TAMING PROMPT-BASED DATA AUGMENTATION FOR LONG-TAILED EXTREME MULTI-LABEL TEXT CLASSIFICATION

Pengyu Xu, Mingyang Song, Ziyi Li, Sijin Lu, Liping Jing, Jian Yu*

Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
{pengyu, mingyang.song, 22251138, 22120406, lpjing, jianyu}@bjtu.edu.cn

ABSTRACT

In extreme multi-label text classification (XMC), labels usually follow a long-tailed distribution, where most labels only contain a small number of documents and limit the performance of XMC. Data augmentation (DA) is a simple but effective strategy to solve such low-resource problems. In this paper, we propose a prompt-based DA method called XDA, which is specifically designed for XMC. First, we employ a soft prompt during the fine-tuning process of the T5 model for label-conditional DA, thereby enabling T5 to augment samples while preserving label-compatibility. Subsequently, XDA performs sample filtering on the augmented samples through the diversity of text and the consistency of labels, which enhances the quality of the DA. In contrast to traditional *sample-level* DA, we propose a *pair-level* DA method by masking the augmented sample-label pairs of head-labels during training, effectively mitigating the long-tailed problem. Comprehensive experiments on benchmark datasets have shown that the proposed XDA outperforms the state-of-the-art counterparts.

***Index Terms*—** extreme classification, multi-label learning, long-tailed, data augmentation, prompt

1. INTRODUCTION

Extreme multi-label text classification (XMC) deals with the problem of predicting the most relevant subset of labels from an enormously large label space. It has a wide range of applications, such as product search [1], document tagging [2], keyword recommendation [3] and so on. Even though various techniques have been proposed for XMC, it is still a challenging task due to the “long-tailed” label distribution [4, 5]. Figure 1 illustrates the long-tailed label distribution in the Wiki10-31K dataset. Only 10% of the labels have more than 10 training samples (i.e., head-labels), while the remaining 90% are long-tail labels with much fewer training samples. In this situation, training classification models for the tail-labels is much more difficult than that for head-labels, which suffers severely from the lack of sufficient training samples.

One immediate approach to address the problem is data augmentation (DA) which can compensate the scarce data for tail-labels [6, 7]. Several existing works [8, 9, 10] resort to applying pre-trained language models (PLMs) for DA in a low-resource setting. However, because of the label co-occurrence, existing approaches of DA struggle in the multi-label scenario [11]. A document usually contains several labels, making the selection for tail-labels no longer independent. As shown in Figure 1, the long-tailed problem is not necessarily eliminated and may even be exaggerated by directly adopting DA methods. Moreover, due to the complex semantics of XMC texts, it is difficult to control the strength of semantic changes after augmentation [9, 12], resulting in a certain amount of low-quality augmented samples.

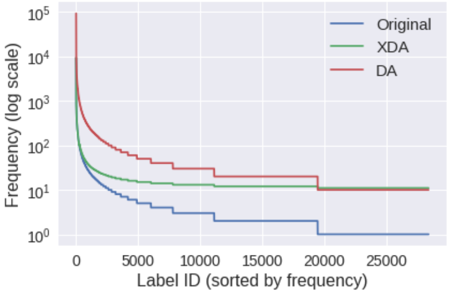

Fig. 1: Wiki10-31K shows a long-tailed distribution of labels (denoted as Original). DA denotes the distribution after augmenting 10 times directly, and XDA denotes the distribution after augmenting 10 times under the XDA framework.

This paper introduces XDA, a novel approach that effectively addresses the aforementioned challenges by incorporating a prompt-based DA method specifically designed for XMC. XDA employs three mechanisms to alleviate the low-quality and long-tailed problems of augmented samples: *Prompt*, *Filter* and *Mask*. Figure 2 shows the pipeline of XDA. First, we use T5 [13] to produce synthetic data conditioned on the labels and the masked text. Our approach enables T5 to augment samples while preserving label-compatibility. Notably, we only allow tuning the additional soft prompts during fine-tuning, which significantly reducing the amount of parameters to be tuned. Then, we perform sample filtering on the augmented samples through the diversity of text and the consistency of labels. At last, in contrast to traditional *sample-level* DA meth-

*Corresponding author

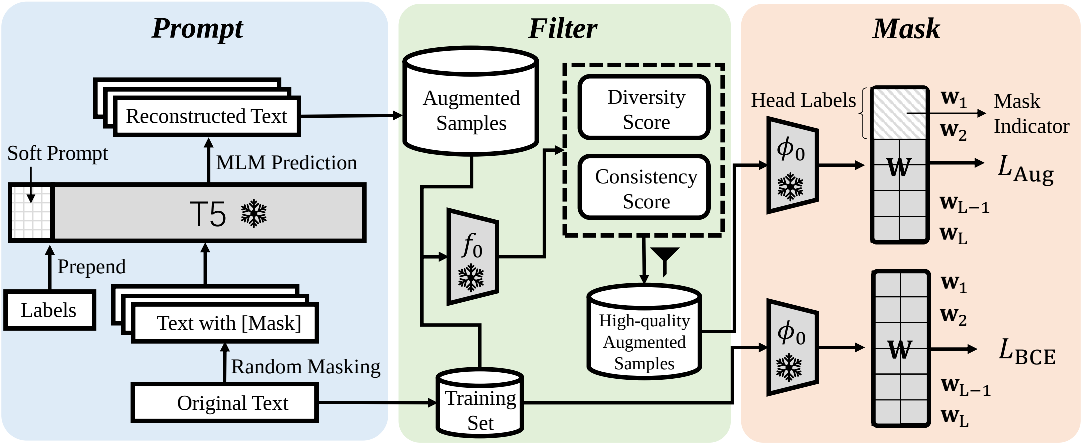

Fig. 2: The architecture of the proposed XDA. The snowflake represents frozen parameters.

ods, we propose a novel masked data augmentation (MDA) approach focuses on *pair-level* DA by masking the augmented sample-label pairs of head-labels during training. This approach exclusively augments positive sample-label pairs for the tail-labels, thereby addressing the label sparsity issue without exacerbating the long-tailed problem, as demonstrated in Figure 1. Our experiments show that XDA outperforms state-of-the-art baselines on three benchmark datasets, especially on tail-labels. Our source code and hyper-parameter settings are released at `https://github.com/stxupengyu/XDA`.

2. METHOD

As depicted in Figure 2, our method XDA consists of three components: *Prompt*, *Filter* and *Mask*. First, we freeze the entire pre-trained model and only allow tuning the additional soft prompts during fine-tuning to produce complete synthetic data conditioned on the output labels. Second, we filter out the low-quality samples in the generated sample set through the diversity of text and the consistency of labels. At last, we mask the head sample-label pairs of the augmented data in the training stage, which mitigates the label sparsity issue.

2.1. Prompt-based Data Augmentation

Prompt-based Learning. Fine-tuning is a conventional approach for adapting pre-trained language models (PLMs) to downstream tasks. However, as PLMs become larger in scale, directly fine-tuning them incurs significant resource consumption. Motivated by recent research [14, 10], we employ the soft prompt technique for fine-tuning. This involves adding a sequence of trainable vectors $\mathbf{P}^j = \{\mathbf{p}_1^j, \cdots, \mathbf{p}_k^j\}$ at each transformer layer. In the training process, we exclusively update the parameters of the soft prompt while keeping all other PLM parameters fixed. The hidden states in the Transformer model is defined as:

$$\mathbf{h}_i^j = \begin{cases} \mathbf{p}_i^j & i \le k \\ \mathbf{e}_i & i > k \wedge j = 0 \\ \mathcal{F}(\mathbf{h}^{j-1})_i & \text{Otherwise} \end{cases} \qquad (1)$$

where $\mathbf{h}_i^j$ is the i-th hidden states at the j-th layer, and $\mathbf{e}_i$ is the word embedding layer. The forward function of the Transformer layer is denoted as $\mathcal{F}(\cdot)$. By adopting the soft prompt, we have adapted the PLM for downstream DA tasks while effectively managing computational costs.

Label-Conditional Data Augmentation. Previous works often use label information as auxiliary assistance during DA[8, 10]. XDA is further utilizing output multi-labels as additional conditioning. Through this method, we strive to enhance the model's proficiency in predicting words in masked positions, taking into account both contextual information and label information. In our XDA process, we generate augmented documents N_a times based on the original ones. Our approach involves merging the text and labels into a single sequence. Then, we randomly mask a specific percentage of the input tokens. Afterward, we utilize the fine-tuned T5 model to predict and fill in the masked tokens to form a new sample.

2.2. Sample Filtering

Pre-Training. Initially, we employ the original training set $\mathcal{P} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ to perform pre-training of the XMC model f. $f(\mathbf{x}, l) = \mathbf{w}_l^\top \phi(\mathbf{x})$, where $\mathbf{w}_l$ is the trainable classifier parameter for the label l and ϕ is the text encoder that maps $\mathbf{x}$ to a feature vector. Our backbone is the same as LightXML [15], which relied on using the multi-layer features of the transformer model as the text representation. We employed binary cross entropy (BCE) L_{BCE} as the loss function. Subsequently, our filtering approach is reliant on the feedback signal derived from the pre-trained f_0.

Diversity. Intuitively, if there is significant semantic difference between an augmented sample $\mathbf{x}_i^j$ and its original corresponding sample $\mathbf{x}_i$, the augmented sample can be considered to have high diversity. Therefore, we utilize KL divergence $D_{\text{KL}}(p(f_0(\mathbf{x}_i^j))||p(\mathbf{y}_i))$ to measure the difference between the two distributions $p(f_0(\mathbf{x}_i^j))$ and $p(\mathbf{y}_i)$, that indicate the diversity of the augmented samples. Consequently, the diversity score $S_{\text{Div}}^{i,j}$ of the augmented sample $\mathbf{x}_i^j$ is defined as:

$$S_{\text{Div}}^{i,j} = D_{\text{KL}}(p(\sigma(\mathbf{W}_0^\top \phi_0(\mathbf{x}_i^j)))||p(\mathbf{y}_i)). \tag{2}$$

Consistency. Low-quality augmented samples often lead to label drift [16, 9], i.e., the predicted labels of augmented samples are inconsistent with those of the original samples. Therefore, we ensure the consistency of labels by filtering out these inconsistent augmented samples. In the multi-label setting, we consider the top-k_i labels of sample $\mathbf{x}_i$, where k_i is the real number of labels, i.e., $k_i = \text{sum}(\mathbf{y}_i)$. And $|\text{Top}_{k_i}(f_0(\mathbf{x}_i^j)) \bigcap \text{Top}_{k_i}(f_0(\mathbf{x}_i))| \leq k_i$, where $\text{Top}_{k_i}(\cdot)$ denotes returns top-k_i indices based on the scores returned by f_0. Our objective is to ensure a close alignment between the top-k_i labels predicted by the augmented sample $\mathbf{x}_i^j$ and those assigned to the original sample $\mathbf{x}_i$. As a consequence, we define the consistency score $S_{\text{Con}}^{i,j}$ of augmented sample $\mathbf{x}_i^j$ as:

$$S_{\text{Con}}^{i,j} = \frac{1}{k_i}|\text{Top}_{k_i}(f_0(\mathbf{x}_i^j)) \bigcap \text{Top}_{k_i}(f_0(\mathbf{x}_i))|. \tag{3}$$

By jointly considering $S_{\text{Div}}^{i,j}$ and $S_{\text{Con}}^{i,j}$, high-quality samples with balanced diversity and consistency can be found.

2.3. Masked Data Augmentation

After performing the aforementioned DA steps, if we directly employ the augmented data for training the XMC model, it will inevitably exacerbate the long-tail problem, as illustrated in Figure 1. Hence, we introduce a novel masked data augmentation (MDA) approach focuses on *pair-level* DA by masking the augmented sample-label pairs associated with the head-labels during the training process. First, we divide the labels to head-labels and tail-labels according to the hyper-parameter, head-to-tail threshold $N_t \in \mathcal{R}^+$. label l is a tail-label if $n^l < N_t$. After that, we obtain a head-label set $\mathcal{H}$ and a tail-label set $\mathcal{T}$. Subsequently, we incorporate the mask indicator $\mathbf{m} = \{m_l\}_{l=1}^L$ into the loss function of the augmented samples to facilitate the implementation of MDA:

$$\begin{aligned} L_{\text{Aug}} = -\sum_{l=1}^{L} & m_l y_l \log(\sigma(f_0(\mathbf{x}, l))) \\ & +(1-y_l)\log(1-\sigma(f_0(\mathbf{x}, l))). \end{aligned} \tag{4}$$

where, m_l is the mask indicator for the label l and $\mathcal{H}$ is the head-label set. $m_l = 0$ if $l \in \mathcal{H}$, and $m_l = 1$ if $l \in \mathcal{T}$. Specifically, during the training process, we keep the encoder $\phi_0(\cdot)$ frozen and solely fine-tune the classifier parameters $\mathbf{W}$. By adding L_{Aug} to L_{BCE}, we obtain a more balanced loss for the classifiers. In summary, MDA exclusively augments positive sample-label pairs for the tail-labels, thereby addressing the label sparsity issue without exacerbating the long-tailed problem.

Datasets	N_{trn}	N_{tst}	L	N_{avg}
Eurlex-4K	15,539	3,809	3,956	20.79
Wiki10-31K	14,146	6,616	30,938	8.52
AmazonCat-13K	1,186,239	306,782	13,330	448.57

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. N_{avg} is the average number of documents per label.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets and Evaluation Metric. We evaluate XDA on 3 public XMC benchmark datasets: Eurlex-4K, Wiki10-31K, AmazonCat-13K. Table 1 contains the statistics of these three datasets. Following the settings of previous works [17, 15], We chose precision at k (P@k) and propensity scored precision at k (PSP@k) [18] as our evaluation metrics for performance comparison. PSP@k is used to evaluate the performance on tail-labels.

Implementation Details. Our model was trained by AdamW [19], and we also used stochastic weight averaging (SWA) [15] with a constant learning rate to avoid overfitting. To optimize GPU memory usage and enhance training, we employ automatic mixed precision (AMP). As for the key hyper-parameters of our proposed method: head-to-tail threshold N_t and times of augmentation N_a, we set $N_t = 50, N_a = 2$ for Eurlex-4K. For Wiki10-31K and AmazonCat-13K, we set $N_t = 10, N_a = 4$ and $N_t = 700, N_a = 2$ respectively.

Baselines. We compare XDA with state-of-the-art (SOTA) XMC methods including the one-versus-all DiSMEC [20]; instance tree based PfastreXML [18]; label tree based Parabel [21], Bonsai [22]; RNN-based AttentionXML [17] and Transformer-based X-Transformer [4], XLNet-APLC [23], LightXML [15], DEPL [5] methods.

3.2. Experimental Results

Main Results. The comparisons of P@k and PSP@k are shown in Table 2. The proposed XDA achieves new SOTA results in **11 out of 12** evaluation columns. (1) Addressing the long-tailed problem significantly enhances the performance of XMC tasks. As the most competitive baseline, DEPL+c also alleviates the data sparsity issue of tail-labels by matching the semantics between documents and augmented label descriptions. Our proposed method, XDA, takes an explicit data augmentation approach, directly mitigating the long-tailed problem in XMC tasks from a distributional perspective, resulting in better overall performance. (2) XDA has remarkable improvement compared to the baselines in tail-labels classification. Since XDA relies on the high-quality augmented data specifically for tail-labels, it is less affected by the dominating training

	Eurlex-4K				Wiki10-31K				AmazonCat-13K			
Methods	P@1	P@5	PSP@1	PSP@5	P@1	P@5	PSP@1	PSP@5	P@1	P@5	PSP@1	PSP@5
DisMEC	83.21	58.73	-	-	84.13	65.94	-	-	93.81	64.06	-	-
PfastreXML	73.14	50.54	-	-	83.57	59.10	-	-	91.75	63.68	-	-
Parabel	82.12	57.89	-	-	84.19	63.37	-	-	93.02	64.51	-	-
Bonsai	82.30	58.35	-	-	84.52	64.69	-	-	92.98	64.46	-	-
AttentionXML	85.12	61.01	44.20	53.87	86.46	67.98	14.49	16.54	95.53	67.0	53.94	76.43
X-Transformer	85.46	60.79	37.85	51.81	87.12	66.69	13.52	15.63	95.75	67.22	51.42	75.57
XLNet-APLC	86.83	61.94	42.21	52.88	88.99	69.79	14.43	16.47	94.56	64.59	52.55	71.36
LightXML	86.12	61.67	40.54	50.50	87.39	68.21	14.09	15.52	94.61	64.45	50.70	70.13
DEPL	85.38	59.91	45.60	53.52	84.54	64.75	16.30	16.27	94.86	64.55	55.94	76.87
DEPL+c	86.43	62.19	44.60	54.64	87.32	67.39	14.90	16.20	96.16	**67.65**	55.21	75.94
XDA	**87.14**	**63.44**	**46.43**	**56.97**	**89.75**	**70.28**	**16.95**	**17.51**	**96.67**	67.40	**56.36**	**77.11**

Table 2: Comparison of XDA to state-of-the-art methods. The bold phase and underscore highlight the best and second best model performance.

instances. Consequently, the PSP@k scores achieved by XDA surpassed those of the SOTA models. An interesting point is that our proposed method, XDA, achieved larger performance improvements on two datasets with relatively limited resources (where N_{avg} is low), Eurlex-4K and Wiki10-31K. The reason behind this is the presence of a greater number of tail-labels in these datasets, making their long-tailed problem more severe compared to AmazonCat-13K. Our method increases the number of samples for tail-labels and thus alleviates the long-tailed distribution in the data.

	Eurlex-4K		Wiki10-31K	
Method	P@5	PSP@5	P@5	PSP@5
Baseline	61.67	50.5	68.21	15.52
XDA (w/o *Prompt*)	62.59	53.61	68.85	16.96
XDA (w/o *Filter*)	62.84	54.96	68.67	16.6
XDA (w/o *Mask*)	61.55	49.74	67.33	15.27
XDA	63.44	56.97	70.28	17.51

Table 3: Ablation test of XDA on two datasets.

Ablation Study. We analyze the impacts of key components in XDA via ablation test. The complete XDA is compared with the following variants: XDA (w/o *Prompt*), XDA removes prompt-based learning and employs vanilla T5-MLM for DA; XDA (w/o *Filter*), XDA removes sample filtering; XDA (w/o *Mask*), XDA removes the masked data augmentation (MDA) strategy and directly utilizes augmented data for training. Table 3 shows the results evaluated on Eurlex-4K and Wiki10-31K datasets in terms of P@5 and PSP@5. (1) Mask training is crucial for the XDA framework. The performance of XDA (w/o *Mask*) is even worse than the baseline, indicating the significant role of MDA. This is because, if DA is performed directly without considering the issue of imbalance in the XMC task, augmentation even exacerbates the long-tail issue. This leads to overfitting on the head-label samples, resulting in a significant decrease in model performance. (2) Prompt-based learning also notably enhances the performance of the XDA framework. Soft prompt assists XDA in producing domain-specific synthetic data with a certain level of diversity. (3) Filtering also plays a crucial role in XDA's performance. Without removing low-quality synthetic data, there is a decrease in the performance gap.

4. CONCLUSION

In this study, we propose XDA, a prompt-based data augmentation framework specifically designed for extreme multi-label text classification (XMC). To address the issue of low-quality augmented samples, XDA leverages a soft prompt during the fine-tuning process of T5, enabling the augmentation of samples while preserving label-compatibility. Additionally, XDA incorporates sample filtering based on the diversity of text and the consistency of labels for the augmented samples. By applying a *pair-level* data augmentation approach to the tail-labels, XDA mitigates the long-tail problem in XMC tasks. Experimental results on three benchmark datasets demonstrate the effectiveness of our proposed XDA method, particularly in improving the performance on tail-labels.

5. ACKNOWLEDGEMENTS

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).

6. REFERENCES

[1] Wei-Cheng Chang, Daniel Jiang, Hsiang-Fu Yu, Choon Hui Teo, Jiong Zhang, Kai Zhong, Kedarnath Kolluri, Qie Hu, Nikhil Shandilya, Vyacheslav Ievgrafov, et al., "Extreme multi-label learning for semantic matching in product search," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2643–2651.

[2] Pengyu Xu, Mingxuan Xia, Lin Xiao, Huafeng Liu, Bing Liu, Liping Jing, and Jian Yu, "Textual tag recommendation with multi-tag topical attention," *Neurocomputing*, vol. 537, pp. 73–84, 2023.

[3] Lin Xiao, Pengyu Xu, Mingyang Song, Huafeng Liu, Liping Jing, and Xiangliang Zhang, "Triple alliance prototype orthotist network for long-tailed multi-label text classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[4] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon, "Taming pretrained transformers for extreme multi-label text classification," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 3163–3171.

[5] Ruohong Zhang, Yau-shian Wang, Yiming Yang, Donghan Yu, Tom Vu, and Likun Lei, "Long-tailed extreme multi-label text classification by the retrieval of generated pseudo label descriptions," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 1062–1076.

[6] Jiaxin Zhang, Jie Liu, Shaowei Chen, Shaoxin Lin, Bingquan Wang, and Shanpeng Wang, "Adam: An attentional data augmentation method for extreme multi-label text classification," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2022, pp. 131–142.

[7] Pengyu Xu, Lin Xiao, Bing Liu, Sijin Lu, Liping Jing, and Jian Yu, "Label-specific feature augmentation for long-tailed multi-label text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 10602–10610.

[8] Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling, "Do not have enough data? deep learning to the rescue!," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 7383–7390.

[9] Jing Zhou, Yanan Zheng, Jie Tang, Li Jian, and Zhilin Yang, "Flipda: Effective and robust data augmentation for few-shot learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8646–8665.

[10] Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang, "Promda: Prompt-based data augmentation for low-resource nlu tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 4242–4255.

[11] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," in *European Conference on Computer Vision*, 2020, pp. 162–178.

[12] Minyi Zhao, Lu Zhang, Yi Xu, Jiandong Ding, Jihong Guan, and Shuigeng Zhou, "Epida: An easy plug-in data augmentation framework for high performance text classification," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 4742–4752.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[14] Brian Lester, Rami Al-Rfou, and Noah Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3045–3059.

[15] Ting Jiang, Deqing Wang, Leilei Sun, Huayi Yang, Zhengyang Zhao, and Fuzhen Zhuang, "Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 7987–7994.

[16] Ehsan Kamalloo, Mehdi Rezagholizadeh, and Ali Ghodsi, "When chosen wisely, more data is what you need: A universal sample-efficient strategy for data augmentation," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1048–1062.

[17] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu, "Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[18] Himanshu Jain, Yashoteja Prabhu, and Manik Varma, "Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 935–944.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Rohit Babbar and Bernhard Schölkopf, "Dismec: Distributed sparse machines for extreme multi-label classification," in *Proceedings of the tenth ACM international conference on web search and data mining*, 2017, pp. 721–729.

[21] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma, "Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 993–1002.

[22] Sujay Khandagale, Han Xiao, and Rohit Babbar, "Bonsai: diverse and shallow trees for extreme multi-label classification," *Machine Learning*, vol. 109, pp. 2099–2119, 2020.

[23] Hui Ye, Zhiyu Chen, Da-Han Wang, and Brian Davison, "Pretrained generalized autoregressive model with adaptive probabilistic label clusters for extreme multi-label text classification," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10809–10819.