Rank-Aware Negative Training for Semi-Supervised Text Classification

Ahmed Murtadha[*] **Shengfeng Pan**[◇] **Wen Bo**[◇] **Jianlin Su**[◇]
Xinxin Cao[†] **Wenze Zhang**[◇] **Yunfeng Liu**[◇]
[◇] Zhuiyi Technology Co., Ltd. Shenzhen, Guangdong, China
{a.murtadha,nickpan,brucewen,bojonesu,wen,glenliu}@wezhuiyi.com
[†]Northwestern Polytechnical University, Xi'an, Shaanxi, China
caoxinxin@mail.nwpu.edu.cn

Abstract

Semi-supervised text classification-based paradigms (SSTC) typically employ the spirit of self-training. The key idea is to train a deep classifier on limited labeled texts and then iteratively predict the unlabeled texts as their pseudo-labels for further training. However, the performance is largely affected by the accuracy of pseudo-labels, which may not be significant in real-world scenarios. This paper presents a Rank-aware Negative Training (RNT) framework to address SSTC in learning with noisy label settings. To alleviate the noisy information, we adapt a reasoning with uncertainty-based approach to rank the unlabeled texts based on the evidential support received from the labeled texts. Moreover, we propose the use of negative training to train RNT based on the concept that ''the input instance does not belong to the complementary label''. A complementary label is randomly selected from all labels except the label on-target. Intuitively, the probability of a true label serving as a complementary label is low and thus provides less noisy information during the training, resulting in better performance on the test data. Finally, we evaluate the proposed solution on various text classification benchmark datasets. Our extensive experiments show that it consistently overcomes the state-of-the-art alternatives in most scenarios and achieves competitive performance in the others. The code of RNT is publicly available on GitHub.

1 Introduction

The text classification task aims to associate a piece of text with a corresponding class that could be a sentiment, topic, or category. With the rapid development of deep neural networks, text classification has experienced a considerable shift towards pre-trained language models (PLMs) (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Lewis et al., 2020). Overall, PLMs are first trained on massive text corpora (e.g., Wikipedia) to learn contextual representation, followed by a fine-tuning step on the downstream tasks (Li et al., 2021; Chen et al., 2022; Tsai et al., 2022; Ahmed et al., 2022). The improvement of these approaches heavily relies on high-quality labeled data. However, labeling data is labor-intensive and may not be readily available in real-world scenarios. To alleviate the burden of labeled data, Semi-Supervised Text Classification (SSTC) typically refers to leveraging unlabeled texts to perform a particular task. SSTC-based approaches commonly attempt to exploit the consistency between instances under different perturbations (Li et al., 2020).

Earlier SSTC-based approaches adopt various data augmentation techniques via back-translation. They employ consistency loss between the predictions of unlabeled texts and corresponding augmented texts by translating the text into a targeted language and then translating it back to the source language (Miyato et al., 2019; Xie et al., 2020; Chen et al., 2020). However, the performance of these approaches requires an additional neural machine translation (NMT), which may not be accurate and bothersome in real-world scenarios. Recently, SSTC has experienced a shift toward self-training, and PLM fine-tuning (Li et al., 2021; Tsai et al., 2022). The basic idea is to fine-tune PLMs on the labeled data and iteratively employ prediction on the unlabeled data as pseudo-labels for further training. However, the pseudo-labels are treated equally likely to the truth labels and thus may lead to error accumulation (Zhang et al., 2021; Arazo et al., 2020).

In this paper, we propose a Rank-aware Negative Training (RNT) framework to address SSTC

[*]Corresponding author.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 771–786, 2023. https://doi.org/10.1162/tacl_a_00574
Action Editor: Doug Downey. Submission batch: 12/2022; Revision batch: 3/2023; Published 7/2023.

under learning with noisy label settings. To alleviate the domination of noisy information during training, we adopt reasoning with an uncertainty-based approach to rank the unlabeled texts by measuring their shared features, also known as evidential support, with the labeled texts. Eventually, the shared features that serve as a medium to convey knowledge from labeled texts (i.e., evidence) to the unlabeled texts (i.e., inference) are regarded as belief functions to reason about the degree of noisiness. These belief functions are combined to reach a final belief about the text being mislabeled. In other words, we attempt to discard the texts whose pseudo-labels may introduce inaccurate information to the training process.

Moreover, we propose using negative training (NT) (Kim et al., 2019) to robustly train with potential noisy pseudo-labels. Unlike positive training, NT is an indirect learning method that trains the network based on the concept that ‘‘the input sentence does not belong to the complementary label’’, whereas a complementary label is randomly generated from the label space except the label of the sentence on-target. Considering the AG News dataset, given a sentence annotated as *sport*, the complementary label is randomly selected from all labels except *sport* (e.g., *business*). Intuitively, the probability of a true label serving as a complementary label is low and thus can reduce the noisy information during the training process. Finally, we conduct extensive experiments on various text classification benchmark datasets with different ratios of labeled examples, resulting in better performance on the test data. Experimental results suggest that RNT can mostly outperform the SSTC-based alternatives. Moreover, it has been empirically shown that RNT can perform better than PLMs fine-tuned on sufficient labeled examples.

In brief, our main contributions are three-fold:

- We propose a rank-aware negative training framework, namely, RNT, to address the semi-supervised text classification problem as learning in the noisy label setting.
- We introduce reasoning with an uncertainty-based solution to discard texts with the potential noisy pseudo-labels by measuring evidential support received from the labeled texts.
- We evaluate the proposed solution on various text classification benchmark datasets. Our extensive experiments show that it consistently overcomes the state-of-the-art alternatives in most cases and achieves competitive performance in others.

2 Related Work

This section reviews the existing solutions of the SSTC task and learning with noisy labels.

Text Classification. Text classification aims at assigning a given document to a number of semantic categories, which could be a sentiment, topic, or aspect (Hu and Liu, 2004; Liu, 2012; Schouten and Frasincar, 2016). Earlier solutions were usually equipped with deep memory or an attention mechanism to learn semantic representation in response to a given category (Socher et al., 2013b; Zhang et al., 2015; Wang et al., 2016; Ma et al., 2017; Chen et al., 2017; Johnson and Zhang, 2017; Conneau et al., 2017; Song et al., 2019; Murtadha et al., 2020; Tsai et al., 2022). Recently, many NLP tasks have experienced a considerable shift towards fine-tuning the PLMs (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Zaheer et al., 2020; Chen et al., 2022; Tsai et al., 2022; Ahmed et al., 2022). Despite the effectiveness of these approaches, the performance heavily relies on the quality of the labeled data, which requires intensive human labor.

Semi-supervised Text Classification. Partially supervised text classification, also known as learning from Positive and Unlabeled (PU) examples, aims at building a classifier using P and U in the absence of negative examples to classify the Unlabeled examples (Liu et al., 2002; Li et al., 2010; Liu et al., 2011). Recent SSTC approaches primarily focus on exploiting the consistency in the predictions for the same samples under different perturbations. Miyato et al. (2016) established virtual adversarial training that perturbs word embeddings to encourage consistency between perturbed embeddings. Variational auto-encoders-based approaches (Yang et al., 2017; Chen et al., 2018; Gururangan et al., 2019) attempted to reconstruct instances and utilized the latent variables to classify text. Unsupervised data augmentation (UDA) (Xie et al., 2020) performed consistency training by making features consistent between back-translated instances. However, these methods mostly require additional systems (e.g., NMT back-translation), which may be bothersome in real-world scenarios. Mukherjee and Awadallah, (2020) and Tsai et al. (2022) introduced uncertainty-driven self-training-based solutions to select samples and performed self-training on the selected data. An iterative framework (Ma et al., 2021), named SENT, proposed to address distant relation extraction via negative training. Self-Pretraining (Karisani and Karisani, 2021) was introduced to employ an iterative distillation procedure to cope with the inherent problems of self-training. SSTC-based approaches and their limitations are well described by van Engelen and Hoos (2020) and Yang et al. (2022). Recently, S^{2}TC-BDD (Li et al., 2021) was introduced to balance the label angle variances (i.e., the angles between deep representations of texts and weight vectors of labels), also called the margin bias. Despite the effectiveness of these methods, the unlabeled instances contribute equally likely to the labeled ones; therefore, the performance heavily relies on the quality of pseudo-labels. Unlikely, our proposed solution addresses the SSTC task as a learning under noisy label settings problem. Since the pseudo-labels are automatically labeled by the machine, we thus regard them as noisy labels and introduce a ranking approach to filter the highly risky mislabeled instances. To alleviate the noisy information resulting from the filtering process, we use negative training that performs classification based on the concept that ‘‘the input instance does not belong to the complementary label’’.

Learning with Noisy Labels. Learning with noisy data has been extensively studied, especially in the computer vision. The existing solutions introduced various methods to relabel the noisy samples in order to correct the loss function. To this end, several relabeling methods have been introduced to treat all samples equally to model the noisy ones, including directed graphical models (Xiao et al., 2015), conditional random fields (Vahdat, 2017), knowledge graphs (Baek et al., 2022), or deep neural networks (Veit et al., 2017; Lee et al., 2018). However, they were built based on semi-supervised learning, where access to a limited number of clean data is required. Ma et al. (2018) introduced a bootstrapping method to modify the loss with model predictions by exploiting the dimensionality of feature subspaces. Patrini et al. (2017) proposed to estimate the label corruption matrix for loss correction. Another direction of research on loss correction investigated two approaches, including reweighting training samples and separating clean and noisy samples (Thulasidasan et al., 2019; Konstantinov and Lampert, 2019). Shen and Sanghavi (2019) have claimed that the deep classifier normally learns the clean instances faster than the noisy ones. Based on this claim, they consider instances with smaller losses as clean ones. A negative training technique (Kim et al., 2019) was introduced to train the model based on the complementary label, which is randomly generated from the label space except for the label on-target. The goal is to encourage the probability to follow a distribution such that the noisy instances are largely distributed in low-value areas and the clean data are generally distributed in high-value areas to facilitate the separation process. Han et al. (2018) proposed to jointly train two networks that select small-loss samples within each mini-batch to train each other. Based on this paradigm, Yu et al. 2019 proposed updating the network on disagreement data to keep the two networks diverged. In this paper, we leverage a robust negative loss (Kim et al., 2019) for noisy data training.

3 Ranked-aware Negative Training

This section describes the proposed framework, namely, Rank-aware Negative Training (RNT), for semi-supervised text classification. An example of RNT is depicted in Figure 1. Suppose we have a training dataset D consisting of a limited labeled set D_l and a large unlabeled set D_u. We follow the pseudo-labels method introduced by Lee (2013) to associate D_u with pseudo-labels based on the concept of positive training. Simply put, we fine-tune the pre-trained language models (e.g., BERT) on the D_l set. It is noteworthy that we use BERT for a fair comparison, while other models can be used similarly. As the pseudo-labels are not manually annotated, we propose ranking the texts based on their potential for mislabeling to identify and discard the most risky mislabeled texts. Specifically, we first capture the shared information (i.e., we refer to this as the evidential support) between the labeled and unlabeled instances. Then, we measure the amount of support that an unlabeled instance receives from the labeled instances being correctly

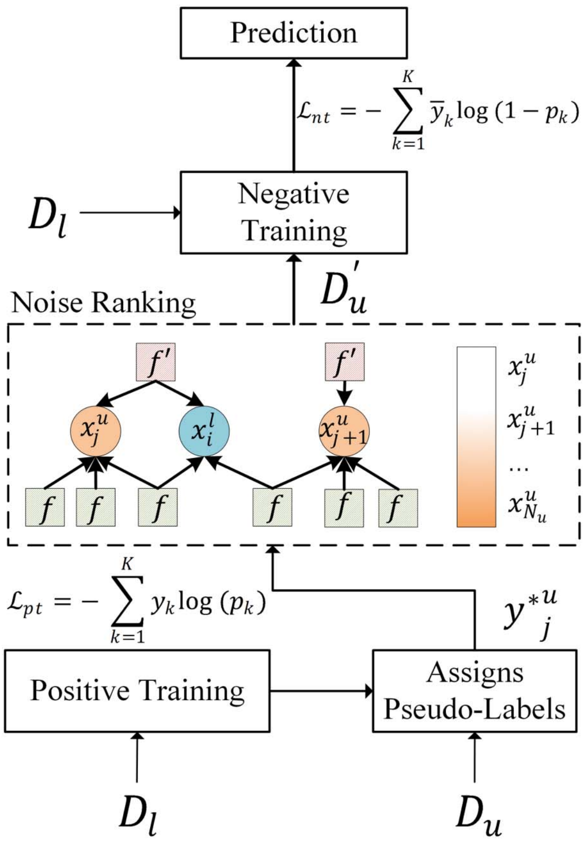


Figure 1: An example of the proposed framework. D_l, D_u, and D_u' denote labeled set, unlabeled set, and filtered unlabeled set, respectively. Briefly, RNT consists of three key steps: (1) Training with PT on limited labeled texts and then iteratively predicting the unlabeled texts as their pseudo-labels; (2) Measuring the evidential support based on the learned embedding space of PT to estimate the degree of noise; and (3) Training with NT on the mixture of clean and filtered data.

labeled. We denote the filtered set as D_u' in Figure 1. Finally, we train on both D_l and D_u' through the concept of the negative training. Next, we describe the framework in detail.

3.1 Task Description

Semi-Supervised Text Classification (SSTC). Let D be the training dataset consisting of a limited labeled set $D_l = \{(x_i^l, y_i^l)\}_{i=1}^{i=N_l}$ and a large unlabeled text set $D_u = \{(x_j^u)\}_{j=1}^{j=N_u}$, where x_i^l and x_j^u denote the input sequences of labeled and unlabeled texts, respectively, and $y_i^l \in \{0,1\}^K$ represents the corresponding one-hot label vector of x_i^l. The goal is to learn a classifier that leverages both D_l and D_u to better generate in the inference step, also known as inductive SSTC.

3.2 Positive and Negative Training

Positive Training (PT). A typical method of training a model with a given input instance and the corresponding labels is referred to as PT. In other words, the model is trained based on the concept that ‘‘the input instance belongs to this label’’. Considering a multi-class classification problem, let $x \in \mathcal{X}$ be an input, $y \in \{0,1\}^K$ be a c-dimension one-hot vector of its label. The training objective $f(x;\theta)$ aims to map the input instance to the k-dimensional score space $f : \mathcal{X} \to \mathbb{R}^k$, where θ is the set of parameters. To achieve this, PT uses the cross-entropy loss function defined as follows:

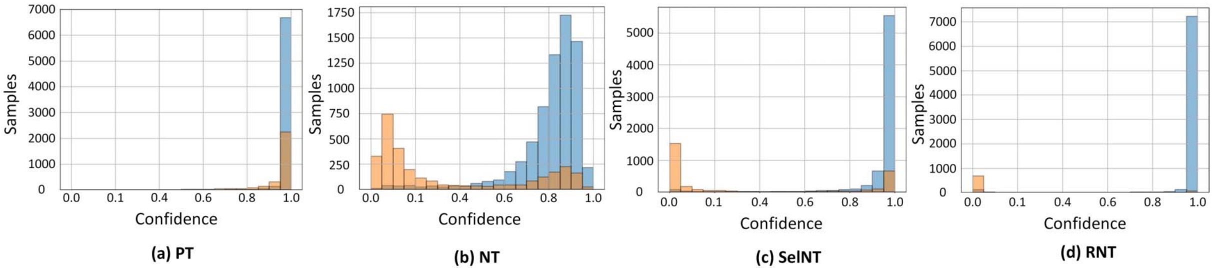
$$\mathcal{L}_{pt} = -\sum_{k=1}^{K} y_k \log(p_k), \qquad (1)$$

where p_k denotes the probability of the k^{th} label. Equation 1 satisfies the claim of PT to optimize the probability value corresponding to the given label as 1 ($p_k \to 1$).

Negative Training (NT). Unlike PT, the model is trained based on the concept that ‘‘the input text does not belong to this label’’. Specifically, given an input text x with a label $y \in \{0,1\}^K$, a complementary label $\bar{y}$ is generated by randomly sampling from the label space except y (e.g., $\bar{y} \in \mathbb{R} \backslash \{y\}$). The cross-entropy loss function of NT is defined as follows.

$$\mathcal{L}_{nt} = -\sum_{k=1}^{K} \overline{y_k} \log(1 - p_k). \qquad (2)$$

To illustrate the robustness of PT and NT against noisiness, we train both techniques on the AG News dataset corrupted with randomly 30% of symmetric noise (i.e., associating the instance with a random label). In terms of confidence (i.e., the probability of the true class), we illustrate the histogram of the training data after PT and NT in Figure 2. As can be seen, with PT in Figure 2(a), the confidence of both clean and noisy instances increases simultaneously. With NT in Figure 2 (b), in contrast, the noisy instances yield much lower confidence compared to the clean ones and thus discourages the domination of noisy data. After NT training, we train the model with only the samples having NT confidence over $\frac{1}{K}$, where K denotes the number of classes. We refer to this



(a) PT

(b) NT

(c) SelNT

(d) RNT

Figure 2: A histogram of PT, NT, and RNT data training distribution conducted on AG News dataset with random 30% noisy-labels, in which blue represents the clean data and orange indicates the noisy data. SelNT further trains the model with only the samples having NT confidence over $\frac{1}{K}$, where K denotes the number of classes.

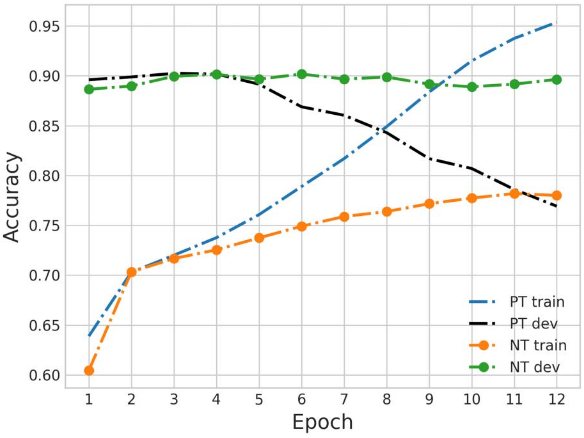


Figure 3: A comparison between PT and NT techniques trained on the AG News dataset corrupted with randomly 30% of symmetric noise. The accuracy of PT on the clean Dev data increases in the early stage. However, overfitting to the noisy training examples results in gradual inaccurate performance on the clean Dev data.

process as Selective NT (SelNT), as illustrated in Figure 2 (c) (Kim et al., 2019). We also depict the distribution of proposed RNT in Figure 2(d), which demonstrates the improvement of RNT in terms of noise filtering. In terms of performance, as shown in Figure 3, the accuracy of PT on the Dev data increases in the early stage. However, the direct mapping of features to the noisy labels eventually leads to the overfitting problem and thus gradually results in inaccurate performance on the clean Dev data.

3.3 Noise Ranking

We begin by extracting the shared features (i.e., evidential support) between the evidences (i.e., the labeled texts) and the inference (i.e., the unlabeled texts). Then, we adopt a reasoning with uncertainty approach to measure the evidential support. The instance with higher evidential support is regarded as a less potential noisy instance. An illustrative example is shown in Figure 4. Next, we describe the process in detail.

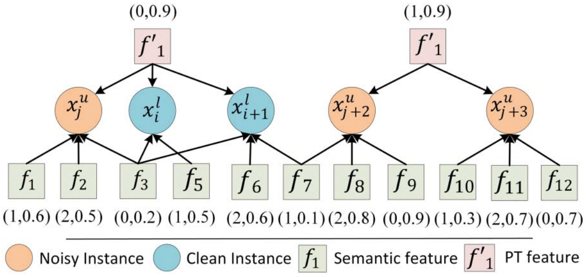


Figure 4: An illustrative example of the evidential support. The instances x_j^u and x_i^l exhibit a similar $\mathcal{L}_{cos}$ value in response to class 0 (i.e., $f_3(0, 0.2)$). The approximate PT's confidence of 0.9, represented by $f_1'(0, 0.9)$, further strengthens this similarity. Consequently, the instance x_j^u is considered less noisy due to the higher degree of evidential support it receives.

3.3.1 Feature Generation

Recall that RNT begins by training on the labeled data using the PT technique. Consequently, we rely on the learned latent space of PT to generate various features with three properties, including automatically generated, discriminating, and high-coverage, as follows.

Semantic Distance. For each instance $x_i \in \{D_l, D_u\}$, we recompute its semantic relatedness to each label $y_i \in \mathcal{Y}$ based on the Angular Margin (AM) loss (Wang et al., 2018).

The AM loss adds a margin term to the Softmax loss based on the angle (i.e., cosine similarity) between an input sample's feature vector and the actual class's weight vector. Notably, the margin term encourages the network to learn feature representations that are well-separated and distinct for different classes. As a result, the angle between the feature vectors of an input sample and different classes becomes an essential factor in estimating the degree of noisiness. For clarity, we first describe the AM loss with respect to angles. Given a training example $(x_i; y_i)$, it can be formulated as:

$$\mathcal{L}_{cos}(x_i, y_i, \phi) = -\sum\nolimits_{k=1}^{K} y_{ik} \log(\frac{e^{s(cos(\theta_{ik} - y_{ik} m))}}{\sum_{j=1}^{K} e^{s(cos(\theta_{ij} - y_{ij} m))}}), \quad (3)$$

where ϕ denotes the model parameters and $cos(.)$ stands for the cosine similarity, which can be read as the angular distance between feature vectors and the class weights. Given an unlabeled instance x_j^u, we recompute its AM loss to each class $y_i \in \mathcal{Y}$ as follows:

$$\mathcal{L}_{cos}(x_j^u, y_i, \phi) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} y_{ik} \log(\theta_{jn}), \quad (4)$$

where N is the number of samples (e.g., 5) from D_l labeled with y_i (i.e., class on-target) and θ_{jn} denotes the cosine similarity between x_j^u and x_n^l (i.e., the deep representations of the PT classifier). The intuition behind this feature is that an unlabeled instance x_j^u, which receives close amount of support from different classes, is regarded as potentially mislabeled. We denote this feature as f and its value consists of the corresponding class y_i as well as the value of $\mathcal{L}_{cos}$. To enable valuable shared knowledge between instances, $\mathcal{L}_{cos}$ is approximated to one digit (e.g., $\mathcal{L}_{cos}(x_j^u, 1) = 0.213 \approx 0.2$). Considering the illustrative example in Figure 4, x_j^u and x_i^l approximately share the same $\mathcal{L}_{cos}$ in response to class 0, (i.e., $f_3(0, 0.2)$).

PT Confidence. Instances with extreme confidence (i.e., close to 1) are generally considered to have a low risk of being mislabeled (Hou et al., 2020). To incorporate the class distribution of PT into the evidential support measurement process, we introduce a new feature, denoted as f', whose value consists of the predicted class and its corresponding probability. Considering the illustrative example in Figure 4, x_j^u and x_i^l share f_1' (i.e., $f_1'(0, 0.9)$), which can be read as both instances are related to the class 0 based on PT classifier with 0.9 confidence.

3.3.2 Evidential Support Measurement

Now can capture shared knowledge between the labeled and unlabeled instances (i.e., the evidential support). We leverage Dempster Shafer Theory (DST) (Yang and Xu, 2013) to address evidential support measurement as reasoning with uncertainty. The goal is to estimate the degree of noisiness for an unlabeled instance by combining its evidence from multiple sources of uncertain information (i.e., PT and semantic features). To achieve this, DST applies Dempster's rule, which combines the mass functions of each source of evidence to form a joint mass function. It is noteworthy that DST has been widely used for various purposes of reasoning (Liu et al., 2018; Wang et al., 2021; Ahmed et al., 2021). The basic concepts of DST are:

- **Proposition**. It refers to all possible states of a situation under consideration. Two propositions are defined: ‘‘clean instance’’, denoted by C, and ‘‘unclean instance’’, denoted by U. Let proposition be $X = \{C, U\}$ and a power set of X be $2^X = \{\emptyset, C, U, X\}$.
- **Belief function**. It associates each $E \in 2^X$ with a degree of belief (or mass), which satisfies $\sum_{E \in 2^X} m(E) = 1$ and $m(\emptyset) = 0$. Different belief functions for various evidences are defined (i.e., the generated features).

Given an unlabeled instance x_j^u and its semantic feature f, we estimate the evidential support that x_j^u receives from labeled instances that share f by the belief function:

$$m_f(E) = \begin{cases} (1-d_f)\max\{P(f), 1-P(f)\} & E = \{C\} \\ (1-d_f)\min\{P(f), 1-P(f)\} & E = \{U\} \\ d_f & E = \{C, U\} \end{cases} \quad (5)$$

where d_f denotes the degree of uncertainty of f, and $P(f)$ is the division of the number of positive instances (i.e., the labeled instances with the same class of the feature on-target f) by all labeled instances shared f. Consider the illustrative example in Figure 4, $f_3(0, 0.2)$ (i.e., semantically related to class 0 with approximated similarity of 0.2), suppose that the positive instances x_i^l and x_{i+1}^l are annotated with class 0, then $P(f_3) = 1.0$. Equation 5 can be read as the more extreme the value of $P(f)$ (i.e., close to 0 or 1) is, the more evidential support the element of C should receive from the feature f. Similarly, we use Equation 5 to estimate the evidential support $m_{f'}(E)$ that x_i^u receives from f'. Note that d_f represents the impact that a given feature may have on the final degree of belief in terms of evidential support measurement. The lower the value, the greater the impact. Note that both types of features are generated based on the latent space of the PT classifier that we believe in its semantic representation as it is trained on the labeled data. Therefore, we empirically set d_f to a small unified value (i.e., 0.2 in our experiments).

The overall evidential support of $E = \{C\}$ that x_j^u receives from its observations is estimated by combining the estimated beliefs as follows:

$$m(E) = m_{f_1}(E) \oplus \ldots \oplus m_{f_n}(E) \oplus m_{f'}(E), \quad (6)$$

where $m(E)$ represents the total amount of evidential support that x_j^u receives, and the combination is computed from the two sets of belief functions, $m_{f_1}(E)$ and $m_{f_2}(E)$, as follows:

$$m_{f_1}(E) \oplus m_{f_2}(E) = \frac{1}{1-U} \sum\nolimits_{E' \cap E'' = E} m_{f_1}(E') m_{f_2}(E''), \quad (7)$$

where E' and E'' denote the power set 2^X elements and $U = \sum_{E' \cap E'' = \emptyset} m_{f_1}(E') m_{f_2}(E'')$ is a measure of the amount of conflict between E' and E''. In words, given the element of $E = \{C\}$, we multiply the combinations of E' and E'' such that $E' \cap E'' = C$ and thus can be regarded as a measure for the amount of support from $\{C\}$. For time complexity, each iteration takes $O(n \times n_f)$ time with n instances and n_f the number of the generated features. Thus, the time complexity can be represented by $O(n^2 \times n_f)$.

3.4 Training Procedure

Now that we can measure the evidential support, we then rank the instances of D_u and select the less risky instances as the filtered set, denoted as $D'_u = \{(x_j^u, y_j^u)\}_{i=1}^{j=N_f}$. Note that the value of N_f is fine-tuned using the Dev set (please refer to Section 4.3 for more details). Finally, we combine both sets D_l and D'_u together for the final NT training, as illustrated in Figure 1. The training procedure can be explained by the following steps. We first generate pseudo-labels using the PT technique from Eq. 1. Then, we apply DST to filter the highly risky instances. Finally, we adopt NT technique, Eq. 2, to alleviate the noisy information during the training. Furthermore, to improve the convergence after NT, we follow Kim et al. (2019) by training only with the instances whose confidence is over $\frac{1}{K}$, denoted as SelNT in Figure 2(c).

4 Experimental Setup

4.1 Dataset

We validate the performance of the proposed RNT on various text classification benchmark datasets (Table 1). In particular, we rely on AG News (Zhang et al., 2015), Yahoo (Chang et al., 2008), Yelp (Zhang et al., 2015), DBPedia (Zhang et al., 2015), TREC (Li and Roth, 2002), SST (Socher et al., 2013a), CR (Ding et al., 2008), MR (Pang and Lee, 2005), TNEWS, and OCNLI (Xu et al., 2020). For the AG News, Yelp, and Yahoo datasets, we follow the comparative approaches by forming the unlabeled training set D_u, labeled training set D_l, and development set by randomly drawing from the corresponding original training datasets. For the other datasets, we split the training set into 10% and 90% for D_l and D_u, respectively. Note that we utilize the original test sets for prediction evaluation.

4.2 Comparative Baselines

For fairness, we only include the semi-supervised learning methods that were built upon the contextual embedding models (e.g., BERT):

- **PLM** is a pre-trained language model directly fine-tuned on the labeled data. We compared to BERT (Devlin et al., 2019; Cui et al., 2021) and RoBERTa (Liu et al., 2019);
- **UDA** (Xie et al., 2020) is an SSTC method based on unsupervised data augmentation with back translation. We use German and English languages for back-translation of English and Chinese, datasets, respectively;
- **UST** (Mukherjee and Awadallah, 2020) introduces select samples by information gain and utilizes cross-entropy loss to perform self-training;

Dataset	#Class	Train		#Dev	#Test	Length	Language	Task	Metric
		#Lab	#Unlab						
AG News	4	10k	20k	8k	7.6k	100	English	Topic	Macro-F1
Yelp	5	10k	20k	10k	5k	256	English	Sentiment	Macro-F1
Yahoo	10	10k	40k	20k	60k	256	English	Topic	Macro-F1
DBPedia	14	10k	20k	10k	70k	160	English	Topic	Macro-F1
TREC	6	5.4k	NA	1.1k	500	30	English	Question	Macro-F1
SST	{2,5}	6.9k	NA	871	1.8k	50	English	Sentiment	Macro-F1
CR	2	3k	NA	378	372	50	English	Sentiment	Macro-F1
MR	2	6.9k	NA	1.7k	2k	50	English	Sentiment	Macro-F1
TNEWS	15	53.3k	NA	10k	10k	128	Chinese	Topic	Accuracy
OCNLI	3	50k	NA	3k	3k	128	Chinese	NLI	Accuracy

Table 1: The statistics of benchmark datasets, where #Lab and #Unlab denote the number of labeled and unlabeled texts, respectively. Note that for datasets with NA, we split #Lab into 10% and 90% for #Lab and #Unlab, respectively.

- **S^{2}TC-BDD** (Li et al., 2021) is an SSTC method that addresses the margin bias problem by balancing the label angle variances.

4.3 Experimental Settings

- **Hyper-parameters.** We use 12 heads and layers and keep the dropout probability to 0.1 with 30 epochs, learning rate of $2e^{-5}$ and 32 batch size. To guarantee the re-productivity without manual effort, we rely on the Dev set to automatically set the value of N_f (i.e., the number of instances in D'_u). First, the ranked Dev set is split into small proportions (i.e., max is 10). Then, m is set to proportions that meet the condition $\lambda = max(p) - st(p)$, where p is a vector that represents the accuracy of RNT on each proportion and st denotes the standard deviation. For example, $\theta = 0.2$ means that D'_u consists of the first 20% of the ranked D_u, as shown in Figure 5. We set the number of negative samples to $K - 1$, where K is the number of classes in the labeled training set.
- **Metrics.** We use the accuracy metric on Clue datasets, including TNEWS and OCNLI, and Macro-F1 scores for all other datasets.

5 Evaluation and Results

We describe the evaluation tasks and report the experimental results in this section. The evaluation criteria are: (I) Is RNT able to rank the instances of being mislabeled?; (II) Can the filtered data enhance the performance of the clean test data?

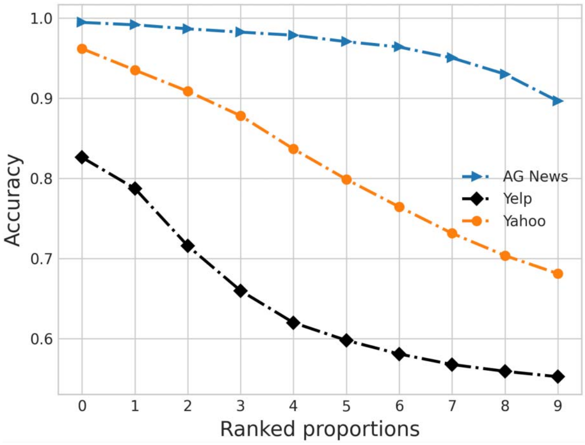


Figure 5: Ranking evaluation on Dev sets with $N_l = 1k$. The ranked Dev set is first split into 10 proportions equally-likely. Then, each proportion is inferred (i.e., calculate its accuracy) independently. The accuracy gradually drops as the noisy texts increase. In our experiment, we choose the proportions whose instances meet λ Section 4.3 for further NT training.

5.1 Results

We use the Dev set to select the best model and average three runs with different seeds. The experimental results are reported in Tables 2, 3, and 4, from which we have made the following observations.

- **Compared to the baselines**, RNT gives the best results compared to its alternatives in most cases and achieves competitive performance in others. We also observe that SSTC-based approaches comfortably outperform the PLM fine-tuning when training

PLM	Model	AG News			Yelp			Yahoo			DBPedia
		30	1k	10k	30	1k	10k	30	1k	10k	30
BERT-Base	Fine-tuning	84.1±0.9	87.8±0.3	90.5±0.2	42.2±1.7	53.2±0.8	58.6±0.5	63.2±0.5	67.1±0.3	70.8±0.2	97.1±0.9
	UDA	85.7±0.3	88.3	90.6	44.6±1.2	55.0	57.6	66.4±0.5	66.6	70.4	98.5±0.6
	UST	**87.2**±0.6	88.6	90.8	44.8±1.1	54.2	57.7	**66.5**±0.3	67.5	71.1	98.4±0.6
	S^{2}TC-BDD	86.9±0.7	88.9	90.7	**45.9**±1.4	55.0	58.6	66.2±0.6	68.0	70.9	**98.8**±0.7
	RNT (Ours)	86.7±0.3	**89.4**±0.1	**91.9**±0.1	44.9±1.2	**56.6**±0.6	**60.2**±0.1	66.2±0.3	**69.1**±0.2	**72.7**±0.1	98.2±0.4
RoBERTa-Base	Fine-tuning	84.9±0.7	88.5±0.3	91.0±0.2	53.7±1.6	57.8±0.7	62.5±0.4	66.6±0.4	68.3±0.5	72.3±0.2	98.1±0.5
	RNT (Ours)	**86.9**±0.4	**89.6**±0.1	**92.2**±0.2	**53.9**±1.4	**60.0**±0.5	63.8±0.1	**67.2**±0.4	**69.6**±0.2	**73.7**±0.1	98.4±0.2
RoBERTa-Large	Fine-tuning	86.5±0.4	89.1±0.2	91.8±0.2	56.2±1.3	62.3±0.6	66.0±0.4	67.8±0.3	70.3±0.3	73.7±0.1	98.3±0.3
	RNT (Ours)	**87.8**±0.3	**89.8**±0.1	**92.6**±0.1	**58.3**±0.9	**63.1**±0.4	**66.8**±0.2	**68.9**±0.2	**71.2**±0.2	**74.3**±0.1	98.8±0.2

Table 2: Comparative results with the state-of-the-art alternatives on 30 examples per label and $N_l \in \{1k, 10k\}$. The results of $N_l \in \{1k, 10k\}$ are retrieved from S^{2}TC-BDD (Li et al., 2021), while the others are our implementations. The scores consists of the average of three runs, and the best scores are in bold.

PLM	Model	TREC		SST-2		SST-5		CR		MR	
		30	10%	30	10%	30	10%	30	10%	30	10%
BERT-Base	Fine-tuning	78.7±1.6	87.1±1.0	76.9±1.6	85.2±0.8	33.2±1.4	39.0±1.1	74.7±1.2	85.8±0.9	66.6±1.4	80.7±0.7
	UDA	83.5±1.1	91.2±0.7	79.9±1.3	85.6±0.3	33.6±1.1	40.6±0.8	81.0±0.7	87.7±0.6	**72.9**±0.9	81.0±0.1
	UST	83.3±1.2	**92.1**±0.8	78.7±1.0	85.6±0.4	33.9±1.1	40.8±0.7	**82.7**±0.8	87.8±0.3	71.1±1.0	81.0±0.3
	S^{2}TC-BDD	81.2±1.3	91.2±0.9	81.1±1.2	85.7±0.5	34.6±1.3	39.6±0.5	82.3±0.9	87.6±0.7	72.1±0.9	80.0±0.6
	RNT (Ours)	**85.2**±1.1	91.4±0.7	**83.8**±1.3	**87.6**±0.4	**35.9**±1.2	**42.3**±0.9	82.6±0.9	**89.3**±0.4	71.5±1.0	**82.4**±0.3
RoBERTa-Base	Fine-tuning	84.2±1.3	92.1±0.7	85.0±0.9	89.5±0.4	39.3±1.0	47.6±0.7	86.5±1.2	91.1±0.7	71.2±1.4	84.9±0.5
	RNT (Ours)	**86.7**±0.8	**93.2**±0.4	**87.7**±0.7	**90.7**±0.4	**40.5**±0.6	**49.6**±0.4	**88.9**±0.6	**92.5**±0.2	**75.8**±0.7	**86.4**±0.2
RoBERTa-Large	Fine-tuning	88.9±1.1	92.5±0.6	87.7±1.0	92.3±0.7	40.5±0.8	51.0±0.6	89.7±0.9	91.8±0.8	82.4±1.2	88.3±0.6
	RNT (Ours)	**89.6**±0.6	**94.0**±0.4	**89.6**±0.6	**93.2**±0.5	**42.8**±0.5	**52.4**±0.3	**92.3**±0.6	**92.6**±0.3	**85.9**±0.6	**88.4**±0.3

Table 3: Comparative results with the state-of-the-art alternatives with 30 samples per label and $N_l = 10\%$ of the labeled texts. Note that all results are the average of three runs with different seeds. Best scores are in bold.

Model	TNEWS	OCNLI
	10%	10%
Fine-tuning	53.9	62.6
UDA	52.3	63.8
UST	54.3	63.7
S^{2}TC-BDD	53.4	64.5
RNT (Ours)	**54.6**±0.4	**65.2**±0.3

Table 4: Comparative results on Chinese datasets based on initial weights from RoBERTa-Large (Cui et al., 2020). The best scores are in bold.

with scarce labeled data (e.g., $N_l = 30$); however, the same performance is expected when N_l is increased (e.g., $N_l \in \{1k, 10k\}$), but it was not supported by the experiments. Furthermore, experimental results demonstrate that RNT is not sensitive to the number of classes compared to SSTC-based alternatives. For instance, UDA (Xie et al., 2020) can perform better on the binary datasets, as shown in Table 3.

- **Compared to the PLM** fine-tuned on the labeled data, RNT comfortably overcomes PLM by considerable margins. For example, the Macro-F1 scores of RNT with $N_l = 30$ are even about 2.6%, 2.7%, and 3.0% on AG News, Yelp and Yahoo datasets, respectively. Moreover, we also observe that RNT can perform better than PLM fine-tuned on sufficient labeled data (e.g., $N_l = 10k$).

5.2 Mislabeling Filtering Evaluation

To evaluate the ability of RNT in mislabeling filtering, we conduct experiments on the Dev sets of AG News, Yelp, and Yahoo datasets as follows. We first associate the instances with the corresponding pseudo-labels (i.e., inferring using the PT classifier). Then, we require RNT to rank them based on their evidential support received from the clean training set (i.e., $N_l = 1k$). Since we have access to the true labels of the Dev set, we can evaluate the performance of the filtering process. Specifically, we divide the ranked Dev set into ten equally-likely proportions (note that we keep the same order of ranking) and calculate the accuracy of each proportion separately (i.e., comparing the pseudo-labels with the true labels). The proportions, as shown in Figure 5, are significantly correlated with the extent of mislabeling. In

Dataset	N_l	Full Dev		Filtering		
		Acc	F1	Prop	Acc	F1
AG News	1k	88.1	88.1	70%	95.8	95.2
	10k	91.9	91.9	70%	98.2	97.9
Yelp	1k	53.8	53.2	30%	68.8	64.2
	10k	60.5	60.2	30%	75.9	72.5
Yahoo	1k	67.1	67.0	30%	89.6	72.6
	10k	72.0	71.2	40%	91.0	83.6

Table 5: Filtering evaluation on Dev sets. Prop, Acc, and F1 denote proportion (i.e., the ratio of filtered texts), accuracy, and Macro-F1, respectively.

	AG News	Yelp	Yahoo
BERT (fine-tune)	87.8	53.2	67.1
S^{2}TC-BDD	88.9	55.0	68.0
RNT w/o ranking	88.1	55.0	67.9
RNT	89.4	56.6	69.1

Table 6: The impact of noise filtering to the overall performance. Note that the number of labeled data is set to $N_l = 1k$. Removing noise ranking from RNT leads to a noticeable performance drop; however, it still performs better than BERT-Based fine-tuned on labeled data and achieves competitive scores comparable to S^{2}TC-BDD.

other words, the accuracy score gradually drops as the mislabeled instances increase and vice-versa. Note that we report the accuracy due to the imbalance labels in the proportions. Moreover, we report the performance of both the full Dev set and the filtered set in Table 5.

5.3 The Impact of Noise Filtering

To assess the impact of noise filtering on the overall performance of RNT, we remove DST and conduct experiments on the AG News, Yelp, and Yahoo datasets. The experimental results presented in Table 6 show that removing noise ranking from RNT causes a performance drop of 1.3, 1.6, and 1.2 on the AG News, Yelp, and Yahoo datasets, respectively. This demonstrates the efficacy of a well-designed noisy ranking in improving text classification performance. Furthermore, we observe that even without noise filtering, RNT outperforms PLM fine-tuning and achieves competitive results compared to other alternatives. This supports the adoption of NT for noisy data.

5.4 The Effect of DST

To validate the contribution of DST on the final performance in terms of mislabeled instances filtering, we implement two variants, namely, RNT Pure and RNT PT-conf, as follows. The RNT Pure is trained on D_l and D_u as a whole without any filtering mechanism, while RNT PT-conf uses the PT confidence to filter the instances in D_u that do not meet the predefined threshold (i.e., 0.9 in our experiments). In other words, instead of DST, we rely on the PT confidence to discard the instances close to the boundary. Empirically, we conduct experiments on AG News, Yelp, and Yahoo datasets with various $N_l = \{30, 1k, 10k\}$. The comparative results are shown in Table 7, from which we made the following observations. Overall, RNT can mostly give the best performance, and the improvements are significant, especially with less limited data (e.g., $N_l = 30$). RNT Pure performs worse due to the absence of a filtering mechanism. RNT PT-conf can achieve competitive performance with sufficient labeled data (e.g., $N_l = 10k$) even in terms of uncertainty. However, it gradually drops with the decrease of labeled data. Intuitively, these results are expected as the performance of the PT classifier heavily relies on the amount of labeled data. In brief, the ablation study empirically supports the contribution of DST to the performance of RNT.

5.5 Denoising Evaluation

Recall that the ultimate goal of DST is to estimate the score of unlabeled instances being mislabeled by the PT classifier. To evaluate the ability of DST to denoising, we adopt a perturbation strategy that has been used widely in the literature (Belinkov and Bisk, 2018; Sun and Jiang, 2019). We randomly pick 30% of the Dev data as the noisy instances. For each instance, we randomly select 30% of the words to be perturbed as follows. Specifically, we apply four kinds of noise: (1) swap two letters per word; (2) delete a letter randomly in the middle of the word; (3) replace a random letter with another in a word; (4) insert a random letter in the middle of the word.

The evaluation results of denoising are reported in Table 8, from which we made the following observations. (1) A considerable margin exists

Model	AG News			Yelp			Yahoo		
	30	1k	10k	30	1k	10k	30	1k	10k
RNT Pure	82.9±0.8	88.1±0.2	91.3±0.2	42.6±1.7	55.0±0.8	59.8±0.3	65.1±0.7	67.9±0.2	71.9±0.1
RNT PT-conf	83.7±1.3	**89.7**±0.2	91.7±0.1	42.2±1.6	54.6±0.6	60.1±0.2	65.4±1.0	68.1±0.4	72.3±0.1
RNT (Ours)	**86.7**±0.3	89.4±0.1	**91.9**±0.1	**44.9**±1.2	**56.6**±0.6	**60.2**±0.1	**66.2**±0.3	**69.1**±0.2	**72.7**±0.1

Table 7: The effect of DST on the performance of RNT. All variants are jointly trained on D_l and D_u using PT and NT. RNT Pure is trained on all instances in D_u without any filtering mechanism, while RNT PT-conf uses the PT-based confidence to filter the instances in D_u that does not meet a predefined threshold (i.e., 0.9 in our experiments).

Dataset	Full Dev		Denoising accuracy
	clean	noise	
AG News	91.92	84.74	87.53
Yelp	60.49	57.14	74.12
Yahoo	71.96	68.46	83.25

Table 8: Denoising evaluation with $N_l = 10k$ and 30% of noise instances. Full Dev denotes the performance on the clean and noisy Dev sets. Denoising indicates the ability of RNT to identify the clean instances.

between the performance of the PT classifier on the clean and noise data, demonstrating the impact of the generated noise. (2) Despite the well-recognized challenge of denoising in NLP, our proposed solution can mostly identify clean instances. (3) Even though the performance can be deemed considerable, noisy information may still exist in the filtered data; therefore, we use NT for further training.

6 Conclusion and Future Work

In this paper, we proposed a self-training semi-supervised framework, namely, RNT, to address the text classification problem in learning with noisy label settings. RNT first discards the high risky mislabeled texts based on reasoning with uncertainty theory. Then, it uses the negative training technique to reduce the noisy information during training. Our extensive experiments have shown that RNT mostly outperformed SSTC-based alternatives. Despite the robustness of negative training, clean samples that have identical distributions with test data are subjected to complementary labels. Consequently, both clean and potentially noisy samples contribute equally to the final performance. A combination of both positive and negative training strategies in a unified framework can remedy the abundance of noisy samples; however, this needs further investigation.

Acknowledgments

We extend our gratitude to the TACL action editor and the anonymous reviewers for their valuable feedback and insightful suggestions, which have significantly contributed to the improvement of our work.

References

Murtadha Ahmed, Qun Chen, Yanyan Wang, Youcef Nafa, Zhanhuai Li, and Tianyi Duan. 2021. DNN-driven gradual machine learning for aspect-term sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: ACL/IJCNLP, Findings*, pages 488–497. https://doi.org/10.18653/v1/2021.findings-acl.43

Murtadha Ahmed, Shengfeng Pan, Bo Wen, Jianlin Su, Wenze Zhang, and Yunfeng Liu. 2022. BERT-ASC: Auxiliary-sentence construction for implicit aspect learning in sentiment analysis. *CoRR*, abs/2203.11702. https://doi.org/10.48550/arXiv.2203.11702

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 *Proceedings of the International Joint Conference on Neural Networks, IJCNN*, pages 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207304

Kyungjune Baek, Seungho Lee, and Hyunjung Shim. 2022. Learning from better supervision: Self-distillation for learning with noisy labels. In *Proceedings of the 26th International Conference on Pattern Recognition, ICPR*, pages 1829–1835. https://doi.org/10.1109/ICPR56361.2022.9956388

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations, ICLR*.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI*, pages 830–835.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 2147–2157. https://doi.org/10.18653/v1/2020.acl-main.194

Mingda Chen, Qingming Tang, Karen Livescu, and Kevin Gimpel. 2018. Variational sequential labelers for semi-supervised learning. In *Proceedings of the Empirical Methods in Natural Language Processing, EMNLP*, pages 215–226. https://doi.org/10.18653/v1/D18-1020

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the Empirical Methods in Natural Language Processing, EMNLP 2017*, pages 452–461. https://doi.org/10.18653/v1/D17-1047

Qianben Chen, Richong Zhang, Yaowei Zheng, and Yongyi Mao. 2022. Dual Contrastive learning: Text classification via label-aware data augmentation. *CoRR*, abs/2201.08702. https://doi.org/10.48550/arXiv.2201.08702

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann LeCun. 2017. Very Deep Convolutional Networks for Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 1107–1116. https://doi.org/10.18653/v1/E17-1104

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: EMNLP, Findings*, pages 657–668. https://doi.org/10.18653/v1/2020.findings-emnlp.58

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514. https://doi.org/10.1109/TASLP.2021.3124365

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4171–4186.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM*, pages 231–240. https://doi.org/10.1145/1341531.1341561

Jesper E. van Engelen and Holger H. Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440. https://doi.org/10.1007/s10994-019-05855-6

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A. Smith. 2019. Variational pretraining for semi-supervised text classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 5880–5894. https://doi.org/10.18653/v1/P19-1590

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. In *Proceedings of the Neural Information Processing Systems*, volume 31.

Boyi Hou, Qun Chen, Zhaoqiang Chen, Youcef Nafa, and Zhanhuai Li. 2020. r-HUMO: A risk-aware human-machine cooperation framework for entity resolution with quality guarantees. *IEEE Transactions on Knowledge and Data Engineering*, 32(2):347–359. `https://doi.org/10.1109/TKDE.2018.2883532`

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. `https://doi.org/10.1145/1014052.1014073`

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 562–570. `https://doi.org/10.18653/v1/P17-1052`

Payam Karisani and Negin Karisani. 2021. Semi-supervised text classification via self-pretraining. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining WSDM*, pages 40–48. `https://doi.org/10.1145/3437963.3441814`

Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. 2019. NLNL: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*, pages 101–110. `https://doi.org/10.1109/ICCV.2019.00019`

Nikola Konstantinov and Christoph Lampert. 2019. Robust learning from untrusted sources. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3488–3498.

Dong-Hyuns Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. 2018. CleanNet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5447–5456. `https://doi.org/10.1109/CVPR.2018.00571`

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 7871–7880. `https://doi.org/10.18653/v1/2020.acl-main.703`

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 and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, pages 5044–5053. `https://doi.org/10.18653/v1/2021.acl-long.391`

Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020. DivideMix: Learning with noisy labels as semi-supervised learning. In *Proceedings of the 8th International Conference on Learning Representations, ICLR*.

Xiaoli Li, Bing Liu, and See-Kiong Ng. 2010. Negative training data can be harmful to text classification. In *Proceedings of Empirical Methods in Natural Language Processing, EMNLP*, pages 218–228.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics, COLING*. `https://doi.org/10.3115/1072228.1072378`

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167. `https://doi.org/10.1007/978-3-031-02145-9`

Bing Liu, Wee Sun Lee, Philip S. Yu, and Xiaoli Li. 2002. Partially supervised classification of text documents. In *Proceedings of the Nineteenth International Conference ICML*, pages 387–394.

Tao Liu, Xiaoyong Du, Yong-Dong Xu, Minghui Li, and Xiaolong Wang. 2011. Partially supervised text classification with multi-level examples. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence, AAAI*. https://doi.org/10.1609/aaai.v25i1.7969

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Zhunga Liu, Quan Pan, Jean Dezert, Jun-Wei Han, and You He. 2018. Classifier fusion with contextual reliability evaluation. *IEEE Transactions on Cybernetics*, 48(5):1605–1618. https://doi.org/10.1109/TCYB.2017.2710205, PubMed: 28613193

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, pages 4068–4074. https://doi.org/10.24963/ijcai.2017/568

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuanjing Huang, and Yaqian Zhou. 2021. SENT: Sentence-level distant relation extraction via negative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, pages 6201–6213. https://doi.org/10.18653/v1/2021.acl-long.484

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 *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3355–3364.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv e-prints*, page arXiv:1605.07725. https://doi.org/10.48550/arXiv.1605.07725

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993. https://doi.org/10.1109/TPAMI.2018.2858821, PubMed: 30040630

Subhabrata Mukherjee and Ahmed Hassan Awadallah. 2020. Uncertainty-aware Self-training for Few-shot Text Classification. In *Proceedings of the Annual Conference on Neural Information Processing Systems NeurIPS*.

Ahmed Murtadha, Qun Chen, and Zhanhuai Li. 2020. Constructing domain-dependent sentiment dictionary for sentiment analysis. *Neural Computing & Applications*, 32(18):14719–14732. https://doi.org/10.1007/s00521-020-04824-8

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics ACL*, pages 115–124. https://doi.org/10.3115/1219840.1219855

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 (CVPR)*, pages 2233–2241. https://doi.org/10.1109/CVPR.2017.240

Kim Schouten and Flavius Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering TKDE*, 28(3):813–830. https://doi.org/10.1109/TKDE.2015.2485209

Yanyao Shen and Sujay Sanghavi. 2019. Learning with bad training data via iterative trimmed loss minimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5739–5748.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013a. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the*

Association for Computational Linguistics, ACL, pages 455–465.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of Empirical Methods in Natural Language Processing, EMNLP*, pages 1631–1642.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *CoRR*, abs/1902.09314. https://doi.org/10.1007/978-3-030-30490-4_9

Yifu Sun and Haoming Jiang. 2019. Contextual text denoising with masked language model. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT)*, pages 286–290. Hong Kong, China. https://doi.org/10.18653/v1/D19-5537

Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. 2019. Combating label noise in deep learning using abstention. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6234–6243.

Austin Cheng-Yun Tsai, Sheng-Ya Lin, and Li-Chen Fu. 2022. Contrast-enhanced semi-supervised text classification with few labels. In *Proceedings of the Thirty-Sixth Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI*, pages 11394–11402. https://doi.org/10.1609/aaai.v36i10.21391

Arash Vahdat. 2017. Toward Robustness against Label Noise in Training Deep Discriminative Neural Networks. In *Proceedings of the Neural Information Processing Systems NeurPIS*, volume 30.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge J. Belongie. 2017. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6575–6583. https://doi.org/10.1109/CVPR.2017.696

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5265–5274. https://doi.org/10.1109/CVPR.2018.00552

Yanyan Wang, Qun Chen, Jiquan Shen, Boyi Hou, Murtadha Ahmed, and Zhanhuai Li. 2021. Aspect-level sentiment analysis based on gradual machine learning. *Knowledge-based Systems, KBS*, 212:106509. https://doi.org/10.1016/j.knosys.2020.106509

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 606–615. https://doi.org/10.18653/v1/D16-1058

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2691–2699. https://doi.org/10.1109/CVPR.2015.7298885

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of the Annual Conferenceon Neural Information Processing Systems*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING*, pages 4762–4772. https://doi.org/10.18653/v1/2020.coling-main.419

Jian-Bo Yang and Dong-Ling Xu. 2013. Evidential reasoning rule for evidence combination. *Artificial Intelligence*, 205:1–29. https://doi.org/10.1016/j.artint.2013.09.003

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. 2022. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20. https://doi.org/10.1109/TKDE.2022.3220219

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the Annual Conference on Neural Information Processing Systems NeurIPS*, pages 5754–5764.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, volume 70, pages 3881–3890.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7164–7173.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Proceedings of the Annual Conference on Neural Information Processing Systems, NeurIPS*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115. https://doi.org/10.1145/3446776

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Annual Conference on Neural Information Processing Systems*, pages 649–657.