Enhancing Multi-Label Text Classification under Label-Dependent Noise: A Label-Specific Denoising Framework

Pengyu Xu, Liping Jing*, Jian Yu
Beijing Key Lab of Traffic Data Analysis and Mining
Beijing Jiaotong University, Beijing, China
pengyu@bjtu.edu.cn

Abstract

Recent advancements in noisy multi-label text classification have primarily relied on the class-conditional noise (CCN) assumption, which treats each label independently undergoing label flipping to generate noisy labels. However, in real-world scenarios, noisy labels often exhibit dependencies with true labels. In this study, we validate through hypothesis testing that real-world datasets are unlikely to adhere to the CCN assumption, indicating that label noise is dependent on the current labels. To address this, we introduce a label-specific denoising framework designed to counteract label-dependent noise. The framework initially presents a holistic selection metric that evaluates noisy labels by concurrently considering loss information, ranking information, and feature centroid. Subsequently, it identifies and corrects noisy labels individually for each label category in a fine-grained manner. Extensive experiments on benchmark datasets demonstrate the effectiveness of our method under both synthetic and real-world noise conditions, significantly improving performance over existing state-of-the-art models.

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). In noisy multi-label classification, most existing methods rely on the class-conditional noise (CCN) assumption (Li et al., 2022b; Xia et al., 2023; Xie and Huang, 2023; Song et al., 2024). This assumption posits that label noise originates from independent label flipping for each category, with each category having a distinct flipping probability.

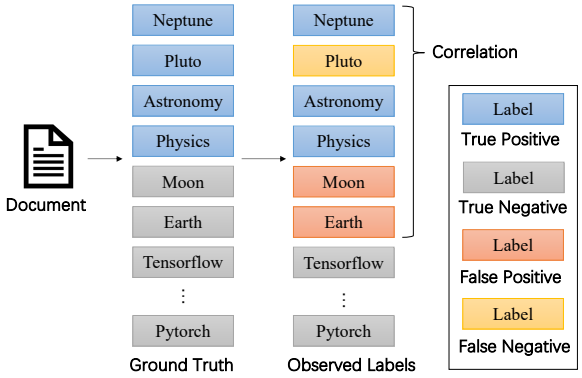


Figure 1: An illustration of noisy multi-label text classification.

However, in practice, noisy labels often exhibit a certain degree of correlation with the true labels (Cui et al., 2020; Xie and Huang, 2022). As shown in Figure 1, due to category ambiguity, lack of expert knowledge, or the influence of attention shift (Wu et al., 2023), annotators are more likely to mislabel the current document as "Moon" or "Earth" instead of unrelated labels such as "Pytorch" or "Tensorflow".

In this paper, our first contribution (Section 2) is to propose a theoretical hypothesis test on the real-world dataset Riedel (Chen et al., 2023) to demonstrate that real-world multi-label noise is less likely to be CCN, and more likely to be label-dependent noise (LDN), where the occurrence of label noise depends on the positive labels associated with the current sample. To mitigate noisy multi-label text classification under LDN, our second contribution (Section 2) is to introduce for generating controllable LDN and analyze the characteristics of training under LDN. Our third contribution (Section 3) is the proposal of a **L**abel-Sp**e**cific **D**enoising (LeD) framework to address LDN. LeD initially introduces a holistic selection metric (HSM), incor-

* Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 5674–5688
November 12-16, 2024 ©2024 Association for Computational Linguistics

porating ranking-enhanced loss (REL) and centroid distance (CD) to assess labels from both output and feature perspectives. Following this, LeD employs the HSM to conduct label-specific noise identification and correction for each label category. The superior performance of LeD is verified in extensive experiments, under LDN with varying noise fractions, including on real-world benchmarks.

2 From CCN to LDN

In this section, we introduce the problem setting of noisy multi-label classification from traditional class-conditional noise (CCN) assumption to our proposed label-dependent noise (LDN) assumption. 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 $[L] = \{1, ..., L\}$. Additionally, proofs of theorems can be found in Appendix A.1.

2.1 Preliminaries

Considering a multi-label classification problem, the input of training stage includes N instances $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, each of which consists of an input vector $\mathbf{x}_i$ and output labels $\mathbf{y}_i = (Y_{i,1}, Y_{i,2}, ..., Y_{i,L}) \in \{0,1\}^L$ related to the input. Here L is the total number of candidate labels. In real-world scenarios, it is often not possible to directly observe the true labels $\mathbf{y}$. Instead, we have an observable distribution of noisy labels $\tilde{\mathbf{y}}$ and a noisy training set $\tilde{\mathcal{D}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^N$. In noisy multi-label classification, our goal is to predict proper labels for each unseen instance by only using the noisy training set.

2.2 Class-Conditional Noise Assumption

The class-conditional noise assumption is commonly used in previous works (Chen et al., 2019; Li et al., 2022b; Xie and Huang, 2023; Chen et al., 2023) on noisy multi-label classification.

Definition 1. *(Noise transition matrix) In multi-label classification, the random variables $\tilde{Y}_{\cdot,j}$ and $Y_{\cdot,j}$ for the label j are related through a noise transition matrix $\mathbf{T}^j \in [0,1]^{2*2}, j \in [L]$. Generally, the transition matrix depends on instances (feature $\mathbf{x}$ and labels $\mathbf{y}$), i.e., $\mathbf{T}^j_{k,l}(\mathbf{x}, \mathbf{y}) = P(\tilde{Y}_j = k | Y_j = l, \mathbf{x}, \mathbf{y})$, where k and $l \in \{0,1\}$.*

Definition 2. *(Class-conditional noise) Under the class-conditional assumption, the transition matrix is assumed class-conditional and instance-independent, i.e., $\mathbf{T}^j_{k,l}(\mathbf{x}, \mathbf{y}) = \mathbf{T}^j_{k,l} = P(\tilde{Y}_j = k | Y_j = l)$.*

As illustrated in Figure 1, in the context of noisy multi-label learning, there exist two types of noise: false positives and false negatives. Among them, false positives often exhibit a strong label correlation with the ground truth $\mathbf{y}$. Therefore, we argue that real-world multi-label noise should at least be *label-dependent*, i.e., $\mathbf{T}^j_{k,l}(\mathbf{y}) = P(\tilde{Y}_j = k | Y_j = l, \mathbf{y})$. To underscore the importance of moving beyond the CCN assumption, which pertains to label-independent noise (LIN), we theoretically validate its significance through hypothesis testing.

Definition 3. *(Label flip) Given the noisy dataset $\tilde{\mathcal{D}}$, consider randomly sampling a validation set $\tilde{\mathcal{D}}' = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^n$, and assume we also know the clean labels $\{\mathbf{y}_i\}_{i=1}^n$ corresponding to the validation set. We define the random variable $Z_j = \{\tilde{Y}_j | Y_j = 0\}$, where Z_j represents the event that the j^{th} class of the current sample flip from a negative label to a positive label. Here, $Z_j \in \{0,1\}$, with a flip probability of $\mathbf{T}^j_{1,0}(\mathbf{x}, \mathbf{y})$.*

Theorem 1. *If CCN assumption holds, then for $\forall j_0, j_1 \in [L]$, random variables Z_{j_0} and Z_{j_1} are independent.*

Next, we verify the independence of Z_{j_0} and Z_{j_1} through hypothesis testing. The null hypothesis H_0 and the corresponding alternative hypothesis H_1 are defined as follows:

$$\mathbf{H_0}: Z_{j_0}, Z_{j_1} \text{ are independent;}$$
$$\mathbf{H_1}: Z_{j_0}, Z_{j_1} \text{ are dependent.}$$

By applying the chi-square test to the real-world noisy multi-label benchmark Riedel (Chen et al., 2019, 2023), the hypothesis testing results show that $\chi^2 = 940.5$ with a p-value of $1.5e^{-206}$, indicating that the result is highly statistically significant. Thus, the null hypothesis $\mathbf{H_0}$ is rejected with the significance value. Hypothesis "$\mathbf{H_1}$: Z_{j_0}, Z_{j_1} are dependent." is accepted. Therefore, based on Theorem 1, we can conclude that the CCN assumption does not hold on Riedel. Please refer to Appendix A.2 for more details.

2.3 Label-Dependent Noise

Now both the intuition and theoretical evidence imply that multi-label noise should be dependent on labels. As presented in Definition 4, we can model label-dependent mislabelling among given labels.

Definition 4. *(Label-dependent noise) Under the*

label-dependent assumption, the transition matrix is $\mathbf{T}^j_{k,l}(\boldsymbol{x},\boldsymbol{y}) = \mathbf{T}^j_{k,l}(\boldsymbol{y}) = P(\tilde{Y}_j = k|Y_j = l,\boldsymbol{y})$

CCN can be seen as a degenerated case of LDN assumption such that all instances have the same noise transition matrix. By assuming LDN, we can better model the label correlation of real-world multi-label noise, as depicted in Figure 1. Note that LDN is also a special case of instance-dependent noise (IDN) (Chen et al., 2021). Its noise transition depends on the labels corresponding to each instance.

Previous works on noisy multi-label classification under CCN assumption often conducted experiments on synthetic noise with varying noise fractions (Li et al., 2022b; Chen et al., 2023). They 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 $\mathbf{T}^j_{1,0}$and $\mathbf{T}^j_{0,1}$ respectively, thereby generating noise labels of false positives and false negatives. Similarly, it is desired to easily generate LDN with any noise fraction for any given benchmark dataset. To stimulate the development of theory and methodology, we propose a novel LDN generator.

As shown in Figure 1, the LDN assumption primarily manifests in the generation of false positive noise. for the generation of false negative noise labels, we adopt the method used in previous studies (Li et al., 2022b), which involves a fixed transition probability $\mathbf{T}^j_{0,1} = \mathbf{T}_{0,1} = \rho_+$. For the generation of false positive noise, we follow a label dependency approach, meaning that true negative labels with strong label correlation to the ground truth are more likely to be flipped to false positive labels (Liang et al., 2023). We simulate the label flips based on a label correlation matrix $\mathbf{C}$, which can be obtained through the label co-occurrence matrix (Su et al., 2022). Each element C_{j_0,j_1} of $\mathbf{C}$ is defined as:

$$C_{j_0,j_1} = \frac{c_{j_0,j_1}}{\sum_{j=1}^{L} c_{j_0,j}}, j_0, j_1 \in [L] \tag{1}$$

$$c_{j_0,j_1} = \begin{cases} 0 & j_0 = j_1 \\ \sum_{i=1}^{N} Y_{i,j_0} \cdot Y_{i,j_1} & j_0 \neq j_1 \end{cases} \tag{2}$$

The probability $\mathbf{T}^j_{0,1}(\mathbf{y})$ of a true negative label j transitioning to a false positive label should be related to the current set of positive labels for the sample.

$$\mathbf{T}^j_{1,0}(\mathbf{y}) = \rho_- * p^j(\mathbf{y}, \mathbf{C}) \tag{3}$$

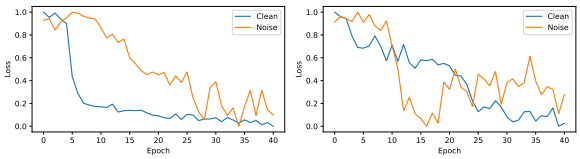
$$p^j(\mathbf{y}, \mathbf{C}) = \begin{cases} 0 & y_j = 1 \\ \sum_{j_0:y_{j_0}=1} \frac{C_{j_0,j}}{\sum_{j_0:y_{j_0}=1} 1} & y_j = 0 \end{cases} \tag{4}$$

Here, ρ_- controls the extent of negative labels transitioning. $p^j(\mathbf{y}, \mathbf{C})$ denotes the probability of a negative label j transitioning to a positive label given the current set of labels and the label correlation matrix. The label noise is label-dependent since it takes into account the label set of each instance.

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 false positive (FP) labels would be much greater than the number of false negative (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., 2022b), 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. The algorithm of LDN generation can be found in the Appendix B.1.

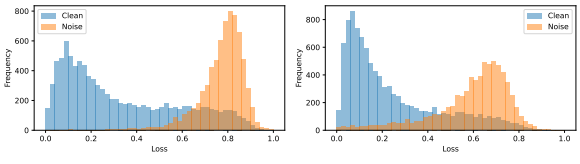
2.4 Characterizations of Training with LDN

In noisy label learning, a simple yet effective method to identify label noise is to utilize the memorization effect (Arazo et al., 2019). This effect highlights that DNNs tend to learn simple and general patterns before memorizing the noise, inspiring many sample selection based approaches (Lu et al., 2023; Song et al., 2024). Existing methods have confirmed that this approach can also be applied under CCN conditions (Li et al., 2022b; Song et al., 2024). However, can this method be used in NMLTC under LDN conditions? Here, we conduct an empirical study to compare the performance of these two types of noise. We generate 40% LDN noise and conduct experiments on the AAPD dataset. Simultaneously, we generate 40% CCN noise for comparison. In all experiments presented in this paper, the DNN model and training hyperparameters we use are consistent.



(a) CCN (b) LDN

Figure 2: Individual loss curves on AAPD.



(a) CCN (b) LDN

Figure 3: Loss distributions on AAPD.

LDN is harder to identify Figure 2 shows the changes in loss values for randomly sampled clean and noisy labels during training (under both CCN and LDN conditions). If we use the small loss criterion (Li et al., 2022b; Song et al., 2024) to filter noisy samples, we find that noisy samples are relatively easier to identify under CCN. During the early epochs of training, clean and noisy labels exhibit a significant difference. However, under LDN, clean and noisy labels are not easily distinguishable. In addition, we carried out a quantitative analysis, as shown in Figure 3. We illustrate the contrast in distribution of loss between labels affected by noise versus clean ones. It can be seen that under CCN noise, the overlap between the clean region and the noisy region is small, indicating that it is easier to identify the noise. In contrast, under LDN, it is relatively difficult to identify the noise due to the larger overlap. The reason is that LDN is very similar to the correct labels, making it prone to overfitting and thus difficult to distinguish from the correct labels. Additional observations can be found in Appendix B.2.

3 Method

Previous work (Han et al., 2018; Northcutt et al., 2021; Song et al., 2024) has identified noisy labels based on the "memorization effect", using the loss values from early epochs of deep learning models. However, as shown in section 2, this approach faces challenges with LDN due to its similarity to the true labels. Therefore, in this section, we propose a **L**abel-Spe**c**ific **D**enoising (LeD) framework. This framework considers various aspects of neural network training, including loss, ranking, and feature space neighbors, which provide a more comprehensive reflection of the likelihood that a label is noisy. Specifically, we introduce a holistic selection metric (HSM) that includes ranking-enhanced loss (REL) and centroid distance (CD). We then identify noisy labels for each label category from the perspective of each individual label category. We use a Gaussian mixture model (GMM) to identify noisy labels among positive and negative labels for each category, resulting in a noise probability for each label. Based on these noise probabilities, we refine the original labels in a fine-grained manner. The corrected labels are then used for retraining the model. The overall framework is shown in Figure 4.

3.1 Noisy Multi-Label Text Classification

The goal of noisy multi-label text classification (NMLTC) is to learn a function f that maps the input instance $\mathbf{x}_i$ and a label l to a relevance score $\hat{Y}_{i,j} = f(\mathbf{x}_i, j)$. We constructed the scoring function f by combining a text encoder ϕ and a multi-label classifier ψ. Following the approach of previous works (Su et al., 2022; Tan et al., 2024; Chai et al., 2024), we employed a BERT-based text encoder 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

$$L_{i,j} = -(\tilde{Y}_{i,j}\log(\hat{Y}_{i,j}) + (1-\tilde{Y}_{i,j})\log(1-\hat{Y}_{i,j})). \quad (5)$$

The notation $L_{i,j}$ represents the loss value associated with the j-th label for the i-th instance.

3.2 Holistic Selection Metric

Due to the presence of LDN, noisy labels and correct labels appear more similar, making it difficult to effectively identify noisy samples using a single metric, such as loss information. Therefore, we propose to jointly use two metrics from different perspectives: ranking-enhanced loss (REL) and centroid distance (CD). REL fully utilizes the information from prediction confidence, while CD relies on the distance in the feature space.

3.2.1 Ranking-Enhanced Loss

When learning with noisy labels, 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).

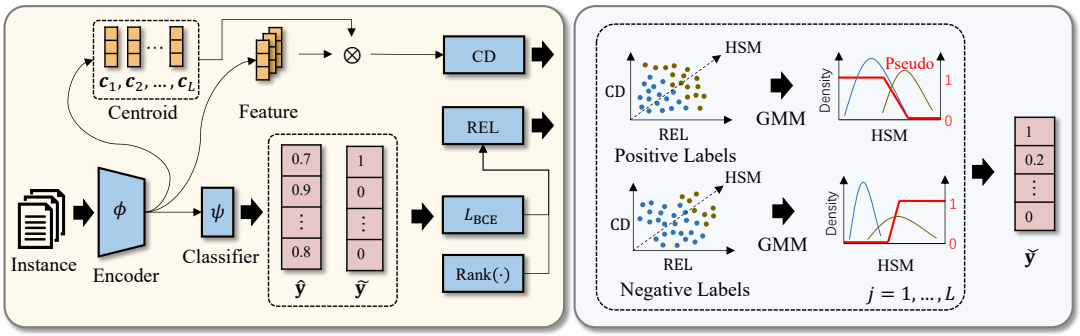


Figure 4: The overall framework of LeD. $\hat{\mathbf{y}}$ represents the predicted labels, $\tilde{\mathbf{y}}$ denotes the observed labels, and $\check{\mathbf{y}}$ stands for the corrected pseudo-labels. CD refers to centroid distance, REL stands for ranking-enhanced loss, and HSM denotes the holistic selection metric.

However, relying solely on the loss value $L_{i,j}$ to identify LDN, we may overlook differences between samples. For instance, some difficult labels may have a relatively high loss $L_{i,j}$, but their prediction $\hat{Y}_{i,j} \in \hat{\mathbf{y}}_i$ could still be correct. Therefore, we propose using the model's predicted ranking of labels as an additional metric. A smaller predicted ranking for a label indicates it is more likely to be clean. Label ranking can reveal distinctions between labels at the sample level. 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 $\text{Rank}(\hat{\mathbf{y}}_i) = (R_{i,1}, R_{i,2}, ..., R_{i,L})$, where $R_{i,j}$ is the rank metric for $\hat{Y}_{i,j}$. To combine ranking information with loss information, we propose the ranking-enhanced loss (REL) by adding an extra weight to the loss information. Thus, REL $E_{i,j}$ can be calculated by:

$$W(\hat{Y}_{i,j}) = \min(\log(R_{i,j}) + 1, \theta), \tag{6}$$

$$E_{i,j} = W(\hat{Y}_{i,j}) \times L_{i,j}. \tag{7}$$

The logarithmic function is used to constrain the scale of the rank values, and a fixed value θ is employed for truncation to ensure that it has a limited impact on the loss information.

3.2.2 Centroid Distance

Although the sample separability achieved through REL is better, the separation metric still relies on model prediction. This reliance means there is still a risk of overfitting the classifier, especially in the case of LDN, where label noise often occurs among similar labels, increasing the likelihood of classifier overfitting. Consequently, this leads to low discrimination between the model predictions of clean and noisy labels. Therefore, solely using REL may not be sufficient to distinguish clean labels from noisy ones when model predictions are close.

Except for separating the samples in the output space, we propose an additional metric computed in the feature space to mitigate the bias introduced by the classifier, as the learned features can handle noise labels better. Specifically, we proposed the centroid distance (CD) metric. for a given sample, we can compute the distance between its feature and the class feature centroid to assess the extent to which the sample's feature differs from its class centroid. To improve the quality of the class feature centroid for distance calculation, we construct the feature centroid by incorporating high-confidence samples from the observed class. The class centroid $\mathbf{c}_j$ based on a high-confidence sample set $\mathcal{H}_j$ is calculated by:

$$\mathbf{c}_j = \frac{1}{|\mathcal{H}_j|} \sum_{i=1}^{|\mathcal{H}_j|} \mathbf{v}_i, \tag{8}$$

$$\mathcal{H}_j = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{S}_j,\ \hat{Y}_{i,j} > h_j\}, \tag{9}$$

where $\mathbf{v}_i$ is the feature of $\mathbf{x}_i$ and $\mathcal{S}_j = \{\mathbf{x}_i | \tilde{Y}_{i,j} = 1\}$. we can use the prediction confidence of class j for sample $\mathbf{x}_i$, e.g., $\hat{Y}_{i,j}$, as the selection criteria compare with the threshold h_j. $\mathcal{H}_j$ is constructed by the samples in $\mathcal{S}_j$ whose corresponding prediction confidence of class $\hat{Y}_{i,j}$ is higher than h_j. The high-confidence threshold h_j is defined as:

$$h_j = \frac{1}{|\mathcal{S}_j|} \sum_{i=1}^{|\mathcal{S}_j|} w_i \times \hat{Y}_{i,j} \tag{10}$$

$$w_i = \max\left(1, \hat{Y}_{i,j} / \bar{Y}_j\right). \tag{11}$$

h_j is calculated by the weighted sum of the prediction confidence of class j for all samples. The

weight w_i increases when a sample's prediction confidence of class j is higher than its class average $\bar{Y}_j$. Thus, the threshold is high enough to ensure the quality of the selected samples. Therefore, the proposed metric CD $D_{i,j}$ can be calculated by $D_{i,j} = -\cos(\mathbf{v}_i, \mathbf{c}_j)$.

To facilitate the integration of these two metrics, we perform min-max normalization on them (Hu et al., 2022), obtaining normalized results $\hat{E}_{i,j}$ and $\hat{D}_{i,j}$ respectively. The linear combination of both metrics results in a new holistic selection metric (HSM) $M_{i,j} = \alpha \cdot \hat{E}_{i,j} + (1-\alpha)\hat{D}_{i,j}$. The combination coefficient α plays a crucial role in determining the balance between the two metrics. By combining the advantages of both metrics, HSM effectively captures both the confidence from model predictions and the robustness from the feature space, thus improving the discrimination between clean and noisy labels.

3.3 Fine-Grained Label-Specific Correction

As previously mentioned, there are two types of noise in noisy multi-labels: false positive and false negative. In this section, we take false positive noise as an example, with the method for handling false negative noise being similar. After obtaining the HSM for each label, we proceed with the identification and fine-grained correction of noisy labels based on each label category. Fine-grained label-specific correction involves re-labeling the noisy dataset based on the HSM values to create a cleaner training set.

If noisy labels are divided globally (without distinguishing between categories), the differences between categories will be ignored. Some difficult categories may all be classified as noise, while some simple categories may not be classified as noise at all. Therefore, we individually identify noisy labels for each class. We first obtain the HSM set $\mathcal{M}_j^+$ corresponding to all positive labels for label j, i.e., $\mathcal{M}_j^+ = \{M_{i,j} | \tilde{Y}_{i,j} = 1\}$. True labels have lower HSM values compared to noisy ones due to DNNs' memorization effect (Arpit et al., 2017; Hu et al., 2023). Therefore, we employ a bi-modal univariate Gaussian mixture model (GMM) for each HSM set using the expectation-maximization (EM) algorithm, resulting in L GMM models for positive labels.

Given the HSM, its clean-label probability is obtained by the posterior probability $P_{\mathcal{G}}(\tilde{Y}_{i,j})$ of the corresponding GMM, Since distinguishing between noisy and clean labels near the decision boundary is challenging, we have employed a fine-grained correction strategy, as opposed to using hard pseudo-labels (Li et al., 2020). The specific approach is as follows:

$$\check{Y}_{i,j} = \begin{cases} 1 - \tilde{Y}_{i,j} & P_{\mathcal{G}}(\tilde{Y}_{i,j}) \leq 0.5 - \epsilon, \\ P_{\mathcal{G}}(\tilde{Y}_{i,j}) & 0.5 - \epsilon < P_{\mathcal{G}}(\tilde{Y}_{i,j}) \leq 0.5 + \epsilon \\ \tilde{Y}_{i,j} & P_{\mathcal{G}}(\tilde{Y}_{i,j}) > 0.5 + \epsilon \end{cases} \tag{12}$$

The corrected pseudo-label $\check{Y}_{i,j}$ is obtained. The implication is that if $P_{\mathcal{G}}(\tilde{Y}_{i,j})$ is large, we consider the label to be likely a clean label and thus keep it unchanged. Conversely, if $P_{\mathcal{G}}(\tilde{Y}_{i,j})$ is small, we consider the label to be likely incorrect and thus perform label flipping. When the value of $P_{\mathcal{G}}(\tilde{Y}_{i,j})$ is close to 0.5, it is difficult to determine the noise situation, so we adopt a soft label approach.

Datasets	N_{trn}	N_{tst}	L	L_{avg}	N_{avg}
MOVIE	105,616	11,736	28	2.1	112
AAPD	54,840	1,000	54	2.4	163
RCV1	23,149	781,028	103	3.2	124

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

4 Experiment

4.1 Experimental Setup

Datasets Following previous work (Chen et al., 2023), we evaluate the proposed model on three synthetic benchmark datasets namely MOVIE, AAPD, and RCV1 with varying LDN fractions. Table 1 contains the statistics of these three benchmark datasets.

Evaluation Metrics For a comprehensive and reliable evaluation, we follow conventional settings (Chen et al., 2019, 2023) and report the following metrics: micro-F1 (mi-F1), macro-F1 (ma-F1) and mean Average Precision (mAP). 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 LeD, we selected the nine most representative baseline models in three groups. (1) MLTC methods: AttXML (You et al., 2019), HTTN (Xiao et al., 2021) and

Noise Rate	$\rho = 0.2$			$\rho = 0.4$			$\rho = 0.6$		
Methods	mi-F1	ma-F1	mAP	mi-F1	ma-F1	mAP	mi-F1	ma-F1	mAP
AttXML	61.89	34.56	51.70	56.72	33.50	44.32	50.16	29.79	40.83
HTTN	61.13	34.45	51.16	56.59	32.82	44.24	49.26	28.58	39.10
LSFA	62.81	37.40	53.26	58.84	33.53	47.87	52.13	29.92	38.45
GCE	65.68	39.65	53.17	61.80	35.95	48.92	52.76	31.37	41.98
WSIC	62.82	38.94	52.82	60.26	35.89	46.89	53.93	31.37	39.38
RTM	64.79	39.15	54.70	62.02	36.26	46.97	53.66	32.18	41.79
MLLSC	63.85	38.48	51.41	60.83	36.46	47.74	53.09	31.57	39.22
Multi-T	65.54	38.84	52.79	60.36	36.06	46.15	52.67	30.87	41.20
nEM	65.37	40.92	54.00	62.45	36.91	48.76	53.15	31.70	42.23
LeD	**66.74**	**42.15**	**55.29**	**63.77**	**37.89**	**49.32**	**56.66**	**34.26**	**44.57**

Table 2: Performance on MOVIE with different LDN ratios.

Noise Rate	$\rho = 0.2$			$\rho = 0.4$			$\rho = 0.6$		
Methods	mi-F1	ma-F1	mAP	mi-F1	ma-F1	mAP	mi-F1	ma-F1	mAP
AttXML	52.16	18.80	43.10	42.65	8.74	32.02	37.49	5.28	28.12
HTTN	55.15	21.16	43.84	46.37	11.98	35.59	40.68	8.69	32.33
LSFA	56.53	22.51	45.90	47.94	11.31	36.32	41.72	9.19	32.90
GCE	54.31	23.59	43.28	47.09	13.20	35.45	42.94	8.74	32.87
WSIC	56.11	23.44	44.56	49.34	13.71	36.63	42.15	9.78	32.67
RTM	54.76	22.20	44.70	49.18	13.39	36.37	42.91	8.82	34.14
MLLSC	55.16	22.67	45.47	47.41	13.59	36.38	41.34	8.41	32.22
Multi-T	56.87	23.16	44.68	49.93	11.78	37.01	43.77	10.61	33.70
nEM	55.46	22.99	**46.43**	48.89	16.16	39.76	43.84	9.89	34.80
LeD	**57.34**	**24.33**	46.30	**50.26**	**17.75**	**41.27**	**45.30**	**11.62**	**36.18**

Table 3: Performance on AAPD with different LDN ratios.

LSFA (Xu et al., 2023). (2) Noisy multi-label learning (NMLL) methods: GCE (Zhang and Sabuncu, 2018), WSIC (Hu et al., 2019), RTM (Patrini et al., 2017), Multi-T (Li et al., 2022b), 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 C.3.

4.2 Experimental Results

Main Results As depicted in Tables 2-4, we have observed the following phenomena: (1) In all cases, our method shows significant improvements compared to other methods. By utilizing a holistic selection metric, we evaluate labels from multiple perspectives, enabling finer-grained identification and correction of noisy labels, which leads to optimal experimental results. (2) Due to overfitting to noisy labels, most existing MLTC methods tend to perform worse compared to NMLL methods. (3) NMLL methods like RTM and Multi-T depend only on loss for noise rate estimation, which is inadequate under LDN. Similarly, nEM and MLLSC are constrained by insufficiently sensitive metrics to detect noisy labels.

Experiments on the Real-world Dataset The Riedel dataset (Chen et al., 2023) is a large-scale real-world NMLTC dataset, containing 53 classes, each representing a different relation. It is derived from the New York Times corpus. The training data consists of 281,270 instances, while the test set includes 3,762 instances. We used the same backbone as nEM to ensure a fair comparison. The results, shown in Table 5, indicate that our method outperforms the best baseline by 10% in terms of the ma-F1 metric, demonstrating the effectiveness of our approach on the real-world NMLTC dataset.

Noise Rate	$\rho = 0.2$			$\rho = 0.4$			$\rho = 0.6$		
Methods	mi-F1	ma-F1	mAP	mi-F1	ma-F1	mAP	mi-F1	ma-F1	mAP
AttXML	71.30	32.10	60.99	65.69	22.23	58.01	64.18	20.20	52.70
HTTN	64.59	27.19	54.87	62.87	20.44	54.16	63.18	19.67	51.78
LSFA	69.67	30.87	59.43	64.55	21.73	56.25	64.23	20.46	53.17
GCE	68.08	27.95	56.93	61.94	20.81	54.78	64.41	21.66	53.70
WSIC	72.69	32.89	63.54	64.36	23.35	57.66	65.02	21.16	53.38
RTM	72.15	33.48	63.41	66.44	21.11	57.75	64.48	20.16	54.11
MLLSC	71.03	32.97	63.18	67.53	20.03	57.85	64.92	19.03	52.56
Multi-T	72.99	31.89	62.75	66.49	19.41	56.65	62.85	19.31	50.95
nEM	71.86	32.93	62.47	67.67	20.92	57.79	63.61	20.71	52.86
LeD	**74.66**	**35.32**	**64.90**	**69.61**	**24.70**	**62.29**	**67.65**	**23.33**	**56.42**

Table 4: Performance on RCV1 with different LDN ratios.

Methods	mi-F1	ma-F1	mAP
AttXML	56.77	33.78	50.81
GCE	55.60	32.95	47.93
WSIC	58.96	35.57	54.06
Multi-T	57.15	34.40	52.87
nEM	59.58	35.70	54.51
LeD	**63.72**	**39.40**	**57.82**

Table 5: Performance comparison on Riedel.

HSM			Dataset		
Loss	REL	CD	MOVIE	AAPD	RCV1
✓			60.15	48.06	68.75
✓	✓		62.15	48.95	69.16
		✓	59.87	47.88	66.29
✓	✓	✓	63.77	50.26	69.61

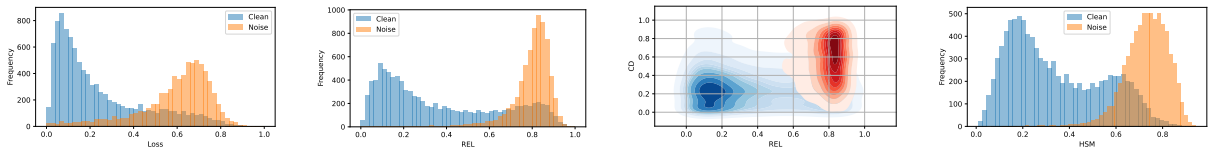
Table 6: Performance comparison of HSM components based on mi-F1 scores across various datasets.

Ablation Study In the following experiments, we aim to analyze the effectiveness of each component of the proposed LeD method on three datasets. The LDN ratio is 0.4. We compare the complete LeD method with the following variants: (a) HSM (Loss): This variant uses only the loss as the metric for identifying noisy labels. (b) HSM (REL): This variant uses the REL metric for identifying. (c) HSM (CD): This variant uses the CD metric for identifying. According to Table 6, we observe that the different components of the HSM metric collectively contribute to enhancing the quality of noise identification. By incorporating instance-level rank information, the model gains the ability to differentiate between different instances, enabling a more accurate distinction between clean and corrupted labels. Additionally, the introduction of the feature-based metric, CD, significantly contributes to noise identification.

Effectiveness of HSM In Figure 5, we present the distributions of the positive label "cs.IT" in the AAPD dataset using HSM(Loss), HSM(REL), and HSM(CD). Firstly, as shown in (a) and (b), loss information demonstrates a certain capability in noise identification. When rank information is introduced, using REL as a metric, the noise identification capability is significantly enhanced (reduced overlapping areas). From (c), we can observe the complementary nature of the prediction-based REL metric and the feature-based CD metric. Finally, in (d), it is evident that combining both metrics in HSM results in a significantly improved noise identification capability (minimal overlapping areas).

5 Related Work

Learning from Noisy Labels In multi-class classification with noisy labels, most approaches leverage the memorization effect of deep neural networks (DNNs) (Arpit et al., 2017), where simple and generalized patterns are learned before overfitting to noisy patterns. Specifically, small-loss instances are likely to be clean instances (Han et al., 2018; Jiang et al., 2018; Wei et al., 2020). Another approach involves sample selection based on feature distributions (Li et al., 2022a, 2023). Recent methods (Hu et al., 2023; Lu et al., 2023) have proposed more comprehensive evaluation metrics to distinguish between clean and corrupted data,



(a) HSM(Loss) (b) HSM(REL) (c) HSM(REL)&HSM(CD) (d) HSM

Figure 5: The visualization of metric distribution on AAPD with 40% LDN noise.

considering aspects such as information throughout the training process and prediction confidence. Inspired by these approaches, we propose a holistic selection metric for noisy MLTC that integrates various noise label characteristics, including multi-label ranking and feature information.

Noisy Multi-Label Learning Noisy Multi-Label Learning is an emerging research topic due to the complexity of noise mechanisms in multi-label settings compared to multi-class problems. Zhao et al. (2021) introduced pre-trained label embeddings for regularization, achieving robust learning. GCE (Zhang and Sabuncu, 2018), WISC (Hu et al., 2019), and MLLSC (Ghiassi et al., 2022) developed robust loss functions by weighting labels. Methods like RTM (Patrini et al., 2017) and Multi-T (Li et al., 2022b) address the estimation problem of noise transition matrices in the multi-label context. The nEM method (Chen et al., 2023) uses latent variable models to model the transition process of noisy labels, achieving robust MLTC. Xia et al. (2023) identifies noisy labels through label correlation. Song et al. (2024) employs the small loss trick for noisy label selection and correction. However, existing methods either assume label noise is entirely random or class conditional, neglecting the fact that label noise is often correlated with current labels in real-world scenarios.

6 Conclusions

In this paper, we first verify that real-world datasets often deviate from the class-conditional noise assumption. Based on this observation, we introduce label-dependent noise (LDN), revealing the characteristics of label-dependent noise and designing a method to generate controllable LDN. Subsequently, we propose a novel label-specific denoising framework to enhance multi-label text classification under label-dependent noise. Extensive experiments on benchmark datasets demonstrate that our method significantly improves performance under both synthetic and real-world noise conditions, outperforming existing state-of-the-art models.

7 Limitations

Our method leverages the memorization effect (Arpit et al., 2017) observed in deep learning models for sample selection and correction. This effect has not been observed in other traditional machine learning methods, limiting the applicability of our framework to deep learning-based approaches only. Although label-dependent noise (LDN) can be considered a special case of instance-dependent noise (IDN) (Chen et al., 2021; Wang et al., 2024), our framework has not been explicitly validated for handling IDN in our experiments. Calculating the HSM for each label can be computationally demanding, particularly with larger label size datasets. This could limit the scalability of our approach for significantly larger datasets.

8 Acknowledgement

This work was partly supported by the National Natural Science Foundation of China under Grant 62176020; the Joint Foundation of the Ministry of Education for Innovation team (8091B042235); the Beijing Natural Science Foundation under Grant L211016; the Fundamental Research Funds for the Central Universities (2019JBZ110); and the State Key Laboratory of Rail Traffic Control and Safety (Contract No. RCS2023K006), Beijing Jiaotong University.

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.

Yuyang Chai, Zhuang Li, Jiahui Liu, Lei Chen, Fei Li, Donghong Ji, and Chong Teng. 2024. Compositional generalization for multi-label text classification: A data-augmentation approach. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17727–17735. AAAI Press.

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.

Zijun Cui, Yong Zhang, and Qiang Ji. 2020. Label error correction and generation through label relationships. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3693–3700. AAAI Press.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Shiming Ge, and Tongliang Liu. 2022a. Selective-supervised contrastive learning with noisy labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 316–325. IEEE.

Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. 2022b. 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.

Chao Liang, Zongxin Yang, Linchao Zhu, and Yi Yang. 2023. Co-learning meets stitch-up for noisy multi-label visual recognition. *IEEE Trans. Image Process.*, 32:2508–2519.

Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. 2023. An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8730–8744. Association for Computational Linguistics.

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.

Yang Lu, Yiliang Zhang, Bo Han, Yiu-Ming Cheung, and Hanzi Wang. 2023. Label-noise learning with intrinsically long-tailed data. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 1369–1378. IEEE.

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.

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.

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.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. 2024. Toward robustness in multi-label classification: A data augmentation strategy against imbalance and noise. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 21592–21601. AAAI Press.

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.

Wei Tan, Ngoc Dang Nguyen, Lan Du, and Wray L. Buntine. 2024. Harnessing the power of beta scoring in deep active learning for multi-label text classification. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 15240–15248. AAAI Press.

Yejiang Wang, Yuhai Zhao, Zhengkui Wang, Wen Shan, and Xingwei Wang. 2024. Limited-supervised multi-label learning with dependency noise. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 15662–15670. AAAI Press.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Tingting Wu, Xiao Ding, Minji Tang, Hao Zhang, Bing Qin, and Ting Liu. 2023. Noisywikihow: A benchmark for learning with real-world noisy labels in natural language processing. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4856–4873. Association for Computational Linguistics.

Xiaobo Xia, Jiankang Deng, Wei Bao, Yuxuan Du, Bo Han, Shiguang Shan, and Tongliang Liu. 2023. Holistic label correction for noisy multi-label classification. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 1483–1493. IEEE.

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, 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.

Ming-Kun Xie and Sheng-Jun Huang. 2022. Partial multi-label learning with noisy label identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(7):3676–3687.

Ming-Kun Xie and Sheng-Jun Huang. 2023. CCMN: A general framework for learning with class-conditional multi-label noise. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):154–166.

Pengyu Xu, Lin Xiao, Bing Liu, Sijin Lu, Liping Jing, and Jian Yu. 2023. 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.

Qian-Wen Zhang, Ximing Zhang, Zhao Yan, Ruifang Liu, Yunbo Cao, and Min-Ling Zhang. 2021. Correlation-guided representation for multi-label text classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3363–3369. ijcai.org.

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.

Wenting Zhao, Shufeng Kong, Junwen Bai, Daniel Fink, and Carla P. Gomes. 2021. HOT-VAE: learning high-order label correlation for multi-label classification via attention-based variational autoencoders. 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 15016–15024. AAAI Press.

A From CCN to LDN

A.1 Proof of Theorem 1

Proof. If the CCN assumption is satisfied, then by Definition 2, the random variable Z_j follows a Bernoulli distribution with parameters $\mathbf{T}^j_{0,1}$. For the joint probability of Z_{j_0} and Z_{j_1}, we have:

$$
\begin{aligned}
&P(Z_{j_0}=l, Z_{j_1}=k) \\
&= P(\tilde{Y}_{j_0}=l, \tilde{Y}_{j_1}=k | Y_{j_0}=0, Y_{j_1}=0) \\
&= P(\tilde{Y}_{j_0}=l | Y_{j_0}=0, Y_{j_1}=0) \\
&\cdot P(\tilde{Y}_{j_1}=k | Y_{j_0}=0, Y_{j_1}=0) \\
&= P(\tilde{Y}_{j_0}=l | Y_{j_0}=0) \cdot P(\tilde{Y}_{j_1}=k | Y_{j_1}=0) \\
&= P(Z_{j_0}=l) \cdot P(Z_{j_1}=k)
\end{aligned}
$$

□

A.2 Hypothesis Test

It is assumed that in the actual data, $n_{k,l}$ represents the number of samples for the joint distribution $\{Z_{j_0}=k, Z_{j_1}=l\}$, where $k,l \in \{0,1\}$. And the total sample size is denoted as $n = \sum_{i=0}^{1}\sum_{j=0}^{1} n_{k,l}$. Under the assumption of $\mathbf{H_0}$, we can estimate the maximum likelihood estimates of the parameters:

$$
\begin{aligned}
\hat{P}(Z_{j_0}=k) &= \frac{n_{k,0}+n_{k,1}}{n}, \\
\hat{P}(Z_{j_1}=l) &= \frac{n_{0,l}+n_{1,l}}{n}
\end{aligned}
$$

Therefore,

$$
\hat{P}(Z_{j_0}=k, Z_{j_1}=l) = \hat{P}(Z_{j_0}=k) \cdot \hat{P}(Z_{j_1}=l)
$$

From this, we can calculate the test statistic:

$$
\chi^2 = \sum_{k=0}^{1}\sum_{l=0}^{1} \frac{(n_{k,l} - n \cdot \hat{P}(Z_{j_0}=k, Z_{j_1}=l))^2}{n \cdot \hat{P}(Z_{j_0}=k, Z_{j_1}=l)}
$$

Now we apply the chi-square test to the real-world noise multi-label benchmark Riedel (Chen et al., 2019). We select the labels "nationality" and "place_lived" from the Riedel validation set as j_0 and j_1, respectively. Hypothesis testing results show that $\chi^2 = 940.5$, with a p-value of $1.5e^{-206}$.

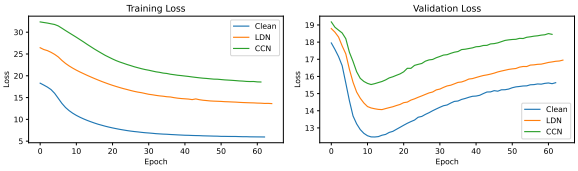
B Label-Denpendent Noise

B.1 LDN Generation Algorithm

Algorithm 1 LDN Generation.

Input: Clean training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, noise fraction parameters ρ_+ and ρ_-.
Output: A dataset with LDN $\tilde{\mathcal{D}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^N$.

Calculate correlation matrix $\mathbf{C}$ by Eq.(1).
for $\mathbf{x}_i, \mathbf{y}_i$ in $\mathcal{D}$ **do**
  for $Y_{i,j}$ in $\mathbf{y}_i$ **do**
    if $Y_{i,j} = 0$ **then**
      Calculate $\mathbf{T}^j_{0,1}(\mathbf{y}_i)$ by Eq.(3).
      $\tilde{Y}_{i,j} \sim \text{Bernoulli}(1, \mathbf{T}^j_{0,1}(\mathbf{y}_i))$
    else
      $\tilde{Y}_{i,j} \sim \text{Bernoulli}(1, 1-\rho_+)$
      Record $\tilde{Y}_{i,j}$
    end if
  end for
  $\tilde{\mathbf{y}}_i = \{\tilde{Y}_{i,1}, \tilde{Y}_{i,2}, ..., \tilde{Y}_{i,L}\}$
  Record $\tilde{\mathbf{y}}_i$
end for
return $\tilde{\mathcal{D}} = \{(\mathbf{x}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^N$.



(a) Training (b) Validation

Figure 6: Loss curves with varying types of 40% noise on AAPD.

B.2 Characterizations of Training with LDN

LDN is easier to fit Figure 6 shows the training and validation loss curves under different noise settings. It can be observed that the training and validation loss is lower under LDN compared to CCN. This suggests that DNNs find it easier to fit LDN. This finding aligns with our intuition since the noisy labels under LDN are closely related to the output labels and can mislead DNNs. In this regard, LDN is more challenging to mitigate because the label-dependent noise significantly confuses DNNs, potentially leading to overfitting.

LDN causes relatively less harm Precisely because LDN is closely related to the output labels, it causes relatively less harm compared to CCN noise, as illustrated in Figure 6b. In CCN, many irrelevant noisy labels significantly affect the model's training, leading to greater harm.

C Experiments

C.1 Datasets

We evaluate the proposed model on three synthetic benchmark datasets and one real-world dataset for noisy multi-label text classification (NMLTC), namely MOVIE, AAPD, RCV1, and Riedel.

- MOVIE: The MOVIE dataset is designed for movie genre classification. It contains movie plots and genre types extracted from the IMDB database and is publicly available [1].
- AAPD: The AAPD dataset (Yang et al., 2018) includes abstracts and corresponding subjects of 55,840 publications in the field of computer science from arXiv [2].
- RCV1: The Reuters Corpus Volume I (RCV1) (Lewis et al., 2004) is a benchmark dataset for text categorization, consisting of newswire articles produced by Reuters from 1996 to 1997 [3].
- Riedel: The Riedel dataset (Chen et al., 2019, 2023) was created by aligning entity pairs from Freebase (a large knowledge graph) with the New York Times (NYT) corpus. The dataset includes 53 relations, with training data from the 2005-2006 corpus and test data from the 2007 corpus [4].

C.2 Evaluation Metrics

Following previous works (Chen et al., 2023), we use three main metrics which are commonly used in MLTC evaluations: micro-F1 (mi-F1), macro-F1 (ma-F1), and mAP.

Micro-F1: This metric is calculated by aggregating the contributions of all classes to compute the average F1 score. It is particularly useful when dealing with imbalanced datasets, as it gives equal weight to each instance. The micro-F1 score is defined as:

$$\text{micro-F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Precision and Recall are computed globally over all instances.

Macro-F1: Unlike micro-F1, macro-F1 calculates the F1 score for each class independently and then takes the average. This metric treats all classes equally, regardless of their frequencies. It is defined as:

$$\text{macro-F1} = \frac{1}{L}\sum_{i=1}^{L} \text{F1}_i$$

where L is the number of classes and F1_i is the F1 score of class i.

Mean Average Precision (mAP): mAP is a measure used to evaluate the ranking quality of the model's predictions. It calculates the average precision across all classes and then averages these values. It is particularly useful for tasks where the order of the predictions matters. The mAP is defined as:

$$\text{mAP} = \frac{1}{L}\sum_{i=1}^{L} \text{AP}_i$$

where AP_i is the average precision of class i.

C.3 Implementation Details

Backbone Given that the Riedel dataset is a multi-instance MLTC dataset, we use the same backbone as nEM (Chen et al., 2023) for fair comparison, i.e., PCNN+ATT. For the other three synthetic datasets, we adopt the pre-trained BERT (Devlin et al., 2019) as the backbone of our model, using the PyTorch implementation from HuggingFace Transformers (Wolf et al., 2019). The maximum document length is set to 512 due to BERT's limitations (Devlin et al., 2019), and documents are either zero-padded or truncated to this length.

All experiments are conducted in a Linux environment with a single Tesla A100 GPU (40GB). Our model is trained using AdamW (Kingma and Ba, 2015). To optimize GPU memory usage and enhance training efficiency, we use automatic mixed precision (AMP).

The training time for the MOVIE and AAPD datasets is approximately 5.4 hours and 7.7 hours, respectively. For the RCV1 and Riedel datasets, the training time is approximately 9.0 hours and 14.5 hours, respectively.

Hyperparameters Regarding the key hyperparameters of our proposed method, the coefficient α and threshold ϵ, we set $\alpha = 0.7$ and $\epsilon = 0.1$ for MOVIE. For AAPD and RCV1, we set $\alpha = 0.7$

[1] https://github.com/davidsbatista/text-classification

[2] https://git.uwaterloo.ca/jimmylin/Castor-data/tree/master/datasets/AAPD/data

[3] http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm

[4] https://github.com/AlbertChen1991/nEM

and $\epsilon = 0.05$. For the Riedel dataset, we set $\alpha = 0.6$ and $\epsilon = 0.15$. We set $\theta = 3$ for all datasets. All experiments are run at least three times with different random seeds, and we report the average values of the results.

D Related Work

Multi-Label Text Classification Early multi-label text classification (MLTC) primarily focused on learning better text representations (Liu et al., 2017) and capturing label correlations (Zhang et al., 2021). Label-specific feature learning (You et al., 2019; Xiao et al., 2019; Ma et al., 2021) introduced label representations to learn specific text representations for different labels, improving label differentiation. Recently, some methods (Su et al., 2022; Xu et al., 2023; Lin et al., 2023) have used contrastive learning to achieve more stable text representations, mitigating the impact of label imbalance. In contrast, we focus on improving MLTC performance in the presence of noisy labels.