efficiently solve for its minimization problem such as incorporating an auxiliary entropy regularization [Cuturi, 2013] and reformulating the minimization with copulas [Chi *et al.*, 2019]. Due to its strict theoretical properties, the Wasserstein distance has been widely applied to various domains, including topic modeling [Li *et al.*, 2020] and semantic distance measure between documents [Kusner *et al.*, 2015].

By analogy to the definition of barycenters in the Euclidean space, the Wasserstein barycenter is a minimizer of a weighted average of squared Wasserstein distances [Agueh and Carlier, 2011]. It can provide an efficient notion for constructing geometric prototypes in the Wasserstein space beyond Euclidean barycenters. Recently, many optimization methods have been proposed to efficiently solve for the minimization problem of Wasserstein barycenters [Schmitz *et al.*, 2018]. Here, we attempt to adopt the label barycenters to regularize the weakly-supervised signals, especially for the scenario with category words available only.

6 Conclusion

In this paper, we concentrate on the weakly-supervised scenario, where the training dataset only contains unlabeled texts and sets of category words for classes. To resolve the task, we develop a novel deep classification model named WTC-WBR, which refines the inaccurate and scarce supervision from the category words by weakly self-training and Wasserstein barycenter regularization. Specifically, we form pseudo-labels by using the category word occurrences of texts. Due to the inaccurate and scarce natures of pseudo-labels, we train the model with the mixtures of pseudo-labels and sharpened predictions. Most importantly, we suggest a Wasserstein barycenter regularization on the deep feature space, enabling to further refine the supervision. We conduct extensive experiments to compare WTC-WBR with the state-of-the-art weakly-supervised, semi-supervised, and supervised methods. Experimental results demonstrate that WTC-WBR is on a par with the existing competitors.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (NSFC) (No.61876071, No.62006094) and Scientific and Technological Developing Scheme of Jilin Province (No.20180201003SF, No.20190701031GH) and Energy Administration of Jilin Province (No.3D516L921421). We appreciate Zihan Wang for valuable discussions and providing pre-trained models.

References

[Agueh and Carlier, 2011] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysisb*, 43(2):904–924, 2011.

[Bogachev and Kolesnikov, 2012] Vladimir Igorevich Bogachev and Aleksandr Viktorovich Kolesnikov. The monge-kantorovich problem: achievements, connections, and perspectives. *Russian Mathematical Surveys*, 67(5):785–890, 2012.

[Chang *et al.*, 2008] Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *AAAI*, pages 830–835, 2008.

[Chen *et al.*, 2020] Jiaao Chen, Zichao Yang, and Diyi Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *ACL*, pages 2147–2157, 2020.

[Chi *et al.*, 2019] Jinjin Chi, Jihong Ouyang, Ximing Li, Yang Wang, and Meng Wang. Approximate optimal transport for continuous densities with copulas. In *IJCAI*, pages 2165–2171, 2019.

[Chizat *et al.*, 2018] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, pages 2292–2300, 2013.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[Druck *et al.*, 2008] Gregory Druck, Gideon Mann, and Andrew McCallum. Learning from labeled features using generalized expectation criteria. In *SIGIR*, pages 595–602, 2008.

[Kusner *et al.*, 2015] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *ICML*, pages 957–966, 2015.

[Lan *et al.*, 2020] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*, 2020.

[Li and Yang, 2018] Ximing Li and Bo Yang. A pseudo label based dataless naive bayes algorithm for text classification with seed words. In *COLING*, pages 1908–1917, 2018.

[Li *et al.*, 2016] Chenliang Li, Jian Xing, Aixin Sun, and Zongyang Ma. Effective document labeling with very few seed words: a topic modeling approach. In *CIKM*, pages 85–94, 2016.

[Li *et al.*, 2018] Ximing Li, Changchun Li, Jinjin Chi, Jihong Ouyang, and Chenliang Li. Dataless text classification: A topic modeling approach with document manifold. In *CIKM*, pages 973–982, 2018.

[Li *et al.*, 2020] Changchun Li, Ximing Li, Jihong Ouyang, and Yiming Wang. Semantics-assisted wasserstein learning for topic and word embeddings. In *ICDM*, pages 292–301, 2020.

[Li *et al.*, 2021] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. A survey on text classification: from shallow to deep learning. *arXiv:2008.00364*, 2021.

[Mekala and Shang, 2020] Dheeraj Mekala and Jingbo Shang. Contextualized weak supervision for text classification. In *ACL*, pages 323–333, 2020.

[Meng *et al.*, 2018] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised neural text classification. In *CIKM*, pages 983–992, 2018.

[Meng *et al.*, 2020] Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. Text classification using label names only: A language model self-training approach. In *EMNLP*, pages 9006–9017, 2020.

[Schmitz *et al.*, 2018] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[Wang *et al.*, 2021] Zihan Wang, Dheeraj Mekala, and Jingbo Shang. X-class: Text classification with extremely weak supervision. In *NAACL*, pages 3043–3053, 2021.

[Xie *et al.*, 2019] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5754–5764, 2019.

[Zhang *et al.*, 2021] Lu Zhang, Jiandong Ding, Yi Xu, Yingyao Liu, and Shuigeng Zhou. Weakly-supervised text classification based on keyword graph. In *EMNLP*, page 2803–2813, 2021.

[Zhou, 2018] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.