Noisy Multi-Label Text Classification via Instance-Label Pair Correction

Pengyu Xu Mingyang Song Linkaida Liu Bing Liu
Hongjian Sun Liping Jing* Jian Yu
Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University, Beijing, China
pengyu@bjtu.edu.cn

Abstract

In noisy label learning, instance selection based on small-loss criteria has been proven to be highly effective. However, in the case of noisy multi-label text classification (NMLTC), the presence of noise is not limited to the *instance-level* but extends to the (instance-label) *pair-level*. This gives rise to two main challenges. (1) The loss information at the pair-level fails to capture the variations between instances. (2) There are two types of noise at the pair-level: false positives and false negatives. Identifying false negatives from a large pool of negative pairs presents an exceedingly difficult task. To tackle these issues, we propose a novel approach called instance-label pair correction (iLaCo), which aims to address the problem of noisy pair selection and correction in NMLTC tasks. Specifically, we first introduce a holistic selection metric that identifies noisy pairs by simultaneously considering global loss information and instance-specific ranking information. Secondly, we employ a filter guided by label correlation to focus exclusively on negative pairs with label relevance. This filter significantly reduces the difficulty of identifying false negatives. Experimental analysis indicates that our iLaCo framework effectively corrects noisy pairs in NMLTC datasets, leading to a significant improvement in model performance.

1 Introduction

Multi-label text classification (MLTC) aims to predict the most relevant labels for each text from a label set. In real applications, noise is inevitably present in the data of MLTC (Snow et al., 2008; Chen et al., 2023). It poses a significant challenge for machine learning models, particularly deep learning models (Frénay and Verleysen, 2014; Arazo et al., 2019). When dealing with learning from noisy labels (LNL), one effective approach to mitigate the impact of noisy data is to identify a clean subset through sample (instance) selection (Hu et al., 2023; Li et al., 2023). By removing data with noisy labels, we can reduce the influence of mislabeled data and improve the learning process. The small-loss criteria is a popular method (Han et al., 2018; Wei et al., 2020; Xia et al., 2022), assuming that samples with lower loss values are likely to be clean. This is based on the observation that deep networks initially learn simple patterns and later tend to overfit to the noisy patterns (Wei et al., 2022; Han et al., 2018; Northcutt et al., 2021).

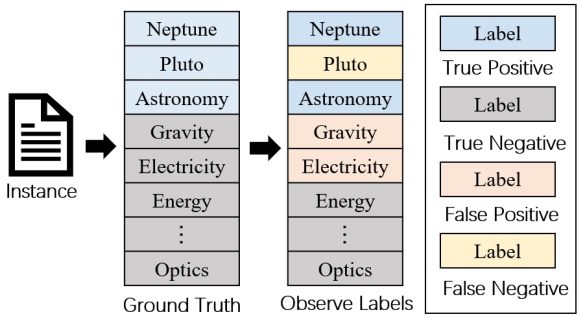


Figure 1: An example of noise in multi-label text classification.

However, what distinguishes noisy multi-label text classification (NMLTC) from typical LNL is that in NMLTC, noise occurs not at the *instance-level* but at the (instance-label) *pair-level*, as depicted in Figure 1. In this scenario, using the small-loss criterion for selection encounters two challenges. (1) The pair-level loss information is global and instance-independent, meaning it does not capture the distinctions between each individual instance. (2) NMLTC exhibits two types of noise: false positives (FP) and false negatives (FN). Specifically, due to the abundance of true negative (TN) pairs in NMLTC, it becomes challenging to identify the FN noise from the large pool of negative pairs.

To address these challenges, we propose a novel **i**nstance-**L**abel pa**i**r **Co**rrection (iLaCo) framework,

* Corresponding author.

aiming to achieve pair selection and correction for noisy multi-label text classification tasks. The framework introduces three key components: holistic selection metric, negative pair filter, and co-correction.

We first designed a holistic selection metric (HSM) to assess the learning difficulty of instance-label pairs. Our HSM consists of both a loss-based metric and a rank-based metric. The loss-based metric captures the global learning difficulty of instance-label pair by considering the associated loss values. A lower loss indicates a lower learning difficulty for the corresponding instance-label pair. On the other hand, the rank-based metric reflects the instance-specific learning difficulty of instance-label pair. In MLTC, predictions rely heavily on the label rank, with lower ranks indicating better memorization of the instance-label pair by the model (Xiao et al., 2021). By incorporating the rank-based metric, we can prevent some instance-label pairs with high loss values but correctly predicted pairs from being classified as corrupted pairs. To ensure stability and consistency, we adopt a time-consistent approach (historical average) (Zhou et al., 2020; Xia et al., 2022), to reflect the learning difficulty of each instance-label pair throughout the entire training process.

To address the challenge posed by the abundance of negative pairs, as depicted in Figure 1, we first observe a strong label correlation between FN labels (i.e., Pluto) and true positive (TP) labels (i.e., Astronomy and Neptune) for the same instance. As a result, we adopt a filter guided by instance-specific label correlation, focusing on negative pairs that exhibit a certain level of label correlation with the positive pairs of each instance. By doing so, we significantly reduce the number of candidate negative pairs for each instance, thereby reducing the difficulty of identifying FN pairs within the negative pairs.

In the final stage, we perform noise estimation and label correction simultaneously for both positive and negative pairs. After obtaining the HSM for positive and negative pairs, we utilize a Gaussian mixture model (GMM) to estimate the noise rates. We then employ a pseudo-labeling approach that combines both soft and hard correction strategies for label correction. It is important to note that in multi-class tasks, after identifying noisy labels, the common practice is to utilize sample selection or sample weighting methods (Li et al., 2023; Hu et al., 2023). However, in the case of instance-label pairs being a binary classification problem, we can directly flip the labels to correct them.

Our contributions can be summarized into four key aspects: (1) We propose the instance-label pair correction (iLaCo) approach, which successfully applies the memorization effect to multi-label text classification for noise reduction. (2) We introduce a holistic selection metric (HSM) that combines the global information of the training process (loss) with the instance-specific information (rank). HSM provides a better reflection of the difficulty in memorizing instance-label pairs. (3) We devise a label correlation-based negative pair filter, which enhances the recognition of false negative pairs by removing most irrelevant true negative pairs through label correlation. (4) The superior performance of iLaCo is validated through extensive experiments on three benchmark datasets.

2 Method

The overall architecture of our model is illustrated in Figure 2. We first employ the standard architecture for multi-label text classification to train the model on the noisy dataset. Subsequently, based on the historical information obtained during the training process, we utilize our proposed iLaCo method to correct the labels and generate pseudo labels. Finally, we retrain the multi-label text classification model using the corrected pseudo labels to obtain the final model. The iLaCo framework consists of three components: the holistic selection metric (HSM), negative pair filter, and co-correction. We present more details of these components in the following sections.

2.1 Noisy Multi-Label Text Classification

In what follows, sets are in calligraphic letters (e.g., $\mathcal{A}$), matrices are in capital bold letters(e.g., $\mathbf{A}$), vectors are in lower-case bold letters (e.g., $\mathbf{a}$), and scalars are in capital or lower-case letters (e.g. A, a). For simplicity, let $[L] = \{1, ..., L\}$. Considering a noisy multi-label text classification (NMLTC) problem, the input of training stage includes N instances $\mathcal{P} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^{N}$, each of which consists of an input vector $\mathbf{x}_i$ and several observed noisy labels $\tilde{\mathbf{y}}_i = (\tilde{Y}_{i,1}, \tilde{Y}_{i,2}, ..., \tilde{Y}_{i,L}) \in \{0, 1\}^L$ related to the input. Here L is the total number of candidate labels. The goal of NMLTC is to learn a function f that maps the input instance $\mathbf{x}_i$ and a label l to a relevance score $\hat{Y}_{i,j} = f(\mathbf{x}_i, j)$. In the testing

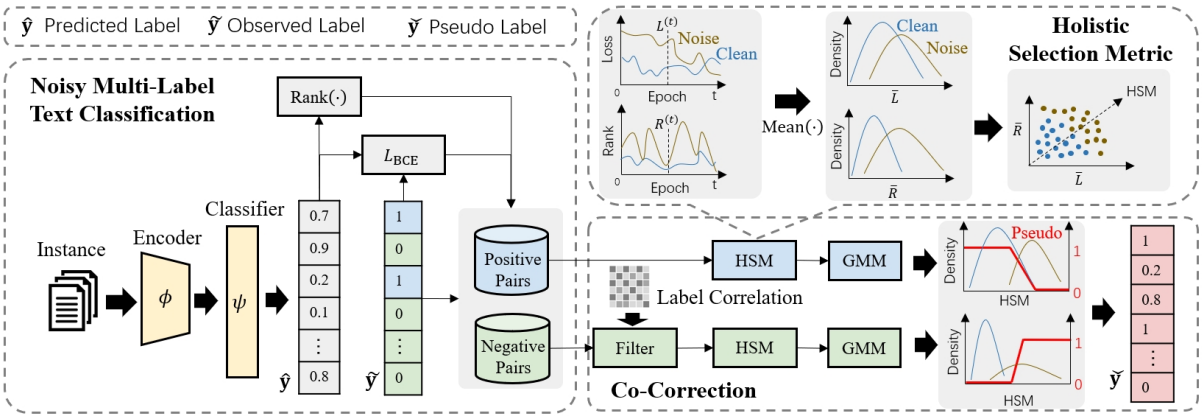


Figure 2: The overall framework of iLaCo.

stage, we aim to recommend the top-k labels with the highest relevance scores for a new instance.

We constructed the scoring function f by combining a text encoder ϕ and a multi-label classifier ψ. Following the approach of previous works (Ma et al., 2021; Xu et al., 2023b), we employed BiLSTM-based text encoders ϕ and adopted a multi-layer MLP as our multi-label classifier ψ. We then employed binary cross entropy (BCE) $L_{\text{BCE}} = \sum_{i=1}^{N}\sum_{l=1}^{L} L_{i,j}$ as the loss function, where

$$L_{i,j} = -(\tilde{Y}_{i,j}\log(\hat{Y}_{i,j}) + (1-\tilde{Y}_{i,j})\log(1-\hat{Y}_{i,j})). \quad (1)$$

The notation $L_{i,j}$ represents the loss value associated with the j-th label for the i-th instance. It can also be interpreted as the loss value corresponding to the instance-label pair with the index $\{i, j\}$.

2.2 Holistic Selection Metric Design

2.2.1 Beyond Small Loss Criteria

When learning with noisy labels (LNL), it is commonly observed that instances with clean labels typically have smaller loss values than those with noisy labels (Han et al., 2018; Northcutt et al., 2021). Such small-loss criteria have been widely adopted for selecting confident examples (Arazo et al., 2019; Wei et al., 2020; Xia et al., 2022). As illustrated in Figure 3 (a), this pattern is also evident in NMLTC. Hence, we can use the loss value $L_{i,j}$ in Equation 1 as a metric for identifying noise in instance-label pair.

However, relying solely on the loss value to reflect whether instance-label pair is well-memorized by the model is not comprehensive enough. This is because the loss and the optimization goal of MLTC are not entirely consistent (You et al., 2019). The final prediction in MLTC depends mainly on the ranking of the label $\hat{Y}_{i,j}$ within $\hat{\mathbf{y}}_i$. Therefore, even if the loss value $L_{i,j}$ is large, $\hat{Y}_{i,j}$ might still be a correct prediction. By introducing instance-specific label rank information, we can better distinguish between hard and noisy pairs For each instance $\mathbf{x}_i$ and its predicted label $\hat{\mathbf{y}}_i$, we can obtain the rank of each label using the rank function Rank$(\cdot)$:

$$\mathbf{r}_i = \text{Rank}(\hat{\mathbf{y}}_i), \quad (2)$$

where $\mathbf{r}_i = (R_{i,1}, R_{i,2}, ..., R_{i,L})$, and $R_{i,j}$ is the rank metric for $\hat{Y}_{i,j}$. As shown in Figure 3 (b), it is evident that rank-based metrics also possess noise identification capabilities. A smaller rank indicates that the label is more likely to be clean. Simultaneously, rank-based metrics that capture instance-specific information complement loss-based metrics that reflect global information, resulting in improved identification performance, as depicted in Figure 2 and Figure 3 (c).

2.2.2 Time Consistency

Let $L_{i,j}^{(t)}, i \in [N], j \in [L], t \in [T]$ denote the loss value corresponding to the j-th label of the i-th instance at the t-th epoch. $R_{i,j}^{(t)}$ represents the instance-specific rank value for the j-th label of the i-th instance at the t-th epoch. Calculating the selection metric directly based on the t-th epoch might lead to unstable results because both the loss values and ranking values exhibit large amplitudes during the optimization process(Zhou et al., 2020; Xia et al., 2022; Hu et al., 2022), as illustrated in Figure 2. Therefore, we mitigate the impact of large amplitudes by averaging over all epochs during the training process, yielding more stable

selection metrics:

$$\bar{L}_{i,j} = \sum_{t=1}^{T} L_{i,j}^{(t)}, \bar{R}_{i,j} = \sum_{t=1}^{T} R_{i,j}^{(t)}. \tag{3}$$

To facilitate the integration of these two metrics, we perform min-max normalization on them (Hu et al., 2022), obtaining normalized results $\hat{L}_{i,j}$ and $\hat{R}_{i,j}$ respectively. The linear combination of both metrics results in a new holistic selection metric (HSM):

$$M_{i,j} = \alpha \cdot \hat{L}_{i,j} + (1-\alpha)\hat{R}_{i,j}. \tag{4}$$

The combination coefficient α plays a crucial role in determining the balance between the two metrics. By combining the advantages of both metrics, HSM not only provides a more comprehensive understanding of the model's behavior but also exhibits consistency in its performance. This consistency contributes to HSM's superior ability to identify noisy labels effectively (see Figure 3 (d)).

2.3 Co-Correction

Due to the characteristics of multi-label learning, there are two types of noise associated with each instance-label pair, namely false positive and false negative. Next, we will employ HSM to perform pair correction for both positive and negative noise.

2.3.1 Positive Pair Correction

Positive instance-label pair refers to the instance-label pair associated with labels for which the observed label is positive. First, we obtain the HSM set $\mathcal{M}^+$ corresponding to all positive instance-label pairs, i.e.,

$$\mathcal{M}^+ = \{M_{i,j} | \tilde{Y}_{i,j} = 1\}, \tag{5}$$

As illustrated in Figure 3 (d), its distribution roughly resembles a bimodal Gaussian mixture. Therefore, we adopt a two-component Gaussian mixture model (GMM) to model the bi-modal distribution (Arazo et al., 2019; Li et al., 2020) of true positive (TP) and false positive (FP) pairs. After training, we could obtain the probability of a pair being corrupted through the posterior probability of HSM distributions. Accordingly, the noise rate σ^+ is estimated as:

$$\sigma^+ = \mathbb{E}_{M_{i,j} \in \mathcal{M}^+}[p(\mu^+ | M_{i,j})], \tag{6}$$

where μ^+ is the Gaussian component with a larger mean, since noisy pairs have typically larger HSM values.

After obtaining the noise rate, we can proceed with pair correction based on the noise rate and the quantiles of $\mathcal{M}^+$. It is worth noting that, in multi-class tasks, after identifying a corrupted label, the usual approach involves sample selection or sample reweighting (Li et al., 2023; Hu et al., 2023). However, for each instance-label pair, as it is a binary classification problem, if the probability of being a TP pair is very low, it is highly likely to be a FP pair. Therefore, we can implement label correction by applying label flipping to obtain pseudo-labels:

$$\check{Y}_{i,j} = \begin{cases} 0 & M_{i,j} > Q_1, M_{i,j} \in \mathcal{M}^+ \\ \tilde{Y}_{i,j} & M_{i,j} \leq Q_1, M_{i,j} \in \mathcal{M}^+ \end{cases}, \tag{7}$$

where $Q_1 = \text{Quantile}(\mathcal{M}^+, 1-\sigma^+)$ denotes the $1-\sigma^+$ quantile of the set $\mathcal{M}^+$. However, since noisy labels and clean labels are challenging to distinguish near the decision boundary (as shown in Figure 2 and Figure 3 (d)), compared to hard pseudo-labels, we have employed a soft-hard combined pseudo-label strategy for label correction. For high-confidence noisy pairs (strong discrimination by HSM), we use a hard pseudo-label for correction. However, for low-confidence noisy pairs, as they are prone to confusion with clean pairs, we adopt a soft pseudo-label for correction. The specific approach is as follows:

$$\check{Y}_{i,j} = \begin{cases} 0 & M_{i,j} > Q_2, M_{i,j} \in \mathcal{M}^+ \\ \frac{M_{i,j}-Q_1}{Q_1-Q_2} & Q_1 < M_{i,j} \leq Q_2, M_{i,j} \in \mathcal{M}^+ \\ \tilde{Y}_{i,j} & M_{i,j} \leq Q_1, M_{i,j} \in \mathcal{M}^+ \end{cases} \tag{8}$$

where, $Q_1 = \text{Quantile}(\mathcal{M}^+, 1-\sigma^+), Q_2 = \text{Quantile}(\mathcal{M}^+, 1-\frac{1}{2}\sigma^+)$. The corrected pseudo-label $\check{Y}_{i,j}$ under positive pairs is obtained. The piecewise function is shown in Figure 2.

2.3.2 Negative Pair Filter and Correction

We also need to collect the corresponding HSM for negative pairs in the observed labels, i.e.,

$$\mathcal{M}^- = \{M_{i,j} | \tilde{Y}_{i,j} = 0\}. \tag{9}$$

However, in practice, this step is intractable. Firstly, there is an excessive number of negative pairs in multi-label learning, significantly increasing the storage burden. More importantly, due to the abundance of TN pairs, the occurrence of FN pairs within them is relatively rare. In such cases, the TN pairs tend to excessively dominate the negative pairs, making it nearly impossible to identify the

FN pairs, as illustrated in Figure 4 (a). Fortunately, as depicted in Figure 1, it is observed that the FN labels are correlated with the positive labels of the current instance. Meanwhile, TN labels mostly lack this kind of label dependency.

Therefore, we adopt a method based on label correlation to filter the massive amount of negative pairs, retaining only those negative labels that have a certain level of label correlation with the current positive labels as our candidate set. i.e.,

$$\mathcal{M}^- = \{M_{i,j} | \tilde{Y}_{i,j} = 0, D_{i,j} > \beta\}, \tag{10}$$

$$D_{i,j} = \sum_{k=1}^{L} \tilde{Y}_{i,k} \cdot C_{k,j}, \tag{11}$$

where $D_{i,j}$ represents the correlation coefficient between the j-th label of the i-th sample and all positive labels of the i-th sample, and β is the threshold. Meanwhile, $C_{k,j}$ denotes the label correlation between the k-th and j-th labels. The label correlation $C_{k,j}$ can be obtained using various methods such as label semantic similarity (Zhang et al., 2021) or the label co-occurrence matrix (Su et al., 2022). In our approach, to better align semantics with the characteristics of the dataset, we opt for the latter method to compute the label correlation matrix $\mathbf{C}$. Each element $C_{k,j}$ of $\mathbf{C}$ is defined as:

$$C_{k,j} = \frac{c_{k,j}}{\sum_{b=1}^{L} c_{k,b}}, k, j \in [L]. \tag{12}$$

$$c_{k,j} = \begin{cases} 0, k = j \\ \sum_{i=1}^{N} \tilde{Y}_{i,k} \cdot \tilde{Y}_{i,j}, k \neq j \end{cases} \tag{13}$$

Therefore, we have obtained a set $\mathcal{M}^-$ composed of high-quality negative pairs. Subsequently, based on the HSM set $\mathcal{M}^-$ corresponding to negative pairs, we also estimate the noise rate σ^- using GMM (similar to Equation 6). After obtaining the noise rate σ^-, we also apply the pseudo-labeling strategy:

$$\check{Y}_{i,j} = \begin{cases} 1 & M_{i,j} > Q_4, M_{i,j} \in \mathcal{M}^- \\ \frac{M_{i,j} - Q_3}{Q_4 - Q_3} & Q_3 < M_{i,j} \leq Q_4, M_{i,j} \in \mathcal{M}^- \\ \tilde{Y}_{i,j} & M_{i,j} \leq Q_3, M_{i,j} \in \mathcal{M}^- \end{cases} \tag{14}$$

where, $Q_3 = \text{Quantile}(\mathcal{M}^-, 1 - \sigma^-), Q_4 = \text{Quantile}(\mathcal{M}^-, 1 - \frac{1}{2}\sigma^-)$. The piecewise function is also can be found in Figure 2.

3 Experiment

3.1 Experimental Setup

Datasets We verify the effectiveness of the proposed method on three synthetic noisy MLTC datasets, i.e. AAPD (Yang et al., 2018), RCV1 (Lewis et al., 2004) and EUR-Lex (Mencía and Fürnkranz, 2008). These datasets are well-known benchmark datasets in the MLTC (Xu et al., 2023b; Ma et al., 2021; Xiao et al., 2019) and NMLTC (Chen et al., 2023) fields. Table 1 contains the statistics of these three benchmark datasets.

Noisy-Label Generation Following previous works (Li et al., 2022; Chen et al., 2023), we randomly flip an element $Y_{i,j}$ in the label vector $\mathbf{y}_i$ from 0 to 1 or 1 to 0 by the probability ρ_- and ρ_+ respectively. In some works (Chen et al., 2023; Ghiassi et al., 2022), it was assumed that $\rho_- = \rho_+$. However, we argue against this approach because in MLTC, the label dimension L is usually much larger than the average number of labels per instance L_{avg}. Therefore, if $\rho_- = \rho_+$, the number of FP labels would be much greater than the number of FN labels. This situation does not accurately reflect the challenges of NMLTC problems. Hence, we adopt the approach proposed in Multi-T (Li et al., 2022), setting $\rho_+ = \rho$ and $\rho_- = \frac{L_{\text{avg}}}{L - L_{\text{avg}}}\rho$. This configuration is designed to ensure that the difference between the number of FP labels and FN labels is relatively small. The noise rate ρ is set to 0.2, 0.4, and 0.6.

Evaluation Metrics For a comprehensive and reliable evaluation, we follow conventional settings and report the following metrics: precision at 5 (P@5) and normalized discounted cumulative gain at 5 (N@5). These metrics have been widely used in literature to evaluate MLTC (Ma et al., 2021; Xiao et al., 2021). Note that only the training set is affected by noise, whereas the evaluation metrics are computed on the clean testing set. The best results are in bold, and the second-best results are in underscore.

Baselines To verify the effectiveness of iLaCo, we selected the nine most representative baseline models in three groups. (1) MLTC Methods: AttentionXML (You et al., 2019), HTTN (Xiao et al., 2021) and LSFA (Xu et al., 2023b). (2) Noisy multi-label learning (NMLL) methods: GCE (Zhang and Sabuncu, 2018), WSIC (Hu et al., 2019), Reweight-T (Patrini et al., 2017), Multi-T (Li et al., 2022),

Datasets	N_{trn}	N_{tst}	D_{vocab}	L	L_{avg}	N_{avg}	W_{trn}	W_{tst}
AAPD	54,840	1,000	69,399	54	2.41	2444.04	163.42	171.65
RCV1	23,149	781,265	47,236	103	3.18	729.67	259.47	269.23
EUR-Lex	11,585	3,865	171,120	3,956	5.32	15.59	1225.20	1248.07

Table 1: Data statistics. N_{trn}, N_{tst} refer to the number of documents in the training and test sets, respectively. D_{vocab} is the vocabulary size of documents. L is the number of labels. L_{avg} is the average number of labels per documents. N_{avg} is the average number of documents per label. W_{trn}, W_{tst} refer to the average number of words per document in the training and test sets, respectively.

Noise Rate	$\rho = 0.2$		$\rho = 0.4$		$\rho = 0.6$	
Methods	P@5	N@5	P@5	N@5	P@5	N@5
AttentionXML (You et al., 2019)	46.28	52.45	42.41	51.33	39.50	48.46
HTTN (Xiao et al., 2021)	45.60	52.04	42.00	50.95	37.93	46.49
LSFA (Xu et al., 2023b)	47.91	54.69	44.25	52.62	40.29	48.77
GCE (Zhang and Sabuncu, 2018)	49.32	55.82	46.08	53.94	41.90	50.49
WSIC (Hu et al., 2019)	47.32	54.57	45.32	53.84	41.32	49.75
Reweight-T (Patrini et al., 2017)	49.88	55.98	45.59	53.91	40.84	49.32
MLLSC (Ghiassi et al., 2022)	48.44	55.08	45.17	54.12	41.57	50.22
Multi-T (Li et al., 2022)	49.23	55.99	46.06	54.77	42.33	50.80
nEM (Chen et al., 2023)	49.89	56.77	46.73	54.37	41.80	50.18
iLaCo	**50.74**	**57.59**	**47.92**	**55.96**	**43.33**	**51.71**

Table 2: Performance on AAPD with different noise ratios.

and MLLSC (Ghiassi et al., 2022). (3) NMLTC method: nEM (Chen et al., 2023). More details about the implementation setting can be found in Appendix A.3.

3.2 Main Results

As depicted in Tables 2-4, we have observed the following phenomena: (1) In most cases, existing MLTC methods tend to perform worse compared to NMLL methods. This is primarily due to the lack of ability to distinguish noisy labels exhibited by these methods. Moreover, these methods often overly prioritize learning head-to-tail knowledge transfer, resulting in overfitting to the noisy labels and subsequently reducing the overall generalization ability of the model. (2) The advantages of NMLL methods are not significant. Methods based on noise transition matrix estimation, such as Reweight-T and Multi-T, are mainly limited by the large number of labels in the MLTC scenario, making it more challenging to model noise transition in high-dimensional spaces. The nEM method, based on probabilistic graphical models, lacks explicit differentiation between positive and negative noise. The MLLSC method explicitly models both positive and negative noise. However, it solely relies on instantaneous DNN output probabilities as a metric, disregarding the potential instability during the training process of MLTC models. (3) In all cases, our method shows significant improvements compared to other methods. Particularly, as the label dimension L of the dataset increases, our method exhibits even greater enhancement. This is mainly due to the fact that as the label dimension increases, the influence of negative pairs becomes more pronounced. However, our method effectively addresses this issue by employing a negative pair filtering approach. Moreover, as the noise ratio increases, our method demonstrates a lower decrease in accuracy compared to other methods. This validates the effectiveness of our approach in accurately recognizing noise within NMLTC scenarios.

3.3 Ablation Study

In the following experiments, we aim to analyze the effectiveness of each component of the proposed iLaCo method on three datasets. To construct the synthetic noise datasets, we use a noise ratio of 0.4. We compare the complete iLaCo method with the

Noise Rate	$\rho = 0.2$		$\rho = 0.4$		$\rho = 0.6$	
Methods	P@5	N@5	P@5	N@5	P@5	N@5
AttentionXML (You et al., 2019)	52.11	85.53	46.30	82.00	43.16	78.28
HTTN (Xiao et al., 2021)	45.76	81.73	44.24	81.03	42.13	78.37
LSFA (Xu et al., 2023b)	50.12	84.51	45.31	81.94	42.89	79.21
GCE (Zhang and Sabuncu, 2018)	48.45	82.24	42.93	81.60	42.38	78.26
WSIC (Hu et al., 2019)	52.01	86.96	45.95	81.81	43.77	78.01
Reweight-T (Patrini et al., 2017)	52.43	86.48	47.32	80.87	41.13	77.11
MLLSC (Ghiassi et al., 2022)	51.95	86.96	47.59	81.67	42.91	77.95
Multi-T (Li et al., 2022)	52.51	87.42	47.07	82.41	43.04	78.94
nEM (Chen et al., 2023)	52.35	87.08	47.74	81.57	41.77	78.36
iLaCo	**54.34**	**88.83**	**49.40**	**85.46**	**45.60**	**82.25**

Table 3: Performance on RCV1 with different noise ratios.

Noise Rate	$\rho = 0.2$		$\rho = 0.4$		$\rho = 0.6$	
Methods	P@5	N@5	P@5	N@5	P@5	N@5
AttentionXML (You et al., 2019)	50.05	57.01	44.19	51.44	38.93	47.82
HTTN (Xiao et al., 2021)	39.98	46.78	36.00	50.88	32.82	40.98
LSFA (Xu et al., 2023b)	48.96	55.53	43.35	51.50	36.67	45.25
GCE (Zhang and Sabuncu, 2018)	49.14	56.17	43.81	51.83	37.97	46.57
WSIC (Hu et al., 2019)	51.96	59.42	46.47	53.77	41.43	49.86
Reweight-T (Patrini et al., 2017)	52.30	59.49	46.25	54.43	40.72	48.57
MLLSC (Ghiassi et al., 2022)	53.27	60.96	47.40	55.23	41.80	50.16
Multi-T (Li et al., 2022)	52.96	60.49	47.13	54.70	41.23	50.09
nEM (Chen et al., 2023)	52.22	58.48	45.37	51.85	37.65	46.63
iLaCo	**53.98**	**61.44**	**48.9**	**56.36**	**44.29**	**53.56**

Table 4: Performance on EUR-Lex with different noise ratios.

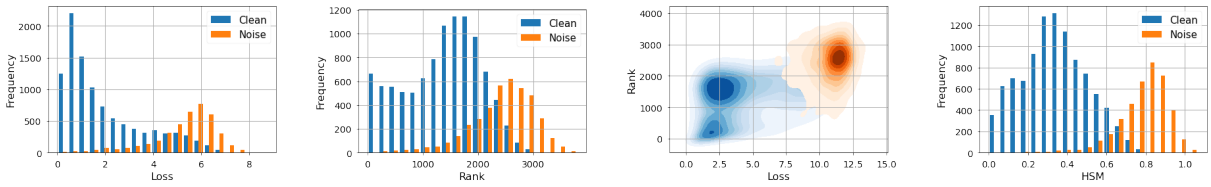
following variants: (a) HSM (loss): This variant utilizes the instantaneous loss value as the selection metric. (b) HSM (rank): This variant employs the instantaneous rank value as the selection metric. (c) HSM (cons.): This variant transforms the instantaneous selection metric into a time consistency metric. (d) Filter: This variant applies a Filter based on label correlations to filter out a large number of negative pairs. Through these comparisons, we aim to assess the impact and effectiveness of each component in improving the performance of iLaCo on the given datasets.

Component analysis According to Table 5, we observe the following: (1) The Filter component significantly improves the model's performance. This improvement is attributed to the effective removal of a substantial number of irrelevant labels from the negative pair set through the Filter component. As a result, the identification of noisy labels within the negative pair set becomes more manageable for us. (2) The different components of the HSM metric collectively contribute to enhancing the quality of noise identification. By incorporating instance-specific rank information, the model gains the ability to differentiate between different instances, enabling a more accurate distinction between clean and corrupted labels. Additionally, the aggregation of information from multiple epochs allows the model to obtain a more consistent selection metric, further enhancing its capability to identify noise.

Effectiveness of HSM. In Figure 3, we present the distributions of positive pairs using HSM(loss), HSM(rank) and HSM. Firstly, as shown in (a) and (b), both the loss and rank metrics demonstrate certain capabilities in identifying noise. From (c),

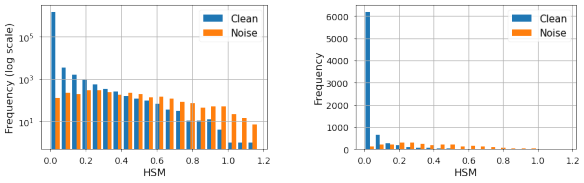
				AAPD		RCV1		EUR-Lex	
HSM (loss)	HSM (rank)	HSM (cons.)	Filter	P@5	N@5	P@5	N@5	P@5	N@5
✓				46.82	54.79	47.46	83.72	45.41	52.7
✓			✓	46.89	55.23	47.74	84.13	47.2	54.85
✓	✓		✓	47.86	55.76	48.7	84.98	47.55	55.38
✓	✓	✓	✓	47.92	55.96	49.4	85.46	48.9	56.36

Table 5: Components ablation study on 40% noise.



(a) HSM(loss) (b) HSM(rank) (c) HSM(loss)&HSM(rank) (d) HSM

Figure 3: The visualization of metric distribution on EUR-Lex with 40% noise.



(a) Before Filter (b) After Filter

Figure 4: Comparison of the HSM distributions under negative pairs before and after Filter on EUR-Lex with 40% noise. Note that the presence of noise is only visible in (a) when using a logarithmic scale.

we can observe the complementary nature of the loss and rank metrics, which motivated us to combine them into a more comprehensive metric for improved accuracy in noise identification. Finally, in (d), it is evident that the combination of both metrics in HSM leads to a significantly enhanced noise identification capability.

Effectiveness of Filter. Figure 4 showcases the histograms of HSM for TN labels and FN labels, both before and after applying our proposed Filter. It can be observed that with the implementation of the Filter, FN labels can be effectively separated. However, without the use of the Filter, it is almost impossible to distinguish between these two types of labels. This demonstrates the effectiveness of our proposed Filter in improving the separation and identification of clean and noisy pairs under negative pairs.

4 Related Work

Multi-Label Text Classification The most common approach for addressing multi-label text classification (MLTC) is to use the identical document representation to train classification models (Liu et al., 2017). Consequently, label-specific feature learning (You et al., 2019; Xiao et al., 2019; Ma et al., 2021), which focuses on capturing the unique characteristics of each label, has shown promise in enhancing label discrimination. Some works (Xiao et al., 2021; Xu et al., 2023b) have also explored transfer learning from head labels to tail labels to mitigate the adverse effects of label long-tail distribution. Recently, there has been growing interest in MLTC under noisy settings (Chen et al., 2019, 2023). The nEM method (Chen et al., 2023) models the transition process of noisy labels using latent variable models to achieve robust MLTC. In this paper, we extend the memorization effect (Arpit et al., 2017) to noisy MLTC for the first time, and propose an instance-label pair correction method.

Learning from Noisy Labels In order to mitigate the influence of data noise, sample selection is an effective approach. The small-loss criterion (Arpit et al., 2017) is the most widely used criterion. MentorNet (Jiang et al., 2018) and MILD (Hu et al., 2023) propose new metrics based on information throughout the training process to distinguish between clean and corrupted data. Approaches such as GCE (Zhang and Sabuncu, 2018), WISC (Hu et al., 2019), and MLLSC (Ghiassi et al.,

2022) tackle multi-label noise learning by introducing robust loss functions or regularization methods. Reweight-T (Patrini et al., 2017) corrects models by estimating the noise transition matrix, while Multi-T (Li et al., 2022) leverages label correlations in multi-label learning to identify label noise, leading to better estimation of the noise transition matrix. Motivated by MILD, we propose a holistic selection metric for noisy MLTC that integrates global training information with instance-specific training information.

5 Conclusions

In this paper, we propose a method for instance-label pair correction that combines the historical loss information and rank information from the training process to identify and correct positive and negative noise in noisy multi-label text classification tasks. Our experiments yield compelling results, highlighting the superiority of our model compared to existing state-of-the-art (SOTA) baselines for multi-label text classification and noisy multi-label classification.

6 Limitations

There are still some limitations to our work. 1) This work utilizes the memorization effect (Arpit et al., 2017) in deep learning for sample selection and correction, which has not been observed in other traditional machine learning methods. Therefore, the proposed method is not applicable to such learning methods. 2) Since our method models the training process at the instance-label pair level, it possesses the ability to recognize instance-dependent noise (Chen et al., 2021). However, our work has not been validated on instance-dependent noise yet, which could be an area for future exploration.

7 Acknowledgement

This work was partly supported by the Fundamental Research Funds for the Central Universities (2019JBZ110); the National Natural Science Foundation of China under Grant 62176020; the National Key Research and Development Program (2020AAA0106800); the Beijing Natural Science Foundation under Grant L211016; and Chinese Academy of Sciences (OEIP-O-202004).

References

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 312–321. PMLR.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR.

Junfan Chen, Richong Zhang, Yongyi Mao, Hongyu Guo, and Jie Xu. 2019. Uncover the ground-truth relations in distant supervision: A neural expectation-maximization framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 326–336. Association for Computational Linguistics.

Junfan Chen, Richong Zhang, Jie Xu, Chunming Hu, and Yongyi Mao. 2023. A neural expectation-maximization framework for noisy multi-label text classification. *IEEE Trans. Knowl. Data Eng.*, 35(11):10992–11003.

Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. 2021. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 11442–11450. AAAI Press.

Benoît Frénay and Michel Verleysen. 2014. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, 25(5):845–869.

Amirmasoud Ghiassi, Robert Birke, and Lydia Y. Chen. 2022. Multi label loss correction against missing and corrupted labels. In *Asian Conference on Machine Learning, ACML 2022, 12-14 December 2022, Hyderabad, India*, volume 189 of *Proceedings of Machine Learning Research*, pages 359–374. PMLR.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In

Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 8536–8546.

Chuanyang Hu, Shipeng Yan, Zhitong Gao, and Xuming He. 2023. MILD: modeling the instance learning dynamics for learning with noisy labels. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 828–836. ijcai.org.

Hengtong Hu, Lingxi Xie, Xinyue Huo, Richang Hong, and Qi Tian. 2022. Vibration-based uncertainty estimation for learning from limited supervision. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXX*, volume 13690 of *Lecture Notes in Computer Science*, pages 160–176. Springer.

Mengying Hu, Hu Han, Shiguang Shan, and Xilin Chen. 2019. Weakly supervised image classification through noise regularization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11517–11525. Computer Vision Foundation / IEEE.

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 *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2309–2318. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397.

Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. 2022. Estimating noise transition matrix with label correlations for noisy multi-label learning. In *NeurIPS*.

Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. 2023. DISC: learning from noisy labels via dynamic instance-specific selection and correction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 24070–24079. IEEE.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 115–124. ACM.

Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. 2021. Label-specific dual graph neural network for multi-label text classification. 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 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3855–3864. Association for Computational Linguistics.

Eneldo Loza Mencía and Johannes Fürnkranz. 2008. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML/PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II*, volume 5212 of *Lecture Notes in Computer Science*, pages 50–65. Springer.

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

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2233–2241. IEEE Computer Society.

Youri Peskine, Damir Korencic, Ivan Grubisic, Paolo Papotti, Raphaël Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4054–4063. Association for Computational Linguistics.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. 2023. Zero-shot multi-label topic inference with sentence encoders and llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16218–16233. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical*

Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 254–263. ACL.

Xi'ao Su, Ran Wang, and Xinyu Dai. 2022. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 672–679. Association for Computational Linguistics.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13723–13732. Computer Vision Foundation / IEEE.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. 2022. Learning with noisy labels revisited: A study using real-world human annotations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. 2022. Sample selection with uncertainty of losses for learning with noisy labels. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 466–475. Association for Computational Linguistics.

Lin Xiao, Pengyu Xu, Liping Jing, and Xiangliang Zhang. 2022. Pairwise instance relation augmentation for long-tailed multi-label text classification. *CoRR*, abs/2211.10685.

Lin Xiao, Pengyu Xu, Mingyang Song, Huafeng Liu, Liping Jing, and Xiangliang Zhang. 2023. Triple alliance prototype orthotist network for long-tailed multi-label text classification. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2616–2628.

Lin Xiao, Xiangliang Zhang, Liping Jing, Chi Huang, and Mingyang Song. 2021. Does head label help for long-tailed multi-label text classification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14103–14111. AAAI Press.

Pengyu Xu, Mingxuan Xia, Lin Xiao, Huafeng Liu, Bing Liu, Liping Jing, and Jian Yu. 2023a. Textual tag recommendation with multi-tag topical attention. *Neurocomputing*, 537:73–84.

Pengyu Xu, Lin Xiao, Bing Liu, Sijin Lu, Liping Jing, and Jian Yu. 2023b. Label-specific feature augmentation for long-tailed multi-label text classification. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 10602–10610. AAAI Press.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3915–3926. Association for Computational Linguistics.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5812–5822.

Ximing Zhang, Qian-Wen Zhang, Zhao Yan, Ruifang Liu, and Yunbo Cao. 2021. Enhancing label correlation feedback in multi-label text classification via multi-task learning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1190–1200. Association for Computational Linguistics.

Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8792–8802.

Tianyi Zhou, Shengjie Wang, and Jeff A. Bilmes. 2020. Curriculum learning by dynamic instance hardness. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

A Appendix

A.1 Datasets

We evaluate the proposed model on three benchmark datasets for MLTC, which are AAPD, RCV1 and EUR-Lex.

- AAPD (Yang et al., 2018) collects the abstract and the corresponding subjects of 55840 publications in the field of computer science from the arXiv.
- Reuters Corpus Volume I (RCV1) (Lewis et al., 2004) comprises more than 80K news items that have been manually sorted into 103 classes.
- EUR-Lex (Mencía and Fürnkranz, 2008) is a collection of documents about European Union law belonging to 3956 subjects. The public version contains 11585 training instances and 3865 testing instances.

A.2 Evaluation metrics

Following previous works (You et al., 2019; Xiao et al., 2019; Xu et al., 2023b), we use two main metrics which are commonly used in MLTC evaluations: the precision at k ($P@k$) and normalized discounted cumulative gain at k ($N@k$).

$P@k$ The precision of the top-k labels is defined as:

$$P@k = \frac{1}{k}\sum_{l=1}^{k} \boldsymbol{y}_{rank(l)} \tag{15}$$

where $\boldsymbol{y} \in \{0,1\}^L$ is the ground truth label vector, and $rank(l)$ is the index of the l-th highest predicted label.

$N@k$ $N@k$ is an evaluation metric that takes into account the return order. The value ranges from 0 to 1, and the higher the better. $N@k$ is defined as follows:

$$DCG@k = \sum_{l=1}^{k} \frac{\boldsymbol{y}_{rank(l)}}{log(l+1)} \tag{16}$$

$$N@k = \frac{DCG@k}{\sum_{l=1}^{min(k,||\boldsymbol{y}||_o)} \frac{1}{log(l+1)}} \tag{17}$$

where $||\boldsymbol{y}||_o$ counts the number of relevant labels in the ground truth label vector $\boldsymbol{y}$. Note that $N@k$ is a metric for ranking, meaning that the order of top-k prediction is considered in $N@k$ but not in $P@k$.

Most MLTC works (You et al., 2019; Ma et al., 2021; Xiao et al., 2022, 2023; Xu et al., 2023a) do not use Average Precision (AP), Recall, F1-Score as evaluation metrics. This is because AP, Recall, F1-Score metrics are more suitable for cases with a small label space, where it is easier to directly predict the target labels. MLTC usually involves scenarios with a large label space, making it difficult to directly predict the target labels. In such cases, most methods primarily focus on providing a predicted ranking of labels, selecting the top-k labels for prediction, rather than predicting the number of labels. Therefore, most methods cannot compute (or are not suitable for) AP, Recall, F1-Score evaluation metrics.

A.3 Implementation Details

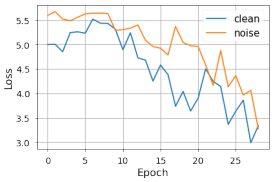
For all three datasets, we used the most frequent words in the training set as a limited-size vocabulary (below 500,000). We truncated each text after 500 words for efficiency. All experiments are carried out in a Linux environment with a single Tesla V100 GPU (32G). To ensure a fair comparison, we employ the same backbone as iLaCo for all the noisy multi-label learning methods. Our model was trained by Adam (Kingma and Ba, 2015) with the learning rate of 1e-3. We also used stochastic weight averaging (You et al., 2019) with a constant learning rate to enhance the performance. As for the key hyper-parameters of our proposed method: coefficient α and threshold β, we set $\alpha = 0.7, \beta = 0.05$ for AAPD. For RCV1 and EUR-Lex, we set $\alpha = 0.7, \beta = 0.05$ and $\alpha = 0.7, \beta = 0.001$ respectively. All experiments are run at least three times with different random seeds, and we report the average values of results.

A.4 Variations of Training Process

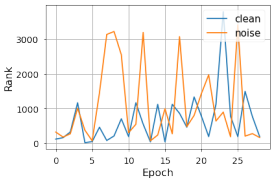
As shown in the Figure 5, this illustrates the variations in two samples based on a loss-based metric and a rank-based metric during the training process. It can be observed that if we choose instantaneous selection metrics, it is not conducive to our selection of clean samples. This also motivates our choice of time-consistent selection metrics.

A.5 Computation Cost

Our algorithm can be divided into three steps. Firstly, we need to train a preliminary MLTC model to obtain the loss (and rank) information for the instance-label pairs. Then, we use the information to correct the original observed labels. Finally,



(a) Loss-based metric



(b) Rank-based metric

Figure 5: Variations of the selection metrics for noisy pairs and clean pairs during the training process.

we retrain the model using the corrected labels. The computational cost of the algorithm mainly depends on the model training and retraining processes in the first and third steps. Regarding space complexity, our method follows existing neural network architectures and does not increase the model size additionally. As shown in the Table 6, we compare the training time and model size with typical baseline methods on EUR-Lex (40% noise). Our experimental results indicate that the training time and model size of our proposed method are in a comparable range with other approaches.

Method	**Performance**	
	Training Time (hours)	**Model Size (GB)**
AttentionXML	0.55	0.27
LSFA	1.59	0.44
MLLSC	0.51	0.29
iLaCo	0.94	0.29

Table 6: Comparison of different methods

A.6 Discussion of LLMs

The application of Large Language Models (LLMs) to mitigate multi-label noise is a promising subject. However, existing LLMs in the multi-label domain are currently mainly focused on zero-shot/few-shot scenarios(Peskine et al., 2023; Sarkar et al., 2023). This is because the performance of LLMs on domain-specific multi-label data is challenging to compare with models trained on domain data.