Efficient Two-stage Label Noise Reduction for Retrieval-based Tasks

Mengmeng Kuang[†], Weiyan Wang[§], Zhenhong Chen[†], Lie Kang[†] and Qiang Yan[†]
[†] Tencent Holdings Ltd.
[§] Hong Kong University of Science and Technology
mengkuang@tencent.com,wwangbc@cse.ust.hk,{hollischen,liekang,rolanyan}@tencent.com

ABSTRACT

The existence of noisy labels in datasets has always been an essential dilemma in deep learning studies. Previous works detected noisy labels by analyzing the predicted probability distribution generated by the model trained on the same data and calculating the probabilities of each label to be regarded as noise. However, the predicted probability distribution from the whole dataset may introduce overfitting, and the overfitting on noisy labels may induce the probability distribution of clean and noisy items to be not conditional independent, making identification more challenging. Additionally, label noise reduction on image datasets has received much attention, while label noise reduction on text datasets has not. This paper proposes a noisy label reduction method for text datasets, which could be applied at retrieval-based tasks by getting a conditional independent probability distribution to identify noisy labels accurately. The method first generates a candidate set containing noisy labels, predicts the category probabilities by the model trained on the rest cleaner data, and then identifies noisy items by analyzing a confidence matrix. Moreover, we introduce a warm-up module and a sharpened cross-entropy loss function for efficiently training in the first stage. Empirical results on different rates of uniform and random label noise in five text datasets demonstrate that our method can improve the label noise reduction accuracy and end-to-end classification accuracy. Further, we find that the iteration of the label noise reduction method is efficient to high-rate label noise datasets, and our method will not hurt clean datasets too much.

CCS CONCEPTS

• **Information systems** → *Information retrieval*; **Data cleaning**.

KEYWORDS

Data cleaning, noisy labels, text dataset, retrieval-based tasks

ACM Reference Format:
Mengmeng Kuang[†], Weiyan Wang[§], Zhenhong Chen[†], Lie Kang[†] and Qiang Yan[†]. 2022. Efficient Two-stage Label Noise Reduction for Retrieval-based Tasks. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), February 21–25, 2022, Tempe, AZ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3488560.3498482

1 INTRODUCTION

The ranking is crucial for search engines, which significantly impacts user experiences [15]. Traditionally, Learning to Rank (LTR) formulates a supervised learning problem to train ranking models on query-document pairs and relevant labels. Therefore, the successes of LTR reasonably rely on large-scale datasets with high-quality labels. However, it is widely known that large datasets often have the label quality problem [21], i.e., label noise. It is hard to avoid any mistakes in labeling such an extensive dataset for human annotators. Furthermore, human annotators may disagree with actual user preferences due to subjectivity [24]. The presence of label noise inherent to training instances has been reported to deteriorate the performance of even the best classifiers in a wide variety of classification tasks [7, 8].

An extensive collection of work has arisen to address the label uncertainty in large datasets, which could be divided into three categories: (i) noisy label reduction, (ii) robust loss functions, and (iii) latent label modeling. In the first category, the reduction strategies usually depended on the probabilities predicted by the model trained on the whole dataset. However, the probabilities predicted by the model trained on the uncertainty dataset were not conditional independent. Besides, it will bring the excessive detection problem that some hard samples [10] may be detected as noise. So a not-so-perfect method might cause serious data lacking. In the second category, many weighted loss functions were performed to reweight the noisy or clean labels to make the neural network learn better. The changes in objective function may not work well on the label noise with random label flapping since the improved loss functions could only grasp the distribution of the uniform label flapping label noise. In the last category, additional neural network layers or models were introduced to identify the potential noisy labels. Unfortunately, the modeling strategies on the latent labels decayed seriously when the noise rate of the dataset increased. Also, the method of interaction between multiple models is very easy to reach a convergence consensus.

To remedy the limitations, we gain motivations from the following aspects. First, the training and testing set must be separated to get a conditional independent probability distribution on the latent labels. Second, a higher confidence threshold for calculating the confidence matrix is needed. Third, a sharped cross-entropy loss function was proved to make the distribution of noisy and clean labels more discriminate. Finally, label noise reduction based on the confidence matrix is more effective for high-rate label noise.

WSDM '22, February 21–25, 2022, Tempe, AZ, USA

ACM ISBN 978-1-4503-9132-0/22/02…$15.00
https://doi.org/10.1145/3488560.3498482

To this end, this paper propose a novel label noise reduction strategy consisting of two stages: (i) noise candidate set generation and (ii) confidence matrix calculation. In order to get a more reliable noise candidate set, we introduce a warm-up module that will provide a high-performance initial prediction model and some robust noise judgment rules. We also perform a sharpened cross-entropy loss function for the deep neural network training to maximize the loss distribution difference between clean and noisy labels.

We provide extensive experiments on five experts-reviewed text datasets used for information retrieval or text classification datasets with artificially injected label noise. It shows that our method can make the datasets with different label noise rates get the best end-to-end classification accuracy and label noise detection accuracy compared with baselines.

Overall, the contributions of this work are presented as follows:

- we perform an efficient strategy to generate a noise candidate set and obtain conditional independent classification probabilities equipped by a warm-up module and the sharpened cross-entropy loss function.
- we introduce the confidence matrix with better thresholds calculated on the conditional independent probabilities to reduce the excessive detection over other label noise reduction strategies.
- the comprehensive experiments validate the effectiveness of our work.

2 BACKGROUNDS

In this section, we discuss the backgrounds of our work. We start with the formal definition and the categories of the noisy label. Then we summarize all previous works related to this topic and their drawbacks.

2.1 Problem Definition

Theoretically, we require to find out and separate the item (x), which has the true label (y^t) differing from its observed one (y^o). However, the theoretical true label is unknown, so some strategies for calculating the label prediction ($\hat{y}$) are needed. After getting predictions, we need to determine the number of instances containing noisy labels by strategies for every category. The numbers filtered by some strategies form the confidence matrix (CM) in the shape of $|C| \times |C|$ where C is the category set. The label noise reduction task estimates the confidence matrix precisely and classifies the noisy instances supervised by it.

2.2 Label noise categories

Some authors [13, 27] categories the noisy label flipping into two parts: uniform label flipping and random label flipping. Uniform label flipping implies a clean label is exchanged with another label sampled uniformly at random. In contrast, Random label flipping means a clean label is swapped with another label from the given number of labels sampled randomly over a unit simplex. The uniform and random label flipping construct the label noise datasets in this work.

2.3 Related Works

Different techniques have been proposed to deal with noisy labels. The main contributions in recent years can be summarized into the following three categories: (1) noisy label reduction, (2) robust loss functions, and (3) latent label modeling.

Noisy label reduction. [19] sought data-centric ways to find the noisy labels, which starts from the assumption that label noise is class-conditional, depending only on the actual latent class, not the data [1]. It thoroughly analyzed the direct estimation of the joint distribution between noisy and actual labels. However, the distribution of predicted probabilities is not conditional independent, and excessive detection problems are two main limitations. From another viewpoint, [18] proposed a method of removing the samples whose pseudo labels differ from the observed labels. The model adopted the moving average method to predict the pseudo-label, which is also known as self-ensembling. Although this method was compatible with other semi-supervised losses, it does not perform well with high-rate label noise. Additionally, there is a type of unsupervised approach named outlier removal [25, 31]. While outliers were not necessarily mislabeled and removing them presented a challenge [6]. The above methods had a common dilemma: serious data lacking, which would ultimately decrease the classification accuracy.

Robust loss functions. [12] weighted the calculated cross-entropy loss to make noisy labels get a low weight, and clean labels get high weight. The weight generation came from a mentor network, which could be regarded as a function with the historical information of the loss. Additionally, [11] followed the same idea of [12] with some semi-supervised learning strategies introduced, such as pseudo-labels and mixups [32]. Besides, [34] demonstrated that mean absolute error (MAE) treats samples equally, while cross-entropy (CE) tilts towards samples that are difficult to identify. Therefore, MAE is more robust than CE for noisy labels, but CE fits fast with higher accuracy. So the authors proposed a generalized cross-entropy (GCE) loss function, which combines the advantages of MAE and CE. However, very similar classes would be mislabeled since they had similar GCE losses. Moreover, [29] pointed out that the traditional cross-entropy loss (CE) is problematic since simple classes would be easier to learn and be overfitted under high-rate noise. Therefore, a new symmetric cross-entropy (SCE) loss function was proposed, which was the weighted sum of (i) the exchange of a label, (ii) the predicted value of a label, and (iii) the original cross-entropy loss. However, the SCE loss function could not work well with too much random label noise on the dataset. Nevertheless, [16] provided a dimensionality-driven explanation of the generalization of neural networks in the presence of label noise. Based on this finding, they proposed a new training strategy, termed Dimensionality-Driven Learning (D2L), that avoided the dimensionality expansion stage of learning. But it still had not remedied the limitation mentioned above.

Latent label modeling. [7] attempted to add two softmax output layers for the neural networks, one for soft classification and the other for predicting the noisy label based on the actual label and the input features. Furthermore, Co-teaching [9] recorded the cross-entropy loss of each sample during the training process of two DNNs, and then divided the samples into two parts according to the losses. A model selected the part with minor loss to feed the other model for the next epoch training. Nevertheless, the training time is too long, and the two networks easily reach convergence consensus. Next, [4] presented the Iterative Noisy Cross-Validation (INCV) method to select a subset of samples, which had a much lower noise rate than the original dataset, resulting in a more stable training process of DNNs. Moreover, they automatically estimated the noise rate of the selected set, which makes INCV more practical for industrial applications. Furthermore, DivideMix [14] continued the thought of co-teaching. The difference was that after picking out clean and noise samples, the noise samples were treated as unlabeled samples and trained through the MixMatch [2] method. Although it can significantly improve the robustness of DNN against noisy labels, when the noise rate of the data set is extremely high, it is difficult to apply to mini-batch training. To adapt the text datasets, [13] proposed to add a nonlinear noise model to the basic CNN structure. The noise model could absorb most of the label noise and help the base model learn better sentence representations through proper initialization and regularization. Unfortunately, the modeling strategies on the latent labels decayed seriously when the noise rate of the dataset increased.

3 METHODS

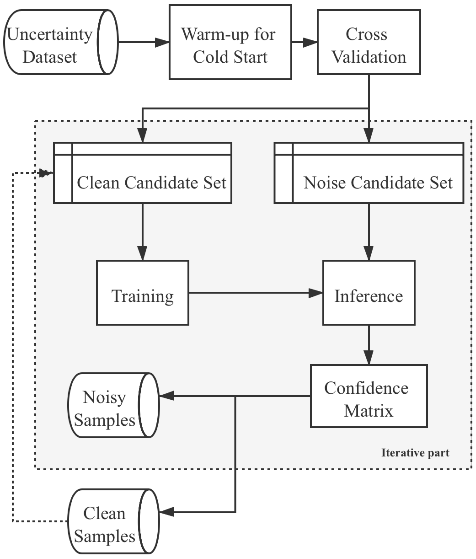

Figure 1: The framework of the label noise reduction pipeline. (The dotted line in the figure indicates that our framework can be iterated by combining the identified clean samples to the clean candidate set.)

In this section, we acquaint the intention of the efficient method for finding noisy labels. Overall, we propose an efficient noise reduction method consisting of two stages:

(1) Noise candidate set generation
(2) Confidence matrix calculation

Fig. 1 illustrates the whole workflow of the noise reduction framework. The framework starts with a warm-up step that trains all the data for several epochs, providing a starting model for generating the noise candidate set. Then in the noise candidate set generation stage, we perform cross-validation to divide the entire dataset into noise candidate set and clean candidate set. After that, we only train on the clean candidate set and then predict the noise candidate set to get the class probabilities. Then in the next stage, the confidence matrix is calculated based on the predicted probabilities to estimate the number of noisy labels and prunes data samples with low confidence.

We organize the rest parts of this section as follows: The details of the noise candidate set generation stage are stated in Sec. 3.1. In Sec. 3.1.1, we discuss the warm-up module as a start point for the noise candidate set generation. Then Sec. 3.1.2 describes the sharpened cross-entropy loss function, which distinguishes the clean label distribution from the noisy one. Finally, we introduce the second stage of confidence matrix calculation in Sec. 3.2.

3.1 Noise candidate set generation

Deep learning is the process of learning a probability distribution $p(\hat{y} = c; x)$ from the feature-label pairs D with minimizing errors. We donate by θ_t the network parameters for M_t obtained after t steps of mini-batch training.

Furthermore, let acc_t and $loss_t$ be accuracy and loss on the validation set with parameter θ_t in cross-validation. We define the following condition of a noise candidate event e_t on the t-th mini-batch training:

$$prev\_acc - acc_t > \lambda_{acc}$$
$$loss_t - prev\_loss \geq \lambda_{loss}$$

The thresholds λ_{acc} and λ_{loss} are learned before the generation process, which will be introduced later in Sec. 3.1.1. In other words, to some extent, those training samples (gathered by mini-batch) may be noisy when the training performance is worse with their occurrence [3, 28]. The process is illustrated by Algorithm 1.

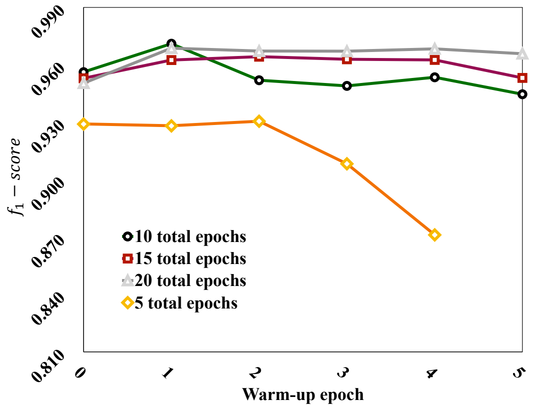

Figure 2: The performance of f_1-score comparison of various warm-up epochs in the entire process of label noise reduction.

Algorithm 1 Noise candidate set generation

```
Require:
    warm-up model $M$
    training data $D$
    noise candidate set $N$
    cross-validation fold $F$
    thresholds $\lambda_{acc}$, $\lambda_{loss}$
Ensure:
for $f$ in range($F$) do
    generate training data $D_f$, testing data $D_{f'}$;
    $prev\_acc, prev\_loss$ initiation;
    while training on $D_f$ do
        generate mini-batch $D_f^t$;
        $M_f^t$ = train($D_f^t$);
        $acc, loss = M_f^t(D_{f'})$;
        if $prev\_acc$ - $acc > \lambda_{acc}$ and
        $loss$ - $prev\_loss \geq \lambda_{loss}$ then
            append $D_f^t$ to $N$;
        end if
        $prev\_acc, prev\_loss = acc, loss$;
    end while
end for
return $N$
```

3.1.1 Warm-up module. As mentioned above, the dilemma of model cold start, including the initial model and thresholds, is our concern. We propose to utilize all the items in the data set to train a general classification model by some warm-up epochs before generating the two sets. This scheme can guarantee the start point of generating the noise candidate set has a relatively high level, and the thresholds (λ_{acc} and λ_{loss}) are calculated from the average shifts from this stage. We conduct experiments to reveal the relationships between detection f_1-scores and the epoch number of the warm-up, which is shown later in Fig. 2. The warm-up epoch ranges from 0 to 5 in different total epochs like 5, 10, 15, and 20. For class-conditional label noise, the network would quickly overfit to noise during warm-up and produce over-confident predictions, which leads to most instances having near-zero normalized loss [14]. Therefore, the epoch of the warm-up module should not be too large, which is in line with the curves in Fig. 2.

3.1.2 Sharpened cross-entropy loss function. Another key to making the noise candidate set generation more accurate is the sharpened cross-entropy loss function employed in the training procedures. The cross-entropy loss function can be expressed by Equ. (1).

$$loss = -\frac{1}{N} \times \sum_{j=1}^{N} \sum_{i=1}^{|C|} y_j^i \times \log(pred_j^i), \tag{1}$$

where N is the number of items, C is the category number and $pred_j$ is the predicted probability vector for x_j.

Nevertheless, this calculation may bring difficulty when a lot of noisy labels exist. As shown in Fig. 3, the normalized loss values of clean and noisy labels in the tenth epoch of a training experiment follow almost the same distribution when uniform and random label noise existing.

Hence, inspired by DivideMix [14], we introduce a similar sharpening formula by a temperature as Equ. (2) to the predicted probabilities.

$$pred_j = \frac{(pred_j)^T}{\sum (pred_j)^T}. \tag{2}$$

We perform another training experiment on a dataset with uniform and random label noise and record the sharpened cross-entropy loss values in the tenth epoch to generate the distributions of noisy and clean items. The distributions of items with noisy and clean labels in Fig. 4 are much different by modifying the sharpened loss function after several epochs of training. The proposed sharpened loss function can significantly alter the loss distribution for noisy samples while keeping the loss value miniature for most clean ones. It is worth noting that the distribution change is not limited to a certain kind of label noise.

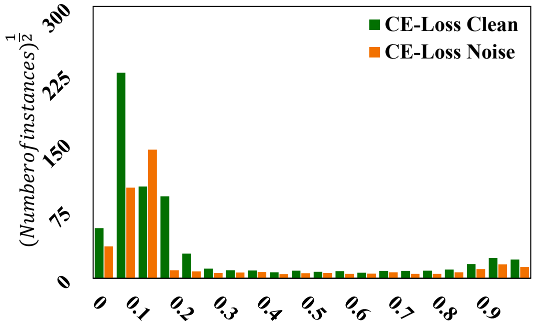

Figure 3: Normalized distribution of cross entropy loss.

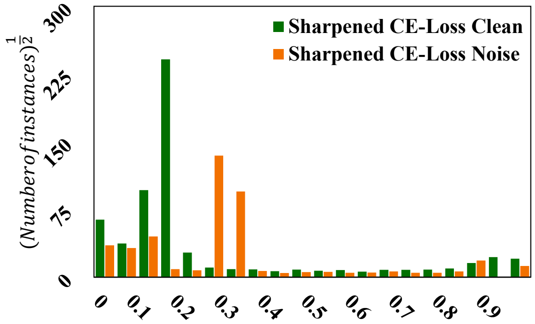

Figure 4: Normalized distribution of sharpened cross entropy loss.

3.2 Confidence matrix calculation

Following the noise candidate set generation, another stage, namely confidence matrix calculation, needs to be implemented to prune and release the noisy labels. Inspired by confident learning, we have the Equ. (3) for calculating every cell in the confidence matrix (CM) by the probabilities of the noise candidate items. Unlike the

Datasets	Size	Words	Label balance
SST-2	67,350	10.41	almost balanced
QNLI	104,744	36.46	balanced
QQP	363,847	22.13	unbalanced
Private	106,863	42.34	balanced
AG's News	120,000	38.31	balanced

Table 1: The statistical characteristics of the datasets.

confident threshold calculated in the confident learning strategy, we make a stricter restriction toward it. This operation aims to reduce the problem of excessive detection that confident learning faced [20].

$$CM_{\hat{y}=i,y^o=j} = |D_{y^o=j;p(\hat{y}=i;x)\geq t_i}|, \tag{3}$$

where threshold t_j is in the form as Equ. (4) and i, j belong to category set C.

$$t_i = |\frac{1}{X_{y^o=j,\hat{y}=i}}| \times \sum p(\hat{y}=i;x), \tag{4}$$

where $i = \arg\max p(\hat{y}=c;x)$.

The number $CM_{\hat{y}=i,y^o=j}$ in the matrix stands for the number of items that the observed label j should be labeled as i. For an observed label j, the number of noisy labels should be $n_j = \sum_{i=1}^{|C|} CM_{\hat{y}=i,y^o=j}$ and our method recognizes n_j items with low confidence in $p(\hat{y} = j;x)$ as noise.

4 EXPERIMENTS

4.1 Metrics

Because our experiment is divided into standard bi-category and multi-category of the end-to-end classification, the macro-average f_1-score is a relatively recognized evaluation metric. At the same time, we compare the label noise reduction module separately. Noisy label detection can also be regarded as a bi-classification problem. Noisy labels are regarded as positive samples, while clean labels are negative samples. Therefore, the f_1-score is a good metric. In order to prove that our label noise reduction module can effectively reduce the over-detection problem, we introduce the detection rate (that is, the proportion of noise samples detected by the approach) as another metric, and the closer the detection rate is to the predefined noise rate, the better.

4.2 Datasets

We implement and validate our method on five datasets: (i) the Stanford Sentiment Treebank (*SST-2*) dataset [26], (ii) the Question Natural Language Inference (*QNLI*) dataset [22], (iii) the Quora Question Pairs (*QQP*) [1] dataset. (iv) a private dataset used in information retrieval and (v) AG's News topic [33] classification dataset. These datasets can be utilized for retrieval-based tasks like document ranking and question answering. Table 1 summarizes their statistical characteristics from three aspects: (i) the size of the training data, (ii) average word number, and (iii) whether the quantity of various labels is balanced or not.

Following is a detailed description of these datasets:

1. The *SST-2* dataset is a corpus with fully labeled parse trees that allows for a complete analysis of the compositional effects of language sentiment, widely used for sentiment analysis research.
2. The *QNLI* dataset is converted from the Stanford Question Answering Dataset (SQuAD 1.0) [23], which is adopted for determining whether the question and sentence are entailed or not.
3. The *QQP* dataset is a collection of question-answer pairs in the community Q&A website Quora.
4. The *Private* dataset is in the form of query-document pairs, and the label represents the relation level between the query and the corresponding document. The dataset is randomly extracted from actual daily service retrieval records [2]. Experts have checked the notations to ensure that the labels are correct.
5. The *AG's News* topic classification dataset is constructed by [33] from a collection of more than 1 million news articles, which gathered from more than 2000 news sources by ComeToMyHead [3] in more than one year of activity.

4.3 Baselines

For the end-to-end learning with noisy labels comparison, we compare our proposed two-stage label noise reduction framework with six representative baseline models [4]. Such methods can be categorized into different classes as follows (the detailed discussion of these methods is presented in Sec. 2.3),

- *Label noise reduction*: Confident Learning (**CL**) (implemented with Co-teaching [5]) and *BERT* pre-trained model enhanced CL (**CL:BERT**);
- *Robust loss functions*: **Cross-Entropy** loss function and **D2L**;
- *Latent label modeling*: **Co-teaching** and **INCV**.

Since **Co-teaching**, **INCV**, and **D2L** strategies were designed just for Image datasets, we reimplemented them with the BERT pre-trained language model to tackle the text datasets. The reimplementation method is replacing the original CNN-based classification models with the same linear classifier based on BERT language representations.

4.4 Implementation details

In the experimental setting, we randomly introduce 10%, 20%, and 30% noisy labels by choosing labels for them in the three binary classification datasets, *SST-2*, *QNLI* and *QQP* datasets for the random label noise, while using a fixed noise label transfer strategy to misplace the randomly selected labels for the uniform label noise in *Private* and *AG's News* datasets.

The hyperparameters of training the baselines are set as their defaults. When training our method, the batch size is set to 256, while the mini-batch size is set to 1 when generating the noise candidate

[1] https://dl.fbaipublicfiles.com/glue/data/QQP-clean.zip

[2] We have removed the sensitive and private information about users.

[3] ComeToMyHead is an academic news search engine that has been running since July 2004.

[4] Other methods mentioned in Sec. 2.3 are not used for comparison because they are not open-sourced, or they are only designed for image datasets.

[5] The implementation is based on https://pypi.org/project/cleanlab/

set. The numbers of epochs for the warm-up module, noise candidate set generation, and final training on the clean candidate set are 1, 10, and 5, respectively. Fig. 5 demonstrates the experimental results on a series of epochs in the noise candidate set generation stage. The temperate for sharpening the cross-entropy loss is set to 2. Five-fold cross-validation is applied in the noise candidate set generation stage. The classifier and sentence embedding method used in this work are based on the pure BERT pre-trained models ([5] for Chinese and [30] for English) with padding size 64. All the experiments are performed on an NVIDIA Tesla V100 GPU.

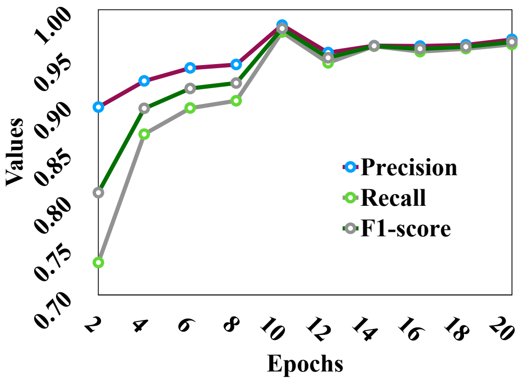

Figure 5: The precisions, recall rates and f_1-scores of various epochs in the noise candidate set generation stage.

5 RESULTS

In this section, we first show the effect of our label noise reduction method in the end-to-end classification task in Sec. 5.1 and present the effect of our label noise reduction method separately in Sec. 5.2. In Sec. 5.3, we explore the feasibility of iterations on our label noise reduction method. Then detecting noisy labels for the five clean datasets and the datasets with real label noise are illustrated in Secs. 5.4 and 5.5 respectively. Finally, we analyze the training efficiency of our method in Sec. 5.6.

5.1 Effect of the end-to-end classification

We evaluate the performance of our method for each dataset in the presence of uniform and random label noise and compare the performance with the baselines. We illustrate the results of end-to-end classification with noisy labels in Table 2, which describes the classification f_1-scores of our method and other baselines on two kinds of label noise in five text datasets with different noise rates.

From the results, we can summarize that:

- The results align with the study from the quantitative measurements of f_1-scores compared with the Cross-Entropy classification. The proposed approach is significantly better than the baselines for both types of label noise for all datasets. Meantime, we observe a gain of approximately 1% - 5% on f_1-scores compared with baselines in the presence of extreme label noise.
- It can be observed from the experimental results that compared with the one-stage noise adaptable training methods (**Co-teaching**, **INCV**, and **D2L**), the two-stage label noise detecting methods (**CL**, **CL:BERT** and **Our method**) can receive a better classification model. The present study confirms the findings that removal of label errors before training is a practical approach for learning with noisy labels [4].
- Our method improved the performance of label noise reduction based methods in both kinds of label noise. Moreover, latent label modeling based methods (**Co-teaching** and **INCV**) preformed well over robust loss function based methods (**Cross-Entropy** and **D2L**) in random label noise.
- Our method achieves the best results on datasets with high-rate (i.e., 30%) label noise. However, the accuracy of the latent label modeling based methods decreased relatively seriously, which is also consistent with our previous analysis.

5.2 Effect of the label noise reduction

Since our proposed learning with noisy labels method relies on the efficient noise label reduction method, it is necessary to compare the accuracy of the label noise reduction module. Meanwhile, we are more concerned about the detection rate in the noise label reduction stage because it is the metric to determine the retention size. Tables 3 and 4 illustrate the detection f_1-scores and the detection rates of our method, sub-modules of our method and a baseline in the same experimental setting like previous, respectively.

From the results, we can confirm that:

- Our proposed label noise reduction module can mostly get the best detection f_1-score and gain more than 10% improvements compared with the baseline (**CL:BERT**).
- The sharpened loss operation can increase by 1%-13% on f_1-score compared with the **cross-entropy** loss function, which was not sharpened.
- The threshold for calculating the confidence matrix (**ConMatrix**) can reduce the excessive detection by 1% - 6% compared with **CL:BERT** from the comparisons on detection rate.

The experimental results proved our expectation that our label noise reduction module can effectively find the noisy labels in text datasets, guaranteeing our end-to-end classification accuracy.

5.3 Feasibility of iteration on label noise reduction

Since the noisy label reduction stage is the focus of our work, can it be iterative all the time? We perform iterative label noise reduction experiments on *SST-2* and *QQP* datasets, and Table 5 summaries the average f_1-scores of the end-to-end classification under different original noise rates. Our results demonstrate that a certain number of iteration performs better on a high-rate label noise dataset. However, when the number of iteration is larger, all results are worse. It is worth discussing these exciting facts revealed by the results. Combined analysis with Sec. 5.2, we think it is because when the amount of noisy labels is relatively small, the accuracy of label noise reduction (i.e., detection f_1-score) is correspondingly lower, which means that when the dataset with very little label noise, it may not improve the cleanliness of the remaining samples after reduction.

	Random									Uniform					
	SST-2			*QNLI*			*QQP*			*Private*			*AG's News*		
Noise rate	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%
CL	0.8835	0.8469	0.8345	0.8316	0.8341	0.8112	0.8548	0.8634	0.8346	0.9336	0.9202	0.9072	0.9154	0.8767	0.8450
CL:Bert	$\underline{0.8842}$	$\underline{0.8574}$	$\underline{0.8435}$	0.8455	0.8394	0.8261	$\underline{0.8749}$	0.8666	0.8381	0.9459	$\underline{0.9333}$	0.9252	$\underline{0.9179}$	0.8718	0.8633
Cross-Entropy	0.8329	0.7908	0.7342	0.8212	0.7963	0.7296	0.8471	0.8202	0.7623	0.9181	0.8954	0.7988	0.8996	0.8732	0.8455
D2L	0.8654	0.8392	0.8296	0.8491	0.8295	0.7949	0.8658	$\underline{0.8730}$	0.8314	0.9440	0.9325	0.9175	0.9084	0.8815	0.8736
Co-teaching	0.8643	0.8309	0.8248	0.8550	0.8385	0.8274	0.8562	$\underline{0.8730}$	0.8341	$\underline{0.9488}$	0.9311	$\underline{0.9257}$	0.9068	$\underline{0.8915}$	0.8763
INCV	0.8684	0.8539	0.8260	$\underline{0.8561}$	$\underline{0.8410}$	$\underline{0.8289}$	0.8649	0.8589	$\underline{0.8382}$	0.9408	0.9263	0.9087	0.9103	0.8887	$\underline{0.8823}$
Our method	0.8909*	0.8879*	0.8776*	0.8565*	0.8477*	0.8485*	0.8906*	0.8861*	0.8715*	0.9547*	0.9331	0.9283*	0.9267*	0.8996*	0.8829*
Impr.	0.76%	3.56%	4.04%	0.05%	0.80%	2.36%	1.79%	1.50%	3.97%	0.62%	-	0.28%	0.96%	0.91%	0.07%

Table 2: The average f$_1$-scores for different noise rates of uniform label noise in two multi-label datasets (the Private dataset and AG's News dataset), and three bi-label datasets (SST-2, QNLI and QQP). * stands for the best f$_1$-scores and the underline stands for the best f$_1$-scores in baselines. "Impr." presents the improvement of our method over the best baseline.

	Random									Uniform					
	SST-2			*QNLI*			*QQP*			*Private*			*AG's News*		
Noise rate	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%
CL:Bert	0.7004	0.8145	0.8654	0.6334	0.7687	0.8310	0.6184	0.7510	0.8146	0.7165	0.8150	0.8609	0.8578	0.9095	0.9372
ConMatrix	0.7450	0.8476	0.8897	0.6883	0.8000	0.9280	0.6884	0.8037	0.8547	0.6432	0.9544	0.8923	0.9460	0.9575	0.9646
Cross-Entropy	0.8391	0.8989	0.9243	0.7045	0.9030	0.9419	0.7994	0.8574	0.8894	0.8771	0.9424	0.9524	0.9560	0.9702	0.9665
Our method	0.9202	0.9548	0.9688	0.8357	0.9157	0.9696	0.8243	0.8945	0.9243	0.9203	0.9748	0.9801	0.9604	0.9718	0.9758

Table 3: The noisy label detection f$_1$-scores for different noise rates of *SST-2, QNLI, QQP, Private* and *AG's News* datasets.

	Random									Uniform					
	SST-2			*QNLI*			*QQP*			*Private*			*AG's News*		
Noise rate	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%
CL:BERT	0.1577	0.2454	0.3313	0.1756	0.2585	0.3412	0.1744	0.2608	0.3400	0.1484	0.2404	0.3290	0.1165	0.2159	0.3138
ConMatrix	0.1464	0.2366	0.3266	0.1565	0.2405	0.3141	0.1587	0.2478	0.3343	0.0880	0.1954	0.3233	0.1048	0.2048	0.3039
Cross-Entropy	0.1201	0.2094	0.3004	0.1537	0.2141	0.3048	0.1267	0.2132	0.2980	0.1037	0.2074	0.2889	0.1039	0.2035	0.3029
Our method	0.1111	0.2092	0.3070	0.1232	0.2206	0.3030	0.1210	0.2116	0.2999	0.1016	0.2004	0.2978	0.1035	0.2029	0.3022

Table 4: The noisy label detection rates for different noise rates of *SST-2, QNLI, QQP, Private* and *AG's News* datasets.

	SST-2			*QQP*		
Noise rate	10%	20%	30%	10%	20%	30%
Iteration 1	0.8909	0.8879	0.8776	0.8906	0.8861	0.8715
Iteration 2	0.8824	0.8819	0.8784	0.8831	0.8824	0.8811
Iteration 3	0.8890	0.8761	0.8643	0.8870	0.8801	0.8718

Table 5: The average f$_1$-scores of the iterative noisy label reduction experiments on *SST-2* and *QQP* datasets with different original noise rates.

Datasets	*SST-2*	*QNLI*	*QQP*	*Private*	*AG's News*
Number	157	357	111	59	827
Percentage	2.33%	3.41%	0.03%	0.10%	6.89%

Table 6: The number and percentage of the detected items in the noisy label detection experiments on clean datasets of *SST-2, QNLI, QQP, Private* and *AG's News* datasets.

5.4 Label noise reduction on clean datasets

A label noise reduction method should not hurt when there is no or little label noise. Combining Tables 3 and 4, we can witness that our reduction method can effectively identify noisy labels from the high detection f$_1$-scores and the precise detection rates. We still want to figure out whether our identification method will over-detect when there is no or less label noise.

We conduct label noise reduction experiments on the five experts-reviewed clean datasets. The experimental results can be seen in the Table 6. On the five clean datasets, the over-detection percentage is relatively low. In these datasets, there are two detection percentages less than 1% and two less than 5%. It proves that our label noise reduction method has a relatively low over-detection rate, reflecting that the reduction method is effective.

Query	Document	Label
Tax Service Office	Insurance and Fund || Provident fund and Medical insurance inquiry || Provident fund query, personal medical insurance query, payment record query	3
Fresh milk cake	Festive flower delivery || Festive flower delivery || Chinese Valentine's Day flowers reservation, Valentine's Day gifts, Valentine's Day flowers, Teacher's Day flowers, Mother's Day flowers reservation, Father's Day flowers reservation, Women's Day flowers reservation, birthday flowers, anniversary flowers, White Day flowers reservation	2
Social Security Card Bank	Appointment || New card application, progress inquiry || Make an appointment, handle affairs, receive a card	1
Recycler	Water heater recovery || Water holding device recovery platform || Home appliance recycling, old home appliance recycling	0

Table 7: The cases containing real label noise detected by our method in *PrivateV0* dataset. The label stands for different relationship level and the bigger, the more relevant. "||" was used to separate different fields in the document.

5.5 Label noise reduction on text datasets with real label noise

In order to verify the detection effect of our reduction method on the dataset with real label noise, we perform label noise reduction experiments on no expert-reviewed *Private* dataset, which we call *PrivateV0*. The *PrivateV0* contains 12,000 query-document pairs with relationship annotations, and every document is composed of three fields: service name, service description, and service tag. Our method exposes 743 samples from the dataset, which may contain potential noisy labels, and 230 are real noise after reviewing. Four cases of different relationship annotations are presented in Table 7.

Moreover, we select a famous text dataset, the Large Movie Review Dataset (v1.0) [17], for further label noise reduction experiments. The Large Movie Review Dataset provides 25,000 highly polar movie reviews from IMDb for training and 25,000 for testing for binary sentiment classification. We perform a noise reduction experiment on the 25,000 training data, and our method guesses 961 records with noisy labels. After manual inspection, we find that more than 700 of them, including the cases in Table 8, are indeed noisy. From the cases, these wrong labels are very confusing, such as with seductive nouns that express wrong feelings after analysis, which ensures that our label noise reduction method can effectively detect some artificially annotated wrong labels in real text datasets.

5.6 The training efficiency of the label noise reduction

The last concern is the efficiency of the label noise reduction method. We constructed label noise reduction experiments with different sizes of datasets in the same task, and Fig. 6 shows the time consumed. Although the data loading time, the noise candidate set generation time, and the confidence matrix calculation time increase correspondingly with the size of the dataset, the data loading time and the confidence matrix calculation time are still relatively short compared to the amount of data. The time of noise candidate set generation is relatively large, but it yields to the law of linear growth with the amount of data.

Review	Label
I just don't understand why this movie is getting beat-up in here. Jeez. It is mindless, it isn't polished and it is (as I am reading) wasted on some. The cast of this movie plays their characters to the 'T' (If you watched Permanent Midnight and became a Ben Stiller fan then yes you will be disappointed). These are misunderstood, well-intentioned misfits trying to save the city/world with nothing but grit and determination. The problem is they don't realize their limits until the big showdown and that's the point! This is 3 times the movie that The Spy Who Shagged Me was yet gets panned by the same demographic group, likely the same people who feel the first AP movie pales in comparison to the sequel. I just don't get it. The jokes work on more then one level; if you didn't get it I know what level you're at.	positive
Working with one of the best Shakespeare sources, this film manages to be creditable to it's source, whilst still appealing to a wider audience.<br /><br />Branagh steals the film from under Fishburne's nose, and there's a talented cast on good form.	negative

Table 8: Some cases detected by our label noise reduction module from the Large Movie Review Dataset

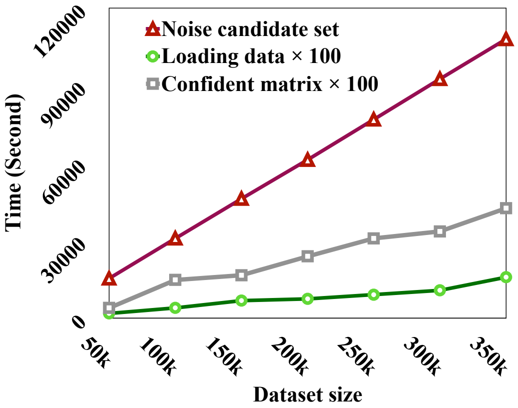

Figure 6: The running time of different modules in the label noise reduction method in different sizes of data.

6 CONCLUSION

In this paper, we presented an innovative method for finding corrupted labels in text classification datasets for retrieval-based tasks. Our effort was constructed on noise candidate set generation and confidence matrix calculation. We advanced the generation methodology by warming up through training from scratch for several epochs and sharpening the cross-entropy loss function. Comprehensive experiments of end-to-end classification and label noise reduction on five text datasets with uniform and random label noise confirmed that our method could efficiently learn with the noisy labels and detecting the label noise.

REFERENCES

[1] Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine Learning* 2, 4 (1988), 343–370.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249* (2019).

[3] Gavin C Cawley and Nicola LC Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* 11 (2010), 2079–2107.

[4] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. 2019. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*. PMLR, 1062–1070.

[5] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. Association for Computational Linguistics, Online, 657–668. https://www.aclweb.org/anthology/2020.findings-emnlp.58

[6] Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25, 5 (2013), 845–869.

[7] Jacob Goldberger and Ehud Ben-Reuven. 2016. Training deep neural-networks using a noise adaptation layer. (2016).

[8] Gaurav Gupta, Anit Kumar Sahu, and Wan-Yi Lin. 2019. Learning in Confusion: Batch Active Learning with Noisy Oracle. *arXiv preprint arXiv:1909.12473* (2019).

[9] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872* (2018).

[10] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2553–2561.

[11] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. 2020. Beyond synthetic noise: Deep learning on controlled noisy labels. In *International Conference on Machine Learning*. PMLR, 4804–4815.

[12] 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.

[13] Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew Nokleby. 2019. An effective label noise model for dnn text classification. *arXiv preprint arXiv:1903.07507* (2019).

[14] Junnan Li, Richard Socher, and Steven CH Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394* (2020).

[15] Tie-Yan Liu. 2011. *Learning to Rank for Information Retrieval*. Springer. https://doi.org/10.1007/978-3-642-14267-3

[16] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. 2018. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*. PMLR, 3355–3364.

[17] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. http://www.aclweb.org/anthology/P11-1015

[18] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. 2019. Self: Learning to filter noisy labels with self-ensembling. *arXiv preprint arXiv:1910.01842* (2019).

[19] Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research* 70 (2021), 1373–1411.

[20] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749* (2021).

[21] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1944–1952.

[22] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822* (2018).

[23] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[24] Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Found. Trends Inf. Retr.* 4, 4 (2010), 247–375. https://doi.org/10.1561/1500000009

[25] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural computation* 13, 7 (2001), 1443–1471.

[26] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.

[27] Yi Sun, Yan Tian, Yiping Xu, and Jianxiang Li. 2019. Limited gradient descent: Learning with noisy labels. *IEEE Access* 7 (2019), 168296–168306.

[28] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159* (2018).

[29] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 322–330.

[30] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[31] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. 2015. Learning discriminative reconstructions for unsupervised outlier removal. In *Proceedings of the IEEE International Conference on Computer Vision*. 1511–1519.

[32] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

[33] 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.

[34] Zhilu Zhang and Mert R Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*.