Improving Semi-Supervised Text Classification with Dual Meta-Learning

SHUJIE LI, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
GUANGHU YUAN, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
MIN YANG, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
YING SHEN, School of Intelligent Systems Engineering, Sun Yat-sen University, Guangdong, China
CHENGMING LI, School of Intelligent Systems Engineering, Sun Yat-sen University, Guangdong, China
RUIFENG XU, Harbin Institute of Technology (Shenzhen), Shenzhen, China
XIAOYAN ZHAO, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

The goal of semi-supervised text classification (SSTC) is to train a model by exploring both a small number of labeled data and a large number of unlabeled data, such that the learned semi-supervised classifier performs better than the supervised classifier trained on solely the labeled samples. Pseudo-labeling is one of the most widely used SSTC techniques, which trains a teacher classifier with a small number of labeled examples to predict pseudo labels for the unlabeled data. The generated pseudo-labeled examples are then utilized to train a student classifier, such that the learned student classifier can outperform the teacher classifier. Nevertheless, the predicted pseudo labels may be inaccurate, making the performance of the student classifier degraded. The student classifier may perform even worse than the teacher classifier. To alleviate this issue, in this paper, we introduce a dual meta-learning (**DML**) technique for semi-supervised text classification, which improves the teacher and student classifiers simultaneously in an iterative manner. Specifically, we propose a meta-noise correction method to improve the student classifier by proposing a Noise Transition Matrix (NTM) with meta-learning to rectify the noisy pseudo labels. In addition, we devise a meta pseudo supervision method to improve the teacher classifier. Concretely, we exploit the feedback performance from the student classifier to further guide the teacher classifier to produce more accurate pseudo labels for the unlabeled data. In this way, both teacher and student classifiers can co-evolve in the iterative training process. Extensive experiments on four benchmark datasets highlight the effectiveness of our DML method against existing state-of-the-art methods for semi-supervised text classification. We release our code and data of this paper publicly at https://github.com/GRIT621/DML.

This work was partially supported by the National Key Research and Development Program of China (2022YFF0902100), the National Natural Science Foundation of China (62376262), Shenzhen Science and Technology Innovation Program (KQTD20190929172835662), Shenzhen Basic Research Foundation (JCYJ20210324115 614039 and JCYJ20200109113441941).
Authors' addresses: S. Li, G. Yuan, M. Yang (Corresponding author), and X. Zhao, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Avenue, University Town, Xili, Nanshan District, Shenzhen, Guangdong Province, Shenzhen, Guangdong, 518055, China; e-mails: ustclsj@mail.ustc.edu.cn, gh.yuan@siat.ac.cn, min.yang@siat.ac.cn, xy.zhao@siat.ac.cn; Y. Shen (Corresponding author) and C. Li, School of Intelligent Systems Engineering, Sun Yat-sen University, 66 Gongchang Road, Guangming District, Shenzhen, Guangdong Province, Shenzhen, Guangdong, 518107, China; e-mails: shenying@mail.sysu.edu.cn, lichengming@mail.sysu.edu.cn; R. Xu, Harbin Institute of Technology (Shenzhen), Harbin Institute of Technology, Shenzhen, University Town, Xili, Nanshan District, Shenzhen, Guangdong Province, Shenzhen, Guangdong, 518055, China; e-mail: xuruifeng@hit.edu.cn.

ACM 1046-8188/2024/04-ART109
https://doi.org/10.1145/3648612

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Applied computing** → **Document management and text processing;**

Additional Key Words and Phrases: Semi-supervised text classification, pseudo labeling, noise transition matrix, meta learning, consistency regularization

ACM Reference Format:
Shujie Li, Guanghu Yuan, Min Yang, Ying Shen, Chengming Li, Ruifeng Xu, and Xiaoyan Zhao. 2024. Improving Semi-Supervised Text Classification with Dual Meta-Learning. *ACM Trans. Inf. Syst.* 42, 4, Article 109 (April 2024), 28 pages. https://doi.org/10.1145/3648612

1 INTRODUCTION

Supervised learning has made significant progress in text classification, where a large number of labeled training examples are available. However, it is difficult, if not impossible, to collect sufficient labeled data in many real-world scenarios due to annotation cost, safety issues or privacy concerns. In addition, the data distribution shift often makes a well-learned model less accurate as time passes [1]. It is difficult to train and generalize a unified supervised model across diverse domains with a small number of labeled samples because the latent spaces encode very different information [2]. Therefore, it is necessary to explore effective deep learning techniques that can learn from a limited number of labeled training examples. In recent years, semi-supervised learning has drawn noticeable attention in text classification. How to effectively learn essential and generalizable feature patterns from a small set of labeled training data still remains the greatest challenge faced by semi-supervised learning.

Pseudo-labeling is one of the most widely used **semi-supervised text classification (SSTC)** approaches, which pseudo-labels the unlabeled samples by employing a teacher classifier trained on the combination of labeled and previously pseudo-labeled samples, and this process is iteratively repeated in a self-training cycle [3]. Although pseudo-labeling has made noticeable progress in SSTC, how it learns effective text classifiers from a limited number of labeled training data is still a long-standing problem. In particular, the confirmation bias issue [4] is ubiquitous, where the noisy pseudo labels could be overly utilized. It could be problematic to learn a high-quality student classifier based on possibly incorrect pseudo labels, especially when the noisy rate is high since the deep neural networks tend to overfit the noisy labeled data. To mitigate such a confirmation bias issue, the quality of the pseudo labels predicted by the teacher classifier should be improved.

So far, some confidence-based filtering methods [5, 6] have been proposed to select pseudo labels with high confidence by manually defining a threshold. However, it is difficult to define an optimal one-size-fits-all threshold value manually in advance for different datasets, since the threshold value could depend on various factors such as the data complexity and the training stage. In addition, the selected pseudo-labeled samples may still be misclassified with high confidence, leading to accumulated errors and making the student classifier obtain worse performance in the process of iterative learning. Furthermore, the incorrectly pseudo-labeled samples tend to be ambiguous samples that appear in minor classes, resulting in the class imbalance problem that can dramatically degrade the performance of classifiers.

In parallel, some studies [7, 8] exploited the **Noise Transition Matrix (NTM)** to mitigate the side-effects of noisy labels. The optimized NTM is expected to bridge the gap between the clean ground-truth labels and noisy pseudo labels [9], thus it helps to learn a noise-tolerated classifier for pseudo-labeling. Although previous methods can explicitly model the generation process of the noisy labels, it is non-trivial to estimate the NTM for the noise distribution of pseudo labels.

Specifically, a large number of clean labeled data per class is required to guide the estimation of NTM, making the correction algorithm infeasible to be applied to minor classes with low resources.

To alleviate the aforementioned challenges, in this paper, we propose a **dual meta learning** (**DML**) technique for semi-supervised text classification, which enhances the teacher and student classifiers in pseudo-labeling simultaneously via an iterative manner. On the one hand, we improve the student classifier by proposing a meta noise correction method to mitigate the noisy issues in pseudo labels predicted by the teacher classifier. We devise a meta-learning strategy to learn the NTM in a learning-to-learn manner, which learn to adaptively select the data with confident labels as meta dataset, so as to guide the derivation of NTM. In this way, we can learn the meta-knowledge of underlying noise distribution, thus boosting the generalization ability of the classifier on the unlabeled data. On the other hand, we propose a meta pseudo supervision method to improve the performance of the teacher classifier for predicting more accurate pseudo labels for unlabeled data. In particular, we also exploit the feedback performance from the student classifier to further steer the teacher classifier to produce more accurate pseudo-labeled samples. Specifically, we iteratively optimize the student classifier with the pseudo-labeled samples predicted by the teacher classifier and update the teacher conditioned on the performance of the student classifier. These two classifiers can learn from each other and co-evolve gradually during the training process.

The main contributions of this paper are listed as follows:

— We propose a meta pseudo supervision method to improve the performance of the teacher classifier, which learns from the feedback signal of how well the student classifier performs on the training data by exploiting the meta learning technique. In this way, the teacher classifier can generate more accurate pseudo labels of the unlabeled training samples.
— We devise a meta noise correction method to model the noise distribution of pseudo labels for improving the performance of the student classifier. This method leverages meta learning to learn meta-knowledge of underlying label distribution and thus enhances the generalization ability of the classifier.
— Extensive evaluations are conducted on four benchmark corpora (i.e., AGNews, Yelp, Yahoo and Amazon) for semi-supervised text classification. Experimental results demonstrate that our DML framework achieves substantially better performance than the compared strong baselines, especially on the datasets with confusing class categories.

2 RELATED WORK

2.1 Semi-Supervised Text Classification

Text classification has drawn noticeable attention from both the academia and industry communities since it has been widely applied in many tasks, such as news categorization [10, 11], sentiment analysis [12, 13], and spam detection [14, 15]. Developing efficient **semi-supervised text classification (SSTC)** methods has long been a goal of the NLP community [16, 17] because the unlabeled data tends to be more abundant and readily available than labeled data. So far, plenty of semi-supervised deep neural networks have been proposed to learn essential feature patterns from limited labeled samples and a large number of unlabeled samples.

Pseudo-labeling is one of the most important semi-supervised learning approaches, which trains a teacher model on a limited number of labeled samples to produce pseudo labels of unlabeled data. Then, the ground-truth instances are combined with the pseudo-labeled instances to learn a student which is supposed to outperform the teacher model [5, 6, 18–20]. Many previous works attempted to mitigate the issue of noisy pseudo labels in the target domain by eliminating noisy data based on confidence scores or uncertainty [21–23]. For example, in [24], four methods

are proposed to select clean data for training and evaluate the degradation in classification effectiveness that these noisy labels bring about. In [22], a confidence level was assigned to each unlabeled instance so as to degrade the influences from the instances with unreliable pseudo labels. In [23], the normalized confidence scores were used to choose pseudo labels of unlabeled data with the balanced class distribution. Subsequently, several studies [25, 26] attempted to rectify the pseudo Labels via Noise Transition Matrix. One representative work was proposed by Zhang et al. [25], which estimated the Noise Transition Matrix and learned a classifier with noisy data jointly. Wang et al. [27] proposed a progressive learning strategy that addressed distribution bias by assigning pseudo labels to unlabeled data based on prediction scores.

Consistency regularization and data augmentation are also widely used techniques for semi-supervised learning based on the assumption of smoothness and clustering [28–31]. The key idea behind consistency regularization is to obtain augmentation invariant output distribution, where the classifier is trained to predict a consistent output for each unlabeled sample and the corresponding perturbed variant [7, 32]. For example, Ye et al. [33] added distributional perturbations at the embedding layer by imitating adversarial attacks and combines with contrastive learning to improve performance. Laine and Aila [34] proposed self-ensembling to make the output prediction of a same instance consistent over different epochs. Tarvainen and Valpola [4] introduced a mean teacher algorithm forcing two models (a teacher and a student) to obtain the same output prediction. Chen et al. [7] trained the model to generate consistent labels for different augmented samples by leveraging the weighted average of all predictions. For better capturing the correlation between discrete features of the embedding layer, Zhang et al. [35] proposed a graph-based regularization method, which encourages the preservation of the latent discrete spatial structure between the embedding layers to obtain better stable properties, thereby reducing overfitting.

To enhance the robustness of the classifier, Miyato et al. [36] applied adversarial perturbations to the input samples and leveraged the consistency constraint of them as regularization term. Xie et al. [37] introduced back translation as the data augmentation technique for text classification. Du et al. [38] presented a graph-based semi-supervised learning technique that tackled the problem of noisy labels by integrating the maximum correntropy criterion.

2.2 Meta-learning Techniques

Meta-learning aims at training a model that is expected to rapidly adapt to novel tasks by leveraging a few datapoints and learning episodes [39]. Concretely, meta-learning is commonly regarded as learning to learn, where an outer meta learner updates an inner base learner such that the learned model optimizes the outer objective [40, 41]. MAML [39] is a representative meta learning method, which learns good parameter initialization for fast adaptation by updating the model on a support set and then evaluating it on a query set. Nichol et al. [42] extended the first-order MAML by ignoring second-order derivatives and also included Reptile by repeatedly sampling and training on a task so as to make the initialization move towards the trained weights on that task.

Beyond that, meta-learning can be utilized to learn the model parameters for data selection, which learns the model with balanced performance over the attributes and derives more generalizable parameters than manually tuned ones [43, 44]. For example, Wu et al. [45] utilized a meta-learner to learn prediction and training policies jointly for different categories. Shu et al. [41] proposed meta-weight-net that adaptively learns an explicit weighting function from data to boost the robustness of the deep model with biased training data. Pham et al. [46] proposed a student-teacher framework, in which the teacher model produced the pseudo labels based on an efficient meta-learning and the student model.

Meta-learning also studies how to extract knowledge from past training to fast adapt or generalize to new tasks or scenarios with limited samples [39, 47]. For example, Liang et al. [43] proposed an aspect-attention meta-learning framework that leverages external knowledge to construct aspect-attention and aspect-contrastive representations to match target aspects. Fu et al.[48] proposed a temporal graph metric learning framework, which can accurately classify different temporal graphs and be adapted to discover new subspaces for unseen classes.

Different from aforementioned methods, we propose a dual meta learning method for semi-supervised text classification, which simultaneously improves the teacher classifier to produce high-quality pseudo labels for unlabeled instances and rectifies the noisy pseudo labels to further enhance the student classifier. Our method takes advantage of meta pseudo supervision, consistency regularization and meta noise transition for semi-supervised learning.

This manuscript can be regarded as a significant extension of our previous conference paper published on SIGIR 2022 [49]. In this journal extension, we have made substantial new contributions. First, we propose a novel dual meta learning (called **DML**) framework to enhance the performance of semi-supervised text classification from the perspectives of both teacher and student classifiers via a unified framework. By enabling simultaneous learning and co-evolution between the enhanced teacher and student classifiers, our framework facilitates their mutual improvement throughout the training process. Second, we propose a **meta noise correction (MNC)** approach to address the issue of noisy pseudo labels generated by the teacher classifier. By incorporating MNC, we can learn the meta-knowledge of underlying noise distribution, thus boosting the generalization ability of the student classifier on the unlabelled data. Third, we conduct a series of comprehensive experiments to thoroughly evaluate the effectiveness of our DML method. Experimental results demonstrate that the proposed DML method in this journal extension substantially outperforms the **Dual Pseudo Supervision (DPS)** method in our previous conference paper. Overall, the proposed methodology, accompanied by extensive experimental validation, highlights the substantial improvements achieved by our proposed DML method over the DPS method proposed in [49].

3 OUR METHODOLOGY

3.1 Problem Definition

Given a training dataset $D = D_l \cup D_u$, we let $D_l = \{x_i^l, y_i^l\}_{i=1}^N$ be a set of N limited labeled samples and $D_u = \{x_j^u\}_{j=1}^M$, where N and M indicate the numbers of unlabeled and labeled samples, respectively. Here, M is much larger than N. In the semi-supervised learning setting, both D_l and D_u are drawn from the same distribution as D. x_i^l (or x_i^u) represents the input sequence of the i-th labeled (or unlabeled) sample and y_i^l is the one-hot label vector of x_i^l, and for using mini-batch SGD to optimize the loss, we sample small batches m, n from each of the large datasets D_l, D_u. The goal of this paper is to learn a well-performing classifier by using both labeled instances and unlabeled instances as the training data. The proposed model contains two networks: a teacher classifier $\mathcal{T}$ and a student classifier $\mathcal{S}$. Formally, we utilize $\mathcal{T}(x^k; \theta_{\mathcal{T}})$ to represent the soft prediction by the teacher model, where $k \in \{l, u\}$ and x^k denotes a labeled instance or an unlabeled instance. $\theta_{\mathcal{T}}$ represents the parameters of the teacher model. For student model, similarly, we utilize $\mathcal{S}(x^k; \theta_{\mathcal{S}})$ to represent the soft prediction by the student classifier, where $\theta_{\mathcal{S}}$ represents the set of parameters of the student.

3.2 Overview of Our Methodology

The overview of our DML framework is illustrated in Figure 1. The architecture of our method consists of a teacher classifier and a student classifier, where the teacher classifier produces pseudo

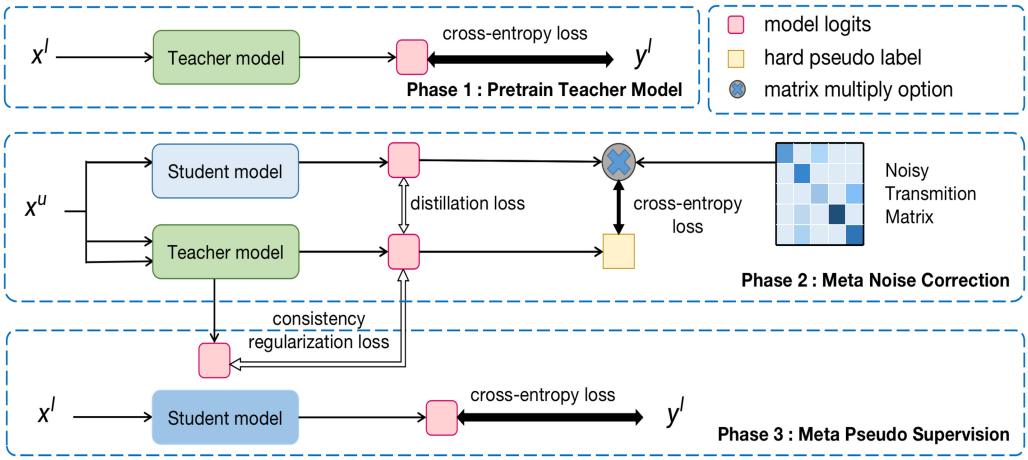


Fig. 1. The primary phases of our approach. In the first stage, the teacher model is initialized with labeled data. In the second stage, meta-learning is used to help the student update, which is our proposed Meta Noise Correction. In the third stage, the meta-learning method is used to help the teacher update, which is our proposed Meta Pseudo Supervision. The solid two-way arrows represent the two to do the cross entropy loss, and the hollow two-way arrows represent the two to do the KL divergence loss.

labels for unlabeled instances which are then combined with the labeled instances to train an effective student classifier.

The training process of the proposed DML method can be divided into four phases. First, we pre-train the teacher classifier on the initial labeled data. Second, we propose a meta noise correction method to model the noise distribution of pseudo labels to boost the performance of the student classifier. Third, we devise a meta pseudo supervision method to learn a reliable teacher classifier by learning from the feedback performance of the student classifier. The second and third phases iteratively repeat until convergence.

3.3 Pre-training the Teacher Classifier

We employ BERT [50] as our base text encoder to obtain the contextual representations of the input sequence because of its impressive performance on many NLP tasks. Formally, for an input sequence x^l in the labeled training set, we use BERT to produce the contextual representation as:

$$\mathbf{H}^l = \text{BERT}([\text{CLS}]x^l[\text{SEP}]) \tag{1}$$

where $\mathbf{H}^l = [\mathbf{h}^l_1, \ldots, \mathbf{h}^l_{|H|}]$ represents the output hidden representations of the BERT model. [CLS] and [SEP] are special tokens. $\mathbf{h}^l_i$ denotes the i-th item in $\mathbf{H}^l$. The hidden state of the classification token [CLS] is represented as $\mathbf{h}^l_{\text{CLS}}$ (i.e., $\mathbf{h}^l_{\text{CLS}} = \mathbf{h}^l_1$). Here, the vector $\mathbf{h}^l_{\text{CLS}}$ can be regarded as the initial contextual representation of the input sequence for output prediction.

Then, we feed $\mathbf{h}^l_{\text{CLS}}$ into a **multi-layer perceptron (MLP)** layer followed by a softmax layer to predict the label distribution:

$$\mathcal{T}(\mathbf{x}^l, \theta_{\mathcal{T}}) = \text{softmax}(W^o \text{MLP}(\mathbf{h}_{\text{CLS}}) + b^o) \tag{2}$$

where $\mathcal{T}(\mathbf{x}^l, \theta_{\mathcal{T}})$ denotes the predicted sentiment distribution by the teacher classifier. W^o and b^o are learnable parameters.

Given a dataset with N labeled training instances $(x_i^l, y_i^l)_{i=1}^N$, the parameters of the teacher classifier are trained to minimize the standard cross-entropy loss function as follows:

$$\mathcal{L}_{\text{CE}}^{\mathcal{T}}(\mathcal{T}(x^l;\theta_{\mathcal{T}}), y^l) = -\frac{1}{N}\sum_{i=1}^{N} y_i^l \log[\mathcal{T}(x_i^l, \theta_{\mathcal{T}})], \tag{3}$$

$$\mathcal{L}_{\text{UDS}} = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i|y_1,\ldots,y_{i-1};G,E), \tag{4}$$

where $x^l = \{x_1^l,\ldots,x_N^l\}$ and $y^l=\{y_1^l,\ldots,y_N^l\}$. Here, y_i^l and $\mathcal{T}(x_i^l,\theta_{\mathcal{T}})$ denote the gold and predicted probabilities, respectively. $\theta_{\mathcal{T}}$ represents the set of learnable parameters of the teacher classifier $\mathcal{T}$.

3.4 Meta Noise Correction

After pre-training the teacher classifier on the originally labeled samples, we use the learned teacher classifier to predict pseudo labels for the unlabeled instances. However, the predicted pseudo labels could be noisy. When inaccurate pseudo labels dominate the unlabeled data, the student model may get even more inferior results than the teacher classifier by learning from noisy pseudo-labeled samples. It is challenging to learn a high-quality text classifier by using the noisy pseudo-labeled data, especially when the noisy rate is high, since the deep models tend to overfit the noisy pseudo-labeled data.

To mitigate the impact of noisy pseudo labels of the unlabeled data, we improve the student classifier by proposing a meta noise correction method. In particular, we devise a meta-learning strategy to learn the Noise Transition Matrix (NTM) in a learning-to-learn manner. The meta noise correction method learns to adaptively select the data with confident labels as meta set, which can steer the derivation of NTM. In this way, we can learn the meta-knowledge of underlying noise distribution, thus boost the generalization capability of the classifier on unlabeled samples.

3.4.1 The Loss Functions of the Student Classifier.

The Knowledge Distillation Loss. For the student classifier, we hope that the model can not only learn from pseudo soft labels $\mathcal{T}(x^u;\theta_{\mathcal{T}})$ of the unlabeled data x^u but also encourage the consistence between the output distribution logits $\mathcal{S}(x^u;\theta_{\mathcal{S}})$ produced by the student classifier and $\mathcal{T}(x^u;\theta_{\mathcal{T}})$ produced by the teacher classifier. Therefore, we devise a knowledge distillation loss $\mathcal{L}_{\text{KD}}^u$ which stimulates the output prediction $\mathcal{S}(\mathbf{x}^u;\theta_{\mathcal{S}})$ of the student classifier to mimic the output prediction $\mathcal{T}(\mathbf{x}^u;\theta_{\mathcal{T}})$ of the teacher classifier. Formally, we calculate the knowledge distillation loss $\mathcal{L}_{\text{KD}}^u$ by minimizing the KL-divergence between the output distributions of the student and teacher classifiers as follows:

$$\mathcal{L}_{\text{KD}}^u(\theta_{\mathcal{S}},\theta_{\mathcal{T}}) = \text{KL}\big(\mathcal{S}(x^u;\theta_{\mathcal{S}})\big\|\mathcal{T}(x^u;\theta_{\mathcal{T}})\big) \tag{5}$$

where KL$(\cdot)$ represents the KL-divergence between two variables.

The Cross-Entropy Loss. For the pseudo-labeled data x^u, the logit $\mathcal{T}(x^u;\theta_{\mathcal{T}})$ produced by the teacher classifier can be converted into the hard pseudo label $\widetilde{y}^u$ (one-hot vectors) by performing the hardsoftmax operation. Then, we can learn a Noise Transition Matrix (NTM) $\mathbf{T}$ to rectify the potentially biased output logit $\mathcal{S}(x^u;\theta_{\mathcal{S}})$ and thus improve the performance of the student classifier. Mathematically, the cross-entropy loss of the student classifier is defined as follows:

$$\mathcal{L}_{\text{CE}}^{\mathcal{S}}(\mathcal{S}(x^u;\theta_{\mathcal{S}})\mathbf{T}, \widetilde{y}^u) = -\frac{1}{M}\sum_{i=1}^{M} \widetilde{y}_i^u \log[\mathcal{S}\left(x_i^u;\theta_{\mathcal{S}}\right)\mathbf{T}]. \tag{6}$$

where $\mathcal{L}_{\mathrm{CE}}^{\mathcal{S}}(\cdot)$ indicates the cross-entropy loss of the student classifier between the ground-truth and predicted labels rectified by T. $\mathcal{S}(x^u;\theta_S)$ represents the output logits of the student classifier on the pseudo-labeled data x^u. $\theta_{\mathcal{S}}$ represents the parameters of the student model. $\widetilde{y}_i^u$ indicates the one-hot (hard) vector of the pseudo label predicted by the teacher classifier for sample x_i^u. $\mathbf{T}$ is the Noise Transition Matrix.

3.4.2 Updating NTM and the Student Classifier with Meta Learning. We define the value of the i-th row and the j-th column of the matrix (i.e., $\mathbf{T}_{ij}$) as the conditional probability of transitioning from the label i to the noisy label j, i.e., $\mathbf{T}_{ij} = p(\widetilde{y} = j|y = i), \forall i, j$. Once the NTM $\mathbf{T}$ is determined, we can leverage it to correct the loss function of the student classifier from noisy labels.

Since NTM turns the clean posterior probability to the noise posterior probability, the loss makes the distribution of the noise posterior probability $p(\widetilde{y}^u = j|x^u)$ and the pseudo label to be close, instead of narrowing the distance between the pseudo label and the clean posterior probability $p(y^u = i|x^u)$. That is, NTM bridges clean data and noisy labels as follows:

$$\begin{aligned} p(\widetilde{y}^u = j \mid x^u) &= \sum_{i=1}^{C} \mathbf{T}_{ij} p(y^u = i \mid x^u) \\ \Rightarrow p(\widetilde{y}^u \mid x^u) &= p(y^u \mid x^u)\mathbf{T} \end{aligned} \tag{7}$$

In this way, although we train the student classifier with noisy pseudo-labeled data, we can recover the desired estimation of class posterior probability $p(y^u|x^u)$.

To explore the inter-class noise transition probabilities and correct the problems caused by noise accumulation, we adopt the meta-learning method to alternatively train the NTM $\mathbf{T}$ on a meta set with clean labels and optimize the sentiment classifier on the pseudo-labeled data corrected by the previously estimated NTM $\mathbf{T}$. Concretely, we treat the limited labeled instances D_l as a meta set with clean labels. The meta set represents the meta-knowledge of underlying label distribution of the originally labeled instances. Formally, we first leverage the meta set to estimate the NTM $\mathbf{T}$ and then utilize the pseudo-labeled data to update the student network $\mathcal{S}$ as follows:

$$\begin{aligned} \mathbf{T}^* &= \arg\min_{\mathbf{T}\in[0,1]^{c\times c}} -\frac{1}{N}\sum_{i=1}^{N} y_i^l \log[\mathcal{S}\left(x_i^l; \theta_{\mathcal{S}}^*(\mathbf{T})\right)] \\ \text{where } \theta_{\mathcal{S}}^*(\mathbf{T}) &= \arg\min_{\theta_S} -\frac{1}{M}\sum_{i=1}^{M} \tilde{y}_i^u \log[\mathcal{S}\left(x_i^u, \theta_S\right)\mathbf{T}] \end{aligned} \tag{8}$$

where the optimal NTM $\mathbf{T}^*$ is achieved by minimizing cross-entropy loss of the student classifier on the originally labeled data. The optimal student classifier $\theta_{\mathcal{S}}^*(\mathbf{T})$ is obtained by minimizing the cross-entropy loss on the pseudo-labeled data with noisy pseudo labels, which depends on the value of $\mathbf{T}$. That is, $\theta_{\mathcal{S}}^*$ is a functional operator with the learned NTM $\mathbf{T}$. In this way, the student classifier is able to avoid the incorrect supervision signals and enhance the generalization capability of the classifier.

Updating the NTM $\mathbf{T}$. Figure 2 demonstrates the training process of the student network and the NTM $\mathbf{T}$. This is an alternating process, which contains three phases: virtual optimization, meta optimization, actual optimization. We optimize the NTM $\mathbf{T}$ via the virtual and meta optimization phases, while the actual optimization phase is adopted to update the student network with fixed $\mathbf{T}$.

In the virtual optimization phase, at the $t+1$-th iteration, we firstly copy the student network $\mathcal{S}$ with parameters $\theta_S^{(t)}$ to a meta network $\mathcal{S}_{meta}$ with parameters $\theta_{\mathcal{S}_{meta}}^{(t)}$. Then, the parameters $\theta_{\mathcal{S}_{meta}}^{(t+1)}$ of the meta network are learned by moving the current parameters $\theta_{\mathcal{S}_{meta}}^{(t)}$ along the gradient

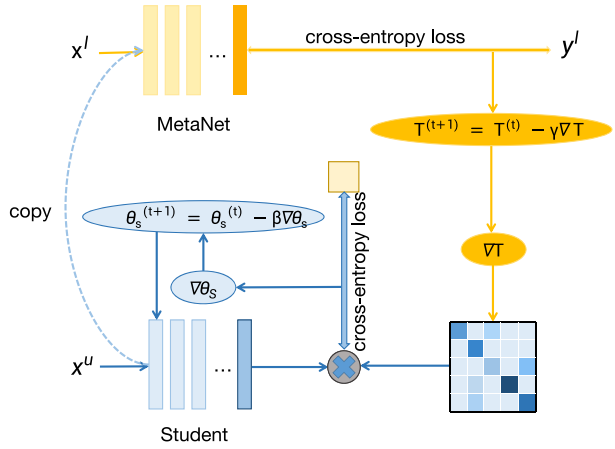


Fig. 2. The learning process of the student network $\mathcal{S}$ and the NTM $\mathbf{T}$ with meta learning.

descent direction of rectified loss function. Mathematically, the process of virtually updating the student model is defined as follows:

$$\theta_{\mathcal{S}_{meta}}^{(t+1)} = \theta_{\mathcal{S}_{meta}}^{(t)} - \gamma \nabla_{\theta_{\mathcal{S}_{meta}}} \mathcal{L}_{\text{CE}} \left(\mathcal{S}_{meta}(x^u; \theta_{\mathcal{S}_{meta}}^{(t)}) \mathbf{T}, \tilde{y} \right) \tag{9}$$

where γ is the learning rate of the meta noise correction method. Since the above virtual optimization is a "virtual" step, it indicates that the parameters in the student model are not actually updated in this stage. The purpose of virtual optimization is to learn updated parameters $\theta_{\mathcal{S}_{meta}}^{(t+1)}$ so as to facilitate the meta optimization of the NTM $\mathbf{T}$.

In the meta optimization phase, we learn the updated NTM $\mathbf{T}^{(t+1)}$ based on the meta network $\theta_{\mathcal{S}_{meta}}^{(t+1)}$ and the NTM $\mathbf{T}^{(t)}$ learned at the t-th iteration. Formally, we define the updating process of the NTM $\mathbf{T}^{(t+1)}$ as follows:

$$\mathbf{T}^{(t+1)} = \mathbf{T}^{(t)} - \gamma \nabla_{\mathbf{T}} \mathcal{L}_{\text{CE}} \left(\mathcal{S}_{meta}(x^l; \theta_{\mathcal{S}_{meta}}^{(t+1)}(\mathbf{T}^{(t)})), y^l \right) \tag{10}$$

where γ denotes the learning rate of the meta noise correction module. The key idea beyond meta optimization is to learn an optimal NTM $\mathbf{T}^{(t+1)}$ with the low empirical risk and excellent generalization capability. After the back propagation process, the NTM $\mathbf{T}^{(t+1)}$ may contain negative values. Since the Noise Transition Matrix expresses the probability of transition from one class to another, where the value $\mathbf{T}_{ij}$ with the row i representing the true class and the column j representing the predicted class that is likely to be transitioning, we normalize the rows to make sure the transition probabilities of class i to be summed up to 1.

In the actual optimization phase, we optimize the parameters of the student classifier on the noisy pseudo-labeled samples. The overall loss of the student model is formed by combining the knowledge distillation loss $\mathcal{L}_{\text{KD}}^{u}$ and the cross-entropy loss $\mathcal{L}_{\text{CE}}^{\mathcal{S}}$. Concretely, after obtaining the updated NTM $\mathbf{T}^{(t+1)}$, the parameters $\theta_{\mathcal{S}}^{(t+1)}$ of the student model can be updated by:

$$\theta_{\mathcal{S}}^{(t+1)} = \theta_{\mathcal{S}}^{(t)} - \beta \nabla_{\theta_{\mathcal{S}}} \left(\mathcal{L}_{CE}^{\mathcal{S}}(\mathcal{S}(x^u; \theta_{\mathcal{S}}^{(t)}) \mathbf{T}, \tilde{y}^u) + \mathcal{L}_{\text{KD}}(\theta_{\mathcal{S}}^{(t)}, \theta_{\mathcal{T}}^{(t)}) \right) \tag{11}$$

where β denotes the learning rate of the student model. By performing the alternating optimization strategy, both the NTM $\mathbf{T}$ and the student model parameters θ_S can be gradually ameliorated until convergence.

3.5 Meta Pseudo Supervision

The core idea behind the pseudo-labeling based semi-supervised learning with the teacher-student framework is to predict reliable pseudo labels of unlabeled samples. In most previous studies, an alternative two-stage pipeline is adopted in each training iteration, i.e., creating pseudo labels via the teacher classifier and training the student classifier with the generated pseudo labels. However, the student classifier is notoriously brittle, obtaining even worse performance than the teacher classifier if the generated pseudo labels are inaccurate [19].

In addition to the meta noise correction method for improving the student classifier, we also propose a meta pseudo supervision method to improve the performance of the teacher classifier so as to produce high-quality pseudo labels for the unlabeled data. In particular, we leverage the feedback performance from the student classifier to facilitate the teacher classifier to produce high-quality pseudo labels of unlabeled samples by exploiting meta-learning techniques. Moreover, we also employ a consistency regularization term to enhance the robustness of the teacher classifier. In this way, we can learn a more reliable teacher classifier, which provides high-quality pseudo-labeled samples for the student classifier and thus boosts the overall text classification performance.

3.5.1 Consistency Regularization. We compute the consistency regularization term that minimizes the KL divergence between two output representations of the same input instance produced by the teacher classifier. Since the dropout technology randomly turns off different network neurons, we pass each instance into the same teacher classifier twice to obtain two different logits $\mathcal{T}(x^u;\theta_{\mathcal{T}_1})$ and $\mathcal{T}(x^u;\theta_{\mathcal{T}_2})$. The two output distributions are expected to be consistent such that the reliability of the teacher classifier can be ensured. Formally, we devise a consistency regularization term $\mathcal{L}_{\text{CR}}^{\mathcal{T}}(\theta_{\mathcal{T}_1},\theta_{\mathcal{T}_2})$ that minimizes the KL-divergence between the two output representations produced by the teacher classifier:

$$\mathcal{L}_{\text{CR}}^{\mathcal{T}}(\theta_{\mathcal{T}_1},\theta_{\mathcal{T}_2}) = \text{KL}\left(\mathcal{T}(x^u;\theta_{\mathcal{T}_1})\middle\|\mathcal{T}(x^u;\theta_{\mathcal{T}_2})\right) \tag{12}$$

where $\mathcal{L}_{\text{CR}}$ represents the consistency regularization function, and $\text{KL}(\cdot)$ represents the KL-divergence.

3.5.2 Updating Teacher with Meta Learning. To learn a high-quality teacher classifier, we utilize the feedback performance of the student classifier on clean labeled samples to steer the teacher classifier to predict more superior pseudo labels of unlabeled instances, rather than keep the teacher classifier unchanged during the learning process. In particular, we iteratively update the student classifier based on the pseudo-labeled samples and update the teacher classifier conditioned on the performance of the student classifier. The teacher and student classifiers are trained in an alternating way until convergence. This alternating learning process between the teacher and student classifiers can be regarded as a bi-level optimization process.

In particular, the teacher classifier is optimized by encouraging the pseudo labels of unlabeled samples to be predicted in a manner that, if the student classifier was trained on the pseudo-labeled samples, the performance of the student classifier could be maximised on clean labeled instances. Mathematically, we update the teacher classifier based on the performance of the student classifier on the originally labeled data D_l:

$$\mathcal{L}_{\text{CE}}^{\mathcal{T}}(\mathcal{S}(x^l;\theta_{\mathcal{S}}),y^l) = -\frac{1}{N}\sum_{i=1}^{N} y_i^l \log[\mathcal{S}(x_i^l,\theta_{\mathcal{S}})] \tag{13}$$

The aforementioned loss function is fully differentiable with respect to $\theta_{\mathcal{T}}$ if we adopt the soft pseudo labels predicted by the teacher classifier. The standard back-propagation can be applied to

ALGORITHM 1: The proposed learning Algorithm.

Input: Labeled data D_l, Unlabeled data D_u, batch size n, m, max iterations I.
Output: Classifier parameter $\theta_{\mathcal{S}}^{(I)}$.

```
Initialize model parameters θ_T^(0) and θ_S^(0), and Noise Transition Matrix T^(0).
Pretrain the teacher model with θ_T^(0) on D_l.
for t = 1 to I do
    {x_i^l, y_i^l}_{i=1}^n ← SampleMiniBatch(D_l, n).
    {x_i^u}_{i=1}^m ← SampleMiniBatch(D_u, m).
    θ_{S_meta}^(t-1) ← Load Parameters θ_S^(t-1)
    Update θ_{S_meta}^(t) by Equation (9) on {x_i^u}_{i=1}^m.
    Update T^(t) by Equation (10) on {x_i^l, y_i^l}_{i=1}^n.
    Update θ_S^(t) by Equation (11) on {x_i^u}_{i=1}^m.
    Update θ_T^(t) by Equation (15) on {x_i^l, y_i^l}_{i=1}^n and {x_i^u}_{i=1}^m.
end for
```

obtain the gradient. Nevertheless, we employ hard pseudo labels to optimize the student classifier because leveraging hard pseudo labels would require fewer computational resources. We use the meta learning technique (i.e., MAML [39]) to update the teacher model adaptively:

$$\begin{aligned} \theta_{\mathcal{T}}^{*} &= \arg\min_{\theta_{\mathcal{T}}} -\frac{1}{N}\sum_{i=1}^{N} y_i^l \log[\mathcal{S}\left(x_i^l; \theta_{\mathcal{S}}^{*}(\theta_{\mathcal{T}})\right)], \\ \text{where} \quad \theta_{\mathcal{S}}^{*}(\theta_{\mathcal{T}}) &= \arg\min_{\theta_{\mathcal{S}}} -\frac{1}{M}\sum_{i=1}^{M} \tilde{y}_i^u \log[\mathcal{S}\left(x_i^u, \theta_{\mathcal{S}}\right)] \end{aligned} \tag{14}$$

Formally, after obtaining the parameters $\theta_{\mathcal{S}}^{(t+1)}$ of the student at the $t+1$-th iteration, we can obtain the parameters $\theta_{\mathcal{T}}^{(t+1)}$ as follows:

$$\theta_{\mathcal{T}}^{(t+1)} = \theta_{\mathcal{T}}^{(t)} - \beta \nabla_{\theta_{\mathcal{T}}} \left(\mathcal{L}_{\mathrm{CE}}^{\mathcal{T}}(\mathcal{S}(x^l; \theta_{\mathcal{S}}^{(t+1)}(\mathcal{T})), y^l) + \mathcal{L}_{\mathrm{CR}}^{\mathcal{T}}(\theta_{\mathcal{T}_1}^{(t)}, \theta_{\mathcal{T}_2}^{(t)}) \right) \tag{15}$$

where β denotes the learning rate of the meta pseudo supervision method. The student and teacher classifiers are optimized in an alternating manner until convergence. In summary, the whole dual meta-learning process for optimizing both student and teacher classifiers is provided in Algorithm 1.

4 EXPERIMENTAL SETUP

4.1 Datasets

We evaluate the effectiveness of DML by conducting a series of experiments on four semi-supervised text classification corpora: *AG News* [51], *Yelp* [51] and *Yahoo* [52]. Following prior works [53, 54], Labeled and Unlabeled data are randomly selected from the original training set, the number of Dev data is 2000 times the class and the provided official test sets are also adopted as our test sets. We provide the statistics of these four corpora in Table 1. The labeled/unlabeled/dev columns represent the maximum number of labeled/unlabeled/dev samples used in all our experiments. The above three kinds of samples are randomly selected from the original training sets, ensuring that there are overlapping samples in these three sets.

The AGNews [51] dataset is widely used for text classification, which contains news articles constructed from more than 2,000 news sources by the academic news search engine

Table 1. The Statistics of the Four Benchmark Corpora

Dataset	Class	Labeled	Unlabeled	Dev	Test
AGNews	4	10,000	20,000	8,000	7,600
Yelp	5	10,000	20,000	10,000	50,000
Yahoo	10	10,000	40,000	20,000	60,000
Amazon	5	10,000	100,000	20,000	60,000

ComeToMyHead. This corpus consists of 120,000 training instances and 7,600 testing instances, where each instance is a short text and a class label chosen from four categories.

The Yelp dataset [51] contains user reviews and the corresponding rating-based labels. In particular, each user review has a five-level label, which can be utilized for fine-grained sentiment classification. This corpus is collected by randomly selecting 130,000 training instances and 10,000 testing instances, where each instance has a rating score from 1 to 5. In total, there are 650,000 training instances and 50,000 testing instances.

The Yahoo [52] corpus contains the documents from Yahoo! Answers, which can be used for topic classification. This corpus is collected by utilizing the 10 largest key categories. There are 140,000 training instances and 6,000 testing instances for each class. In total, 1,400,000 training instances and 60,000 testing instances are adopted in the experiments. In this corpus, only the best answers and the corresponding categories are utilized.

The Amazon [51] dataset is constructed by randomly selecting 600,000 training reviews and 130,000 testing reviews for each score category from 1 to 5. In total, there are 3,000,000 training samples and 650,000 testing samples.

4.2 Baselines

We compare our DML method with several strong baseline methods for semi-supervised text classification to estimate the effectiveness of DML, which we describe below:

- The VAMPIRE [16] method leverages a variational autoencoder to pretrain a text encoder and utilizes the learned internal states as features in the downstream classifier. In the experiments, the default parameters in the original paper are utilized.
- The BERT [50] method yields impressive results in text classification. In particular, we adopt the average pooling over the output of the BERT encoder. An MLP layer is then applied to produce the output prediction.
- The VAT [36] method extends the adversarial and virtual adversarial training for text classification by employing perturbations into word embeddings rather than into original input sequences. The recurrent neural network is adopted as the text encoder.
- The UDA [37] method leverages advanced data augmentation methods, e.g., back-translation and RandAugment to substitute the previous noising strategies. The quality of the augmented data has a great impact on the performance of semi-supervised learning.
- The S^2TC-BDD [54] method leverages the margin loss and performs Gaussian linear transformation to obtain balanced label angle variances, so as to improve the performance of semi-supervised learning.
- The RNT [55] framework addresses semi-supervised text classification by utilizing uncertainty-based ranking of unlabeled texts and negative training with complementary labels, resulting in improved performance and outperforming existing alternatives in most scenarios. Since RNT employs some additional data augmentation strategies (e.g., back translation), we adopt the same experimental datasets and their corresponding data division as used in their released code so as to avoid reporting inaccurate results. For a fair comparison,

we adopt BERT$_{base}$ (bert-base-uncased) as the pre-trained language model, denoted as **RNT**$_{BERT}$. Consequently, as shown in Table 3, we only report the results of RNT with fixed unlabeled samples (i.e., 20,000 unlabeled samples) on the AGNews, Yelp and Yahoo datasets, along with different numbers of labeled samples.

4.3 Implementation Details

In the DML framework, the teacher and student classifiers have the same architecture but of independent weights. The proposed DML framework is model-agnostic and can be adopted by different deep neural models. Here, we choose BERT$_{base}$ [50] as our text encoder due to its impressive performance in text classification. We use the dev set to automatically tune the hyperparameters. The max input length is set to 256. We set the batch size to 32 as in RNT [55]. Dropout (dropout rate = 0.7) is applied to mitigate the overfitting problem. The learning rate γ is set to 1e-3. For the experiments using less than 1,000 labeled examples, we tune the learning rate β from {1e-5, 2e-5, 5e-5}, while we tune the learning rate β from {7e-5, 1e-4} when the number of labeled instances is greater than or equal to 1,000.

The SGD [56] optimizer is adopted to optimize the proposed DML method. It is noteworthy that we merely apply the student classifier to produce the results of the testing samples. All experiments are carried out on a 32G Tesla-V100 GPU. Additionally, we carry out experiments using five different random seeds while keeping hyperparameters consistent. We report the average results with standard deviation.

4.4 Evaluation Metrics

We evaluate the proposed DML method and baselines with two evaluation metrics: Micro-F1 and Macro-F1. Micro-F1 takes into account the number of each category in the evaluation, so it is suitable for evaluating the unbalanced data distribution. Macro-F1 averages the precision and recall over all categories regardless of the amount of data in each category.

5 EXPERIMENTAL RESULTS

5.1 Results by Varying Number of Unlabeled Data

Following previous works, we do experiments with a fixed number of labeled instances (i.e., 100), while different numbers of unlabeled instances (i.e., 200, 2,000, 20,000 for AGNews and Yelp; 400, 4,000, 40,000 for Yahoo) are adopted. The experimental results over the four datasets are summarized in Table 2. The proposed DML method consistently and substantially outperforms the baseline methods with different amounts of unlabeled instances. We can see that when labeled data and unlabeled data are very scarce, DML shows great advantages, especially on the Yelp dataset. For example, on the Yelp dataset, the highest improvement can reach 4.0%, and the average improvement can reach 2.68%. VAMPIRE, which is a small-scale model, performs poorly since it does not take full advantage of both unlabeled and labeled samples for training. BERT performs better than VAMPIRE, since BERT is a much bigger model than VAMPIRE and can learn more discriminative text representations than those from the VAE model used in VAMPIRE. The proposed DML performs even better than state-of-the-art baselines by using meta learning to improve the teacher and student classifiers simultaneously.

5.2 Results by Varying Number of Labeled Data

In addition, we conduct extensive experiments by utilizing a fixed number of unlabeled instances (i.e., 20,000 unlabeled instances for AGNews and Yelp, 40,000 unlabeled instances for Yahoo, and 100,000 unlabeled instances for Amazon) along with varying numbers of labeled instances (i.e.,

Table 2. Experimental Results (Micro-F1 and Macro-F1) with Fixed 100 Labeled Samples along with Different Numbers of Unlabeled Samples (i.e., 200, 2,000, 20,000 Unlabeled Samples on AGNews and Yelp; 400, 4,000, 40,000 Unlabeled Samples on Yahoo, and 1,000, 10,000, 100,000 Unlabeled Samples on Amazon)

		Micro-F1(%)			Macro-F1(%)		
Dataset	*Methods/# N*	*200*	*2000*	*20000*	*200*	*2000*	*20000*
AGNews	BERT	85.5	85.6	85.6	85.5	85.5	85.6
	VAMPIRE	32.9	42.1	70.5	21.9	34.1	69.8
	VAT	85.0	87.0	86.8	84.5	87.0	86.7
	UDA	84.4	85.3	85.5	84.3	85.2	85.5
	S^2TC-BDD	85.7	86.3	87.2	85.7	86.4	87.2
	DPS	87.0	87.8	88.7	86.9	87.7	88.8
	DML (Ours)	**88.1± 0.3**	**89.0± 0.2**	**89.5± 0.3**	**88.1± 0.2**	**89.0± 0.3**	**89.5± 0.3**
Yelp	BERT	38.5	39.3	39.9	37.0	37.9	37.1
	VAMPIRE	23.8	21.1	22.7	16.1	12.4	14.4
	VAT	29.9	29.4	24.4	27.8	28.7	19.7
	UDA	39.7	37.9	38.7	34.4	36.2	35.7
	S^2TC-BDD	40.3	41.7	41.7	37.2	38.0	40.3
	DPS	45.4	46.8	48.5	41.7	44.5	46.0
	DML (Ours)	**47.9 ± 1.7**	**49.8 ± 1.2**	**50.5 ± 0.9**	**45.7 ± 1.6**	**46.9 ± 1.3**	**48.2 ± 0.8**
Dataset	*Method /# N*	*400*	*4000*	*40000*	*400*	*4000*	*40000*
Yahoo	BERT	58.2	58.4	58.9	57.1	57.4	57.3
	VAMPIRE	16.2	22.1	38.9	7.4	17.5	35.6
	VAT	51.9	52.3	53.4	52.1	52.4	54.2
	UDA	50.8	55.9	57.6	50.0	55.0	56.7
	S^2TC-BDD	59.3	59.8	61.8	58.6	59.0	59.5
	DPS	60.4	62.0	63.2	59.8	61.1	62.4
	DML (Ours)	**61.0 ± 0.3**	**62.6 ± 0.2**	**64.7 ± 0.2**	**60.5 ± 0.3**	**61.7 ± 0.2**	**64.5 ± 0.2**
Dataset	*Method /# N*	*1000*	*10000*	*100000*	*1000*	*10000*	*100000*
Amazon	BERT	34.1	34.5	35.2	32.0	33.2	33.6
	VAMPIRE	20.2	21.6	22.3	11.4	12.2	12.9
	VAT	28.8	29.7	29.0	26.4	27.5	27.1
	UDA	37.0	38.3	39.5	36.5	37.7	38.1
	S^2TC-BDD	38.2	40.3	40.7	34.3	35.2	35.8
	DPS	39.5	41.2	42.2	33.7	39.2	39.0
	DML (Ours)	**41.9 ± 0.7**	**42.8 ± 0.9**	**44.7 ± 0.8**	**35.7 ± 0.8**	**40.9 ± 0.8**	**42.5 ± 0.7**

The best results are highlighted in boldface.

100, 1,000, 10,000 labeled instances). We summarize the experimental results on the four evaluation datasets in Table 3. Generally speaking, the proposed DML consistently outperforms the baseline methods in terms of both metrics on the four datasets. S^2TC-BDD substantially obtains better performance than the pre-trained models (e.g., BERT and VAMPIRE) in terms of Micro-F1 and Macro-F1 scores by a large margin, especially when the labeled instances are limited. DPS, which is our preliminary method, performs better than other baselines by adaptively updating the teacher classifier to predict more accurate pseudo labels. The RNT_{BERT} method exhibits excellent performance by using 1,000 and 10,000 labeled training instances, comparable to the DPS approach. However, its performance diminishes when trained on the datasets with only 100 labeled instances. This discrepancy may be attributed to the intrinsic characteristics of the RNT method, which incorporates random noise and may prove less efficacious when confronted with a limited number of labeled

Table 3. Experimental Results (Micro-F1 and Macro-F1) with Fixed Unlabeled Samples (i.e., 20,000 Unlabeled Samples on AGNews and Yelp; 40,000 Unlabeled Samples on Yahoo, and 100,000 Unlabeled Samples on Amazon) along with Different Numbers of Labeled Samples (i.e., 100, 1,000, 10,000 Labeled Samples)

		Micro-F1(%)			Macro-F1(%)		
Dataset	*Method /# M*	*100*	*1000*	*10000*	*100*	*1000*	*10000*
AGNews	BERT	83.9	87.8	90.5	84.0	87.8	90.5
	VAMPIRE	70.5	83.3	87.6	69.8	83.3	87.6
	VAT	86.8	88.6	89.8	86.7	88.6	89.7
	UDA	85.5	88.3	90.6	85.5	88.3	90.6
	S^2TC-BDD	87.2	88.9	90.7	87.2	88.9	90.7
	DPS	88.7	89.9	91.8	88.8	89.9	91.8
	RNT_{BERT}	85.8 ± 0.3	89.4 ±0.1	91.9 ± 0.1	85.8 ± 0.3	89.4 ± 0.1	91.9 ± 0.1
	DML (Ours)	**89.5± 0.3**	**90.6± 0.1**	**92.7± 0.1**	**89.5± 0.3**	**90.6± 0.1**	**92.7± 0.1**
	ChatGPT	78.63 *(zero-shot)*			76.62 *(zero-shot)*		
Yelp	BERT	34.4	53.8	58.3	32.4	53.2	58.6
	VAMPIRE	22.7	47.6	55.1	14.4	47.6	55.3
	VAT	24.4	55.1	56.6	19.7	54.8	56.9
	UDA	38.7	55.4	58.0	35.7	55.0	57.6
	S^2TC-BDD	41.7	55.2	58.3	40.3	55.0	58.6
	DPS	48.5	55.4	60.4	45.9	55.7	60.5
	RNT_{BERT}	44.1 ± 1.2	56.9 ±0.6	60.4 ± 0.1	43.2 ± 1.2	56.6 ±0.6	60.2 ± 0.1
	DML (Ours)	**50.5 ± 0.9**	**58.3 ± 0.4**	**62.7 ± 0.1**	**48.2 ± 0.8**	**58.1 ± 0.3**	**62.5 ± 0.1**
	ChatGPT	48.48 *(zero-shot)*			39.53 *(zero-shot)*		
Yahoo	BERT	56.4	67.6	71.3	55.0	67.1	70.8
	VAMPIRE	38.9	54.7	64.4	35.6	54.5	64.4
	VAT	53.4	68.5	70.1	54.2	67.5	69.7
	UDA	57.6	67.2	70.7	57.6	66.6	70.4
	S^2TC-BDD	61.8	68.7	71.3	59.5	68.0	70.9
	DPS	63.2	69.1	72.0	62.4	68.6	71.7
	RNT_{BERT}	62.5 ± 0.3	69.2 ±0.2	72.9 ± 0.1	62.1 ± 0.3	69.1 ±0.2	72.7 ± 0.1
	DML (Ours)	**64.7 ± 0.2**	**70.5± 0.2**	**74.8 ± 0.2**	**64.5 ± 0.2**	**70.3 ± 0.2**	**74.6 ± 0.2**
	ChatGPT	48.49 *(zero-shot)*			42.98 *(zero-shot)*		
Amazon	BERT	35.2	49.8	54.7	33.6	48.9	53.6
	VAMPIRE	22.3	42.5	50.3	12.9	41.2	50.1
	VAT	29.0	50.8	52.1	27.1	49.9	51.8
	UDA	39.5	50.4	54.4	38.1	49.7	52.5
	S^2TC-BDD	40.7	50.3	54.5	35.8	49.9	52.8
	DPS	42.2	52.9	55.1	39.0	52.9	54.7
	DML (Ours)	**44.7± 0.7**	**54.6 ± 0.5**	**58.3 ± 0.3**	**42.5± 0.7**	**54.3± 0.5**	**58.8 ± 0.3**
	ChatGPT	41.92 *(zero-shot)*			37.19 *(zero-shot)*		

The best results are highlighted in boldface.

instances. It can be observed that the Micro-F1 and Macro-F1 scores exhibit a gradual increase as the number of labeled samples in the training dataset progressively grows from 100 to 10,000.

5.3 Results by ChatGPT

Due to the current prominence of large-scale models (LLMs) [57, 58], we also adopt ChatGPT (gpt-3.5-turbo) in the zero-shot setting as a strong baseline. Notably, we adopt the entire test set for AGNews, which has 7,600 samples. However, due to resource constraints imposed by ChatGPT API, we randomly select 15,000 test samples from each of the other three datasets (Yelp, Yahoo and

Table 4. Prompts for ChatGPT Invocation and Test Set Sample Sizes in AGNews, Yelp, Yahoo, and Amazon Datasets

Dataset	Test Sample	Prompt
AGNews	7,600 (full)	Classify the following news into World(1), Sports(2), Business(3) and Sci/Tech(4), only return the number:
Yelp	15,000	Classify the follow review into star 1-5 different sentiment levels from negative to positive, only return the number:
Yahoo	15,000	Classify the follow topic into Society & Culture(1), Science & Mathematics(2), Health(3), Education & Reference(4), Computers & Internet(5), Sports(6), Business & Finance(7), Entertainment & Music(8), Family & Relationships(9) and Politics & Government(10), only return the number:
Amazon	15,000	Classify the follow review into star 1-5 different sentiment levels from negative to positive, only return the number:

Amazon) to evaluate the effectiveness of ChatGPT, similar to previous work [59, 60]. In particular, we formulate prompts for each task corresponding to the respective datasets, as illustrated in Table 4. Although we constrain the prompts only to elicit numerical responses, the output generated by the model may still exhibit variability. For instance, in the case of the Yelp dataset, ChatGPT may return "5 (positive)" with its own interpretation. To address this issue, we extract the numerical value from the response and restrict it to fall within the range of total categories. If a non-numeric value or a value outside the permissible range is yielded, we assign it a value of -1.

As shown in Table 3, ChatGPT exhibits excellent zero-shot performance on sentiment classification. Specifically, ChatGPT in the zero-shot setting performs slightly worse than our DML method with a small number of labeled training samples (i.e., 100 labeled samples) on the Yelp and Amazon datasets. However, noteworthy distinctions emerge in the context of other text classification tasks, particularly evident in the case of a 10-category classification task such as Yahoo. In this case, the performance of ChatGPT is approximately 20% inferior to that of DML. This discrepancy could be attributed to ChatGPT's potential deficiency in specific domain knowledge or fine-grained category distinctions required for accurate classification in such tasks.

5.4 Ablation Study

An ablation study is performed to analyze the impact of each component in DML. We provide the results of the models in terms of removing the meta pseudo supervision component (denoted as w/o MPS), removing the meta noise correction component (called w/o MNC)[1], and removing both MPS and MNC (called w/o MPS+MNC). We report the ablation test results in Tables 5–8 for AGNews, Yelp, Yahoo and Amazon, respectively. In particular, for the AGNews and Yelp datasets, we utilize a combination of 100 labeled instances and 200 unlabeled instances, while for the Yahoo corpora, we employ 100 labeled instances and 400 unlabeled instances and for the Amazon corpora, we employ 100 labeled instances and 1,000 unlabeled instances.

Generally speaking, from the results, we can observe that combining both meta pseudo supervision and meta noise correction achieves the best performance on the four datasets. Concretely, the results decrease sharply when removing the meta pseudo supervision (DPS). This is within our

[1]The reported results of DPS and DML w/o MPS are different since they use different experimental settings such as the learning rate and batch sizes. Specifically, for the DML method, we explore a broader range of hyperparameter combinations in order to investigate its performance thoroughly.

Table 5. Ablation Test Scores on AGNews with 100 Labeled Samples and 200 Unlabeled Samples

Model	Micro-F1(%)	Macro-F1(%)
DML	**88.1**	**88.1**
DML w/o pretrain	87.0	86.9
DML w/o MPS	86.7	86.7
DML w/o MNC	87.0	86.9
DML w/o MPS+MNC	85.4	85.3

"MPS" means meta pseudo supervision and "MNC" means meta noise correction.

Table 6. Ablation Test Results (Micro-F1 and Macro-F1 Scores) on Yelp with 100 Labeled Samples and 200 Unlabeled Samples

Model	Micro-F1(%)	Macro-F1(%)
DML	**47.9**	**45.7**
DML w/o pretrain	47.2	45.0
DML w/o MPS	46.3	44.6
DML w/o MNC	45.4	41.7
DML w/o MPS+MNC	44.8	40.7

"MPS" means meta pseudo supervision and "MNC" means meta noise correction.

Table 7. Ablation Test Results (Macro-F1 and Micro-F1) on Yahoo with 100 Labeled Samples and 400 Unlabeled Samples

Model	Micro-F1(%)	Macro-F1(%)
DML	**61.0**	**60.5**
DML w/o pretrain	60.3	59.8
DML w/o MPS	59.7	59.5
DML w/o MNC	60.4	59.8
DML w/o MPS+MNC	57.4	57.2

"MPS" means meta pseudo supervision and "MNC" means meta noise correction.

Table 8. Ablation Test Results (Macro-F1 and Micro-F1) on Amazon with 100 Labeled Samples and 1,000 Unlabeled Samples

Model	Micro-F1(%)	Macro-F1(%)
DML	**41.9**	**35.7**
DML w/o pretrain	41.1	34.8
DML w/o MPS	41.0	34.2
DML w/o MNC	39.5	33.7
DML w/o MPS+MNC	37.3	36.7

"MPS" means meta pseudo supervision and "MNC" means meta noise correction.

expectation since iteratively updating the teacher classifier based on the feedback signal from the student classifier can improve the quality of the generated pseudo labels of the unlabeled samples. For instance, for the AGNews dataset, the Micro-F1 score drops 1.6% when removing the meta pseudo supervision. In addition, we can observe that the meta noise correction also contributes to the effectiveness of the proposed DML. For example, for the Yelp dataset, the Macro-F1 score drops

by 4.0% when removing the meta noise correction. This verifies the effectiveness of the meta noise correction in improving the student classifier by rectifying the noisy pseudo labels.

We also examine the overall impact of the MNC approach. Our findings indicate that the effectiveness of MNC becomes more pronounced when the boundaries between labels are more blurred. For instance, on hierarchical classification datasets such as Yelp and Amazon, we observe an average decrease in both Micro and Macro scores by approximately 2.45% and 3%, respectively. These results align with our expectations, as MNC excels at estimating the transfer of noise between categories in such scenarios. These ablation results provide valuable insights into the effectiveness of different components in our proposed DML framework, shedding light on the nuanced impacts associated with their integration.

We sought to examine the impact of pre-training by conducting ablation tests in which we removed the pre-training phase (denoted as DML w/o pretrain), utilizing the original BERT model for subsequent operations. Our analysis reveals a slight decrease in the performance of the DML model when the pre-training phase is omitted. This decline in performance signifies the significance of the pre-training phase as a crucial contributing factor. It can be attributed to the limited availability of labeled data, which allows the DML model to focus on acquiring task-specific features. The absence of pre-training potentially hampers the model's ability to learn effective representations from a smaller labeled dataset. Conversely, the pre-trained model serves as a valuable resource by providing an initial state for the DML model, enabling it to adapt more effectively to the specific task at hand. This initialization with pre-trained weights facilitates the transfer of knowledge and promotes faster convergence during training. By leveraging the pre-trained model as a starting point, the DML model benefits from pre-learned representations that capture general language patterns and structures.

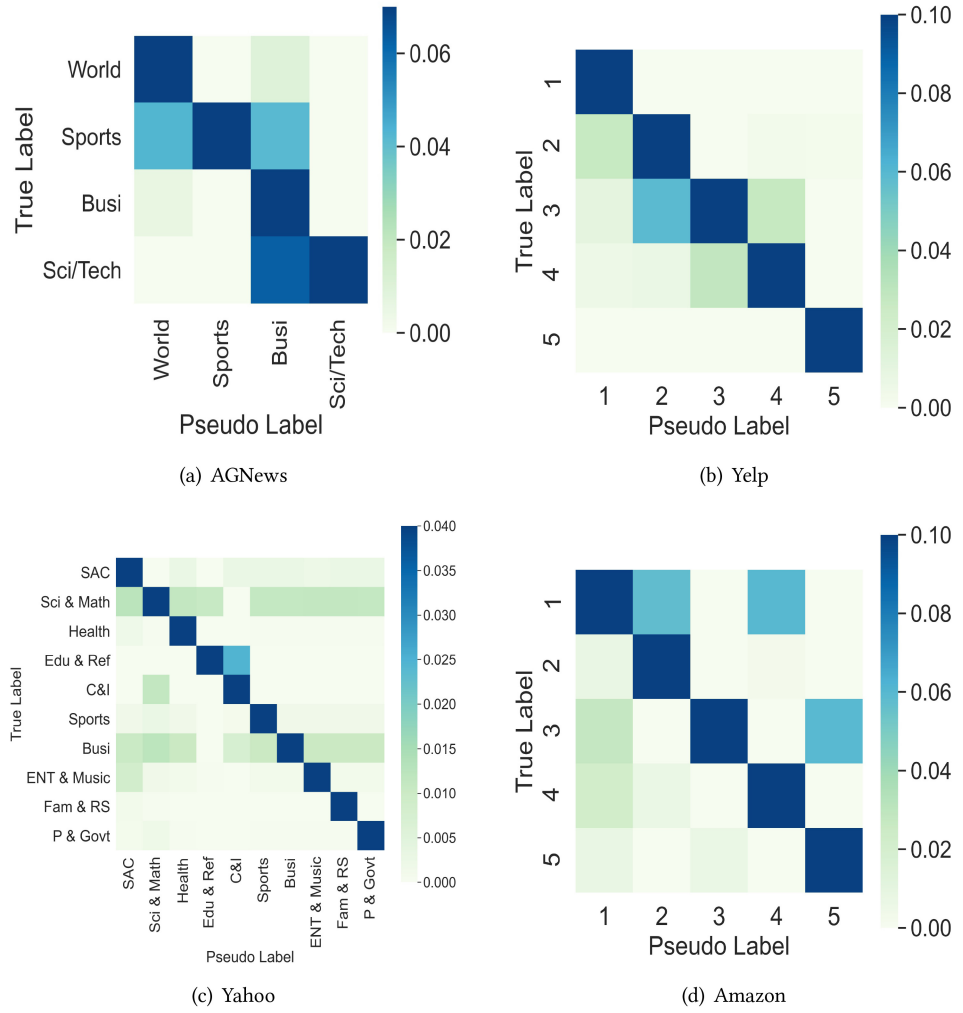
5.5 Qualitative Analysis

To investigate the effectiveness of the proposed DML qualitatively, we utilize several exemplary cases from the four datasets to illustrate the effectiveness of meta noise correction and meta pseudo supervision.

5.5.1 Effectiveness of Meta Noise Correction. We illustrate the learned Noise Transition Matrix (NTM) on four experimental datasets in Figure 3, from which we can observe the similarity relationship between label categories from the learned NTM.

From Figure 3(a) we can observe that the SPORTS category is easily transferred to the WORLD and BUSINESS categories, since the SPORTS news articles are often closely related to those of the WORLD and BUSINESS categories. Taking the instance AG_EX.1 in Table 9 as an example. The true label of AG_EX.1 is SPORTS, but the pseudo label predicted by the teacher classifier is WORLD since many country names are mentioned in the sentence. On the contrary, the label of AG_EX.2 in Table 9 can be easily predicted as WORLD since no sport related words appear in the sentence. Hence, the SPORTS category can be easily misclassified as WORLD and BUSINESS, but not vice versa. In addition, the TECHNOLOGY category is easily transferred to the BUSINESS category, but not vice versa. As shown by AG_EX.3 in Table 9, the words "business" and "consumers" may make the teacher classifier incorrectly predict the BUSINESS category, while the gold label should be TECHNOLOGY by considering the words "computer associates international Inc." and "antispyware software".

From Figure 3(b) we see that the user reviews in the Yelp dataset can be easily shifted from NEUTRAL to NEGATIVE. Generally, the reviews of 1-star and 5-star scores are quite different and determined, while the user's reviews with 2-star and 3-star scores are often ambiguous and easy to be confused. This phenomenon is well illustrated by Yelp_EX in Table 9.



(a) AGNews (b) Yelp

(c) Yahoo (d) Amazon

Fig. 3. Estimated Noise Transition Matrix on the four datasets.

From Figure 3(c) we can observe that the EDUCATION category is easily transferred to the INTERNET category. For example, the pseudo label of Yahoo_EX1 predicted by the teacher classifier is ED & REF, while the rectified label is C& I. This verifies that the estimated NTM can reduce the interference of noisy pseudo labels of unlabeled data. We can also observe that the category FAM& RS is the most different category, which has limited similarity with other categories. Yahoo_EX2 in Table 9 shows that the sentence mentions "family" multiple times and is not related to other categories, which could be a very good anchor data. From Figure 3(d) we observe that the user reviews in the Amazon dataset can be easily shifted to POSITIVE. This phenomenon is aptly exemplified by Ama_EX as shown in Table 9.

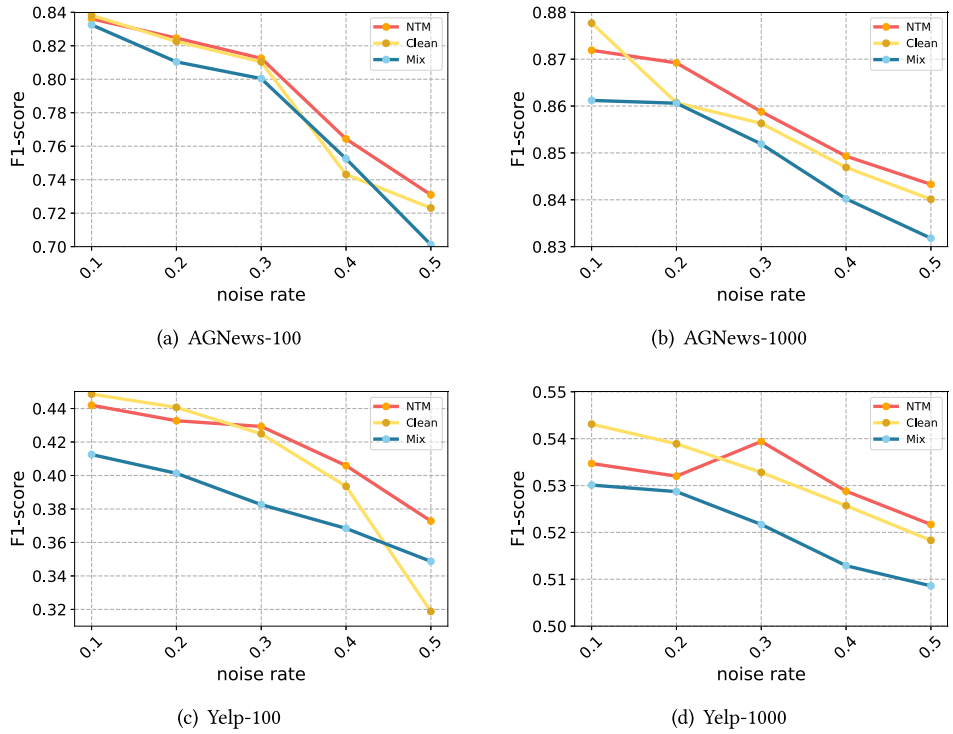
5.5.2 Robust of Noise Transition Matrix. To gain a more intuitive understanding of the Noise Transition Matrix employed in **MNC (meta noise correction)** and to assess its robustness, we introduced a controlled proportion of symmetric noise into the datasets, effectively substituting the noise attributed to pseudo labels. To ensure the elimination of confounding factors that could

Table 9. Some Examples with Gold Labels, Rectified Labels Learned by Meta Noise Correction, and Pseudo Labels Predicted by the Teacher Classifier

	Sentence	Gold label	Rectified label	Pseudo label
AG_EX.1	"**Australian swimmer** Ian Thorpe beat arch-rival Michael Phelps in **the men's 200-meter freestyle** on Monday as **the United States** pursued **China**, **Australia** and **Japan** in the medals table on day three of **the Olympic Games**."	Sports	Sports	World
AG_EX.2	"**UN** tries to keep up dialogue after **Khartoum** irked by **Darfur** criticism."	World	World	World
AG_EX.3	"WASHINGTON - **Computer Associates International** Inc. (CA) said Monday it **acquired** PestPatrol Inc., a firm marketing **antispyware software** to enterprises, small **businesses** and individual **consumers**."	Tech	Tech	Business
Yelp_EX	"Talk about **overpriced**. $18 for a fairly basic pasta with some **obviously** frozen chicken chopped up over it. The pasta itself was **ok, as was the sauce**. The desserts are **pretty good**. But honestly, that is a $10 dish whose price has **been inflated**."	3	3	2
Yahoo_EX1	" What's the longest English word without **a vowel** in it? And what does that word mean if it's not a common word. The longest word without **a vowel** is Rhythm. It is referenced on below **web site** where you can **find** more fun facts."	ED&REF	ED&REF	C&I
Yahoo_EX2	"What is the difference between a close knit **family** and extended **family**? A close knit **family** is simply a **family** that is close–it could apply to a nuclear **family (mom, dad, kids)** or extended **family** (cousins, uncles, grandparents). An extended **family** describes a **family** unit that extends beyond mom, dad, and kids to include cousins, uncles/aunts, and grandparents."	Fam&RS	Fam&RS	Fam&RS
Ama_EX	"You **will not be able to** build a timber frame house, This book vaguely breezed across history and went into various types of timber structures. The mortise and tenon joints that were illustrated were **fascinating but not practical for** a home builder."	1	1	2

The red words are supposed to lead to the incorrect classification of the examples and the green words may facilitate the correct classification of the examples.

potentially impact the experimental outcomes, our study exclusively employed noise-injected labeled data and employed a single model for comparative analysis. Specifically, we randomly sampled 100 and 1,000 labeled data from the AGNews and Yelp datasets respectively, and injected varying proportions of symmetric noise [61] ($\alpha = 10\%, 20\%, 30\%, 40\%, 50\%$) into the training data. In the context of symmetric noise, it is pertinent to note that a sample containing an inaccurate label exhibits independence with respect to both the feature vector and the actual class affiliation of the sample. We reduced the initial number of iteration steps, while keeping the original parameter settings unchanged. Given the noise level α and n labeled data points, we conducted the following experiments to evaluate the effectiveness of the Noise Transition Matrix (NTM) depicted in Figure 4: (1) Only the remaining $(1-\alpha)\cdot n$ clean data except the noise data were used for training, denoted as "***Clean***". (2) Experiments were conducted on the entire set of n labeled data, including both $(1-\alpha)\cdot n$ clean instances and $\alpha \cdot n$ noise-injected instances, denoted as "***Mix***". (3) Building upon the experiment (2), the $\alpha \cdot n$ noise labels were rectified using the Noise Transition Matrix employed in MNC, denoted as "***NTM***".



(a) AGNews-100 (b) AGNews-1000

(c) Yelp-100 (d) Yelp-1000

Fig. 4. Robustness testing with noise injection on 100 and 1,000 labels for AGNews and Yelp.

The results in Figure 4 illustrate the effectiveness of ***NTM*** and the impact of increasing noise levels on the performance of different training methods. It is evident that the ***Mix*** training approach, incorporating noise-labeled data, exhibits the poorest performance. This decline in performance can be attributed to the interference caused by the presence of noise during the ***Mix*** training process, resulting in label misinformation that hampers the learning process of the model. Notably, in scenarios where the availability of samples is limited (i.e., 100), as depicted in Figures 4(a) and 4(c), a substantial increase in noise leads to a sudden drop of over 10% in the evaluation. Conversely, the ***NTM*** method demonstrates comparable performance to ***Clean*** when the noise levels are low. As the noise levels increase and the availability of labeled data decreases, ***NTM*** showcases its distinctive advantage by effectively leveraging the noisy data for repair purposes. Consequently, ***NTM*** surpasses the performance of using solely labeled data. Moreover, as the number of samples increases, the repair effect of ***NTM*** becomes evident at an earlier stage. For instance, in the case of the AGNews dataset, the performance of ***NTM*** slightly surpasses the ***Clean*** at a noise rate of 0.3 when considering 100 labeled samples. However, when 1,000 labeled samples are utilized, the NTM method significantly outperforms the ***Clean*** at a lower noise rate of 0.2. This observed phenomenon can be attributed to ***NTM***'s ability to model and repair noise. By extracting valuable information from noisy data, ***NTM*** enhances the model's robustness and generalization capabilities. The ***NTM*** employed in MNC effectively utilizes the noisy data to mitigate the adverse effects of label noise, thereby improving overall performance in scenarios with limited labeled data.

5.5.3 Adaptability of Meta Noise Correction. The quality of pseudo-labels plays a pivotal role in influencing the performance of the student model. The traditional threshold-based method offers

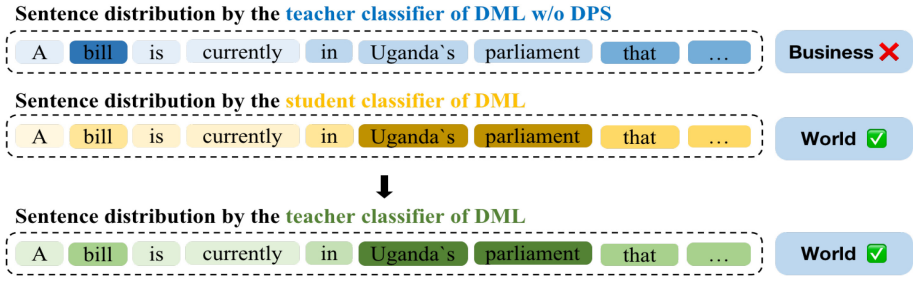
Table 10. Exploring NTM and Threshold-Based Approaches with Varied Pseudo-Label Thresholds (95%, 90%, 80%, and 70%) on Yelp Dataset across Different Data Scales, while Maintaining a Fixed 100 Labeled Data Points and Varying Unlabeled Data Points from 20,000 to 200

M			20,000	2,000	200
Micro-F1(%)	DML		**50.5**	**49.8**	**47.9**
	threshold	95%	49.3 (↓ 1.2)	47.2 (↓ 2.6)	45.6 (↓ 2.3)
		90%	49.4 (↓ 1.1)	47.0 (↓ 2.8)	46.3 (↓ 1.6)
		80%	49.8 (↓ 0.7)	47.4 (↓ 2.4)	45.1 (↓ 2.8)
		70%	49.6 (↓ 0.9)	47.3 (↓ 2.5)	44.9 (↓ 3.0)
Macro-F1(%)	DML		**48.2**	**46.9**	**42.8**
	threshold	95%	47.3 (↓ 0.9)	44.1 (↓ 2.8)	40.1 (↓ 2.7)
		90%	47.5 (↓ 0.7)	43.5 (↓ 3.4)	41.2 (↓ 1.6)
		80%	46.5 (↓ 1.7)	45.0 (↓ 1.9)	38.8 (↓ 4.0)
		70%	46.9 (↓ 1.3)	43.8 (↓ 3.1)	40.4 (↓ 2.4)

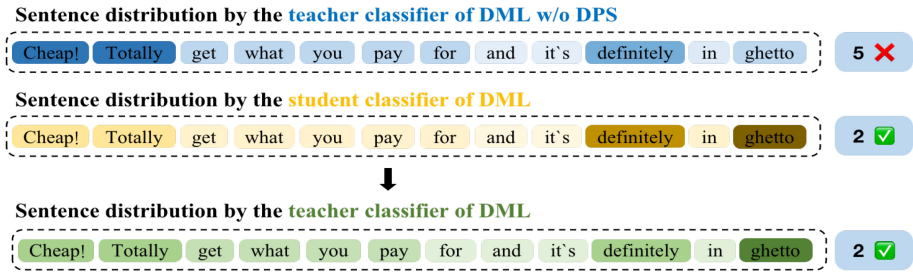
a straightforward approach to filtering pseudo-labels [6, 62]. The core principle of this technique is that a prediction is only considered as a pseudo-label if the model's confidence in predicting an unlabeled sample surpasses a predetermined threshold. To further refine this process, our methodology incorporates the application of a Noise Transition Matrix employed in MNC, which characterizes the noise relationship between authentic labels and pseudo-labels. Leveraging this matrix enables the correction of noise within pseudo-labels, while also facilitating mutual interactions with the classifier, ultimately resulting in enhanced accuracy. We conducted an empirical comparison of both methods on the Yelp dataset across varying data scales, with a fixed 100 labeled instances and a range of 20,000 to 200 unlabeled instances. We also explore different pseudo-label threshold values, including 95%, 90%, 80%, and 70%. The results are presented in Table 10. It is worth noting that in these comparative experiments, we keep all other variables constant, and the only difference lies in whether we used NTM or just different threshold values.

In accordance with the results presented in Table 10, it becomes evident that the Noise Transition Matrix (NTM) yields a highly favorable impact on the performance of the student model, and increasing the threshold value does not necessarily yield improved results. On the contrary, excessively high threshold values are associated with a conspicuous degradation in performance. For instance, when employing a 95% threshold with 200 unlabeled data points, it resulted in a decrease of 2.3% in Micro-F1 and a decrease of 2.7% in Macro-F1. Furthermore, it is worth noting that the optimal threshold value is not fixed across varying data volumes. For instance, when there were 20,000 or 2,000 unlabeled data points, the optimal threshold for the best performance was 80%. However, in the case of 200 unlabeled data points, an 80% threshold resulted in a significant decrease, with a maximum drop of 4% observed in Macro-F1. The remarkable adaptability of NTM allows it to autonomously attain the most favorable results, thus demonstrating its ability to dynamically adjust and perform optimally in response to differing data magnitudes.

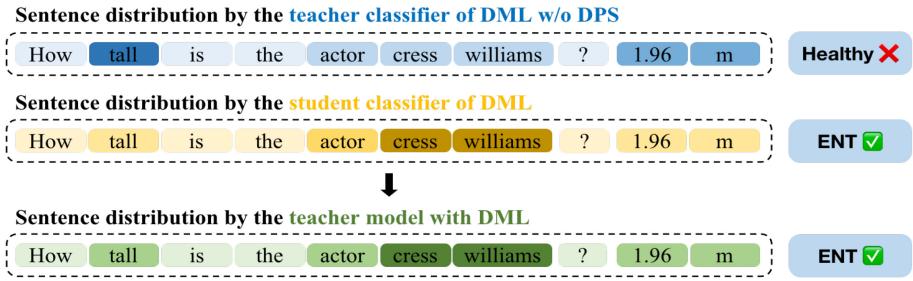
5.5.4 Effectiveness of Meta Pseudo Supervision. We choose four samples from the four experimental datasets to investigate the effectiveness of the meta pseudo supervision method. These samples are correctly predicted by the teacher and student classifiers of DML but incorrectly classified by the teacher classifier of DML w/o MPS. Figure 5 provides the attention weights of the examples.



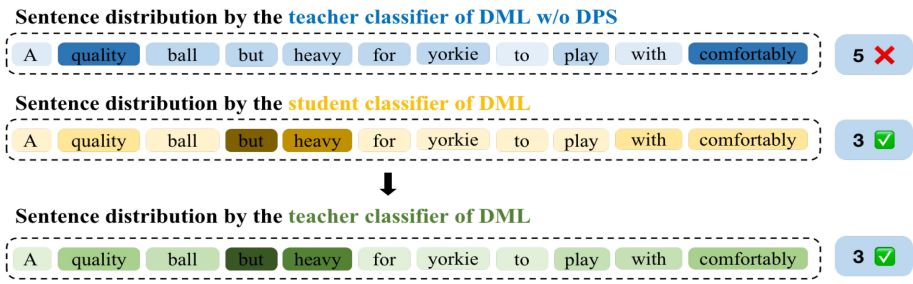
(a) The attention scores for a sentence "*A bill is currently in Uganda's parliament that would strengthen women's rights.*" from the AGNews dataset.



(b) The attention scores for a sentence "*Cheap. Totally get what you pay for. And it's definitely in the ghetto.*" from the Yelp dataset.



(c) The attention scores for a sentence "*how tall is the actor cress williams? 1.96m*" from the Yahoo dataset.



(d) The attention scores for a sentence "*A quality ball but heavy for yorkie to play with comfortably.*" from the Amazon dataset.

Fig. 5. The attention scores during MPS for different sentences from four datasets. The color depth represents the importance degree of the words.

Figure 5(a) shows the attention scores of different methods for a sentence chosen from the AGNews test set. We can observe that the teacher classifier of DML w/o MPS pays much attention to the word "bill" that is closely related to the BUSINESS category, resulting in incorrect prediction. While the student and teacher classifiers of DML can extract the important words "Uganda's parliament" and thus obtain the correct prediction WORLD. This can to some extent verify the effectiveness of the meta pseudo supervision method, making the teacher classifier produce better pseudo labels of unlabeled samples.

Figure 5(b) demonstrates the attention scores of different methods for a sentence chosen from the Yelp test set. We can observe that the teacher classifier of DML w/o MPS tends to focus on the words "Cheap!", "totally" and "definitely", leading to incorrect prediction of 5-star. While both the teacher and student classifiers of DML can detect the discriminative word "ghetto" describing a restaurant and get the correct attitude towards the negative sentiment.

Figure 5(c) illustrates the attention scores of different methods for a sentence chosen from the Yahoo test set. As shown in Figure 5(c), the teacher classifier of DML w/o MPS tends to pay much attention to the words "tall" and "1.96m", which are closely related to physical examination and results in predicting the incorrect category HEALTHY. While both the teacher and student classifiers of DML successfully capture the essential words "actor", "cress" and "williams", resulting in the correct prediction "ENTERTAINMENT". The aforementioned example verifies the effectiveness of meta pseudo supervision and meta noise correction methods.

Figure 5(d) visually represents the attention scores assigned by different algorithms for a specific instance from the Amazon test set. It is noteworthy that in the case of DML w/o MPS, the teacher classifier exhibits a strong emphasis on the words "quality" and "comfortably", which are particularly relevant to the physical examination aspect. Regrettably, this excessive focus on these words results in the misclassification of the sentiment label as category 5. Conversely, the DML model, which incorporates the enhanced teacher and student classifiers, demonstrates the ability to accurately identify the crucial words "but" and "heavy", consequently leading to the correct prediction of category 3. This example serves as compelling evidence that highlights the effectiveness of the meta pseudo supervision employed within our DML framework.

5.6 Hyperparameter Analysis

We analyze the impact of the learning rates γ and β for the meta noise correction and the meta pseudo supervision methods, respectively. We conduct experiments on the Yelp dataset with 100 labeled and 200 unlabeled instances. In the experiments, we report the Macro-F1 and Micro-F1 scores on the Yelp dataset by varying the the learning rates γ and β from 1e-5 to 1e-2.

Figure 6 illustrates the experimental results by varying the student learning rate γ. The proposed DML method obtains the best results when γ =1e-3. As γ increases from 1e-5 to 1e-2, the Macro-F1 and Micro-F1 scores grow gradually till we achieve the optimal values, after which the results decrease sharply. The learning rate β shows the similar trend. Figure 7 demonstrates the experimental results by varying the teacher learning rate β. The proposed DML method obtains the best results when β =1e-4. As β increases from 1e-5 to 1e-2, the Macro-F1 and Micro-F1 scores grow slightly till β =1e-4, after which the results decrease sharply. From the results we can observe that the optimal value of γ is much larger than that of β, since γ is used to update the NTM whose parameters are much smaller than the parameters of the student and teacher classifiers.

6 CONCLUSION

In this paper, we proposed a dual meta learning (DML) technique for semi-supervised text classification (SSTC), which improved the pseudo-labeling via meta learning. The proposed DML attempted to improve the teacher and student classifiers in pseudo-labeling simultaneously.

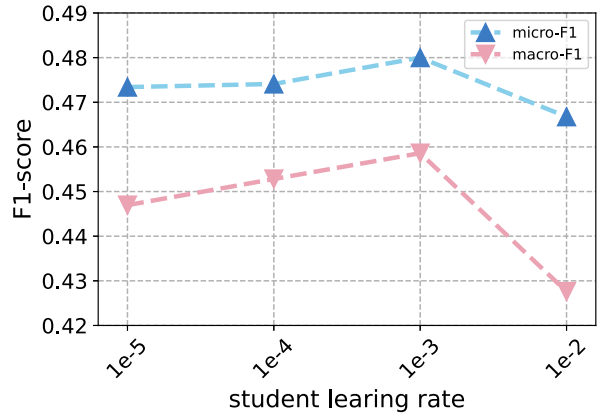


Fig. 6. The macro F1-score and micro F1-score by varying student learning rate γ from 1e-5 to 1e-2. Here, we fix the teacher learning rate and set β=1e-4.

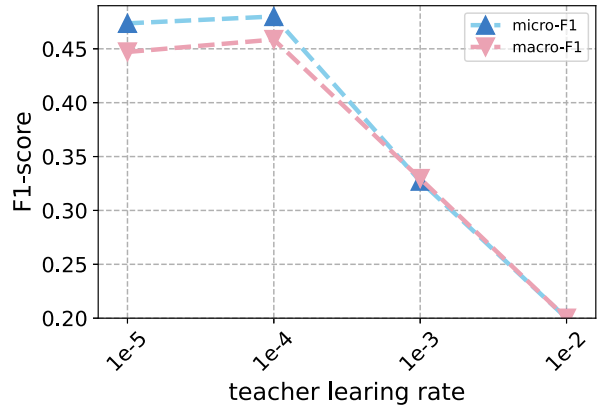


Fig. 7. The macro F1-score and micro F1-score by varying teacher learning rate β from 1e-5 to 1e-2. Here, we fix the student learning rate and set γ=1e-3.

Concretely, we introduced a meta noise correction method to update the student classifier by proposing a Noise Transition Matrix (NTM) with meta-learning to mitigate the noisy issues in pseudo labels. In addition, we also devised a meta pseudo supervision method to improve the teacher classifier, which utilized the feedback performance from the student classifier to further steer the teacher classifier to produce more accurate pseudo labels for the unlabeled data. Extensive experiments on four SSTC corpora showed that the proposed DML method achieved consistently and substantially better performance than the strong baseline methods.

In the future, we plan to explore the prompting techniques to exploit the large-scale pre-trained language models in an effective way. In addition, we also would like to exploit the variational information bottleneck principle to reduce the spurious correlations between the input features and output prediction so as to mitigate the overfitting issue in semi-supervised text classification.

REFERENCES

[1] Lan-Zhe Guo, Zhi Zhou, and Yu-Feng Li. 2020. Record: Resource constrained semi-supervised learning under distribution shift. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1636–1644.

[2] Wenhui Yu, Xiao Lin, Junfeng Ge, Wenwu Ou, and Zheng Qin. 2020. Semi-supervised collaborative filtering by text-enhanced domain adaptation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2136–2144.
[3] Hanrui Wu, Qingyao Wu, and Michael K. Ng. 2021. Knowledge preserving and distribution alignment for heterogeneous domain adaptation. *ACM Transactions on Information Systems (TOIS)* 40, 1 (2021), 1–29.
[4] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NIPS* (2017).
[5] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S. Davis. 2022. Rethinking pseudo labels for semi-supervised object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1314–1322.
[6] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021. FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems* 34 (2021), 18408–18419.
[7] Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *ACL* (2020).
[8] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. 2019. Are anchor points really indispensable in label-noise learning? *Advances in Neural Information Processing Systems* 32 (2019).
[9] Bixiao Zeng, Xiaodong Yang, Yiqiang Chen, Hanchao Yu, and Yingwei Zhang. 2022. CLC: A consensus-based label correction approach in federated learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* (2022).
[10] Tianchi Yang, Linmei Hu, Chuan Shi, Houye Ji, Xiaoli Li, and Liqiang Nie. 2021. HGAT: Heterogeneous graph attention networks for semi-supervised short text classification. *ACM Transactions on Information Systems (TOIS)* 39, 3 (2021), 1–29.
[11] Lia Bozarth and Ceren Budak. 2020. Toward a better performance evaluation framework for fake news classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 60–71.
[12] Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 6280–6285.
[13] Qi Zhang, Jie Zhou, Qin Chen, Qingchun Bai, and Liang He. 2022. Enhancing event-level sentiment analysis with structured arguments. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1944–1949.
[14] Yuan Gao, Maoguo Gong, Yu Xie, and Alex Kai Qin. 2020. An attention-based unsupervised adversarial model for movie review spam detection. *IEEE Transactions on Multimedia* 23 (2020), 784–796.
[15] Prabhat Agarwal, Manisha Srivastava, Vishwakarma Singh, and Charles Rosenberg. 2022. Modeling user behavior with interaction networks for spam detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2437–2442.
[16] Suchin Gururangan, Tam Dang, Dallas Card, and Noah A. Smith. 2019. Variational pretraining for semi-supervised text classification. *arXiv preprint arXiv:1906.02242* (2019).
[17] Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. 2018. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370* (2018).
[18] Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, Vol. 3. 896.
[19] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
[20] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. 2020. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *arXiv preprint arXiv:2001.06001* (2020).
[21] Yang Zou, Zhiding Yu, B. V. K. Kumar, and Jinsong Wang. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision*. 289–305.
[22] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng MaXiaoyu Tao, and Nanning Zheng. 2018. Transductive semi-supervised deep learning using min-max features. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 299–315.
[23] Zhedong Zheng and Yi Yang. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision* 129, 4 (2021), 1106–1120.
[24] Andrea Esuli and Fabrizio Sebastiani. 2013. Improving text classification accuracy by training label cleaning. *ACM Transactions on Information Systems (TOIS)* 31, 4 (2013), 1–28.
[25] Yivan Zhang, Gang Niu, and Masashi Sugiyama. 2021. Learning noise transition matrix from only noisy labels via total variation regularization. In *International Conference on Machine Learning*. PMLR, 12501–12512.

[26] Yang Liu. 2022. Identifiability of label noise transition matrix. *arXiv preprint arXiv:2202.02016* (2022).
[27] Zengmao Wang, Bo Du, and Yuhong Guo. 2019. Domain adaptation with neural embedding matching. *IEEE Transactions on Neural Networks and Learning Systems* 31, 7 (2019), 2387–2397.
[28] Deming Zhai, Hong Chang, Shiguang Shan, Xilin Chen, and Wen Gao. 2012. Multiview metric learning with global consistency and local smoothness. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 3 (2012), 1–22.
[29] Rodrigo G. F. Soares, Huanhuan Chen, and Xin Yao. 2012. Semisupervised classification with cluster regularization. *IEEE Transactions on Neural Networks and Learning Systems* 23, 11 (2012), 1779–1792.
[30] Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 991–1000.
[31] Bingbing Xu, Junjie Huang, Liang Hou, Huawei Shen, Jinhua Gao, and Xueqi Cheng. 2020. Label-consistency based graph neural networks for semi-supervised node classification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1897–1900.
[32] Abulikemu Abuduweili, Xingjian Li, Humphrey Shi, Cheng-Zhong Xu, and Dejing Dou. 2021. Adaptive consistency regularization for semi-supervised transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6923–6932.
[33] Haibo Ye, Xinjie Li, Yuan Yao, and Hanghang Tong. 2022. Towards robust neural graph collaborative filtering via structure denoising and embedding perturbation. *ACM Transactions on Information Systems* (2022).
[34] Samuli Laine and Timo Aila. 2017. Temporal ensembling for semi-supervised learning. *ICLR* (2017).
[35] Yuan Zhang, Fei Sun, Xiaoyong Yang, Chen Xu, Wenwu Ou, and Yan Zhang. 2020. Graph-based regularization on embedding layers for recommendation. *ACM Transactions on Information Systems (TOIS)* 39, 1 (2020), 1–27.
[36] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2018), 1979–1993.
[37] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems* 33 (2020), 6256–6268.
[38] Bo Du, Tang Xinyao, Zengmao Wang, Lefei Zhang, and Dacheng Tao. 2018. Robust graph-based semisupervised learning for noisy labeled data via maximum correntropy criterion. *IEEE Transactions on Cybernetics* 49, 4 (2018), 1440–1453.
[39] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*. PMLR, 1126–1135.
[40] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*. PMLR, 2304–2313.
[41] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in Neural Information Processing Systems* 32 (2019).
[42] Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* (2018).
[43] Bin Liang, Xiang Li, Lin Gui, Yonghao Fu, Yulan He, Min Yang, and Ruifeng Xu. 2022. Few-shot aspect category sentiment analysis via meta-learning. *ACM Transactions on Information Systems (TOIS)* (2022).
[44] Hung-yi Lee, Shang-Wen Li, and Ngoc Thang Vu. 2022. Meta learning for natural language processing: A survey. *arXiv preprint arXiv:2205.01500* (2022).
[45] Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label classification. *arXiv preprint arXiv:1909.04176* (2019).
[46] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. 2021. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11557–11568.
[47] Ruobing Xie, Yalong Wang, Rui Wang, Yuanfu Lu, Yuanhang Zou, Feng Xia, and Leyu Lin. 2022. Long short-term temporal meta-learning in online recommendation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1168–1176.
[48] Dongqi Fu, Liri Fang, Ross Maciejewski, Vetle I. Torvik, and Jingrui He. 2022. Meta-learned metrics over multi-evolution temporal graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 367–377.
[49] Shujie Li, Min Yang, Chengming Li, and Ruifeng Xu. 2022. Dual pseudo supervision for semi-supervised text classification with a reliable teacher. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2513–2518.
[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL* (2019).

[51] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems* 28 (2015), 649–657.
[52] Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *AAAI*, Vol. 2. 830–835.
[53] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation for consistency training. *The 34th Conference on Neural Information Processing Systems* (2019).
[54] Changchun Li, Ximing Li, and Jihong Ouyang. 2021. Semi-supervised text classification with balanced deep representation distributions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. 5044–5053.
[55] Ahmed Murtadha, Shengfeng Pan, Wen Bo, Jianlin Su, Xinxin Cao, Wenze Zhang, and Yunfeng Liu. 2023. Rank-aware negative training for semi-supervised text classification. *arXiv preprint arXiv:2306.07621* (2023).
[56] Léon Bottou. 2012. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*. Springer, 421–436.
[57] Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S. Yu. 2023. A comprehensive evaluation of ChatGPT's zero-shot Text-to-SQL capability. (2023). arXiv:cs.CL/2303.13547
[58] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
[59] Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. *arXiv preprint arXiv:2305.00118* (2023).
[60] Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2023. LLMs as counterfactual explanation modules: Can ChatGPT explain black-box text classifiers? *arXiv preprint arXiv:2309.13340* (2023).
[61] Brendan Van Rooyen, Aditya Menon, and Robert C. Williamson. 2015. Learning with symmetric label noise: The importance of being unhinged. *Advances in Neural Information Processing Systems* 28 (2015).
[62] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems* 33 (2020), 596–608.

Received 15 November 2022; revised 15 November 2023; accepted 6 February 2024