IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 January 2025, accepted 4 February 2025, date of publication 10 February 2025, date of current version 18 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3540090

RESEARCH ARTICLE

Enhancing Domain Generalization Performance in Low-Resource Setting via External Dataset and Pseudo Labeling With Sentence-BERT

JUNHO LEE[1], SEUNGUK YU[1], (Graduate Student Member, IEEE), JINHEE JANG[1], KEUNHYEUNG PARK[1], AND YOUNGBIN KIM[1,2], (Member, IEEE)
[1]Department of Artificial Intelligence, Chung-Ang University, Dongjak, Seoul 06974, South Korea
[2]Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University, Dongjak, Seoul 06974, South Korea

Corresponding author: Youngbin Kim (ybkim85@cau.ac.kr)

This work was supported in part by the Chung-Ang University Research Scholarship Grants, in 2023; and in part by the Institute for Information and Communications Technology Planning and Evaluation (IITP) through Korean Government (MSIT) (Artificial Intelligence Graduate School Program, Chung-Ang University) under Grant 2021-0-01341.

ABSTRACT Recent studies on data augmentation have focused on improving model performance with limited training data within a specific dataset. While the goal is to enhance performance on the dataset itself, this approach also addresses broader challenges, such as enhancing domain generalization. Building on this, we propose the Out-of-Domain Pseudo Labeling (`OOD-PL`) method, a data augmentation technique designed to ensure data diversity and enhance domain generalization of model in low-resource settings. Our approach introduces external data and assigns pseudo labels based on semantic vicinal interpolation with the intended training data. We observed significant improvements in domain generalization across three datasets from different domains. Unlike traditional methods, this approach utilizes other samples as a form of augmentation for the training data. Our method can be flexibly integrated with existing augmentation techniques, and we demonstrated that it performs well even when the available training data is extremely limited. Furthermore, we conducted various in-depth analysis experiments to strengthen the validity of our proposed method and demonstrate its robustness in effectively enhancing domain generalization. As a result, we were able to propose a methodology that overcomes the limitations of using specific datasets, even in situations where their availability is restricted, by leveraging out-of-domain samples.

INDEX TERMS Data augmentation, domain generalization, low-resource NLP.

I. INTRODUCTION

Data augmentation is essential in the area of deep learning, as it not only enhances task-specific performance but also improves generalization capabilities of a model. By synthetically increasing data diversity, data augmentation allows models to learn from a broader range of patterns [15]. When training model with small datasets, the limited variety of the data restricts the model's ability to generalize, driving ongoing efforts to mitigate these limitations through data augmentation process. However, most prior studies [4], [19], [20] have primarily focused on boosting performance by employing data augmentation to small training samples [10].

For text data, it is common practice to utilize pre-trained language models (PLMs) and fine-tune them for specific downstream tasks. While this approach adapts the model effectively to the selected task, in low-resource settings with limited labeled data, it often leads to challenges such as overfitting or the model memorizing the training data [7]. As a result, the model's ability to generalize to unseen data diminishes, along with its performance in domain generalization. To mitigate this, data augmentation techniques have been employed to increase data diversity, thereby enhancing the model's generalization capabilities [17].

The associate editor coordinating the review of this manuscript and approving it for publication was Amin Zehtabian.

Data augmentation methods are generally divided into two categories: rule-based and model-based. Rule-based augmentation generates new data by applying predefined rules or transformations to the existing data [4], [19], [24]. While this approach is cost-effective and easy to implement, it offers limited diversity, as it closely mirrors the original data distribution [22]. On the other hand, model-based augmentation utilizes trained models to enhance data diversity [9], [14]. Although this method is more effective in generating diverse data, it comes with increased complexity and higher computational costs compared to rule-based techniques. To tackle the challenge of acquiring diverse data efficiently, semi-supervised learning methods leveraging unlabeled data have been introduced, with a primary focus on improving performance in low-resource settings [20]. In response, we propose a data augmentation method that ensures data diversity at a low cost, specifically tailored for low-resource environments.

In this study, we introduce an Out-of-Domain Pseudo Labeling (OOD-PL) method that utilizes external data to enhance data diversity in low-resource environments at minimal cost. In our study, a low-resource setting refers to a situation where learning may be insufficient due to the limited amount of a specific dataset. We used the concept of out-of-domain to refer to situations where a dataset with a domain or context different from the specific dataset the model was trained on is being handled [25]. Therefore, when there is a limited amount of data for a particular domain, we aimed to evaluate the model's generalization ability and adaptability by training or evaluating it on a different dataset. Ensuring a balanced distribution within training batches, avoid of spurious correlations, has proven effective for improving domain generalization [18]. However, training exclusively on domain-specific datasets can lead to spurious correlations. To mitigate this issue, we expand the scope and diversity of the training data by identifying semantically similar external data based on the existing datasets.

We utilized the Sentence-BERT model [11] to compute semantic similarity between the original and external data, assigning pseudo labels to the external data based on its similarity to the original data [6]. To mitigate overfitting during the data augmentation process, we utilized the OOD-PL method at the early stopping point during training. Furthermore, we implement a similarity-based sampling strategy to enhance the diversity of the training data. This approach allows the model to learn from a broader range of data once it demonstrates satisfactory performance on the original dataset, thereby improving its generalization ability.

The OOD-PL method enables the model to learn from a diverse range of data by selecting samples based on semantic similarity, gradually incorporating more varied data as training progresses. This approach allows the model to generalize across different domains, avoiding over-specialization to a single dataset. The experimental results from our proposed method indicate that the data augmentation using OOD-PL outperformed various baseline methods in text classification tasks across diverse domains. This suggests that our approach, which leverages external data through semantic similarity, effectively overcomes the limitations of traditional augmentation methods, such as rule-based techniques. In addition to comparing with baseline methods, we conducted further analysis on the combination of data augmentation methods, performance based on model size, the number of data to be augmented, and the methodology of pseudo labeling, thereby demonstrating the superiority of our method.

The contributions of this study are as follows:

- We propose the OOD-PL augmentation method, which utilizes external data to reduce spurious correlations in low-resource environments. It requires no additional cost for data augmentation and can be trained in an unsupervised manner using pseudo labels.
- Applying OOD-PL to three sentiment analysis datasets —each originating from different domains but addressing the same task—resulted in enhanced domain generalization performance. These results demonstrate that the proposed method is not restricted to a single domain and can be effectively utilized across diverse domains.
- Through additional experiments conducted from various perspectives, including the combination of data augmentation methods and performance analysis based on model size, we confirmed the advantages of our proposed OOD-PL method. Consequently, we thoroughly analyzed the ideal conditions for implementing data augmentation using our methodology.

This paper is organized as follows. Section II covers various augmentation techniques and prior research on domain generalization through data augmentation. Next, Section III outlines the approach and framework of the proposed method, and Section IV presents the experimental setup. Then, Section V presents the results and analyses that validate the effectiveness of the proposed method. Section VI and VII respectively cover additional experiments validating the effectiveness of the proposed method, as well as discussing its limitations and potential future work. Finally, Section VIII summarizes our work and discusses avenues for future research.

II. RELATED WORK

A. DATA AUGMENTATION

Data augmentation, which aims to enhance model performance and generalization by expanding or transforming datasets, can be divided into two main approaches: rule-based and model-based methods.

1) RULE-BASED AUGMENTATION

Rule-based augmentation methods provide a simple and cost-effective approach for text data. Easy Data Augmentation (EDA) [19] creates new training samples by performing word replacements, insertions, swaps, and deletions. This method effectively improved performance in tasks such as text classification through simple sentence editing, with-

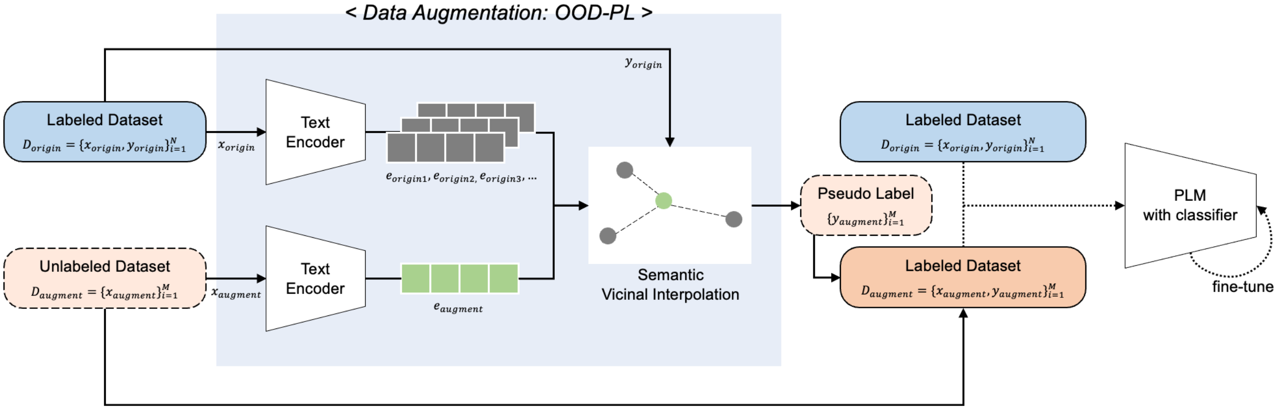
< Data Augmentation: OOD-PL >
y_{origin}
Labeled Dataset
$D_{origin} = \{x_{origin}, y_{origin}\}_{i=1}^{N}$
x_{origin}
Text Encoder
$e_{origin1}, e_{origin2}, e_{origin3}, \dots$
Unlabeled Dataset
$D_{augment} = \{x_{augment}\}_{i=1}^{M}$
$x_{augment}$
Text Encoder
$e_{augment}$
Semantic Vicinal Interpolation
Labeled Dataset
$D_{origin} = \{x_{origin}, y_{origin}\}_{i=1}^{N}$
Pseudo Label
$\{y_{augment}\}_{i=1}^{M}$
Labeled Dataset
$D_{augment} = \{x_{augment}, y_{augment}\}_{i=1}^{M}$
PLM with classifier
fine-tune

FIGURE 1. **Training process using the proposed Out-of-Domain Pseudo Labeling (OOD-PL) augmentation method. We incorporate external datasets during data augmentation, assigning pseudo labels to these datasets through semantic vicinal interpolation with the existing data. In the training process, we strategically employ early stopping and curriculum sampling techniques to effectively utilize the external datasets. Through this approach, we propose the OOD-PL augmentation method that is robust for domain generalization, extending the external datasets as labeled dataset and train them alongside the existing dataset. Here, *N* and *M* represent the total amounts of the labeled and unlabeled datasets, respectively.**

out requiring additional model training, which led to the emergence of various modified versions of the approach. SoftEDA [24] builds on EDA by incorporating soft labels to reduce semantic distortion during augmentation. Arbitrary Easy Data Augmentation (AEDA) [4] introduces randomly selected special characters to generate new samples.

Rule-based augmentation methods have the advantage of quickly augmenting data according to predefined patterns, and their efficiency has been demonstrated through various existing approaches. However, they often struggle to deviate meaningfully from the original data distribution, limiting the model's capacity to learn diverse patterns [22]. Additionally, rule-based methods require heuristic tuning of hyperparameters tailored to each augmentation technique. Recently, methods have emerged to automatically optimize word replacement or insertion probabilities in augmentation process, addressing the issue of semantic distortion that was often criticized in traditional rule-based techniques [26].

2) MODEL-BASED AUGMENTATION

Model-based augmentation methods use pre-trained models to generate diverse data augmentations. Back Translation [14], for instance, creates variations by translating text between different languages while maintaining semantic consistency. Self-Supervised Manifold Based Augmentation (SSMBA) [9], similar to back translation, introduces noise to parts of the data and recovers them through self-supervised learning to generate new samples. Although effective, these methods often come with high computational costs. Unsupervised Data Augmentation (UDA) [20] utilizes unlabeled data to learn consistency across different augmentation techniques, improving performance within the same domain, even in low-resource settings. While it relies on external data, its primary goal is to enhance performance within the original domain. Text AutoAugment [12] automates the search for optimal augmentation strategies, enabling effective domain generalization with limited labeled data, though its performance may be constrained if insufficient validation data is available from external domains. Additionally, various approaches have been developed that leverage language models in multiple ways for text data augmentation. These include abstractly summarizing documents and then expanding them, sampling semantic representations of sentences from diverse distributions, and applying back translation [27], [28].

B. DOMAIN GENERALIZATION

Domain generalization focuses on enabling models with the ability to generalize effectively to unseen data outside the training set. Vicinal Risk Minimization (VRM) [2] addresses this by incorporating data distributions in the vicinity of the training set into the learning process, thereby enhancing the model's robustness to new, unseen data. In data augmentation, the concept of vicinity has been applied through methods such as EDA [19] and SoftEDA [24], which introduce simple rule-based perturbations to text, generating lexically similar variations. These methods help the model learn a range of transformations while maintaining the same labels for augmented data that remain close to the original in the latent space [19]. In contrast, Back Translation [14] generates more diverse expressions by translating sentences between languages, assigning the same labels to semantically similar outputs within the vicinity.

Similarly, SSMBA [9] defines vicinity based on the manifold assumption, assigning identical labels to data reconstructed from noise. Mixup [21] further enhances data diversity by generating new data points through linear interpolation between existing samples. Previous studies have extensively explored techniques for domain generalization through various approaches, such as data augmentation, labeling strategies, and data generation. Recently, more complex methodologies have emerged to address domain adaptation and data imbalance issues, such as simultaneously applying pseudo labeling and synonym generation techniques [29]. Additionally, frameworks have been proposed that use domain metadata to re-adjust domain-specific weights during the testing phase, or adaptively separate robust features without relying on target domain information [30], [31].

In this study, we propose a method OOD-PL that assigns pseudo labels to external data based on the semantic similarity between the external data and the training samples. Unlike traditional rule-based augmentation methods that involve simple sentence editing at the character or word level, our approach considers the semantic meaning of sentences from the model's perspective, providing an effective way to utilize external data. Additionally, our methodology demonstrates that by effectively utilizing external datasets, such as through semantic vicinal interpolation, it is possible to overcome the limitations of insufficient data in situations where a specific dataset cannot be fully leveraged.

III. METHOD

We propose the OOD-PL method for data augmentation, which leverages external data to ensure data diversity at low cost in low-resource environments. The proposed method is based on two inputs: a labeled dataset and an unlabeled dataset from external data sources. The overall structure of the model is shown in Fig. 1, providing a detailed explanation of the proposed method through pseudo-labeling based on these two inputs, as well as the similarity-based diversity sampling strategy designed to progressively train the model with increasingly diverse samples.

A. SEMANTIC VICINAL INTERPOLATION

We observed that, in low-resource environments, training only on labeled data can lead the model to memorize labels and learn domain-specific characteristics, making it vulnerable to domain generalization [7]. Therefore, from a data augmentation perspective, we expanded the data by incorporating additional external data and generating pseudo-labels for the unlabeled data, allowing the model to learn a broader range of patterns compared to the original dataset. In this process, the out-of-domain samples are directly merged with the existing labeled dataset.

Inspired by previous research that maintains the same labels based on semantic similarity [14] and that interpolate between two data points in latent space to estimate soft labels [21], we propose a method to pseudo-label external data using soft labels based on semantic similarity. To achieve this, we employed the Sentence-BERT [11] model to calculate the similarity between labeled and unlabeled data in the semantic space, quantifying the semantic differences between data points. Based on this quantified semantic difference, we derived the semantic vicinity by identifying external data with high similarity to the labeled data. The semantic vicinity, denoted as s, is calculated as follows.

$$y_{origin} = [0, 1], \tag{1}$$

$$y_{augment} = [max((1-s), 0.5), min(s, 0.5)], \tag{2}$$

By calculating the semantic vicinity as shown in Equation (2), we ensured that even when the similarity is low, the label of the comparison target remains unchanged, preserving the labels of the original labeled data as much as possible. To generate accurate pseudo-labels during training, we utilized label information within each batch. As a result, the average semantic vicinity information, calculated based on the number of compared samples, is used to generate the pseudo-labels. As a result, the out-of-domain samples, with appropriate labels assigned, are incorporated into the model's training process.

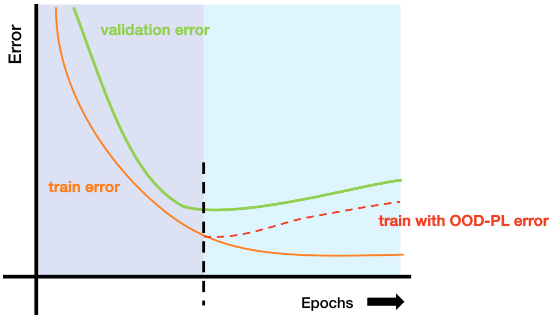

FIGURE 2. The loss convergence patterns observed in experiments based on the use of early stopping. While conventional early stopping halts training once the loss reaches an appropriate value (orange solid line), we instead begin training with external datasets at that point (orange dashed line), maintaining the training difficulty at an optimal level and addressing domain generalization.

B. TRAINING START POINTS

The OOD-PL method proposed in this study expands the training data by incorporating external data into the existing dataset. To prevent a decline in training efficiency due to the influence of out-of-domain samples from the early stages of training, we applied early stopping, a technique commonly used to prevent overfitting and improve generalization.

Unlike the conventional use of early stopping, where training is halted at a certain point, we utilized this point to begin training with the OOD-PL method. This approach ensures that the model sufficiently learns the domain from the existing data first, and then proceeds to learn from diverse out-of-domain data. The goal is to prevent the model from overfitting to labeled data and learning false correlations, ultimately improving generalization performance. The objective behind the proposed training timing and its impact is illustrated in Fig. 2.

C. CURRICULUM SAMPLING

In the process of utilizing external data for training, we also employ curriculum sampling to progressively expose the model to a diverse range of data distributions. This algorithm is designed to gradually increase the semantic diversity of

Algorithm 1 *OOD* Sampling With Curriculum Learning

```
Initialize similarity list $S = []$
Set initial batch iteration $n_0$
Set increment step $\Delta n$
Set maximum batch iteration $n_{max}$
Initialize current batch iteration $n \leftarrow n_0$
Initialize previous validation accuracy $acc_{prev} \leftarrow 0$
while training epoch do
    for each $OOD_i$ sample in $OOD$ do
        Calculate semantic vicinal interpolation similarity
          $s_i$ for $OOD_i$
        Store $s_i$ in S
    end for
    Sort $S$ in ascending order of similarity
    Select top $n$ elements from $S$
    Evaluate current validation accuracy $acc_{current}$
    if $acc_{current} > acc_{prev}$ then
        Increment $n$ by $\Delta n$ up to $n_{max}$
        Update $acc_{current} \leftarrow acc_{prev}$
    end if
    end while=0
```

out-of-domain data throughout the training process. By doing so, we aim to avoid confusion that may arise from arbitrary data sequencing and to reflect the semantic changes within the data in the training order, maximizing training efficiency. We provide pseudo code for the curriculum sampling algorithm above.

First, we initialize the similarity list S and set the number of epochs for training, as well as the total number of curriculum steps. During training, the similarity score s_i for each *OOD* sample is calculated and stored in the similarity list S. Afterward, the list is sorted in descending order, and the top-ranked items are selected so that the external data with the highest semantic similarity is utilized first. With this approach, we calculated the semantic similarity for newly selected *OOD* samples at each step and used the selected data to train the model. If the validation accuracy $acc_{current}$ of the existing data improved within each sample compared to acc_{prev}, it was considered that the model had successfully learned the current level of diversity achieved.

As a result, during our training process, *OOD* samples were selected according to the curriculum learning algorithm, and pseudo-labeling using soft labels was applied to the sampled *OOD* dataset, enabling semantic vicinal interpolation. These datasets were introduced after the early stopping point mentioned earlier, allowing the model to thoroughly learn the domain from the existing data before introducing additional data to improve generalization performance. At this point, we used binary cross-entropy (BCE) loss, and if we let f represent the pre-trained language model and classification layer for model training, the final loss function for the proposed method is as follows.

$$\begin{aligned} L_{OOD\text{-}PL} &= BCE(f(x_{origin}), y_{origin}) \\ &\quad + BCE(f(x_{augment}), y_{augment}), \end{aligned} \tag{3}$$

IV. EXPERIMENTAL SETTINGS

A. DATASETS

In this study, we employed sentiment analysis datasets from various domains that focus on the same task, specifically Stanford Sentiment Treebank (SST-2) [16], Internet Movie Database (IMDB) [8], and Yelp [1]. SST-2 dataset comprises approximately 67,000 training samples and 872 validation samples for sentiment analysis of movie reviews. Similarly, the IMDB dataset focuses on movie reviews, containing 25,000 samples for both training and validation. Notably, 90% of the sentences in the IMDB dataset range from 52 to 2,780 tokens in length, significantly exceeding the maximum sentence length of 268 tokens found in SST-2. The Yelp dataset includes reviews from various commercial sectors, including restaurants, hotels, and shops, with 650,000 training samples and 50,000 evaluation samples. The dataset features ratings ranging from 0 to 4, which correspond to a 1 to 5-star rating system. To ensure consistency in our analysis, we classified ratings of 1 and 2 as negative and ratings of 4 and 5 as positive, while excluding the 3-star reviews.

To integrate *OOD* datasets, we combined WMT-14 [1], a translation dataset including a diverse array of topics and expressions, with AG-News [23], a news classification dataset. In the merging process, we established a minimum sentence length threshold of 100 tokens to exclude relatively short samples compared to the training data.

B. EXPERIMENTAL SETUP

We employed BERT [3] architecture as our pre-trained language model.[1] A single linear layer served as the classification layer. To calculate semantic similarity, we utilized Sentence-BERT [11] architecture.[2] The maximum input sequence length was set to 512 tokens, and we used a batch size of 64. The learning rate was configured at 5e-5, and the network was optimized using the Adam optimizer [5]. Training was conducted for up to 30 epochs, with early stopping implemented to prevent overfitting and to ensure a fair comparison with other augmentation methods. The early stopping patience was set to 5, meaning that training would cease if no performance improvement was observed over five consecutive evaluations.

For comparison with baseline augmentation methods, we applied a two-fold data augmentation approach. Specifically, for the `OOD-PL` method, one *OOD* sample was mixed into each training batch. A total of 2,500 *OOD* samples were selected for similarity comparison, and within each batch, 8 samples were pseudo-labeled based on semantic similarity. The models were implemented using the PyTorch framework and trained on a single GeForce RTX 3090 GPU.

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

[2] https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

TABLE 1. Performance comparison between the augmentation methods and the proposed `OOD-PL` method based on different train and validation datasets. We evaluated domain generalization by training on a minimal portion of a specific dataset and testing it on another dataset. In each cell, the top value represents accuracy, and the bottom value represents empirical risk. The augmentation method with the highest accuracy for each training dataset proportion is highlighted in bold, while the method with the lowest accuracy is underlined.

Valid Dataset	IMDB			Yelp		
Train Dataset: SST-2 (%)	0.5	1.0	5.0	0.5	1.0	5.0
No Augmentation	78.75 0.648	79.32 0.624	82.31 0.412	76.52 1.006	78.50 0.984	81.33 0.929
EDA	50.40 0.699	50.22 0.705	77.87 0.462	51.04 0.694	57.38 0.671	**87.58** 0.380
BT	54.02 0.685	64.94 0.656	82.50 0.413	55.77 0.683	68.81 0.585	86.82 0.320
SSMBA	69.14 0.663	71.24 0.614	81.49 0.433	64.13 0.732	69.17 0.681	84.13 0.537
`OOD-PL` (Ours)	**78.89** 0.670	**80.99** 0.648	**83.17** 0.632	**78.24** 1.028	**82.53** 1.008	85.53 0.786
Valid Dataset	SST-2			Yelp		
Train Dataset: IMDB (%)	0.5	1.0	5.0	0.5	1.0	5.0
No Augmentation	75.32 1.536	79.91 1.319	82.10 1.172	82.33 1.298	87.01 1.115	88.90 1.010
EDA	**78.59** 0.473	81.77 0.425	77.74 0.490	85.17 0.458	**89.72** 0.393	88.84 0.308
BT	78.81 0.458	83.18 0.393	83.58 0.402	85.09 0.369	89.56 0.273	88.84 0.308
SSMBA	78.22 0.463	84.13 0.387	83.30 0.393	**85.22** 0.423	88.29 0.374	87.47 0.315
`OOD-PL` (Ours)	76.32 2.168	**84.67** 1.151	**84.59** 0.903	84.55 1.330	89.20 1.296	**89.84** 1.269
Valid Dataset	SST-2			IMDB		
Train Dataset: Yelp (%)	0.5	1.0	5.0	0.5	1.0	5.0
No Augmentation	82.14 4.669	83.66 4.137	83.08 3.381	80.34 2.377	82.43 1.971	83.27 1.836
EDA	82.67 0.469	82.91 0.405	83.51 0.411	84.36 0.418	86.20 0.331	84.40 0.366
BT	83.59 0.390	**84.28** 0.376	82.05 0.425	**84.46** 0.367	**86.32** 0.351	81.37 0.333
SSMBA	**83.70** 0.412	83.11 0.381	83.88 0.375	83.21 0.403	85.43 0.351	85.67 0.333
`OOD-PL` (Ours)	83.59 5.007	84.14 4.662	**84.81** 3.274	82.33 2.425	84.27 2.375	**85.73** 2.314

C. BASELINES

The proposed approach utilizing *semi-informative sets* can be applied independently of the model selection. Therefore, we selected various active learning baseline methods and compared the differences before and after applying our approach to each model. The baseline methods used in this study are as follows.

- EDA [19] involves four simple operations—synonym replacement, random insertion, random swap, and random deletion—to introduce variability in text data while preserving meaning, aimed at enhancing model performance on small datasets.
- BT [14] is a method in which monolingual target-language sentences are automatically translated into the source language to create synthetic parallel data, which is then used to enhance Neural Machine Translation (NMT) quality without altering the NMT architecture.
- SSMBA [9] uses a corruption function to perturb data off the manifold and a reconstruction function, typically a masked language model, to project the corrupted data back onto the manifold, generating augmented examples that improve robustness for out-of-domain data.

D. EVALUATION METRICS

To evaluate performance on external data, we employed two metrics: accuracy and empirical risk. Our chosen main task was sentiment analysis, a traditional text classification task, so we selected accuracy as the primary metric, as commonly used in previous studies. Additionally, since the domains of the datasets used for training and validation differ, we introduced empirical risk to assess the potential risks associated with training on a different dataset. These metrics were derived from the loss function during predictions on the validation data, measuring the error between the model's predictions and the actual values. The empirical risk for each

TABLE 2. Performance comparison between the augmentation methods and the proposed `OOD-PL` method based on the same train and validation datasets. In each cell, the top value represents accuracy, and the bottom value represents empirical risk. The augmentation method with the highest accuracy for each training dataset proportion is highlighted in bold, while the method with the lowest accuracy is underlined.

Train & Valid Dataset	SST-2			IMDB			Yelp		
Train Dataset (%)	0.5	1.0	5.0	0.5	1.0	5.0	0.5	1.0	5.0
No Augmentation	81.76	84.51	86.37	81.90	81.67	83.70	90.45	91.29	93.98
	1.154	1.161	0.981	1.186	1.192	1.139	1.082	0.925	1.065
EDA	60.75	68.05	77.08	57.75	81.53	87.59	92.63	93.04	94.02
	0.663	0.606	0.545	0.674	0.426	0.336	0.224	0.213	0.190
BT	61.93	71.46	82.23	**82.87**	**86.47**	86.18	**92.98**	**93.78**	**94.33**
	0.650	0.585	0.394	0.420	0.376	0.349	0.199	0.179	0.167
SSMBA	73.10	75.51	80.36	81.30	85.88	87.17	92.39	93.03	93.81
	0.641	0.577	0.491	0.481	0.398	0.339	0.214	0.173	0.201
`OOD-PL` (Ours)	**84.86**	**86.11**	**88.42**	82.50	84.67	**88.90**	92.30	92.54	93.99
	1.159	1.146	1.125	1.935	1.432	1.332	2.372	2.347	2.339

batch, with the number of data points in the batch denoted as n, is mathematically defined as follows.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

$$empirical\ risk = \frac{1}{n}\sum_{i=1}^{n} loss(f(x_i), y_i). \tag{5}$$

Given the sensitivity of low-resource environments to data selection and random seed variation, all evaluation experiments were repeated five times, and comparisons were based on the average performance values.

V. DISCUSSION

A. EXPERIMENTAL RESULTS

To evaluate domain generalization performance, we conducted experiments comparing our approach to baseline data augmentation methods. The general format of the experiment involved using a portion of a specific dataset for training and then measuring performance on different validation datasets. The selected proportions of the training dataset were 0.5%, 1.0%, and 5.0%. We selected each of the three datasets introduced earlier as the training dataset and evaluated performance on the other datasets. In situations where the available dataset for training is limited, we focused on conducting experiments to assess how much the text classification performance can be improved by additionally incorporating out-of-domain samples. The results of these experiments are presented in Table 1.

As shown in the table, our method outperformed other baseline augmentation methods in 10 out of 18 cases. Unlike other augmentation methods, this indicates that training with the out-of-domain samples we specified yielded more favorable results, even though the model was not directly trained on texts identical to the valid dataset. We observed that the proposed methodology performed better in various scenarios, even in restrictive situations where direct in-domain datasets could not be utilized. Notably, when using SST-2 as the training dataset, other augmentation methods generally resulted in scores ranging from the 50s to 60s, which were significantly lower than when no augmentation method was applied. However, despite using only 0.5% to 1.0% of the SST-2 dataset for training, our proposed method demonstrated superior performance on validation datasets from different datasets. In particular, it showed an improvement of nearly 20-30 points compared to *EDA*, which had the lowest performance, and overall, the scores were higher than those without augmentation methods.

When IMDB and Yelp were used as training datasets, our method consistently outperformed other augmentation techniques in most cases. In particular, when the training dataset was maximally utilized at 5.0%, our method consistently achieved the highest performance compared to other methods. In cases where these datasets were used for training, performance was generally lower without utilizing any augmentation methods. While methods such as *EDA*, *BT*, and *SSAMBA* occasionally showed the best performance in certain cases, their results varied significantly depending on how much of the dataset was used for training, often yielding the lowest performance in some cases.

Additionally, when comparing the augmentation methods in terms of *empirical risk*, our proposed method generally showed higher values compared to both the baseline methods and cases where no augmentation was applied. Traditional methods often increase data volume by introducing noise within the existing dataset, while our method leverages external datasets such as WMT-14 and AG-News, allowing the model to learn additional patterns not present in the original dataset. Despite the higher empirical risks compared to other augmentation techniques, our method demonstrated superior accuracy in domain generalization. This suggests that the proposed `OOD-PL` method is not confined to a specific domain but is well-suited for a broader range of data types.

B. TRAINING THE SAME DOMAIN DATASET

Since our proposed `OOD-PL` method incorporates external data, there were concerns about potential performance degradation within the original dataset's domain. To address this issue, unlike the previous experiments designed to assess domain generalization, we conducted experiments where the

TABLE 3. **Performance comparison between the proposed** `OOD-PL` **method combined with existing augmentation methods and the use of a single augmentation method. The train dataset was fixed to IMDB, while different datasets were used as the validation datasets. The augmentation method with the highest accuracy for each training dataset proportion is highlighted in bold.**

Valid Dataset	SST-2			Yelp		
Train Dataset (%)	0.5	1.0	5.0	0.5	1.0	5.0
EDA	78.59	81.77	77.74	85.17	**89.72**	88.84
BT	78.81	83.18	83.58	85.09	89.56	88.84
OOD-PL	76.32	**84.67**	**84.59**	84.55	89.20	**89.84**
EDA +OOD-PL	78.61	77.72	77.63	86.08	85.00	84.73
BT +OOD-PL	**82.91**	81.96	83.30	**88.27**	86.47	86.09

train and validation datasets were identical. The results of this experiment are shown in Table 2.

As seen in Table 1, a similar pattern emerged in this experiment, where using our proposed method led to a relative increase in empirical risk, but also resulted in improved accuracy. In experiments with the IMDB and Yelp datasets, the results from *BT* were notably strong. This may be because, unlike *EDA*, which involves adding or removing characters, *BT* does not introduce additional noise but simply translates the sentence while preserving its original meaning and structure, making it more advantageous when both training and validation are conducted on the same dataset.

When SST-2 was used as both the training and validation dataset, our proposed method consistently outperformed *EDA*, the worst-performing method, by more than 11 points. Even when compared with *BT*, which showed strong results on the IMDB and Yelp datasets, the difference was only 1-2 points, demonstrating that our method achieves comparable performance even when training and validating on the same dataset.

C. COMBINING THE AUGMENTATION METHODS

As demonstrated in previous results, our proposed augmentation method OOD-PL achieved superior domain generalization performance even with a small training dataset. Since it can be independently applied regardless of whether other augmentation methods are used, it is compatible with existing methods. The results of combining our proposed method with the selected baseline methods are presented in Table 3.

As a result, using only our proposed augmentation method without combining it with any others generally yielded the best performance. However, when using just 0.5% of the training dataset, the *BT*+OOD-PL combination showed the highest performance, with a 2-6 points difference compared to other models. In scenarios where very limited data is available or computational resources are insufficient for extensive training, combining our method with existing augmentation techniques, as shown in this case, can positively impact domain generalization even with minimal data.

TABLE 4. **Performance differences based on the criteria for sample selection in curriculum sampling when applying the proposed** `OOD-PL` **method. The experiment used the Yelp dataset as the train dataset and the IMDB dataset as the validation dataset. The best-performing cases, depending on the size of the train dataset, are highlighted in bold, with the values in parentheses indicating the performance differences compared to *Top*-1. Similar performance differences were observed when considering sample selection across other dataset combinations as well.**

Train Dataset	*Top*-1	*Bottom*-1	*Curriculum*-1
1%	83.14	**83.33** (+0.19)	83.25 (+0.11)
5%	85.47	84.30 (-1.17)	**86.25** (+0.78)
10%	84.91	83.11 (-1.80)	**86.00** (+1.09)

D. IMPACT OF EARLY STOPPING

To prevent degrading performance within the original domain of our existing dataset while incorporating external datasets for augmentation, we introduced the concept of early stopping in the proposed OOD-PL method. After reaching this early stopping point, we began augmenting with external datasets. The performance differences observed based on whether early stopping was applied can be seen in Fig. 3.

The experimental results show that incorporating the proposed OOD-PL method at the early stopping point during training led to an average performance improvement of 6 points compared to applying the method from the start of training. In addition to improved performance, the variation was also reduced by 3.1 percentage points, yielding a more stable result with a standard deviation of 5.1 percentage points. These findings suggest that the timing of introducing external datasets, based on the early stopping point, can significantly affect model performance. It allows for sufficient learning on the original data while enhancing the model's domain generalization capabilities.

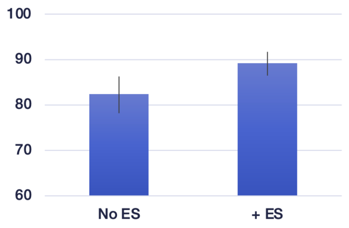

FIGURE 3. **Performance differences based on the use of early stopping (ES) when applying the proposed** `OOD-PL` **method. In this experiment, 5% of the IMDB dataset was used as the training dataset, and the Yelp dataset was used for validation. Similar performance differences were observed when considering early stopping with other dataset combinations as well.**

E. IMPACT OF CURRICULUM SAMPLING

We also utilized semantic similarity to prioritize *OOD* samples with higher similarity levels during the process of leveraging external datasets. The performance differences, depending on how samples were selected in curriculum sampling process, are presented in Table 4. *Top*-1 and *Bottom*-1 represent the selection of the highest and lowest similarity samples, respectively, while *Curriculum*-1 refers to the selection of a single sample based on curriculum sampling.

The experimental results indicate that sampling only the highest and lowest similarity samples generally yielded worse performance compared to the approach that applied curriculum sampling. When only 1% of the train dataset was used, there was no noticeable performance change due to the limited amount of data. However, as the data amount increased to 5% and 10%, learning based on curriculum sampling became more effective. This suggests that as the amount of data available for augmentation increases, the proposed curriculum sampling method can effectively utilize semantically similar data from the external dataset, leading to meaningful improvements in data augmentation.

VI. THREATS TO VALIDITY

In implementing data augmentation using the proposed OOD-PL method, we made several critical decisions regarding the experimental components. Recognizing that these choices can directly affect the results, we consider them potential threats to validity and have thoroughly analyzed how their selection may impact the outcomes.

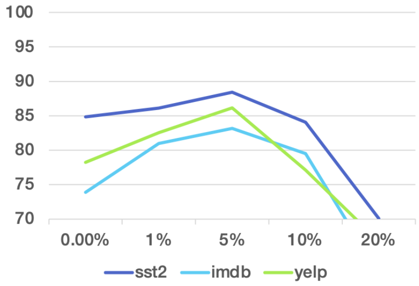

FIGURE 4. Performance differences of the model based on the ratio of *OOD* samples. In this experiment, 5% of the SST-2 dataset was used as the training dataset, while all three datasets were utilized for validation. Similar performance trends were observed when considering different dataset combinations and sample ratios.

A. INTERNAL VALIDITY

To perform data augmentation, we were able to select the number of *OOD* samples to be augmented. To observe any performance differences that might arise from changing this number, we conducted additional experiments. The results of analyzing the model's performance with varying sample sizes can be found in Fig. 4. As a result, when the sample size was increased up to 5%, performance gradually improved. However, beyond this point, as more *OOD* samples were used, performance began to decline. This suggests that an excessive amount of *OOD* samples may negatively impact model training by hindering the model's ability to properly learn from the in-domain data. Even with just 1% of the samples, performance showed notable improvement, but our experiments confirmed that 5% is the most optimal value for data augmentation.

Additionally, as we incorporated external datasets, we considered pseudo-labeling for the selected samples and conducted experiments to observe performance differences based on the labeling method. We considered both hard and soft labeling approaches, and the results of these experiments can be seen in Fig. 5. The soft label approach demonstrated an average performance improvement of about 1 point compared to the hard label method, with relatively lower performance variance, indicating its effectiveness in enhancing the model's generalization ability and stability. This is consistent with the approach taken by SoftEDA [24], which improves performance by assigning soft labels to mitigate semantic damage caused by EDA [19]. As a result, the model was better able to generalize to external datasets and maintain more stable performance.

TABLE 5. Performance differences in data augmentation using the proposed OOD-PL method based on model parameter size. The experiment utilized 1% of the SST-2 dataset as the train dataset, while all three datasets were used for validation.

	DistilBERT-base	BERT-base	BERT-large
Model Parameters	66M	110M	340M
Accuracy on Valid Dataset			
SST-2	57.63	86.11	87.38
IMDB	54.99	80.99	81.85
Yelp	51.60	82.53	85.38

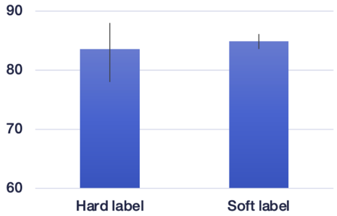

FIGURE 5. Variation in model performance depending on the choice of pseudo-labeling method, with the dashed line representing performance variance. The experiment utilized 1% of the Yelp dataset as the training dataset and the IMDB dataset for validation. Consistent results were observed when considering sample selection across different dataset combinations.

B. EXTERNAL VALIDITY

We chose text classification as the primary task for data augmentation in conjunction with baseline methods. This single-task focus led to a dependent selection of key models and datasets, which were reflected in our experimental setup. To explore whether the proposed OOD-PL method, with early stopping and curriculum sampling, can be broadly applied to other natural language processing tasks, it would be valuable to extend the task scope to areas such as natural language generation. Since our proposed method is a data augmentation technique that leverages external datasets without altering the model architecture, it is expected to be flexibly applicable across different tasks.

We selected SST-2, IMDB, and Yelp, which are commonly used in text classification tasks, for our experiments. While other datasets could also be considered, we chose these because they represent text in the form of reviews from diverse domains. During the training process for each dataset, we used only a small fraction of the training data while setting the validation dataset from a different source. This approach demonstrated the effectiveness of our proposed method in promoting domain generalization.

For our text classification experiments involving data augmentation, we conducted experiments using the BERT model with 110M parameters. Since the size of PLMs can directly influence their performance, we also extended the experiments to include smaller models [13], such as DistilBERT,[3] and larger models [3], such as the BERT large variant.[4] The results of these experiments, examining the impact of PLM size, can be found in Table 5.

When using the smaller 66M-parameter DistilBERT-base model, we observed a performance drop of approximately 30 points compared to the 110M model. However, the BERT-base model with 110M parameters showed only a small performance gap compared to the 340M model. This suggests that, by employing the proposed OOD-PL method, even models of moderate size can achieve high performance. While our experiments focused primarily on BERT-based text encoder models, further experiments with a broader range of model families should be conducted in the future.

VII. LIMITATIONS AND FUTURE WORK

We discuss the limitations of the proposed OOD-PL data augmentation and outline directions for future work. First, since our method directly utilizes external datasets for data augmentation, the choice of these external datasets can significantly influence the performance. Therefore, it is crucial to explore and automate the process of selecting suitable external datasets tailored to the target dataset in order to enhance data augmentation using our method. Further investigation and analysis in this area are necessary.

As mentioned in the previous discussion of threats to validity, we also observed that the performance can vary significantly depending on the proportion of *OOD* samples used in our method. Our prior analysis was based on empirical exploration, and in practice, a more direct investigation is needed to determine the ideal proportion of *OOD* samples when applying our method to data augmentation. In conclusion, to fully optimize the use of our proposed data augmentation method, we plan to conduct further research on the automation of exploring relevant external datasets and experimental settings, which have been determined manually.

VIII. CONCLUSION

In this study, we propose an Out-of-Domain Pseudo Labeling (OOD-PL) augmentation method for efficient domain generalization learning in low-resource environments. In the data augmentation process, we not only use internal data but also strategically incorporate external data through early stopping, ensuring sufficient learning of the existing data before utilizing the external data. Additionally, we employ curriculum sampling to selectively choose the data to be used. Our goal was to enhance the model's domain generalization ability by training reliably on multiple datasets. As a result, we observed significant performance improvements with the proposed OOD-PL method compared to other baseline augmentation methodologies. Notably, even with limited training datasets, combining this method with other augmentation techniques yielded strong performance, highlighting its potential for use in low-resource environments.

Thus, the proposed OOD-PL method enables learning across a broader domain by strategically utilizing external datasets. It can operate independently of existing data augmentation methods, allowing for flexible integration with various augmentation techniques. In our experiments, we observed the advantages of our method when combined with representative augmentation approaches. However, we expect that as more diverse augmentation techniques emerge in future research, our OOD-PL method will continue to perform well in domain generalization, effectively augmenting data. Additionally, research on exploring different hyperparameters and automating the selection of appropriate datasets is considered a direction for future work.

ACKNOWLEDGMENT

(Junho Lee and Seunguk Yu are co-first authors.)

REFERENCES

[1] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna, "Findings of the 2014 workshop on statistical machine translation," in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 12–58. [Online]. Available: https://aclanthology.org/W14-3300

[2] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik, "Vicinal risk minimization," in *Proc. Adv. Neural Inf. Process. Systems. (Neurips)*, vol. 13, Jan. 2000, pp. 416–422. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2000/file/ba9a56ce0a9bfa26e8ed9e10b2cc8f46-Paper.pdf

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[4] A. Karimi, L. Rossi, and A. Prati, "AEDA: An easier data augmentation technique for text classification," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2021, pp. 2748–2754. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.234

[5] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, Dec. 2014, p. 6.

[6] D. H. Lee, "Pseudo-label: The simple and efficient semisupervised learning method for deep neural networks," in *Proc. Workshop Challenges Represent. Learn. (ICML)*, 2013, vol. 3, no. 2, p. 896. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=798d9840d2439a0e5d47bcf5d164aa46d5e7dc26

[7] D. Li and H. Zhang, "Improved regularization and robustness for finetuning in neural networks," in *Proc. Adv. Neural Inf. Process. Systems. (NeurIPS)*, 2021, pp. 27249–27262.

[8] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2011, pp. 142–150. [Online]. Available: https://aclanthology.org/P11-1015.pdf

[9] N. Ng, K. Cho, and M. Ghassemi, "SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1268–1283.

[10] I. Okimura, M. Reid, M. Kawano, and Y. Matsuo, "On the impact of data augmentation on downstream performance in natural language processing," in *Proc. 3rd Workshop Insights From Negative Results (NLP)*, 2022, pp. 88–93. [Online]. Available: https://aclanthology.org/2022.insights-1.12.pdf

[3] https://huggingface.co/distilbert-base-uncased

[4] https://huggingface.co/bert-large-uncased

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410.pdf

[12] S. Ren, J. Zhang, L. Li, X. Sun, and J. Zhou, "Text AutoAugment: Learning compositional augmentation policy for text classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9029–9043. [Online]. Available: https://aclanthology.org/2021.emnlp-main.711.pdf

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[14] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2016, pp. 86–96. [Online]. Available: https://aclanthology.org/P16-1009.pdf

[15] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[16] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642. [Online]. Available: https://aclanthology.org/D13-1170.pdf

[17] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, pp. 8052–8072, Aug. 2023. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9782500

[18] X. Wang, M. Saxon, J. Li, H. Zhang, K. Zhang, and W. Y. Wang, "Causal balancing for domain generalization," in *Proc. ICML Workshop Spurious Correlations, Invariance Stability*, Jan. 2022. [Online]. Available: https://openreview.net/pdf?id=imav8hheb2M

[19] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 6382–6388. [Online]. Available: https://aclanthology.org/D19-1670.pdf

[20] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 6256–6268. [Online]. Available: https://papers.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf

[21] H. Zhang, M. Cissé, Y. Dauphin, and D. López-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2017. [Online]. Available: https://openreview.net/pdf?id=r1Ddp1-Rb

[22] L. Zhang and K. Ma, "A good data augmentation policy is not all you need: A multi-task learning perspective," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2190–2201, May 2023. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9936613

[23] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, Sep. 2015. [Online]. Available: https://papers.nips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf

[24] J. Choi, K. Jin, J. H. Lee, S. Song, and Y. Kim, "SoftEDA: Rethinking rule-based data augmentation with soft labels," in *Proc. Int. Conf. Learn. Represent.*, Feb. 2024. [Online]. Available: https://openreview.net/pdf?id=OiSbJbVWBJT

[25] B. Haddow and P. Koehn, "Analysing the effect of out-of-domain data on SMT systems," in *Proc. 7th Workshop Stat. Mach. Transl.*, Jun. 2012, pp. 422–432. [Online]. Available: https://aclanthology.org/W12-3154/

[26] J. Choi, "AutoAugment is what you need: Enhancing rule-based augmentation methods in low-resource regimes," in *Proc. 18th Workshop Student Res.*, 2024, pp. 1–8. [Online]. Available: https://aclanthology.org/2024.eacl-srw.1/

[27] S. Ghosh, U. Tyagi, S. Kumar, C. K. Evuru, S. Ramaneswaran, S. Sakshi, and D. Manocha, "ABEX: Data augmentation for low-resource NLU via expanding abstract descriptions," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2024, pp. 726–748.

[28] Y. Yao and A. Koller, "Simple and effective data augmentation for compositional generalization," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (Long Papers)*, vol. 1, 2024, pp. 434–449. [Online]. Available: https://aclanthology.org/2024.naacl-long.25.pdf

[29] Y. Lee, D. Lee, and K. Jung, "MILAB at PragTag-2023: Enhancing cross-domain generalization through data augmentation with reduced uncertainty," in *Proc. 10th Workshop Argument Mining*, 2023, pp. 207–211. [Online]. Available: https://aclanthology.org/2023.argmining-1.24/

[30] H. Yao, "Improving domain generalization with domain relations," in *Proc. 12th Int. Conf. Learn. Represent.*, 2024. [Online]. Available: https://iclr.cc/virtual/2024/poster/19136

[31] R. Song, F. Giunchiglia, Y. Li, M. Tian, and H. Xu, "TACIT: A target-agnostic feature disentanglement framework for cross-domain text classification," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 17, pp. 18999–19007, doi: 10.1609/aaai.v38i17.29866.

JUNHO LEE received the B.S. degree in computer engineering and the M.S. degree in artificial intelligence from Chung-Ang University, Seoul, South Korea, in 2020 and 2024, respectively. His research interests include generating natural sentences through multi-modal processing, with interests in natural language processing and machine learning.

SEUNGUK YU (Graduate Student Member, IEEE) received the B.S. degree from the School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea, in 2023. He is currently pursuing the M.S. degree with the Department of Artificial Intelligence, Chung-Ang University. His research interests include natural language processing, especially in Korean, data analytics, and machine learning.

JINHEE JANG received the B.S. degree from the Department of Chinese Language and Literature, Jeju National University, Jeju-si, South Korea, in 2021. She is currently pursuing the M.S. degree with the Department of Artificial Intelligence, Chung-Ang University. Her research interests include deep learning and natural language processing, with a focus on multilingual applications.

KEUNHYEUNG PARK received the B.S. degree from the School of Korean Language and Literature, Chung-Ang University, Seoul, South Korea, in 2022. He is currently pursuing the M.S. degree with the Department of Artificial Intelligence, Chung-Ang University. His research interests include natural language processing and stylometry.

YOUNGBIN KIM (Member, IEEE) received the B.S. and M.S. degrees in computer science and the Ph.D. degree in visual information processing from Korea University, in 2010, 2012, and 2017, respectively. From August 2017 to February 2018, he was a Principal Research Engineer with Linewalks. He is currently an Assistant Professor with the Graduate School of Advanced Imaging Science, Multimedia and Film, Chung-Ang University. His current research interests include data mining and deep learning.