Towards Better Hierarchical Text Classification with Data Generation

Yue Wang[1,2*], **Dan Qiao**[1*], **Juntao Li**[1†], **Jinxiong Chang**[2],
Qishen Zhang[2], **Zhongyi Liu**[2], **Guannan Zhang**[2], **Min Zhang**[1]
[1]School of Computer Science and Technology, Soochow University
[2]Ant Group
ywangnlp@stu.suda.edu.cn

Abstract

Hierarchical text classification (HTC) focuses on classifying one text into multiple labels, which are organized as a hierarchical taxonomy. Due to its wide involution in realistic scenarios, HTC attracts long-term attention from both industry and academia. However, the high cost of hierarchical multi-label annotation makes HTC suffer from the data scarcity problem. In view of the difficulty in balancing the controllability of multiple structural labels and text diversity, automatically generating high-quality data for HTC is challenging and under-explored. To fill this blank, we propose a novel data generation framework tailored for HTC, which can achieve both label controllability and text diversity by extracting high-quality semantic-level and phrase-level hierarchical label information. Experimental results on three benchmarks demonstrate that, compared with existing data augmentation methods, the data generated from our method can bring the most significant performance improvements of several strong HTC models. Extensive analysis confirms that the improvements yielded by our proposed method do correlate to the enhancement of label controllability and text diversity. [1]

1 Introduction

Hierarchical Text Classification (HTC) is a representative multi-label text classification problem, aiming to assign one text with multiple labels in a given label hierarchical taxonomy. HTC is widely involved in realistic scenarios, e.g., news classification (Lewis et al., 2004; Sandhaus, 2008), science paper classification (Kowsari et al., 2017), E-commerce (Gao, 2020). To solve such an important and challenging task, adequate high-quality labeled data is indispensable. However, multi-labeled data annotation is usually expensive, and the hierarchical structure of multi-label further makes such annotation unaffordable. Thus, the primary goal of our approach is to deal with the data scarcity problem. Existing works (Kowsari et al., 2017; Shimura et al., 2018; Banerjee et al., 2019; Zhou et al., 2020; Wang et al., 2022c,d; Jiang et al., 2022) focus on enhancing the model ability to relieve the need for annotated data, but few of them address this problem by automatically generating high-quality data.

Recently, since generative Pre-trained Language Models (PLMs) have achieved surprising performance (Brown et al., 2020; Lewis et al., 2020; Raffel et al., 2020), generative data augmentation has drawn more and more attention (Anaby-Tavor et al., 2020; Schick and Schütze, 2021; Wang et al., 2022a; Meng et al., 2022). With the help of the rich knowledge obtained from the large-scale data in the pre-training stage, generative data augmentation can further improve the quality of the generated data, which can better alleviate the data scarcity problem. However, generating high-quality data for HTC is still an under-explored problem. We argue that two main challenges exist for this problem: the need for label controllability and text diversity. If the meaning of the generated data is out of the control of the given labels, it is usually noisy data assigned with wrong labels, which does little help and even harms HTC models. Besides, the generated data may not improve the generalization ability of HTC models if the expression is similar to the original text. Intuitively, restricted label constraints can ensure label controllability but might impede the text diversity of the generated data, especially in HTC, which has multiple label constraints in a hierarchical structure (Kumar et al., 2021). Therefore, achieving a good balance between controllability and diversity is the key to generating high-quality data for HTC.

To improve label controllability and text diversity of generated data, we propose a novel data

[*] Equal contribution. Work is done during the internship of Yue Wang at Ant Group.

[†] Corresponding author

[1] Our codes are publicly available at https://github.com/wangyuenlp/Data-Generation-for-HTC.

Findings of the Association for Computational Linguistics: ACL 2023, pages 7722–7739
July 9-14, 2023 ©2023 Association for Computational Linguistics

generation framework for HTC, which aims to enrich the input information of the data generation model. Specifically, we first design the semantic-level and phrase-level label information enhanced prompt, consisting of label names and phrases extracted from origin training samples. Although label names can improve label controllability, to a certain extent, only leveraging label names makes the model inevitably generate similar data since the same input label name needs to create hundreds and even thousands of data samples. Thus, the phrases extracted from original train samples can be used as a supplement to label names. To ensure label controllability and text diversity, we need the phrases to represent the critical information that can infer the origin label correctly and simultaneously avoid common expressions. Therefore, to deal with these problems, we propose a hierarchical label information enhanced keyword extractor to extract the most relevant phrase to the label meaning. Finally, to further improve the label controllability, we introduce consistency filtering to filter out the low-quality generated data, which makes the HTC model predict different labels compared to the controlled labels. Experimental results on three benchmarks confirm the effectiveness of our proposed method. Furthermore, both quantitative and qualitative analyses show the superiority of our method over existing data augmentation methods in comparing label controllability and text diversity.

In a nutshell, our contributions are as follows:

- To the best of our knowledge, we are the first to explore the effectiveness of generative data augmentation methods on the HTC problem.
- We propose a data generation framework for HTC, which leverages both semantic-level and phrase-level hierarchical label information to enhance the label controllability and text diversity of the generated data.
- We confirm the effectiveness of our methods over several HTC models on three benchmarks, which can bring 1.39, 1.37, and 1.25 Macro-F1 improvements of BERT-base, RoBERTa-base, and state-of-the-art HTC model HGCLG on WOS, respectively.

2 Related Work

Hierarchical Text Classification Compared to multi-label text classification, due to the unique label hierarchical taxonomy, the previous work on HTC focus on fully using the hierarchical information of the label taxonomy. Kowsari et al. (2017); Shimura et al. (2018); Wehrmann et al. (2018); Banerjee et al. (2019) train different classifiers for different nodes or levels and transfer the knowledge of the parent nodes' classifiers to the child nodes, which are called local approaches (Zhou et al., 2020). Different from local approaches, global approaches treat HTC as a flat multi-label text classification problem. Due to the lack of hierarchical information in the classifier, the global approaches focus on utilizing hierarchical information to further improve performance. Mao et al. (2019) use hierarchical information to enhance the Label Assignment Policy and propose a deep reinforcement learning-based general framework for HTC; Wu et al. (2019) utilize a meta-learner to deal with the complexity and dependencies of different labels; Aly et al. (2019) introduce capsule networks for HTC and utilize label co-occurrence to initialize weight better; Deng et al. (2021) propose text-label mutual information maximization and label prior matching to filter out irrelevant information and learn better hierarchy-aware representations; Rojas et al. (2020) define HTC as a sequence-to-sequence problem and utilize an auxiliary synthetic task and external knowledge to improve performance further. Recently, there are more and more focus on leveraging a structure encoder to model the hierarchical information (Zhou et al., 2020; Chen et al., 2021; Wang et al., 2021a, 2022c,d; Jiang et al., 2022). Different from previous work, we focus on introducing generative data augmentation to address the data scarcity problem of HTC and use a structure encoder to model hierarchical label information to improve the quality of the generated data.

Generative Data Augmentation With the development of text generation models, generative data augmentation gains more and more attention from the community. Existing works apply generative data augmentation methods to various NLP downstream tasks, including multi-class text classification (Malandrakis et al., 2019; Liu et al., 2020; Anaby-Tavor et al., 2020), multi-label text classification (Zhang et al., 2020), text entailment (Vu et al., 2021; Wang et al., 2022b), relation extraction (Lee et al., 2021), sequence labeling (Wang et al., 2022a), intent classification and slot tagging (Lee et al., 2021; Rosenbaum et al., 2022), etc. Besides, there is also a line of work fo-

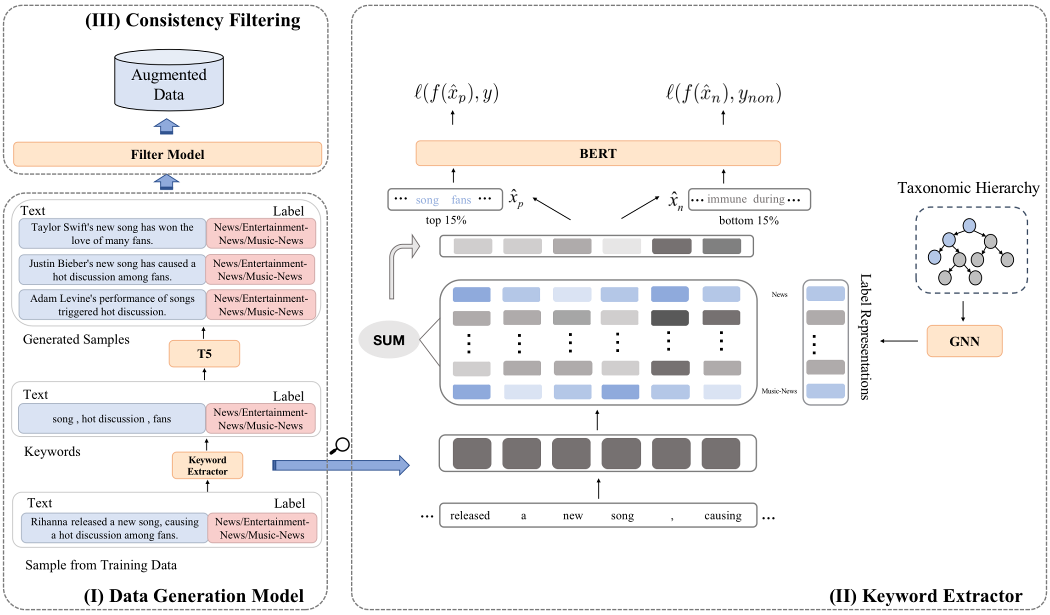
(III) Consistency Filtering
Augmented Data
Filter Model
Text
Label
Taylor Swift's new song has won the love of many fans.
News/Entertainment-News/Music-News
Justin Bieber's new song has caused a hot discussion among fans.
News/Entertainment-News/Music-News
Adam Levine's performance of songs triggered hot discussion.
News/Entertainment-News/Music-News
Generated Samples
T5
Text
Label
song , hot discussion , fans
News/Entertainment-News/Music-News
Keywords
Keyword Extractor
Text
Label
Rihanna released a new song, causing a hot discussion among fans.
News/Entertainment-News/Music-News
Sample from Training Data
(I) Data Generation Model
$\ell(f(\hat{x}_p), y)$
$\ell(f(\hat{x}_n), y_{non})$
BERT
… song fans …
$\hat{x}_p$
top 15%
$\hat{x}_n$
… immune during …
bottom 15%
Taxonomic Hierarchy
SUM
News
Music-News
Label Representations
GNN
… released a new song , causing …
(II) Keyword Extractor

Figure 1: An illustration of our data generation framework for HTC, consisting of **(I).** Data Generation Model, **(II).** Keyword Extractor, and **(III).** Consistency Filtering. For **(I)**, we use the label names and keywords extracted from the training sample to prompt the T5 model to generate new data. For **(II)**, the keyword extractor needs to ensure the HTC model can infer the original label y from the extracted keywords while predicting all-zero label y_{non} when masking the keywords from the original texts. For **(III)**, the filter model removes low-quality generated data whose texts are inconsistent with the meaning of assigned labels.

cused on generating new data without fine-tuning PLMs to solve various tasks on the zero-shot setting, also called dataset generation (Schick and Schütze, 2021; Wang et al., 2021b; Meng et al., 2022; Ye et al., 2022). However, the effectiveness of generative data augmentation methods on the HTC problem is under-explored. Apart from the difference in applied tasks, existing works also differ in model architecture, objective functions, and learning paradigms. Specifically, Malandrakis et al. (2019) propose a conditional variational autoencoders-based controlled text generation model for data augmentation; Liu et al. (2020) utilize reinforcement learning to guide the conditional generation; Vu et al. (2021) leverage data generation model to generate unlabeled synthetic data and conduct self-training; Lee et al. (2021) propose an example extrapolator to generate new labeled synthetic data from existing examples. Wang et al. (2022a) propose a soft prompt-based data generation model for low-resource scenarios. Wang et al. (2022b) unify diverse NLP tasks into a text-to-text format to pre-train a multi-task data generation PLM. In this work, we focus on using generative data augmentation to bring benefits to HTC models in the full-supervised setting and improve the quality of generated data by incorporating hierarchical label information.

3 Method

In this section, we first give the task formulation for HTC and the goal of data generation. Then, we introduce our framework consisting of the semantic-level and phrase-level label information enhanced prompt, hierarchical label information enhanced keyword extractor and consistency filtering.

3.1 Task Formulation

To the HTC task, given the sample (x, y) from the training set D_{Train}, the HTC model M_{HTC} is trained to predict y according to the text x. Here y is a subset of label set Y, which is organized as a directed acyclic graph $G = (Y, E)$, where the edge set E denotes the hierarchical information of Y. The data generation model M_{GEN} aims to make the final HTC model M_{AUG} perform better than M_{HTC}. To achieve this goal, M_{GEN} can utilize D_{Train} to generate D_{GEN}, whose number is usually larger than D_{Train}. The training data of M_{AUG} consists of both D_{Train} and D_{GEN}.

3.2 Overview of Our Framework

Our framework consists of three models: the data generation model (Sec. 3.3), the keyword extractor (Sec. 3.4), and the filter model (Sec. 3.5). We show an illustration of our framework in Figure 1. In general, we first use D_{Train} to train the keyword extractor M_{KE} and the filter model M_{Filter}. Then we use M_{KE} to extract keywords $\hat{x}$ for each instance in D_{Train}. Afterward, to train the data generation model M_{GEN}, we organize the label names and extracted keywords $\hat{x}$ of each instance in D_{Train} as the input sequence and the original text as the target sequence. In the inference stage of M_{GEN}, we utilize the sampling-based decode strategy to generate different texts according to the label names and extracted keywords $\hat{x}$ of instances in D_{Train}. Finally, we use the filter model M_{Filter} to remove low-quality generated data. In the following, we will describe our framework in detail.

3.3 Semantic-Level and Phrase-Level Label Information Enhanced Prompt

To improve label controllability and text diversity, we propose the semantic-level and phrase-level label information enhanced prompt consisting of label names and keywords extracted from the samples of D_{Train}. Both label names and keywords can improve label controllability by providing important label information to M_{GEN} from the semantic and phrase levels. Besides, due to the diversity of phrase expressions under the same semantics, providing keywords extracted from different samples with the same label can also help to generate diverse data. Specifically, given the sample (x, y), we organize the input of the M_{GEN} as *'generate with label: y; generate with keywords: $\hat{x}$'* and the target output as x, where $\hat{x}$ refers the keywords extracted from the x. In the inference stage, we follow the former prompt format to generate candidate data D_{CAND}, whose keywords are also extracted from samples of D_{Train}. We shuffle the order of keywords at both the training and inference stage to avoid the over-fitting problem.

3.4 Hierarchical Label Information Enhanced Keyword Extractor

The quality of keywords is important to balance label controllability and text diversity. Biased phrases irrelevant to the label meaning may mislead the data generation model and harm the label controllability. Excessively unimportant phrases unrelated to the labels of samples may control M_{GEN} to generate stern expression, which may do harm to diversity. To improve the quality of extracted keywords, we propose the hierarchical label information enhanced keyword extractor. For the sample (x, y), we want to extract sub-sequence $\hat{x}$, where $\hat{x}$ consists of the most relevant keywords for the corresponding multi-label y. Specifically, we first use Graphormer (Ying et al., 2021) to model the hierarchical information $G = (Y, E)$ and get the label embedding l_j for each node $j \in Y$. Formally:

$$\{l_1, l_2, \dots, l_n\} = Graphormer(Y, E).$$

Next, for the given input x, we get the text embedding of each token x_i by:

$$\{t_1, t_2, \dots, t_n\} = BERT_{emb}(x),$$

where $BERT_{emb}(x)$ denotes using BERT (Devlin et al., 2019) to encode the given sentence x.

We then utilize the attention mechanism to catch the relevance between token x_i and label y_j as:

$$Q_i = t_i W_Q, K_j = l_j W_K, A_{ij} = \frac{Q_i K_j^T}{\sqrt{d_h}},$$

where $W_Q \in \mathbb{R}^{d_h \times d_h}$ and $W_Q \in \mathbb{R}^{d_h \times d_h}$ are two weight matrices. Afterward, we use Gumbel-Softmax (Jang et al., 2016) to calculate the probability that x_i is the keyword of class y_i by:

$$P_{ij} = \text{gumbel\_softmax}\left(A_{i1}, A_{i2}, \dots, A_{ik}\right)_j,$$

which satisfies $\sum_{j \in Y} P_{ij} = 1$.

Therefore, the relevance score between token x_i and a multi-label y can be calculated as:

$$P_i = \sum_{j \in y} P_{ij}.$$

Intuitively, if $\hat{x}$ contains the most relevant information, a classifier can infer the origin label correctly from $\hat{x}$ and predict null labels from the rest of the sequence. This intuition guides us to design the final loss function $\mathcal{L}$. Specifically, with the calculated P_i for each token x_i, we first select the sequence of the most relevant keywords $\hat{x}_p$ and the sequence of the least relevant keywords $\hat{x}_n$. $\hat{x}_p$ contains the top 15% tokens with the highest P_i while $\hat{x}_n$ contains 15% token with the lowest P_i. For better training, we generate $\hat{x}_p$ and $\hat{x}_n$ from x by masking the rest of the tokens in practice. Finally, we encourage the classifier f to assign $\hat{x}_n$

with the original multi-label y and $\hat{x}_n$ with all-zero label y_{non} by the final loss function:

$$\mathcal{L} = \ell(f(\hat{x}_p), y) + \ell(f(\hat{x}_n), y_{non}),$$

where ℓ denotes the binary cross entropy loss for multi-label classification. We jointly update the parameter of the graph network $Graphormer$ and the text encoder $BERT_{emb}$. After training, for each sentence x, we directly use $\hat{x}_p$ as the extracted keywords $\hat{x}$ for each sample at the inference stage.

3.5 Consistency Filtering

Although the above methods improve the label controllability as far as possible, they may still generate low-quality samples whose meaning is contradicted the assigned label. This phenomenon is inevitable due to the pursuit of text diversity. To filter out these samples, we introduce consistency filtering (Anaby-Tavor et al., 2020). Specifically, we first use the training set D_{Train} to train an HTC model M_{Filter}. Then, we use M_{Filter} to predict the labels of D_{CAND} and compare the predicted labels with their assigned labels. Only the generated samples whose predicted labels by M_{Filter} are consistent with assigned labels will be kept, constituting D_{GEN}. We combine D_{GEN} with D_{Train} to train the target HTC model M_{Aug} from scratch.

4 Experiments

4.1 Datasets and Metrics

We conduct experiments on three HTC benchmarks: Web-of-Science (WOS) (Kowsari et al., 2017), The New York Times Annotated Corpus (NYT) (Sandhaus, 2008), and Reuters Corpus Volume I Version 2 (RCV1-V2) (Lewis et al., 2004). WOS is a science paper classification dataset, while NYT and RCV1-V2 are news classification datasets. We show the statistic of three benchmarks in Table 1. We follow Zhou et al. (2020); Chen et al. (2021); Wang et al. (2022c) to pre-process the data and report the Macro-F1 and Micro-F1 as the metrics of HTC models.

4.2 Baselines

To confirm the effectiveness of our methods, we compare their improvements with different data augmentation methods on multiple representative HTC models. Our baselines can be grouped into two types: HTC and data augmentation methods.

Dataset	Train	Dev	Test	Classes	Avg(y_i)	D
WOS	30,070	7,518	9,397	141	2.0	2
NYT	23,345	5,834	7,292	166	7.6	8
RCV1-V2	20,833	2,316	781,265	103	2.2	4

Table 1: The statistic of three HTC benchmarks. **Train, Dev, Test** refer to the sample numbers of training, development and test set, respectively. **Classes** refers to the number of classes in each dataset. **Avg(y_i)** represents the average number of classes per sample. **D** represents the maximum number of the hierarchical label level.

HTC. We choose three baselines: **(1) BERT** (Devlin et al., 2019), **(2) RoBERTa** (Liu et al., 2019), and **(3) HGCLR** (Wang et al., 2022c). BERT and RoBERTa are popular PLMs, which are widely used as text encoders. HGCLR is a state-of-the-art HTC model which utilizes contrastive learning to incorporate hierarchical label information.

Data Augmentation. We select five strong data augmentation methods for comparisons: **(1) EDA** (Wei and Zou, 2019) is a rule-based data augmentation method that uses four simple editing operations to disturb texts; **(2) Back Translation (BT)** (Sennrich et al., 2016) first translate the texts into other languages and then translate them to the original language; **(3) LAMBADA** (Anaby-Tavor et al., 2020) is a generative data generation method that generates new data according to label names; **(4) GDA** (Zhang et al., 2020) utilize generative PLMs to generate label-invariant perturbations on the texts; **(5) PromDA** (Wang et al., 2022a) is a state-of-the-art generative data augmentation method that uses an efficient fine-tuning technique to train the data generation model.

4.3 Implementation Details

We implement all the data generation and HTC models with the open-sourced toolkit *Transformers*[2] (Wolf et al., 2020). For data generation models, we take the T5 (Raffel et al., 2020) as our backbone PLM and conduct fine-tuning from the publicly available checkpoint *t5-large* [3]. In the fine-tuning stage, we use Adafactor (Shazeer and Stern, 2018) optimizer to update all parameters of T5 and set the learning rate as 1e-3 and batch size as 32. We fine-tune data generation models for 10 k steps, evaluate the perplexity on the development set every 1 k steps, and save the model with the lowest perplexity. In the inference stage,

[2] https://github.com/huggingface/transformers

[3] https://huggingface.co/t5-large

Method	WOS		NYT		RCV1-V2	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
BERT-base (Devlin et al., 2019)	79.87 ± 0.41	85.96 ± 0.23	65.69 ± 1.04	78.20 ± 0.21	66.85 ± 0.80	85.86 ± 0.14
+EDA (Wei and Zou, 2019)	79.82 ± 0.32	85.87 ± 0.29	65.16 ± 0.50	77.17 ± 0.45	67.02 ± 0.12	85.47 ± 0.32
+BT (Sennrich et al., 2016)	80.09 ± 0.52	86.02 ± 0.23	65.43 ± 0.39	77.46 ± 0.08	66.41 ± 0.54	85.38 ± 0.26
+LAMBADA (Anaby-Tavor et al., 2020)	80.38 ± 0.33	86.28 ± 0.24	66.69 ± 0.26	78.18 ± 0.44	66.35 ± 0.57	86.00 ± 0.03
+GDA (Zhang et al., 2020)	80.35 ± 0.37	86.26 ± 0.27	66.20 ± 0.71	78.04 ± 0.40	66.30 ± 0.77	85.59 ± 0.26
+PromDA (Wang et al., 2022a)	80.05 ± 0.29	85.95 ± 0.33	65.94 ± 0.36	77.92 ±0.08	66.60 ± 0.38	85.34 ± 0.35
+Ours	81.26 ± 0.23$^{\uparrow}$	86.77 ± 0.17$^{\uparrow}$	66.85 ± 0.43$^{\uparrow}$	78.34 ± 0.46$^{\uparrow}$	67.41 ± 0.49$^{\uparrow}$	86.06 ± 0.37 $^{\uparrow}$
RoBERTa-base (Liu et al., 2019)	79.94 ± 0.82	86.24 ± 0.45	67.77 ± 0.38	79.55 ± 0.27	**68.63 ± 0.53**	86.81 ± 0.14
+EDA (Wei and Zou, 2019)	80.37 ± 0.58	86.32 ± 0.31	66.50 ± 0.38	78.48 ± 0.39	66.71 ± 1.44	86.10 ± 0.13
+BT (Sennrich et al., 2016)	80.30 ± 0.26	86.38 ± 0.10	67.18 ± 0.67	78.90 ± 0.24	66.67 ± 0.82	86.47 ± 0.28
+LAMBADA (Anaby-Tavor et al., 2020)	80.54 ± 0.35	86.49 ± 0.18	67.75 ± 0.32	79.41 ± 0.38	66.84 ± 0.45	86.47 ± 0.42
+GDA (Zhang et al., 2020)	80.42 ± 0.42	86.57 ± 0.23	67.54 ± 0.36	79.43 ± 0.23	68.03 ± 0.31	86.50 ± 0.10
+PromDA (Wang et al., 2022a)	80.15 ± 0.61	86.04 ± 0.31	67.32 ± 0.38	79.01 ± 0.13	67.43 ± 0.71	86.34 ± 0.17
+Ours	81.31 ± 0.12$^{\uparrow}$	86.96 ± 0.13$^{\uparrow}$	**68.16 ± 0.25**$^{\uparrow}$	**79.64 ± 0.17**$^{\uparrow}$	68.50 ±0.42	**86.84 ± 0.12**$^{\uparrow}$
HGCLR (Wang et al., 2022c)	80.82 ± 0.20	86.63 ± 0.27	66.90 ± 0.60	78.48 ± 0.48	66.64 ± 0.93	86.04 ± 0.12
+EDA (Wei and Zou, 2019)	80.85 ± 0.45	86.78 ± 0.44	66.08 ± 0.36	77.53 ± 0.29	66.72 ± 0.32	85.20 ± 0.19
+BT (Sennrich et al., 2016)	79.63 ± 0.61	85.89 ± 0.38	65.42 ± 0.50	77.37 ± 0.35	67.03 ± 0.57	85.72 ± 0.26
+LAMBADA (Anaby-Tavor et al., 2020)	81.25 ± 0.24	87.08 ± 0.17	66.78 ± 0.42	78.22 ± 0.31	66.55 ± 0.37	85.58 ± 0.06
+GDA (Zhang et al., 2020)	80.99 ± 0.50	86.84 ± 0.19	66.25 ± 0.37	77.99 ± 0.31	65.30 ± 0.49	85.13 ± 0.25
+PromDA (Wang et al., 2022a)	80.60 ± 0.37	86.63 ± 0.23	66.08 ± 0.44	78.02 ± 0.10	66.88 ± 0.34	85.72 ± 0.20
+Ours	**82.07 ± 0.18**$^{\uparrow}$	**87.36 ± 0.08**$^{\uparrow}$	67.69 ± 0.52 $^{\uparrow}$	79.03 ± 0.24$^{\uparrow}$	67.90 ± 0.82$^{\uparrow}$	86.25 ± 0.20$^{\uparrow}$

Table 2: The performance of our proposed methods and all baselines. We report the results based on our implementations, which are the average performance of 5 runs using different random seeds. ***Bold*** indicates the current state-of-the-art performance. ↑ indicates our framework brings improvements on the corresponding baseline. ± denotes the standard deviation.

we use Nucleus Sampling (Holtzman et al., 2019) to generate diverse data. Specifically, we set Top-p as 0.9 and get five independently sampled outputs for one input text. For all data augmentation methods, we augment data 5 times the number of samples in the original training set. For EDA, we use the official implentation [4]. For BT, we use the open-sourced English to France machine translation model *Helsinki-NLP/opus-mt-en-fr* [5] and France to English model *Helsinki-NLP/opus-mt-fr-en* [6]. For HTC models, we use *bert-base-uncased*[7] and *roberta-base*[8] as the initial checkpoints for BERT and RoBERTa respectively and completely follow Wang et al. (2022c) to set hyper-parameters. We report the average performance of HTC models using five different random seeds.

4.4 Main Results

We report the performance of all baselines and our proposed method in Table 2. From the results, we can find that EDA and BT fail to improve the performance of the HTC models significantly, which shows simple perturbations on the original texts cannot help to improve the generalization ability of HTC models in the full-supervised setting. Existing generative data augmentation methods can help some HTC models achieve better performance, but the improvements are marginal and fail to improve performance on NYT and RCV1-V2. This phenomenon shows that generating diverse data satisfied with multiple label constraints for HTC is still challenging. With the help of the data generated from our proposed method, HTC models can achieve better performance in almost all settings. The improvements are more significant than other data augmentation methods. Besides, we also find the improvements of Macro-F1 brought by generative data augmentation methods are better than the Micro-F1, which we speculate results from the long-tail labels suffering from a more severe data scarcity problem. All these results confirm the effectiveness of our proposed methods to deal with the data scarcity problem for HTC.

4.5 Model Ablation

We conduct a model ablation study and report the results in Table 3. We find that only providing label names or keywords to the data generation models harms the performance of HTC models. This result confirms our hypothesis that we need to deliver both semantic-level and phrase-level label information to enhance the ability of the data generation models. We also replace the proposed keyword extractor with yake (Campos et al., 2018a,b, 2020), a state-of-the-art unsupervised keyword extractor. The results show that without the help of the hier-

[4] https://github.com/jasonwei20/eda_nlp

[5] https://huggingface.co/Helsinki-NLP/opus-mt-en-fr

[6] https://huggingface.co/Helsinki-NLP/opus-mt-fr-en

[7] https://huggingface.co/bert-base-uncased

[8] https://huggingface.co/roberta-base

Method	Macro-F1	Micro-F1
BERT	79.87 ± 0.41	85.96 ± 0.23
Ours	**81.26 ± 0.23**	**86.77 ± 0.17**
-remove keywords	80.38 ± 0.33	86.28 ± 0.24
-remove label names	80.21 ± 0.08	86.13 ± 0.14
-replaced with yake	80.73 ± 0.49	86.59 ± 0.20
-replaced with T5-base	80.87 ± 0.12	86.58 ± 0.17

Table 3: Model ablation results. We conduct experiments based on BERT-base on the WOS dataset and report the average performance of 5 different runs. ± denotes the standard deviation.

Method	Self-BLEU	Macro-F1	Micro-F1
EDA (Wei and Zou, 2019)	0.8132	79.82 ± 0.32	85.87 ± 0.29
BT (Sennrich et al., 2016)	0.9905	80.09 ± 0.52	86.02 ± 0.23
LAMBADA (Anaby-Tavor et al., 2020)	0.7184	80.38 ± 0.33	86.28 ± 0.24
GDA (Zhang et al., 2020)	0.6726	80.35 ± 0.37	86.26 ± 0.27
PromDA (Wang et al., 2022a)	0.6915	80.05 ± 0.29	85.95 ± 0.33
Ours	**0.5987**	**81.26 ± 0.23**	**86.77 ± 0.17**

Table 4: Text diversity of different augmentation methods. The lower Self-BLEU score means higher text diversity. We also report the performance of the BERT-base on the WOS dataset with these augmented data. ± denotes the standard deviation.

Method	w/o CF		with CF	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
EDA (Wei and Zou, 2019)	79.87 ± 0.31	85.83 ± 0.27	79.82 ± 0.32	85.87 ± 0.29
BT (Sennrich et al., 2016)	**79.95 ± 0.39**	**85.96 ± 0.23**	80.09 ± 0.52	86.02 ± 0.23
LAMBADA (Anaby-Tavor et al., 2020)	78.51 ± 0.37	85.55 ± 0.29	80.38 ± 0.33	86.28 ± 0.24
GDA (Zhang et al., 2020)	78.49 ± 0.51	85.34 ± 0.33	80.35 ± 0.37	86.26 ± 0.27
PromDA (Wang et al., 2022a)	78.92 ± 0.42	85.45 ± 0.31	80.05 ± 0.29	85.95 ± 0.33
Ours	79.37 ± 0.36	85.69 ± 0.34	**81.26 ± 0.23**	**86.77 ± 0.17**

Table 5: The effect of Consistency Filtering (denoted as **CF**). We report the average performance of the BERT-base on the WOS dataset when using these augmented data and with 5 random seeds. ± denotes the standard deviation.

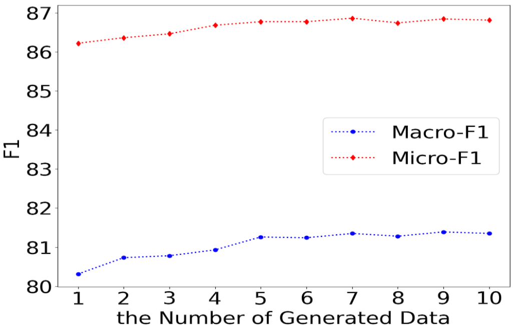

Figure 2: The effect of the number of generated data. We change the number of data generated from our proposed methods and report the performance of the BERT-base on WOS. **The number of generated data** refers to how many times we generate compared to the original training data. All results are the average of 5 runs.

archical label information, the extracted keywords may not be the words most relevant to the assigned labels, which harms the quality of the generated data. Besides, we also replace *t5-large* (770M parameters) with *t5-base* (220M parameters) [9] to study the effect of the backbone generative PLMs. With the decrease in parameters, the performance of our proposed method drops slightly. Despite the drop, it still performs better than other data generation methods, which use *t5-large* as the backbone, compared to the results reported in Table 2. Furthermore, without the help of the unified data generation prompt format, Macro-F1 drops more significantly than Micro-F1, which shows our method can transfer the knowledge from head labels to long-tail labels and thus can mitigate the sparse label distributions.

4.6 The Number of Generated Data

We also study the effect of the number of generated data. The results are shown in Figure 2. We can observe that when the number of generated data is smaller than 5 times the number of original train data, with the increase of the number of generated data, the performance of the HTC model improves. When bigger than 5 times the original training data, no steady growth trend is observed. Therefore, we generate 5 times the original training data in our experiments. Besides, we find that the change of Macro-F1 is more significant than Micro-F1. We speculate it also results from the generated data doing more help to the long-tailed data.

4.7 The Diversity of the Generated Data

We use Self-BLEU (Zhu et al., 2018) as the metric to conduct a quantitative analysis of the diversity of the generated data. Because we want the generated data to diversify as much as possible under the condition of meeting the need for label controllability, we report the Self-BLEU score of filtered data after using consistency filtering. From the results in Table 4, we can observe that the Self-BLEU score of EDA and BT is significantly higher than generative data augmentation methods, and our proposed method achieves the lowest Self-BLEU score. These results further confirm the effectiveness of our proposed method to improve text diversity while ensuring label controllability.

4.8 The Effect of Consistency Filtering

We also study the effect of consistent filtering and report the results in Table 5. EDA and BT achieve no improvements after consistency filtering, which

[9] https://huggingface.co/t5-base

Method	Performance		Total Extra Cost	Filter Model Cost	Data Generation Model Cost	Keyword Extractor Cost
	Macro-F1	Micro-F1				
BERT-base (Devlin et al., 2019)	79.87 ± 0.41	85.96 ± 0.23	-	-	-	-
EDA (Wei and Zou, 2019)	79.82 ± 0.32	85.87 ± 0.29	4 GPU hours on 80GB NVIDIA A100	4 GPU hours on 80GB NVIDIA A100	-	-
BT (Sennrich et al., 2016)	80.09 ± 0.52	86.02 ± 0.23	4 GPU hours on 80GB NVIDIA A100	4 GPU hours on 80GB NVIDIA A100	-	-
LAMBADA (Anaby-Tavor et al., 2020)	80.38 ± 0.33	86.28 ± 0.24	30 GPU hours on 80GB NVIDIA A100	4 GPU hours on 80GB NVIDIA A100	26 GPU hours on 80GB NVIDIA A100	-
GDA (Zhang et al., 2020)	80.35 ± 0.37	86.26 ± 0.27	30 GPU hours on 80GB NVIDIA A100	4 GPU hours on 80GB NVIDIA A100	26 GPU hours on 80GB NVIDIA A100	-
PromDA (Wang et al., 2022a)	80.05 ± 0.29	85.95 ± 0.33	30 GPU hours on 80GB NVIDIA A100	4 GPU hours on 80GB NVIDIA A100	26 GPU hours on 80GB NVIDIA A100	-
Ours	**81.26 ± 0.23**	**86.77 ± 0.17**	34 GPU hours on 80GB NVIDIA A100	4 GPU hours on 80GB NVIDIA A100	26 GPU hours on 80GB NVIDIA A100	4 GPU hours on 80GB NVIDIA A100

Table 6: Computational cost analysis of different data augmentation methods. **Total Extra Cost** represents the total extra computational cost caused by different data augmentation methods compared to when data augmentation is not used, which is the sum of filter model cost, data generation model cost, and keyword extractor cost. To show the trade-off between performance and cost, we also report the performance of the BERT-base on the WOS dataset with these different data augmentation methods. ± denotes the standard deviation.

Original Label:	Medical, Idiopathic Pulmonary Fibrosis
Original Text:	the present study investigated the effects of diesel exhaust (de) on an experimental model of bleomycin (blm)-induced lung injury and fibrosis in mice. blm was intravenously administered to both nrf2(+/+) and nrf2(-/-) c57bl/6j mice on day 0. the mice were exposed to de for 56 days from 28 days before the blm injection to 28 days after the blm injection. inhalation of de induced significant inhibition of airway clearance function and the proinflammatory cytokine secretion in macrophages, an increase in neutrophils, and severe lung inflammatory injury, which were greater in nrf2(-/-) mice than in nrf2(+/+) mice. in contrast, inhalation of de was observed to induce a greater increase of hydroxyproline content in the lung tissues and significantly higher pulmonary antioxidant enzyme mrna expression in the nrf2(+/+) mice than in nrf2(-/-) mice. de is an important risk factor, and nrf2 regulates the risk of a de inhalation induced immune response during blm lung injury and fibrosis in mice.
EDA Aug Data:	the present study investigated the effects of diesel exhaust de on an experimental model part of bleomycin blm induced lung injury and fibrosis in mice blm was intravenously administered to both nrf and nrf c bl j diamond state mice on day the mice were exposed to de for days make from days before the blm injection to days after the blm injection inhalation of de part induced significant inhibition of airway clearance function discrimination and the proinflammatory cytokine secretion in depicted object macrophages an increase in neutrophils and severe lung inflammatory injury which clarence day were canvass greater in nrf mice than in nrf mice in contrast inhalation of de was observed to induce a greater increase of hydroxyproline content in the lung tissues and significantly higher pulmonary antioxidant enzyme mrna expression angstrom in the nrf mice secernment than division in nrf mice de is an open demarcation important risk factor growth and nrf regulates the risk of a de inhalation induced immune response during blm lung injury and fibrosis significant in mice
BT Aug Data:	The effects of diesel exhaust gas (de) on an experimental model of bleomycin (blm)- and fibrosis induced lung lesions in mice were studied. blm was administered intravenously to both nrf2(+/+) and nrf2(-) c57bl/6j mice on day 0. mice were exposed for 56 days from 28 days prior to blm injection at 28 days after blm injection. inhalation of the significantly induced inhibition of respiratory clearance function and proinflammatory cytokine secretion in macrophages, an increase in neutrophils and severe pulmonary lesions, which were higher in nrf2(–) mice than in nrf2(+/+) mice.
Yake Keywords:	present study investigated, diesel exhaust, model of bleomycin, investigated the effects,effects of diesel, experimental model, mice, blm, blm injection, lung, injury, days, present, fibrosis, lung inflammatory injury
Yake Aug Data:	the present study investigated the effects of diesel exhaust (de) on lung fibrosis using a model of blm injection. we observed that de causes lung fibrosis in mice by observing an increase in de lung fibrosis and that de cannot be seen until a week or more after the injection. compared to the model of de, de causes lung fibrosis and fibrosis with similar effect, and also reduces lung fibrosis without affecting lung function. the experimental model of de was able to detect changes in lung fibrosis in mice with de through three weeks, seven weeks and seven months. the data indicate that de causes lung fibrosis in mice mainly through increased de of lung fibrosis. it was also shown that de causes lung fibrosis in mice by increasing de lung fibrosis and that de causes a loss of lung function in mice. these results suggest that de causes lung fibrosis in mice by using de. the present study has the potential to help design treatment programs for de, thus giving consideration to de process the blm pipeline. (c) 2017 the authors. published by elsevier ltd.
Ours Keywords:	bleomycin, lung, injury, fibrosis, intravenously, inhalation, cytokine, neutrophils, inflammatory, pulmonary
Ours Aug Data:	background: respiratory disease may induce acute respiratory distress syndrome (ards) in allergic individuals. recent studies have shown that bronchopulmonary infarction is associated with ards. the purpose of this study is to examine the effect of lung injury on the innate immune system and to identify the underlying mechanisms in a patient with ards. methods: this study was conducted in two hospitals. the first was a single-centre randomized trial comparing two treatment groups: an exposure to a stw 3 tidal cardiopulmonary exercise (tph-ip) and a control group that received regular bleomycin inhalation for 6 weeks. the second was a three-level mixed-effects model that involves three different innate immune cells in response to tph2 and cytokines. the inhalation of the tph2 and tph3 cells was studied by in vitro cytokine profiling. data were analyzed by the chi-square test, pearson correlation coefficients, and poisson's ratio. results: the tph2 and tph3 groups induced different levels of apoptosis, neutrophil and macrophage infiltration, and proliferation. the lung injury induced lower respiratory tract injury than the control group, which were asymptomatic for the dna damage in the tph2 cells. the tph2 cell infusion induced the highest tph2 and cytokine concentrations, and the tph3 cells infiltrating the lung tissue. the lung injury caused dna damage, as confirmed by the cytokine profiling. conclusions: tph2 and tph3 cells inhalation induced higher ards and correlated with the inflammatory response, as determined by the mrna expression and protein levels in vivo. we observed that the neutrophils and macrophages were induced to respond in the same way to tph2. we therefore conclude that tph2 plays a crucial role in the ards through triggering cellular cytokine activation.

Table 7: Case analysis of the extracted keywords and samples generated by different data augmentation methods. Duplicates with the original text have been marked in orange.

shows that simple perturbations may not make texts deviate from the original semantics. Therefore, few noise data with wrong labels exist in the data augmented from EDA and BT. Although generative data augmentation methods achieve significant improvements than EDA and BT after consistency filtering, they even perform worse before consistency filtering. The phenomenon may result from the noise data with wrong labels, which is inevitably due to the pursuit of diversity. These results show that consistency filtering is indispensable for generative data augmentation methods to balance label controllability and text diversity. Besides, compared to existing data generation methods, our proposed method can achieve the best performance with or without the help of consistency filtering, which further confirms its effectiveness.

4.9 Computation Cost Analysis

We conduct a computation cost analysis of different data augmentation methods, which is reported in Table 6. Despite the small extra computational cost, EDA and BT fail to improve the performance of the HTC models significantly. Furthermore, with similar computation costs, other generative data augmentation methods cannot bring as many improvements as our methods. The results show that increasing computational cost is not a sufficient condition for improving performance and further embodies the technical contribution of our method.

4.10 Case Analysis

In Figure 7, we conduct a case analysis on the extracted keywords and generated data. To the data augmented from EDA and BT, there is a lot of overlap with the original text, while generative data augmentation methods rarely. To the extracted keywords, we can observe that yake may draw some words that are irrelevant to the label information, e.g., *'present study investigated, investigated the effects, days, present'*. These excessively useless keywords can not improve label controllability and hinder the diversity of the generated data, which causes the generated data to have very similar semantics to the original sample. With the help of the hierarchical label information and the loss function we designed, the keywords extracted by our methods can pick up the most important information that causes the original text to be annotated as the original label. Therefore, the data generated by our proposed methods can satisfy the label constraints and have differences with the original texts from both semantic and grammatical levels. We show more generated data in appendix C.

5 Conclusion

HTC is an important and challenging task, which usually suffers from the data scarcity problem because of the high cost of hierarchical multi-label annotation. With the development of generative PLMs, generating high-quality data for HTC becomes possible. In this paper, to deal with the data scarcity problem, we explore the effect of generative augmentation methods on HTC models. In order to improve the label controllability and text diversity, we propose a novel data generation framework for HTC, which consists of semantic-level and phrase-level label information enhanced prompt, hierarchical label information enhanced keyword extractor and consistency filtering. Despite the challenge of generating high-quality data for HTC, our proposed framework can achieve a balance between the controllability of multiple hierarchical labels and text diversity and improve the performance of several strong HTC models, which is demonstrated by experimental results on three HTC benchmarks and comprehensive analysis.

6 Limitations

Despite the effectiveness of our proposed method, it still has two main limitations: **(1).** Generative data augmentation methods need to use the original HTC training set to fine-tune the backbone generative PLMs. Then, they need to go through an inference stage to generate data. Both the training and inference stage need more GPU resources, which increase carbon emissions. Although the data generation is usually complete offline and does not improve the time cost of online progress, we leave how to relieve the need for GPU resources as future directions. **(2).** Although we conduct experiments on three widely used HTC benchmarks, the language of all these benchmarks is English, which has limited morphology. The effectiveness of our proposed method on language with varied morphology needs to be further confirmed.

7 Ethics Statement

The cases shown in this paper are generated automatically from different data augmentation methods, which do not represent the viewpoints of the authors. Due to social bias and lack of professional knowledge, the data generated from generative PLMs may contain misleading and toxic information, which needs to be addressed before being applied to realistic scenarios. Besides, all data generated from our proposed methods are only for scientific research. Finally, we provide comprehensive details of our model implementation and upload the source code, which guarantees the reproducibility of our experimental results.

Acknowledgements

This work is supported by the National Science Foundation of China (NSFC No. 62206194), the Natural Science Foundation of Jiangsu Province, China (Grant No. BK20220488), and JSSCBS20210661. This work is also supported by Beijing Academy of Artificial Intelligence (BAAI) and Ant Group through Ant Innovative Research Program.

References

Rami Aly, Steffen Remus, and Chris Biemann. 2019. Hierarchical multi-label classification of text with capsule networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 323–330.

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

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6295–6300.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018a. A text feature based automatic keyword extraction method for single documents. In *European conference on information retrieval*, pages 684–691. Springer.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018b. Yake! collection-independent automatic keyword extractor. In *European Conference on Information Retrieval*, pages 806–810. Springer.

Haibin Chen, Qianli Ma, Zhenxi Lin, and Jiangyue Yan. 2021. Hierarchy-aware label semantics matching network for hierarchical 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 (Volume 1: Long Papers)*, pages 4370–4379.

Zhongfen Deng, Hao Peng, Dongxiao He, Jianxin Li, and Philip Yu. 2021. HTCInfoMax: A global model for hierarchical text classification via information maximization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3259–3265, Online. Association for Computational Linguistics.

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.

Dehong Gao. 2020. Deep hierarchical classification for category prediction in e-commerce system. In *Proceedings of The 3rd Workshop on e-Commerce and NLP*, pages 64–68.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Ting Jiang, Deqing Wang, Leilei Sun, Zhongzhi Chen, Fuzhen Zhuang, and Qinghong Yang. 2022. Exploiting global and local hierarchies for hierarchical text classification. *arXiv preprint arXiv:2205.02613*.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 364–371. IEEE.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. *Advances in Neural Information Processing Systems*, 34:14542–14554.

Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *arXiv preprint arXiv:2102.01335*.

David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9031–9041.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nikolaos Malandrakis, Minmin Shen, Anuj Goyal, Shuyang Gao, Abhishek Sethi, and Angeliki Metallinou. 2019. Controlled text generation for data augmentation in intelligent artificial agents. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 90–98.

Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. 2019. Hierarchical text classification with reinforced label assignment. 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)*, pages 445–455.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *arXiv preprint arXiv:2202.04538*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Kervy Rivas Rojas, Gina Bustamante, Arturo Oncevay, and Marco Antonio Sobrevilla Cabezudo. 2020. Efficient strategies for hierarchical text classification: External knowledge and auxiliary tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2252–2257.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022. LINGUIST: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 218–241, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus ldc2008t19.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. 2018. Hft-cnn: Learning hierarchical category structure for multi-label short text categorization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 811–816.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. Strata: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731.

Xuepeng Wang, Li Zhao, Bing Liu, Tao Chen, Feng Zhang, and Di Wang. 2021a. Concept-based label embedding via dynamic routing for hierarchical 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 (Volume 1: Long Papers)*, pages 5010–5019.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022a. 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)*, pages 4242–4255.

Yufei Wang, Jiayi Zheng, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, and Daxin Jiang. 2022b. Knowda: All-in-one knowledge mixture model for data augmentation in few-shot nlp. *arXiv preprint arXiv:2206.10265*.

Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. 2022c. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7109–7119.

Zihan Wang, Peiyi Wang, Tianyu Liu, Yunbo Cao, Zhifang Sui, and Houfeng Wang. 2022d. Hpt: Hierarchy-aware prompt tuning for hierarchical text classification. *arXiv preprint arXiv:2204.13413*.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021b. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*.

Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. 2018. Hierarchical multi-label classification networks. In *International conference on machine learning*, pages 5075–5084. PMLR.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6382–6388.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label 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)*, pages 4354–4364.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34:28877–28888.

Danqing Zhang, Tao Li, Haiyang Zhang, and Bing Yin. 2020. On data augmentation for extreme multi-label classification. *arXiv preprint arXiv:2009.10778*.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.

Noisy	Method	Macro-F1	Micro-F1
0%	BERT	79.87 ± 0.41	85.96 ± 0.23
0%	Ours	81.26 ± 0.23	86.77 ± 0.17
10%	BERT	78.92± 0.50	85.81 ± 0.44
10%	Ours	80.99± 0.49	86.64 ± 0.41

Table 8: The performance when there are noisy labels in the original data. We conduct experiments based on BERT-base on the WOS dataset and report the average performance of 5 different runs. **Noisy** refers to the percentage of noisy labels. ± denotes the standard deviation.

A Human Evaluation for Filter Model

To further confirm the performance of the M_{Filter} model, we conduct a human evaluation of 200 sampled generated data and compare the consistency between human judgments and model predictions. According to the human annotation, 200 sampled generated data consist of 191 correct samples and 9 noisy samples. Our filter model keeps 187 out of 191 correct samples and only 2 noisy samples, which confirms the effectiveness of the M_{Filter} model.

B Noisy Label

To examine the performance when there are noisy labels in the original data, we add 10% noisy data in the original training set of WOS and conduct experiments on BERT-base. The results show that the performance keeps almost unchanged with our proposed method but drops significantly without our method. We hypothesize the stability comes from the "Semantic-Level and Phrase-Level Label Information Enhanced Prompt," which can unify the format of model inputs to keep the generalization ability obtained in the pre-training stage and utilize both the information of label names and keywords to avoid disturbing from noisy label names as much as possible.

C Case Analysis

We show more data generated by our proposed methods in Table 9, 10 and 11. We can observe that the data generated by our proposed method is diverse at both semantic and grammatical levels and satisfies the meaning of the original labels.

Original Label:	CS, Cryptography
Original Text:	a new method for obtaining random bijective s-boxes based on discrete chaotic map is presented. the proposed method uses a discrete chaotic map based on the composition of permutations. the obtained s-boxes have been tested on the number of criteria, such as bijection, nonlinearity, strict avalanche criterion, output bits independence criterion, equiprobable input/output xor distribution and maximum expected linear probability. the results of performance test show that the s-box presented in this paper has good cryptographic properties. the advantage of the proposed method is the possibility to achieve large key space, which makes it suitable for generation of s-boxes for larger values of n. also, because this method uses discrete chaotic map based on the composition of permutations which has finite space domain, there is no need for discretization of continuous values of chaotic map, so the process of generation of s-boxes is not affected by approximations of any kind.
Augmented Text 1:	in this paper, we aim to propose a novel model to predict the performance of the random number generator (rn) in cryptographic protocols. the model consists of three steps. the first step is to produce a logistic map describing the input space; the second step is to introduce a random number matrix representing the input space. then, the generator is constructed by reusing the logistic map. the second step is to utilize the xor function of the input data to obtain the information about the generator. the results show that the proposed model can describe the sensitivity of the proposed model against some of the weaknesses of the rn model in cryptography and in cryptographic protocols. furthermore, this model makes the model very robust against some kinds of attacks. this work is the first step towards the analysis of the rn model's performance. the results of the proposed model are compared with those obtained from an analysis of the existing models. the proposed model is a promising alternative to the current schemes which seek for the reduction of the computational costs by employing a multivariate function. (c) 2016 elsevier b.v. all rights reserved.
Augmented Text 2:	this paper describes a new algorithm that produces random and non-random matrices by varying the type of input data. the new technique is used to achieve a low complexity mapping of the input data, which makes it very suitable for implementation in real physical systems. the experimental results of the proposed approach are presented as follows: the first one uses a linear matrix function to represent the input data, while the second one uses a non-random matrix that captures the independent distribution of input data. the third one uses a linear matrix function to capture the sensitivity of the input data. in order to analyze the performance of the proposed method, an experimental setup is constructed by varying the type of input data and the input data. the new method can be applied to several fields, like cryptography, information security, etc., for the identification of the hidden hidden data. the new method is presented as a way to enhance the security level of a given setup.
Augmented Text 3:	with the increasing popularity of rfid-based cryptography techniques, it is becoming more attractive to use them in applications such as biometrics and neurobiology. one promising approach is to exploit the emergence of quantum-inspired random number generators (qisr) with the aim of enhancing the security of the qisr network. in this paper, we investigate the performance of two qisr systems based on the qisr constructs. firstly, we build a model of the qisr by modifying the dqr, i.e. the first one, composed by six independent qisrs, to determine the qisr axiom. then, we establish a linear correlation between the derived qisrs and the parameters of the proposed model. based on these correlations, the model is transformed to the standard qisr model. using the proposed model, we show that the performances of these two schemes are quite comparable. according to our results, the two schemes can be regarded as promising alternative approaches to generate the qisr network. we conclude that using the proposed model makes the qisr system promising for achieving the goal of generating a small but effective network.
Augmented Text 4:	the main aim of the paper is to evaluate the performance of the entropy-based fuzzy multi-objective matrix (muxm) based on multivariate linear equations in which the input is a linear map whose matrix is generated by a subspace. the paper makes use of an evolutionary algorithm for generating recursive matrix in the presence of input a/k. a model that adopts a set of feasible solutions is presented. to quantify the invariance of the matrix, a graphical model has been employed, where the matrix has a recursive function that makes the n-norm less linear. the resulting map can thus be used as a nonlinear objective function in which the matrix can be reduced to an equivalent hamiltonian matrix of the same size. an example is given, where the matrix is transformed into a permutative matrix, whose dimension is the sum of the squares of n - 1 and the square of the sum of the squares of the matrix. the paper makes the case that the matrix is equivalent to the squares of the matrix, where n is the number of variables that can be represented by the matrix. moreover, the obtained entropy-based fuzzy multi-objective matrix is practically applicable for obtaining the random matrix of the matrix. a sensitivity analysis on the parameters is also presented, and the resulting results are in good agreement with those of the existing experimental results of similar models.
Augmented Text 5:	this paper proposes a model for assessing the performance of an integrated energy-efficient routing (ier) protocol in bijective switching networks. the model is based on the use of the taylor series dynamic routing (tdr) with fixed points and distributed compensation mechanisms. the nonlinearity of the tdr protocol makes it suitable for high-power systems. furthermore, it has a very small linear number of operational points, which makes it suitable for small-scale operation such as power systems with a small number of participants. as an application example, we present a philipino implementation of the ieee 14-bus system. the obtained results show that a small number of participants can be identified as successful in the proposed method and the optimal routing parameters can be determined according to the availability of a good routing network. the use of a forward-backward shift from the tdr protocol to the forward routing makes the proposed method more effective for small-scale networks. the benefits of the proposed method are as follows: first, it makes it possible to remove the dependency on the router set-point; second, it reduces the number of participants, which makes it suitable for microcontrollers. lastly, it makes it possible to obtain a closed-loop control scheme which makes it possible to select and control the number of agents. the presented model is applied to the ieee 14-bus system to investigate the ieee 14-bus system's performance and the effectiveness of its implementation in relation to the conventional solutions. compared with the conventional solution, the proposed method shows an average reduction in energy consumption by 84% and improvement in the energy efficiency by 36% and 44%. the proposed approach is a feasible and promising alternative for the energy-efficient routing of networks with low-power systems. (c) 2017 elsevier b.v. all rights reserved.

Table 9: Case study of the samples generated by our proposed methods.

Original Label:	Medical, Ankylosing Spondylitis
Original Text:	ankylosing spondylitis (as) is a chronic systemic inflammatory disease. via autoimmune mediators, as can damage the auditory system similar to other systems. otoacoustic emission studies in as patients showed that the damage that causes hearing loss was in the outer hair cells. the medial olivocochlear (moc) reflex is used to evaluate the moc efferent system (moes), which includes the outer hair cells. the aim of this study was to evaluate the presence of subclinical damage in the inner ear with the aid of the moc reflex test in as patients with no hearing complaints. thirty-four patients with as and a control group of 30 healthy volunteers with similar demographic characteristics were evaluated in the study. otoacoustic emission responses, moc reflex results, and frequency-specific and total suppression findings were compared between the groups. the relationship between clinical and laboratory findings for the as patients, and the moc reflex data were also investigated. reduced moc reflex response (p = 0.04) and suppression (p = 0.019) were detected in as patients. when the clinical and laboratory findings for the as patients and the moc reflex test results were compared, a significant correlation was found only between the moc reflex and the erythrocyte sedimentation rate. the results showed that as can damage the inner ear, especially the moes, and can reduce the moc reflex response without clinical hearing loss.
Augmented Text 1:	introduction: the incidence of spondyloarthritis (spa) has been increasing in recent years, and the disease activity is aggravated in patients with spa. this study was conducted to explore the influence of the inflammatory markers il-10 and il-12 on spa patients' clinical and radiological findings, and to explore the underlying molecular mechanisms of spa. methods: sixty-seven patients with spa (39 males and 38 females) were included in the study. all patients underwent a clinical and laboratory evaluation. they were clinically evaluated based on ph, the study included the evaluation of the ph, the assessment of the c-reactive protein (crp), and the arthrocyte sedimentation rate (esr). the results were compared to the international classification of diseases, 9th revision (icd-9) scores. results: the results showed that the frequency of esr was higher in the patients with spa than in the control group (p 0.001), and the frequency was higher in the spa group than in the control group (p 0.001). the esr score of the spa group was higher than that of the control group (p 0.001). the prp, crp, and esr levels of the patients with spa showed no significant differences between the groups. compared to the control group, the crp and esr values of the patients with spa were higher than that in the control group (p 0.001, respectively). we also performed a comparison of the crp and crp scores of the patients with spa. conclusions: the study showed that the presence of il-10 and il-12 levels increases the risk for spa.
Augmented Text 2:	introduction: the study was performed to investigate the correlation between the inflammatory markers scdc50 and the rheological parameters of ankylosing spondylitis (as) patients. materials and methods: we retrospectively included 70 as patients and 70 healthy controls, matched according to the egyptian criteria. all patients had at least one of the evaluated laboratory findings: 1) erythrocyte sedimentation rate (esr) in the right hemisphere, 2) somatosensory pathway (lps in the right hemisphere, 3) sa in patients with peripheral neuropathy, 4) mri signal in the left hemisphere (lmi), 5) ankylosing spondylitis (ssa) patients, 6) c-reactive protein (crp) in the right hemisphere (rho), and 7) molecule in the left hemisphere. the patients were divided into four groups according to the presence or absence of the clinical signs of as. all the patients underwent evaluations for the presence of ankylosing spondylitis based on the scdc50, lps, crp, ankylosing spondylitis (spa), or the absence of any clinical signs of the disease. results: the mean values of the scdc50 and the rheological parameters in the left hemisphere were not different according to the clinical signs. the rheological parameters of the left hemisphere were not significantly different according to the presence of the clinical signs of as patients. however, the lps and the crp parameters of the right hemisphere were statistically different according to the presence or absence of the clinical signs of as. compared with the lps group, the lps group showed an increase in the scdc50 values and a decrease in the crp values of all the patients with as compared with the control group. the levels of cytochrome p450 2a1 (cyp2a1), alpha-smooth muscle actin (alpha-sma), and the rheological parameters of the egyptian hemisphere were found to be higher in the as
Augmented Text 3:	objectives: the objective of this study was to evaluate the impact of arthroscopic hearing loss on patients with inflammatory rheumatic diseases (ird). methods: the study comprised 2 groups of patients who were treated with hy-procedural therapy of ird, including those with ankylosing spondylitis (as), diffuse ird (dii), and focal ird (fii). patients were clinically assessed with a modified chi-square test and paired t-test. analysis of a subset of the data set (n = 636) was performed to analyse the frequency peaks of acoustic signals in the resonant frequencies (mean +/- sd: 2.85) in the acoustic group and from the dii-positive group (n = 634). results: the microscopic examination of the patients with spondylitis showed that otoacoustic signals in the vestibular system are decreased, whereas acoustic signal is increased. the percentages of patients with dii and fii patients were mainly found in the otoacoustic group (mean +/- sd: 4.8%). the percentages of patients with dii and fii patients were not significantly different from controls (mean +/- sd: 0.35 +/- 0.08, p = 0.023). after controlling for the influence of the sympathetic modulation, the percentage of patients with dii were not significantly different (mean +/- sd: 1.4%; p = 0.032). the percentages of patients with otoacoustic modulation tended to be higher in patients with dii than in controls (mean +/- sd: 1.41%; p = 0.002). conclusions: the study did not show an increased proportion of patients with ird who had hearing loss associated with the acoustic modulation. this suggests that the presence of inflammatory mediators associated with the pain associated with the hearing loss, is not significantly different from the effect of the disease itself.
Augmented Text 4:	background: the aim of this study was to investigate the differences in the etiology of otoacoustic dyspnoea (otoacoustic dyspnoea) between patients with active ankylosing spondylitis (as) and healthy controls (hcs). methods: in this prospective, observational study, 58 as patients and 59 hcs were prospectively recruited. the medical records of all participants were analysed for the presence of the presence of the following: bmi, total qtl, eosinophil count, mrna and protein levels; and the presence of otoacoustic echogenicity. results: the median age of the patients was 35.9 years (range: 14-62 years). most patients (81.5%) were male, with mean (sd) age 23.8 (5.2) years. the mean etiology of as was found to be naive (66.6%), with mild (28.3%), moderate (22.4%) and severe (26.8%) clinical manifestations. the most common clinical manifestation was etiology of as (39.3%). all the patients showed recurrent etiologies of as (88.4%). otoacoustic echogenicity was found in 32 (34.7%) of the patients, being not observed in 37 (28.9%) of them. the moc showed a significant decrease in frequency of the erythrocyte sedimentation rate (esr) values in as patients. conclusions: the results of this study show that as patients and hcs have similar clinical phenotypes with high frequency of the etiology of as. our study can provide baseline data for the investigation of the role of the etiology of as and of its possible mechanisms.
Augmented Text 5:	background: spondyloarthritis (sa) is a chronic inflammatory rheumatic disease of the axial skeleton. fibroblasts (fibroblasts) of the neck are common to the disease. in the past two decades, fibroblasts of patients with sa have been widely studied. to date, there has been little insight into the underlying molecular pathophysiology of sa in the west. objective: we aimed to investigate the role of fibroblasts in the pathogenesis of sa in a cohort of patients with sa. methods: in this retrospective observational study, 30 consecutive patients with sa and 21 healthy individuals were included. clinical and laboratory data were collected; fibroblasts recombinant fibroblasts were analysed by a sybr amplification kit. results: in a subset of 14 patients, fibroblasts were identified as the primary cell type in the most abundant cells. fibroblasts from the most abundant cell type in the patients were found to be more abundant. additionally, the distribution of fibroblasts was altered in many patients. fibroblasts from the most abundant cell type (m2-m3) were identified as the primary cell type of fibroblasts. the most abundant cells of the fibroblasts were oligodendrocyte glycoprotein (og) and cd29. the expression of fibroblast neoplastic cell marker, erythrocyte sedimentation rate (esr), and the expression of the transcription factor cd44 induced by fgf-2 was detected in two patients. sa patients exhibited an increase in the number of fgf-2 mrna transcripts with the highest mrna levels after fgf-2 treatment. the frequency of sa cell differentiation was not different between the two groups. the frequency of cd29 expression was higher in the fgf-2 group than that in the other two groups. fgf-2 was higher in the og group and the frequency of cd34 expression was higher than that of the other patients. conclusion: fgf-2 is a marker for sa and was found to be higher in patients with sa than in healthy

Table 10: Case study of the samples generated by our proposed methods.

Original Label:	CS, Cryptography
Original Text:	with ever increasing load and insufficient growth of generation and transmission capacities, the operating constraints of modern integrated power system are increasing. it has led to ever increasing dynamic problems of low frequency oscillations in the system that needs to be detected and damped out quickly and efficiently. in electrical power networks small oscillations appear from time to time. these oscillations concern the quantities determining the equilibrium point of the system, and following which, system stability and system behaviors are influenced. the objective of our study is to check the static stability of the high voltage power to small perturbations of electrical network. in this context, after linearizing the system, power system stabilizer (pss) has so far been extensively utilized to mitigate these problems. this paper presents a novel and efficient approach for the optimal tuning of power system stabilizer parameters (pss) using a genetic algorithm (ga) with the eigenvalue-based objective function. the proposed approach is implemented and examined in a system with a single machine connected to an infinite bus via a transmission line. the results of this technique have been verified by eigenvalue analysis and time-domain simulations. the obtained results were evaluated and compared with ones obtained by ordinary gas. eigenvalue analysis and nonlinear system simulations demonstrated the effectiveness of the proposed approach in damping the electromechanical oscillations and enhancing the system dynamic stability.
Augmented Text 1:	this paper presents a method of determining the electromechanical parameters of the piezoelectric elastomer system (pees) of an electrical network driven by a power source, i.e., a main voltage source. the approach considers the elastomer system as a nonlinear system composed of several electrical components, the mechanical components of which are represented by mechanical parts, and the electrical components. a typical example is taken from the literature wherein the electromechanical components are modeled as a constant eigenvalue of the power supply voltage. this study was conducted on a simulated example of an existing commercial pee system using a voltage source, thereby determining parameters of its mechanical component. the obtained results show that the use of the pees can reduce the nonlinear problems encountered in determining the electrical network's performance even when the electrical network is sufficiently weak. the present method is also very effective in identifying the most efficient way to improve the efficiency of the system. a case study was performed on a 1.2 mhz pees system using a power system controller. the obtained results are in good agreement with the simulation results. (c) 2014 elsevier ltd. all rights reserved.
Augmented Text 2:	a problem we encountered in the development of an effective electrical network for solving problems such as the one experienced by a client for the repair of a bus damaged by an electrical short circuit is the one that is insufficient to supply the electricity for the generating station. the problem is formulated as a nonlinear programming problem with three steps: (a) selecting a suitable analytical model to represent the system, (b) determining the parameters of the model, and (c) optimizing the model through the use of an efficient regression technique. we have used the mathematical language matlab/simulink in the electrical network and implemented it into a matlab/simulink environment. we have evaluated the performance of the proposed network model through simulations in terms of the dynamic performance of the proposed model. the obtained results clearly indicate that the proposed model is a promising tool for identifying and assessing the influence of the system parameters on the performance of the electrical network.
Augmented Text 3:	in recent years, many scientific disciplines have focused on developing energy-efficient strategies as a means to control costs, while avoiding environmental constraints. one such example is the field of electromechanical transient response of the electrical network, where a voltage source is connected to a damping device. this is the case of the system experiencing an excitation of an electrical generator and its oscillating rotor. these systems are subject to an enormous electrical load and this has been evaluated by various studies, including simulations, in order to understand their influence on system dynamics, as well as to understand their relationship with the electrical behaviour of these systems. this work introduces a network-based methodology to investigate the influence of the system parameters on system dynamics, with the goal of determining the ultimate value of the systems that can be optimal for control. simulations of a typical manufacturing facility are conducted in order to characterize the system. the results obtained show that the system can be improved by up to 45% in reducing the overall system power consumption by up to 50%. in addition, a comprehensive analysis of the existing network model is performed to show the relevance of these parameters in determining the optimal solution. (c) 2016 published by elsevier ltd.
Augmented Text 4:	determining the optimal performance of a power system can be done by reducing the quantities and constraints that the system has to consider in its design. in this paper, a nonlinear electrical network system with multiple energy harvesters and multi-generation sources is evaluated in order to determine the most cost-efficient way of powering a power system. the nonlinear electrical network system has to consider the effects of all the components that are involved in the process of power conversion: the source, the converter, the loads and the electrical network. the main contribution of this paper is to investigate the influence of the system's parameters and their interaction on the power system performance. a simplified model that combines both an electrical network and a fluid-structure-network model is presented. the model is built using the kalman filter, a well-known mathematical technique for non-linear systems. to estimate the optimum performance, a modified inverse-transform (idre) method using the first and second order shift operator is used. moreover, a constraint based on the nonlinear electrical network is used, called the zero-crossing operator, to be adopted in the problem. the parameters are estimated using a multi-step optimization algorithm based on the power system model. the results indicate that in order to find an optimal solution of the network, it is necessary to understand the influence of the parameters in the network and to be able to take into account the influences of the parameters. (c) 2016 elsevier b.v. all rights reserved.
Augmented Text 5:	the article presents an approach to determine the most cost-efficient way of resolving the problem of energy management of high power grid systems such as, for example, the distribution network or even the entire electrical network. the main objective is to reach the equilibrium point between the nominal and actual value of the active components of the network, thus reaching the most cost-efficient ones. to reach the optimum solution, the paper first proposes a nonlinear recursive method based on the interaction of the kinematic and electromechanical parameters. it then moves on to use an iterative algorithm based on the recursive least squares technique to find the optimal solutions of the optimal problem. furthermore, it uses the ahysteresis model to evaluate the effectiveness of the proposed method. the effectiveness of the proposed method has been evaluated through simulations on a 125kw (ems) power grid connected to a distribution network for a nominal size of 5000mw. the results show that the proposed approach can reduce the energy consumption by up to 4%. (c) 2016 published by elsevier ltd.

Table 11: Case study of the samples generated by our proposed methods.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.