UNIGEN: Universal Domain Generalization for Sentiment Classification via Zero-shot Dataset Generation

Juhwan Choi[1], **Yeonghwa Kim**[1], **Seunguk Yu**[1], **Jungmin Yun**[1] and **YoungBin Kim**[1,2]

[1]Department of Artificial Intelligence, Chung-Ang University
[2]Graduate School of Advanced Imaging Sciences, Multimedia and Film, Chung-Ang University
{gold5230, movie112, bokju128, cocoro357, ybkim85}@cau.ac.kr

Abstract

Although pre-trained language models have exhibited great flexibility and versatility with prompt-based few-shot learning, they suffer from the extensive parameter size and limited applicability for inference. Recent studies have suggested that PLMs be used as dataset generators and a tiny task-specific model be trained to achieve efficient inference. However, their applicability to various domains is limited because they tend to generate domain-specific datasets. In this work, we propose a novel approach to universal domain generalization that generates a dataset regardless of the target domain. This allows for generalization of the tiny task model to any domain that shares the label space, thus enhancing the real-world applicability of the dataset generation paradigm. Our experiments indicate that the proposed method accomplishes generalizability across various domains while using a parameter set that is orders of magnitude smaller than PLMs.

1 Introduction

As the size and performance of pre-trained language models (PLMs) increase, generation of new data by using PLMs has attracted the attention of many researchers (Anaby-Tavor et al., 2020; Kumar et al., 2020; Yoo et al., 2021). While scholars have applied this method to solve data augmentation problems, in recent studies, they have started to explore zero-shot dataset generation settings (Meng et al., 2022; Ye et al., 2022a, 2023). This novel approach first generates training data from a PLM based on a specific prompt and trains a tiny task model (TAM) by using the dataset generated in the first step. This strategy facilitates effective distillation of the knowledge pertaining to the desired task from the PLM and helps train the TAM without the need for guidance from human-annotated data, thereby enabling zero-shot learning and achieving low-cost inference compared to the case in which PLMs are used directly for inference.

However, the approaches proposed thus far have relied on domain-specific prompts, for example, "*The movie review in positive sentiment is:*." Because the data generated using this prompt are related only to the domain of movie reviews, the TAM trained on these data has limited generalization ability across other domains. This is the primary limitation of the TAM-based approach compared to prompt-based zero-shot learning that directly uses PLMs (PROMPTING), which allows for generalizability across diverse domains. This restricts the real-world applicability of the TAM-based approach because it requires many separately trained TAMs for various domains. Moreover, as the costs of dataset generation and TAM training increase, the cost-efficiency of the TAM-based approach may decrease. Hence, a novel strategy is desired to effectively distill the domain generalizability of large-scale PLMs into TAMs while maintaining the cost-efficiency of TAMs.

Meanwhile, the existing approaches to domain generalization often require multiple source domains (Wang et al., 2022; Zhou et al., 2022). This requirement limits the application of these methods because it is difficult to gather the required data from multiple domains. Although the concept of single-domain generalization, which achieves domain generalizability by using data from only one source domain, has been proposed in recent computer vision studies, such a concept is yet to be explored for natural language processing (Qiao et al., 2020; Wang et al., 2021).

In this study, we propose a simple but effective method called UNIGEN to solve the problem of domain generalizability between PLMs and TAMs. Table 1 presents a comparison between UNIGEN and the existing approaches. UNIGEN first focuses on generating a domain-invariant training dataset that is not restricted to specific domains. This allows TAMs to achieve domain generalizability without the need for multiple source domains.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1–14

	Learning without Human-annotated Data	Domain Generalizability	Light Inference	Handling Noise of Generated Data
Task-specific Fine-tuning	✗	✗	✓	
Previous Domain Generalization (Tan et al., 2022)	✗	✓	✓	
PROMPTING	✓	✓	✗	
ZEROGEN (Ye et al., 2022a)	✓	✗	✓	✗
PROGEN & SUNGEN (Ye et al., 2022b; Gao et al., 2023)	✓	✗	✓	✓
UNIGEN (Ours)	✓	✓	✓	✓

Table 1: Comparison between previous approaches and UNIGEN.

We extend domain generalization strategies based on supervised contrastive learning (Khosla et al., 2020), as suggested in a previous work (Tan et al., 2022). Moreover, we employ additional tactics such as momentum encoder (He et al., 2020) and denoised memory bank, in addition to the method suggested by the previous work (Tan et al., 2022). Furthermore, because the PLM-based dataset generation method can generate noisy data (Ye et al., 2022b; Gao et al., 2023; Zou et al., 2024), we propose a pseudo-relabeling-based additional denoising method.

Our experiments show that UNIGEN achieves generalizability across various domains and outperforms PROMPTING. This indicates that smaller TAMs can be used universally in various domains, thereby reducing the costs of PROMPTING, dataset generation, and TAM training.

Our contributions are summarized as follows:

- We propose UNIGEN, a universal domain generalization strategy by using zero-shot dataset generation.
- We develop a pseudo-relabeling-based method for denoising the generated data.
- Our extensive experiment reveals that the TAM trained using UNIGEN has domain generalizability, and it can outperform the PLM with considerably fewer parameters.

2 Related Work

2.1 Dataset Generation for Efficient Zero-shot Learning

The evolution of PLMs in terms of parameter size and performance has facilitated zero-shot learning through the use of well-designed prompts (Radford et al., 2019; Brown et al., 2020). However, it is expensive to directly deploy these massive models into daily services because the process requires numerous rounds of inference. Dataset generation mitigates this problem through the generation of training datasets by using PLMs and training a small TAM on the generated datasets (Meng et al., 2022; Ye et al., 2022a). This TAM is deployed in downstream tasks to reduce inference costs and improve performance compared to PROMPTING.

However, mere generation, that is, ZEROGEN, yields noisy data, such as incorrectly labeled data or irrelevant data (Ye et al., 2022b; Gao et al., 2023). PROGEN (Ye et al., 2022b) proposed to alleviate this problem by adding examples based on in-context feedback. Meanwhile, SUNGEN (Gao et al., 2023) proposed to re-weigh the generated samples during training using noise-robust loss. Additionally, a concurrent study suggested to leverage multiple PLMs as data generator and assign weight to generated samples in single training procedure, different from SUNGEN (Zou et al., 2024).

In this work, we propose a novel approach to extend dataset generation for universal domain generalization that is not restricted to specific training source data, as well as a pseudo-relabeling-based method to denoise the generated dataset.

2.2 Methods for Learning from Noisy Data

Researchers have explored various methods to mitigate noisy label data, which is wrongly labeled from ground-truth labels (Song et al., 2023). A relevant study in this field defined two types of noisy labels and evaluated the effectiveness of various methods with respect to BERT model (Agro and Aldarmaki, 2023). Another study proposed to leverage GPT-4 to provide the guidance to noisy labeled data (Wang et al., 2023). However, they suffer from the necessity of massive LLMs that demand cost. Moreover, these studies primarily focused on the human-crafted noisy label, rather than the noisy label of data generated by PLMs.

In this work, we suggest a straightforward method to handle noisy data based on pseudo-relabeling, particularly designed for synthetic data.

2.3 Domain Generalization for Text Classification

Domain generalization aims to improve the generalization ability in the target domain by employing source data from multiple domains to mitigate the domain shift problem (Wang et al., 2022; Zhou et al., 2022). This domain shift can be observed in natural language processing tasks, such as restaurant reviews and reviews of consumer electronics. For example, *long waiting time* in a restaurant's reviews can represent a negative sentiment about the restaurant, while *long battery life* in a laptop's reviews can represent a positive sentiment of the laptop (Tan et al., 2022).

Previous studies to alleviate domain shift in text classification have focused primarily on domain adaptation setting, for which training data are needed in the target domain (Chen and Cardie, 2018; Ye et al., 2020; Guo et al., 2020). Recently, researchers have explored the application of domain generalization to natural language processing tasks. A representative study applied supervised contrastive learning (Khosla et al., 2020) to achieve domain generalizability in text classification tasks (Tan et al., 2022).

In this work, we extend an existing method for domain generalization to generate datasets, including the adoption of momentum encoder (He et al., 2020), in addition to proposing a denoising memory bank to further enhance its effectiveness and handle noisy data.

3 Method

3.1 Preliminaries

3.1.1 Dataset Generation

First, we briefly explain the concept and notation of the preliminary dataset generation method, that is, ZEROGEN (Ye et al., 2022a). ZEROGEN aims to create a synthetic dataset $\mathcal{S}_{syn} = (\mathcal{X}_{syn}, \mathcal{Y}_{syn})$ by using a large-scale PLM $\mathcal{P}$ and task-specific prompt $\mathcal{T}_{task}$. For a text classification problem, a desired pseudo-label y_{syn} is first sampled from the uniform distribution across every class. Next, y_{syn} is passed to the prompt $\mathcal{T}_{task}$ to construct $\mathcal{T}_{task}(y_{syn})$, that is, the final prompt for $\mathcal{P}$. Thereafter, synthesized input data x_{syn} are generated using $\mathbf{x}_{syn} \sim \mathcal{P}(\cdot|\mathcal{T}_{task}(y_{syn}))$. Finally, $\mathcal{S}_{syn}$ is composed of these pairs of generated $(\mathbf{x}_{syn}, y_{syn})$. Notably, the domain of $\mathcal{S}_{syn}$ is defined by the structure of $\mathcal{T}_{task}$. For example, a $\mathcal{T}_{book}$ = "*The book review in <y> sentiment is:* " would harness $\mathcal{P}$ to generate $\mathbf{x}_{syn}$ about book reviews. The TAM is trained on the generated $\mathcal{S}_{syn}$ and deployed for inference instead of directly using PLMs with PROMPTING.

3.1.2 Supervised Contrastive Learning

Supervised contrastive learning (Khosla et al., 2020) is a variant of contrastive learning (Chen et al., 2020) that utilizes label values. It allows for explicit pulling of the representation of positive (i.e., same class) samples to the anchor representation while pushing negative representations away from the anchor. Studies have reported that this characteristic is valuable for domain generalization, which aims to group the representations of different domains (Kim et al., 2021; Tan et al., 2022). The supervised contrastive loss is expressed as follows:

$$\mathcal{L}_{SCL} = -\sum_{\mathbf{z}_i \in B} \frac{1}{|P(i)|} log \frac{exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau_{SCL})}{\sum_{\mathbf{z}_a \in A(i)} exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau_{SCL})} \quad (1)$$

where $\mathbf{z}$ denotes an encoded representation, and $\mathbf{z_i}$ is an anchor. $P(i) \equiv \mathbf{z}_j \in B, y_j = y_i$ is the set of positive samples for each anchor i, and $\mathbf{z}_p$ symbolizes a positive representation from $P(i)$. $A(i) \equiv \mathbf{z}_j \in B, j \neq i$ refers to the union of every sample, except the anchor, including positive and negative samples. $\mathbf{z}_a$ indicates each representation from $A(i)$. B denotes a mini-batch, and τ_{SCL} is the temperature of supervised contrastive learning.

Although supervised contrastive learning is effective, the introduction of a memory bank and momentum encoder may augment the advantages of the method (Wu et al., 2018; He et al., 2020). The potency of contrastive learning is often influenced by the size of B because a larger B may introduce more diverse negative samples. However, increasing the size of B can introduce concerns related to memory consumption. A memory bank is a mechanism that fulfills this demand for a greater number of negative samples by storing previously processed samples within the dictionary M. Memory-efficient contrastive learning can be achieved using this dictionary with the current batch, that is, establishing a union of B and M instead of solely using B to construct $P(i)$ and $A(i)$. Momentum encoder is another technique that smooths the process of updating the representations

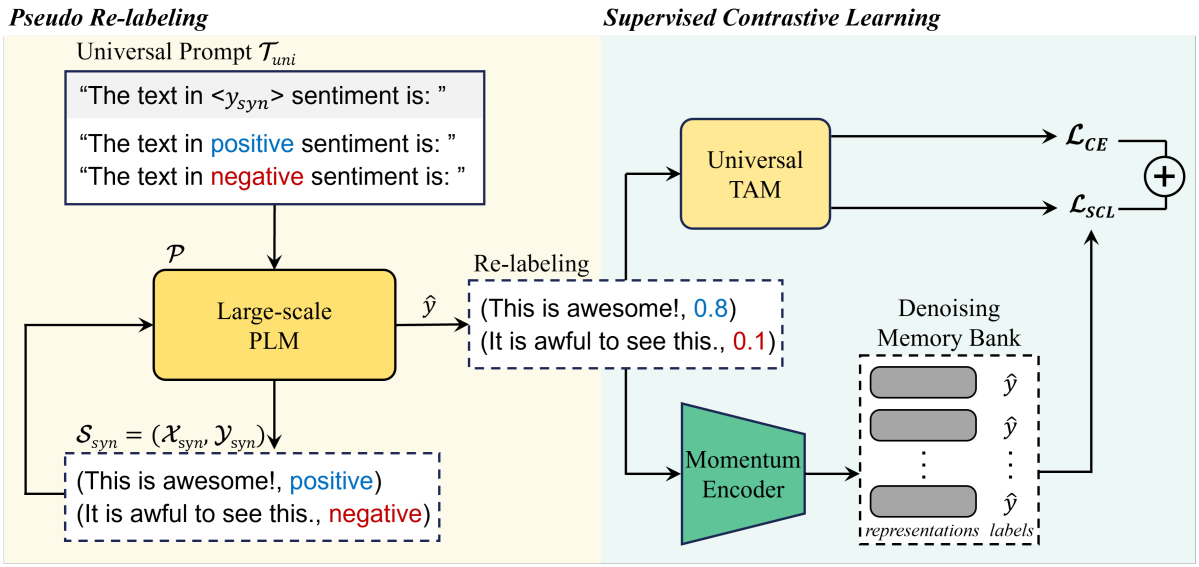

Figure 1: Overall framework for generating a dataset and training a TAM using UNIGEN.

stored in M. The momentum encoder θ_k is trained by momentum update, $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$, where m is a coefficient for momentum update, and θ_q is a normal encoder that is updated through backpropagation. By using the momentum encoder, the representations in M are processed by θ_k.

3.2 UNIGEN

To build a TAM that can be applied universally to various target domains, UNIGEN generates a domain-invariant dataset by using the universal prompt $\mathcal{T}_{uni}$, instead of task-specific $\mathcal{T}_{task}$. Consider "*The text in <y> sentiment is:*" as an example of $\mathcal{T}_{uni}$. Next, the final input prompt for $\mathcal{P}$ is constructed as $\mathcal{T}_{uni}(y_{syn})$. The synthesized input data x_{syn} are generated by following the same process as that of ZEROGEN:

$$\mathbf{x}_{syn} \sim \mathcal{P}(\cdot|\mathcal{T}_{uni}(y_{syn})) \tag{2}$$

This configuration of prompt design allows us to generate a sentence with the desired label without being restricted to any specific domain. Therefore, it steers $\mathcal{P}$ to generate various sentences within a predefined label space. This domain-invariant data generation allows the TAM trained using UNIGEN to learn the domain-invariant characteristics of the desired label space, thereby resulting in generalizability across the domains that share the label space. Supervised contrastive loss is applied along with conventional cross entropy loss to aid this process. The training loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{SCL} \tag{3}$$

where α is a hyperparameter that balances the ratio between the two losses.

3.3 Handling Noisy Data through Relabeling

However, the application of $\mathcal{T}_{uni}$ instead of $\mathcal{T}_{task}$ might lead to the generation of noisy sentences, which was noted as a drawback of ZEROGEN. This is because $\mathcal{T}_{uni}$ does not have a specific topic to guide the generation process. Furthermore, a previously developed approach to effectively mitigate this problem is applied in the training phase but not the generation phase. Therefore, there is scope to improve the quality of $\mathcal{S}_{syn}$ (Gao et al., 2023). This problem highlights the necessity to use a denoising scheme in the generation procedure. In the present work, we propose a pseudo-relabeling-based denoising process for dataset generation. In a previous study, the approach of relabeling the generated data and assigning soft labels for data augmentation was proposed (Yoo et al., 2021). Herein, we first perform pseudo-relabeling by using $\mathcal{P}$:

$$\ell(y_i|\mathbf{x}_{syn}) = \mathcal{P}(\mathcal{M}(y_i)|\mathcal{T}_{uni}(\mathbf{x}_{syn})) \tag{4}$$

where $\mathcal{M}(\cdot)$ denotes a verbalizer that transforms each label y_i into a word. We share $\mathcal{T}_{uni}$ between this process and the generation process. These logit values yielded by $\mathcal{P}$ are normalized using the softmax function with the temperature τ_{RE} :

$$\hat{y}_i = p(y_i|\mathbf{x}_{syn}) = \frac{exp(\ell(y_i|\mathbf{x}_{syn})/\tau_{RE})}{\sum_j exp(\ell(y_j|\mathbf{x}_{syn})/\tau_{RE})} \quad (5)$$

Finally, we assign $\hat{y}_i$ instead of the predefined y_{syn} to the generated $\mathbf{x}_{syn}$. This provides two distinct advantages: (1) because $\hat{y}_i$ is a soft label rather than a hard label, it contains richer information about $\mathbf{x}_{syn}$, such as the degree of the desired label, which enhances the effectiveness of training (Szegedy et al., 2016). (2) Because it relabels the generated $\mathbf{x}_{syn}$ and replaces the predefined y_{syn}, it can solve the noisy label issue, which results in the generation of $\mathbf{x}_{syn}$ that does not correspond to the designated y_{syn}, as pointed out in previous work (Gao et al., 2023). We validate the effectiveness of this relabeling strategy in the ablation study described in Section 4.5.1.

Furthermore, we discard $\mathbf{x}_{syn}$ if its pseudo-label $\hat{y}_i$ does not exceed the threshold T_{RE} to enhance the quality of $\mathcal{S}_{syn}$. This guarantees that only those data that have the desired degree of each label are maintained.

3.4 Denoising Memory Bank

In addition to the relabeling strategy, we propose a denoising memory bank mechanism to further alleviate the issue of noisy data. We first use SUNGEN (Gao et al., 2023) that learns weights of each training sample $\mathbf{w}$ for loss function within the training process to assign small weights to noisy data by employing a noise-robust loss function. We aim to ensure that the memory bank M contains clean samples, rather than noisy samples. We utilize the weights $\mathbf{w}$ learned from the noise-robust loss function for this purpose. In the process of updating M, we store only those samples whose weights are larger than the threshold T_{MB}. This organization of the memory bank ensures the exclusion of noisy samples from the comparison, resulting in higher-quality negative and positive samples (Robinson et al., 2021).

4 Experiment

4.1 Experimental Setup

In this section, we briefly explain the experimental setup used herein to validate the effectiveness of UNIGEN. We employ seven different sentiment classification datasets in our main experiment. Among them, IMDB (Maas et al., 2011), SST-2 (Socher et al., 2013), and Rotten Tomatoes (Pang and Lee, 2005) are datasets comprising movie reviews. Meanwhile, the Amazon (McAuley and Leskovec, 2013) dataset consists of customer reviews of various products, and the Yelp (Zhang et al., 2015) dataset is composed of restaurant reviews. CR (Ding et al., 2008) is another customer review dataset focusing on consumer electronics. Lastly, Tweet (Rosenthal et al., 2017) is composed of messages from Twitter. This configuration allows us to evaluate the ability of UNIGEN, which can be applied to various domains without providing any prior information or domain-specific training. Following the previous study, we adapted long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and DistilBERT (Sanh et al., 2019), and we included RoBERTa (Liu et al., 2019) as our TAM. We compared our approach to ZEROGEN and SUNGEN, as well as to PROMPTING using GPT2-XL (Radford et al., 2019), to ensure a fair comparison. We did not include other larger PLMs in the experiments because the previous work discovered that larger PLMs did not offer performance gains (Ye et al., 2022a). We report the average of the performance results obtained across five different random seeds.

4.2 Comparison with Task-specific TAMs

Table 2 presents a comparison between the experimental results of UNIGEN and PROMPTING and task-specific TAMs trained by ZEROGEN and SUNGEN. The comparison results suggest that UNIGEN can generalize across various domains using a *single* model *without* requiring any prior information about the test domain. Nonetheless, UNIGEN underperformed compared to the task-specific baselines in each domain. However, the primary benefit of UNIGEN lies in its unique domain generalizability while using orders-of-magnitude fewer parameters than PLMs. Additionally, its training procedure is more efficient than those of other TAM training strategies. As can be inferred from Table 3, SUNGEN generates and synthesizes 1,000k data for each task domain. This means that 5,000k data would be required for our experiment, which involves five different domains, in addition to individual denoising processes for finding the best weights of the samples in each of these domains. By contrast, UNIGEN is not limited by such restrictions and requires only a single data generation and denoising process, as well as a single training process. This is extremely beneficial when a novel test

Model Test Domain	#Param	Training Domain	Setup	SST-2	IMDB Movie	Rotten	Amazon Products	Yelp Restaurant	CR Electronics	Tweet Tweet	Average
GPT2-XL	1.5B	-	PROMPTING	82.15	70.26	77.56	79.06	78.04	80.30	80.38	78.25
LSTM	7M	Movie	ZEROGEN	75.11	66.39	69.85	67.24	70.25	69.32	63.43	68.80
			SUNGEN	**78.79**	**69.97**	**73.76**	72.15	73.21	70.39	66.84	**72.16**
		Products	ZEROGEN	64.26	61.82	60.13	70.32	67.78	69.46	62.29	65.15
			SUNGEN	67.83	63.87	63.46	**74.43**	73.71	73.35	63.51	68.59
		Restaurant	ZEROGEN	67.41	63.01	62.74	68.73	75.51	69.23	66.35	63.28
			SUNGEN	69.15	66.62	64.56	73.22	**79.56**	70.12	67.43	70.09
		Electronics	ZEROGEN	64.69	59.13	60.20	66.34	67.72	72.50	60.25	64.40
			SUNGEN	68.38	64.33	63.25	72.61	73.01	**76.18**	66.78	69.22
		Tweet	ZEROGEN	61.84	60.17	59.43	64.13	63.68	65.02	74.10	64.05
			SUNGEN	66.57	63.96	64.21	69.36	71.68	72.57	**81.29**	69.95
		-	UNIGEN	64.15	60.02	60.51	63.82	63.20	69.61	70.32	64.52
DistilBERT	66M	Movie	ZEROGEN	80.06	69.13	74.73	73.02	72.77	73.59	74.83	74.02
			SUNGEN	**82.43**	**70.59**	**76.37**	74.13	73.56	75.14	75.96	75.45
		Products	ZEROGEN	71.04	64.99	65.57	74.54	71.89	74.57	71.93	70.65
			SUNGEN	72.35	65.95	66.84	**76.92**	74.98	75.84	73.01	72.27
		Restaurant	ZEROGEN	77.32	65.47	68.86	74.01	77.94	74.89	73.74	73.18
			SUNGEN	78.93	67.12	69.92	74.93	**80.67**	76.06	75.28	74.70
		Electronics	ZEROGEN	73.77	66.14	66.78	72.38	73.21	78.82	74.58	72.24
			SUNGEN	74.49	67.19	68.29	73.49	75.34	**80.49**	75.37	73.52
		Tweet	ZEROGEN	73.98	66.58	67.43	72.88	71.86	75.68	80.86	72.75
			SUNGEN	75.12	67.53	69.06	73.64	72.73	78.17	**82.46**	74.10
		-	UNIGEN	77.67	67.81	73.16	75.06	74.81	79.86	81.41	**75.68**
RoBERTa	110M	Movie	ZEROGEN	84.38	73.03	78.38	77.38	76.83	77.36	77.94	77.90
			SUNGEN	**85.24**	**74.09**	**79.19**	78.56	77.61	78.21	79.72	78.95
		Products	ZEROGEN	79.14	71.16	70.92	79.94	75.79	76.35	80.17	76.21
			SUNGEN	81.51	71.28	72.67	**81.50**	77.76	78.55	81.94	77.87
		Restaurant	ZEROGEN	82.87	70.71	69.58	78.61	81.47	76.43	79.51	77.03
			SUNGEN	83.65	71.40	71.05	79.42	**82.72**	77.60	80.92	78.11
		Electronics	ZEROGEN	76.82	69.42	67.89	75.02	76.53	81.24	76.51	74.78
			SUNGEN	77.51	71.23	68.77	76.91	78.33	83.49	79.03	76.47
		Tweet	ZEROGEN	78.43	68.31	72.25	78.09	74.61	79.08	82.96	76.25
			SUNGEN	82.19	70.62	73.21	79.84	76.27	81.46	83.25	78.12
		-	UNIGEN	84.86	72.24	78.82	80.79	79.15	**86.37**	**87.89**	**81.45**

Table 2: Experimental results of UNIGEN and baselines across various datasets and training domains. The performance of TAM, which is superior to that of PROMPTING, is underlined, and the best result in each test dataset within the group for each TAM is presented in boldface.

	Amount of generated data	Number of trained TAMs
ZEROGEN	1,000k	5
SUNGEN	5,000k	5
UNIGEN	1,000k	1

Table 3: Amount of data generated for training TAMs by using each method, and number of trained TAMs per method.

domain is introduced, where ZEROGEN and SUNGEN necessitate a separate procedure for the new domain, but UNIGEN directly reuses the already trained TAM.

Notably, the performance of the LSTM-based TAM trained using UNIGEN was significantly lower than that of ZEROGEN and SUNGEN. This implies that while a small-sized TAM can be trained effectively for a single, specific domain, but suffers from generalizing to a universal domain that requires a broad understanding of generated data, as evidenced by detailed study in Appendix E.

Accordingly, the performance of the TAM trained using UNIGEN improves significantly as the model size increases. For instance, the DistilBERT-based TAM trained using UNIGEN exhibited the best average performance against each task-specific baseline. This is particularly remarkable as it outperformed the SUNGEN baseline in the movie domain, which has three in-domain datasets, giving it an inherent advantage for average performance. Moreover, the RoBERTa-based TAM trained using UNIGEN not only yielded the best average performance against these baselines but also outperformed PROMPTING in every domain. This result indicates that it can surpass the zero-shot performance of its PLM counterpart (e.g., GPT2-XL) while using less than 10% of the number of parameters and securing the domain generalizability of the PLM, extending the achievement of the previous study that leveraged small TAMs in single domain (Ye et al., 2022a).

RoBERTa	DVD	Electronics	Kitchen	Book	Average
PROMPTING w/ GPT2-XL	77.73	78.71	81.64	80.27	79.59
UNIGEN	78.14	80.68	82.31	80.93	80.52
SUPERVISED (Tan et al., 2022)	91.40	95.10	95.05	93.25	93.70

Table 4: Experiments conducted using multi-domain review dataset. The experimental result of SUPERVISED was reported in a previous study (Tan et al., 2022) with the memory bank size of 64.

4.3 Comparison with Supervised Domain Generalization Method

Next, we analyzed the performance of UNIGEN against that of a domain generalization method that uses human-annotated data (Tan et al., 2022). For this purpose, we used a multi-domain review dataset comprising four domains: DVD, books, kitchen and housewares, and consumer electronics (Blitzer et al., 2007). Following the previous study, we split the dataset into 1,600 training data and 400 testing data for each domain. Table 4 presents the comparison results. These results suggest that UNIGEN can be applied to various domains, and its performance is superior to that of its PLM counterpart. Notably, the SUPERVISED baseline relies on three source domains with human-annotated data to generalize to a target domain, while UNIGEN is based on zero-shot dataset generation and does not require any human-annotated data, which greatly improves its real-world applicability.

4.4 Domain Generalizability of UNIGEN

To intuitively examine the domain generalizability of UNIGEN, we plotted the T-SNE (Van der Maaten and Hinton, 2008) visualization of the features interpreted by the RoBERTa-based TAM trained using UNIGEN. Figure 2 depicts the visualization results. These results suggest that the single TAM classified the given data from every domain *without* explicit training or prior information about the domains, thus demonstrating the unique efficiency of UNIGEN.

Table 5 presents examples of the sentences generated using UNIGEN. These examples showcase that UNIGEN can generate domain-invariant sentences with the designated labels. By training TAMs on these data, it is possible to distill the domain generalizability of PLMs into TAMs.

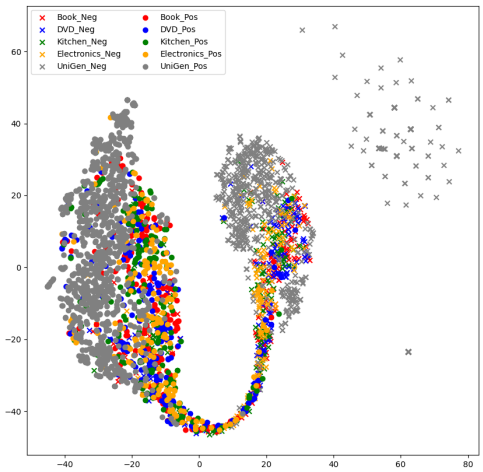

Figure 2: T-SNE visualization of the encoded representation of the RoBERTa model trained using UNIGEN. The model was trained only on the data generated using UNIGEN, which is shown in gray color. We used the test set of the multi-domain review dataset.

4.5 Ablation Study

This section describes the ablation studies conducted to offer rationales for the engineering choices made in this study. We used the DistilBERT-based TAM for these experiments.

4.5.1 Effectiveness of Relabeling Strategy

First, we performed an ablation study to validate the effectiveness of the relabeling strategy discussed in Section 3.3. We compared the basic approach that uses soft labels to the two other options. The first option utilizes the pseudo-relabeling process, but it assigns hard labels instead of soft labels. In other words, it only reflects the decision emanating from the PLM, not the probability. The second option completely excludes the relabeling process. While this option would generate the dataset faster than the other options, it might generate text with noisy labels, as already discussed in previous works (Ye et al., 2022a,b; Gao et al., 2023).

The experimental results are presented in the second and third rows of Table 6. They suggest that the use of soft labels offers practical benefits in terms of performance. This finding is consistent with that of a previous study in which the strength of soft labels was demonstrated (Yoo et al., 2021; Fang et al., 2024). Therefore, according to the results of this ablation study, relabeling the generated data with the assignment of soft labels is effective for mitigating the issue of noisy labels.

Positive Examples	Labels
You are a person who is hardworking, honest, and reliable. You have a good sense of humor, and you love being in charge.	$[0.19, 0.81]$
You are beautiful, you are powerful, you are amazing.	$[0.29, 0.71]$
In a city full of great ideas and creativity, I've met a few people who have done things you wouldn't believe.	$[0.26, 0.74]$
The American Dream is alive in this great city. As a new generation of American heroes begins to realize their own American Dream.	$[0.24, 0.76]$
Negative Examples	**Labels**
No one likes it. Nobody wants it. It is a disgrace.	$[0.7, 0.3]$
The company is no longer in business and has ceased operations.	$[0.71, 0.29]$
Please don't use this feature to communicate with customers	$[0.74, 0.26]$
Do not buy from this seller.	$[0.79, 0.21]$

Table 5: Examples of the data generated using UNIGEN.

DistilBERT	SST-2	IMDB	Rotten	Amazon	Yelp	CR	Tweet	Average
UNIGEN	77.67	67.81	73.16	75.06	74.81	79.86	81.41	**75.68**
UNIGEN w/ Hard Relabeling	77.18	67.18	72.37	72.91	72.95	78.14	80.39	74.45
UNIGEN w/o Relabeling	76.34	66.58	71.78	70.63	70.97	76.59	79.62	73.22
UNIGEN w/o Denoising MB	77.06	67.13	72.04	74.69	73.66	78.47	80.84	74.84
UNIGEN w/o SCL	75.53	66.10	69.63	71.43	69.58	77.22	79.31	72.69
Combined Prompts	74.19	63.16	71.08	73.62	72.93	78.05	78.02	73.01

Table 6: Results of ablation studies on methodological choices in Section 4.5.1, 4.5.2, and 4.5.3.

DistilBERT	SST-2	IMDB	Rotten	Amazon	Yelp	CR	Tweet	Average
UNIGEN w/ GPT2-XL	77.67	67.81	73.16	75.06	74.81	79.86	81.41	**75.68**
UNIGEN w/ Gemma-2b	71.50	69.40	67.04	76.48	76.89	77.24	52.03	70.08
UNIGEN w/ Qwen2-1.5B	66.37	63.19	63.76	71.69	72.44	66.06	63.49	66.71
UNIGEN w/ Phi-1.5	74.98	68.35	70.82	73.86	75.11	71.82	84.01	74.13

Table 7: Results of ablation studies on comparison between various PLMs in Section 4.5.4.

4.5.2 Effectiveness of Supervised Contrastive Learning and Denoising Memory Bank

Second, we conducted a comparison to investigate the effectiveness of supervised contrastive learning, which was discussed in Section 3.1.2, and denoising memory bank, which was discussed in Section 3.4. The results of the comparison are presented in fourth and fifth rows of Table 6. Intuitively, if the quality of each of the data in the dataset is given as a weight, it would be effective to employ only high-quality samples for comparing contrastive learning rather than utilizing all data, regardless of their quality. The experimental result in the fourth row demonstrated that the use of a denoising memory bank yielded a performance gain, which was consistent with our intuition. Similarly, the result in the fifth row suggests that supervised contrastive learning plays a crucial role in UNIGEN.

4.5.3 Comparison with Combined Domain-specific Datasets

Third, we compared the performance of the TAMs trained with two different synthetic datasets. The first uses the synthetic dataset generated with the prompt of UNIGEN, and the second uses the concatenation of datasets generated with five different domain-specific prompts used in the other experiments. For this experiment, we only differentiated the synthetic dataset used for training and set every other configuration identical, such as the usage of pseudo-relabeling and denoised memory bank, as well as other hyperparameters. The result of the ablation study is presented in the last row of Table 6. The result indicates that the model trained with the dataset generated by the universal prompt in UNIGEN demonstrated better average performance. This suggests that the broad understanding of the label space offered by the synthetic dataset generated by UNIGEN plays an important role in domain generalization.

4.5.4 Comparison between PLMs for Data Generation

Lastly, we evaluated the performance of TAMs trained using various PLMs. Initially, we utilized GPT2-XL as the PLM for data generation. In this experiment, we extended the evaluation by incorporating more recent models as data generators. Specifically, we compared the performance of TAMs trained with UNIGEN using Gemma-2b (Team et al., 2024), Qwen2-1.5B (Yang et al., 2024), and Phi-1.5 (Li et al., 2023), which are more recent models with parameter sizes comparable to GPT2-XL. All other configurations, aside from the PLM used for data generation, were kept consistent with the original GPT2-XL-based TAM.

Table 7 presents the results of this experiment. Interestingly, the findings suggest that employing more recent PLMs does not necessarily lead to better performance in UNIGEN. The TAM trained

with GPT2-XL, our original choice for data generation, achieved the highest average performance. This aligns with previous studies, which indicate that using larger PLM does not always result in superior outcomes (Ye et al., 2022a). However, despite using identical hyperparameters and prompts to ensure a fair comparison, it is important to recognize that optimal hyperparameters, such as top-k, top-p, and τ_{RE}, as well as the prompt configurations, may vary for each PLM. Future research could focus on developing a unified framework to optimize hyperparameters and prompts for each PLMs, akin to methods like AutoAugment (Cubuk et al., 2019; Ren et al., 2021).

5 Conclusion

In this study, we proposed UNIGEN in an attempt to achieve universal domain generalization. UNIGEN successfully transferred the domain generalizability of PLMs into orders-of-magnitude smaller TAMs. Moreover, human annotation was not required for UNIGEN, which significantly reduced the burden of acquiring labeled data from multiple source domains. Our relabeling method and denoising memory bank offered additional performance gains. Furthermore, our extensive experiments demonstrated that UNIGEN outperformed PROMPTING, facilitating light inference while preserving the domain generalizability of PLMs.

Although we explored an interesting framework for zero-shot, lightweight domain generalization, the performance of UNIGEN appears weaker than those of baseline models that are trained on each domain in several cases. It is desirable to achieve a higher level of performance than those of the in-domain baselines, which we will attempt in future work. To this end, the generation of small task-specific data for additional training of the TAM trained using UNIGEN is a possible approach, especially when a downstream task domain is introduced. By employing TAMs that are pre-trained using UNIGEN as a warm start, high performance could be achieved in the target domain with a small amount of task-specific data, which would reduce the total amount of data generated compared to that when individually training each TAM by using ZEROGEN or SUNGEN from scratch. Another possible approach may involve combining UNIGEN with the concept of test-time learning (Jeong et al., 2023). Similar to the first strategy, it may generate small amounts of test domain-specific data given test-time data as in-context examples. We are committed to exploring these possible strategies, which will enhance the effectiveness of UNIGEN.

Limitations

The primary limitation of UNIGEN is its relatively weaker in-domain performance than those of baselines that are trained with domain-specific datasets. While it is beneficial for its smaller parameter set and lower inference cost while maintaining the domain generalizability of PLMs, there exists a tradeoff between in-domain performance and efficiency, unlike ZEROGEN and SUNGEN. Therefore, a method for further enhancing the performance of UNIGEN should be explored, as stated in the Conclusion section. A possible solution is a proper prompt designed for UNIGEN because the quality of the generated sentences is affected by prompt design. Even though we adapted an effective prompt designed in a previous work (Ye et al., 2022a), a more effective prompt for UNIGEN that aims to generate diverse and general expressions could exist.

Ethics Statement

The data generated by the PLM may contain biased sentences, which may offend the readers. This can be attributed to the potential bias of PLMs (Liu et al., 2022). These generated biased sentences do not reflect the views of the authors.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1C1C1008534), and Institute for Information & communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program, Chung-Ang University).

References

Maha Agro and Hanan Aldarmaki. 2023. Handling realistic label noise in bert text classification. In *Proceedings of ICNLSP*, pages 11–20.

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 AAAI*, pages 7383–7390.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*, pages 440–447.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*, pages 1877–1901.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, pages 1597–1607.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of NAACL*, pages 1226–1240.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of CVPR*, pages 113–123.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of WSDM*, pages 231–240.

Tianqing Fang, Wenxuan Zhou, Fangyu Liu, Hongming Zhang, Yangqiu Song, and Muhao Chen. 2024. On-the-fly denoising for data augmentation in natural language understanding. In *Findings of EACL*, pages 766–781.

Jiahui Gao, Renjie Pi, Lin Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning. In *Proceedings of ICLR*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In *Proceedings of AAAI*, pages 7830–7838.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of CVPR*, pages 9729–9738.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Hwang, and Jong Park. 2023. Test-time self-adaptive small language models for question answering. In *Findings of EMNLP*, pages 15459–15469.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of EMNLP*, pages 4163–4174.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of NeurIPS*, pages 18661–18673.

Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. 2021. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of ICCV*, pages 9619–9628.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings AACL 2020 Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

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*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of ACL*, pages 142–150.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of RecSys*, pages 165–172.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Proceedings of NeurIPS*, pages 462–477.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of EMNLP*, pages 188–197.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115–124.

Fengchun Qiao, Long Zhao, and Xi Peng. 2020. Learning to learn single domain generalization. In *Proceedings of CVPR*, pages 12556–12565.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shuhuai Ren, Jinchao Zhang, Lei Li, Xu Sun, and Jie Zhou. 2021. Text autoaugment: Learning compositional augmentation policy for text classification. In *Proceedings of EMNLP*, pages 9029–9043.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *Proceedings of ICLR*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of SemEval*, pages 502–518.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2023. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of CVPR*, pages 2818–2826.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Domain generalization for text classification with memory-based supervised contrastive learning. In *Proceedings of COLING*, pages 6916–6926.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(86):2579–2605.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072.

Song Wang, Zhen Tan, Ruocheng Guo, and Jundong Li. 2023. Noise-robust fine-tuning of pretrained language models via external guidance. In *Findings of EMNLP*, pages 12528–12540.

Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. 2021. Learning to diversify for single domain generalization. In *Proceedings of ICCV*, pages 834–843.

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 EMNLP (Demo Track)*, pages 38–45.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of CVPR*, pages 3733–3742.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In *Proceedings of EMNLP*, pages 7386–7399.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of EMNLP*, pages 11653–11669.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022b. Progen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of EMNLP*, pages 3671–3683.

Jiacheng Ye, Chengzu Li, Lingpeng Kong, and Tao Yu. 2023. Generating data for symbolic language with large language models. In *Proceedings of EMNLP*, pages 8418–8443.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. In *Findings of EMNLP*, pages 2225–2239.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of NeurIPS*.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415.

Tianyuan Zou, Yang Liu, Peng Li, Jianqing Zhang, Jingjing Liu, and Ya-Qin Zhang. 2024. Fusegen: Plm fusion for data-generation based zero-shot learning. *arXiv preprint arXiv:2406.12527*.

A Prompt for Each Domain

Domain	Prompt
Movie	The *movie review* in [positive/negative] sentiment is:
Products	The *product review* in [positive/negative] sentiment is:
Restaurant	The *restaurant review* in [positive/negative] sentiment is:
Electronics	The *electronics product review* in [positive/negative] sentiment is:
Tweet	The *tweet* in [positive/negative] sentiment is:
UNIGEN & PROMPTING	The *text* in [positive/negative] sentiment is:

Table 8: The prompt used for each domain in ZEROGEN and SUNGEN, as well as the prompt used for UNIGEN and PROMPTING.

B Implementation Detail

For UNIGEN, we first generated 1,000k data from the 1.5B GPT2-XL model as $\mathcal{P}$ by using the prompt $\mathcal{T}_{uni}$ "*The text in positive/negative sentiment is:* ", which is a slightly modified version of the best prompt suggested in a previous study. Top-k and top-p were set to 40 and 0.9 during the generation procedure, respectively. The soft relabeling process was performed using a τ_{RE} of 0.1. After obtaining the soft labels of each of the generated samples, we filtered them using T_{RE} of 0.2. This required the largest value from the soft labels to be larger than the sum of the uniform distribution and T_{RE}, for instance, 0.7 in binary classification with T_{RE} of 0.2. As an example, the sentence corresponding to the soft label $[0.64, 0.36]$ was discarded because it did not exceed the threshold.

After generation, we followed the bi-level optimization approach proposed in SUNGEN to cleanse the generated dataset and find the sample weights for 50 epochs. The outer learning rate was set to 5e-2, and we randomly sampled 50k data for each outer validation process. Then, we selected 200k data with high weights, which represent high-quality data, to train the TAMs.

We used a one-layer bi-LSTM model for the LSTM-based TAM and the distilbert-base-uncased and roberta-base from Transformers (Wolf et al., 2020) for the DistilBERT-based TAM and RoBERTa-based TAM, respectively. We trained the LSTM-based TAM for 5 epochs with the learning rate of 1e-3 by using the Adam (Kingma and Ba, 2015) optimizer. The DistilBERT-based TAM was trained for 3 epochs with a learning rate of 2e-5 by using the Adam optimizer. The RoBERTa-based TAM was trained for 3 epochs with a learning rate of 2e-5 by using the Adam optimizer. During the training process, α for supervised contrastive learning loss was set to 0.5, with a projection size of 256. The temperature τ_{SCL} was set to 0.2, and the memory bank size M was set to 64. The coefficient m for updating the momentum encoder was set to 0.999, and the threshold of the denoising memory bank T_{MB} was set to 0.8. The dataset generation and training procedures were executed using on a single NVIDIA A100 40GB GPU. Please refer to attached source code for further details.[1]

C Extensibility of Relabeling Strategy

DistilBERT	SST-2	IMDB	Rotten	Amazon	Yelp	CR	Tweet	Average
ZEROGEN	80.06	69.13	74.73	73.02	72.77	73.59	74.83	74.02
ZEROGEN w/ Hard Relabeling	80.72	69.25	73.98	73.41	73.18	73.76	74.91	74.17
ZEROGEN w/ Soft Relabeling	81.79	70.40	75.32	73.65	73.31	74.72	75.14	**74.90**

Table 9: Experimental result on the extensibility of relabeling strategy. We trained the TAM using ZEROGEN based on the movie domain.

We examined the extensibility of the relabeling strategy discussed in Section 3.3. We applied two different options for relabeling, namely assigning hard labels and soft labels to ZEROGEN. Table 9 summarizes the results. These results suggest that the relabeling strategy is beneficial for the performance of the TAM trained using ZEROGEN. Therefore, filtering the generated data through the relabeling strategy is an extensive strategy for enhancing zero-shot learning methods based on dataset generation. Furthermore, the assignment of soft labels was more beneficial compared to the assignment of hard labels, which is consistent with the results of the ablation study described in Section 4.5.1. We will further investigate the relabeling-based approach to enhance ZEROGEN and SUNGEN in future works.

D Additional Experiment on Domain Generalizability

To further reveal the domain generalizability of UNIGEN, we conducted an additional experiment on Amazon Review dataset (Ni et al., 2019). We used 5-core data for 29 domains and reported the performance of PROMPTING using GPT2-XL (Radford et al., 2019) and RoBERTa-based TAM trained by UNIGEN. The result in Table 10 demonstrates the performance of UNIGEN that is comparable with PROMPTING, with parameters less than 10%. Additionally, this experiment showcases the universality of UNIGEN, the characteristics that distin-

[1] https://github.com/c-juhwan/unigen

Domain	PROMPTING	UNIGEN
Fashion	93.29	91.16
Beauty	95.63	92.87
Appliances	68.27	79.10
Arts, Crafts and Sewing	91.05	92.08
Automotive	91.07	88.23
Books	89.18	91.26
CDs and Vinyl	82.44	86.42
Cell Phones and Accessories	90.47	88.65
Clothing, Shoes and Jewelry	91.83	90.80
Digital Music	93.72	90.62
Electronics	88.68	88.34
Gift Cards	94.03	92.50
Grocery and Gourmet Food	92.31	91.09
Home and Kitchen	92.11	91.53
Industrial and Scientific	91.07	92.34
Kindle Store	89.49	92.76
Luxury Beauty	90.03	91.82
Magazine Subscriptions	85.97	89.64
Movies and TV	86.39	88.19
Musical Instruments	90.72	90.20
Office Products	91.74	89.60
Patio, Lawn and Garden	89.96	87.87
Pet Supplies	90.60	89.91
Prime Pantry	93.64	88.15
Software	82.55	83.39
Sports and Outdoors	88.63	90.36
Tools and Home Improvement	87.41	88.90
Toys and Games	91.54	92.02
Video Games	85.79	86.07
Average	89.30	89.51

Table 10: The result of the experiment on the Amazon Review dataset.

guish UNIGEN from previous ZEROGEN and SUNGEN. Compared to previous methods that would require 29 separately trained TAMs to conduct this experiment, UNIGEN only used one single TAM to perform the experiment, which exhibits the real-world applicability of UNIGEN.

E Additional Study on the Performance of UNIGEN on Small-sized TAMs

We found that UNIGEN suffers to exhibit its performance on the LSTM model from the experiment in Table 2. To further investigate this phenomenon, we expand our experiment into two different small-sized TAMs: TextCNN (Kim, 2014) and TinyBERT (Jiao et al., 2020). Table 11 showcases the result of the additional experiment. In the case of TextCNN-based TAM, baseline methods such as ZEROGEN and SUNGEN demonstrated comparable or slightly higher performance compared to that of LSTM-based TAM. Nonetheless, TextCNN-based TAM trained on UNIGEN reported slightly worse performance compared to LSTM-based TAM despite increased parameter size. We hypothesize that this phenomenon is owing to the architecture of TextCNN, which leverages CNN layers that have fixed window size, leading to limited ability to understand the context of diverse expression generated by UNIGEN. On the contrary, TinyBERT-based TAM trained on UNIGEN exhibited the best average performance among the baselines. Furthermore, its average performance is comparable to DistilBERT-based TAM despite a much smaller parameter size. It is noteworthy that TinyBERT is also a model that has a general understanding of the language through knowledge distillation from BERT. Through this investigation, we reveal that the pre-trained knowledge of the TAM aids the successful training of the TAM through UNIGEN.

Model Test Domain	#Param	Training Domain	Setup	SST-2	IMDB Movie	Rotten	Amazon Products	Yelp Restaurant	CR Electronics	Tweet Tweet	Average
GPT2-XL	1.5B	-	PROMPTING	82.15	70.26	77.56	79.06	78.04	80.30	80.38	78.25
LSTM	7M	Movie	ZEROGEN	75.11	66.39	69.85	67.24	70.25	69.32	63.43	68.80
			SUNGEN	**78.79**	**69.97**	**73.76**	72.15	73.21	70.39	66.84	**72.16**
		Products	ZEROGEN	64.26	61.82	60.13	70.32	67.78	69.46	62.29	65.15
			SUNGEN	67.83	63.87	63.46	**74.43**	73.71	73.35	63.51	68.59
		Restaurant	ZEROGEN	67.41	63.01	62.74	68.73	75.51	69.23	66.35	63.28
			SUNGEN	69.15	66.62	64.56	73.22	**79.56**	70.12	67.43	70.09
		Electronics	ZEROGEN	64.69	59.13	60.20	66.34	67.72	72.50	60.25	64.40
			SUNGEN	68.38	64.33	63.25	72.61	73.01	**76.18**	66.78	69.22
		Tweet	ZEROGEN	61.84	60.17	59.43	64.13	63.68	65.02	74.10	64.05
			SUNGEN	66.57	63.96	64.21	69.36	71.68	72.57	**81.29**	69.95
		-	UNIGEN	64.15	60.02	60.51	63.82	63.20	69.61	70.32	64.52
CNN	10M	Movie	ZEROGEN	74.34	67.91	70.22	68.69	71.03	70.89	64.77	69.69
			SUNGEN	**76.98**	**68.97**	**73.49**	73.04	73.97	71.55	69.43	**72.49**
		Products	ZEROGEN	63.46	62.13	60.35	70.94	68.34	72.34	65.71	66.18
			SUNGEN	65.89	63.27	61.97	**73.98**	72.81	74.02	67.38	68.47
		Restaurant	ZEROGEN	67.76	64.18	62.16	70.17	76.65	71.27	65.43	68.23
			SUNGEN	68.86	65.62	64.96	73.20	**77.87**	72.43	68.36	70.19
		Electronics	ZEROGEN	65.05	63.04	62.13	67.19	69.50	73.66	63.23	66.26
			SUNGEN	67.43	65.13	63.25	70.82	72.79	**77.42**	67.19	69.15
		Tweet	ZEROGEN	60.56	60.68	61.33	64.91	64.37	66.86	75.62	64.90
			SUNGEN	65.12	61.56	63.42	66.45	68.46	68.71	**80.17**	67.70
		-	UNIGEN	62.31	60.48	61.82	61.08	61.63	68.24	65.95	63.07
TinyBERT	14.5M	Movie	ZEROGEN	78.95	68.37	71.34	70.59	71.35	71.18	68.94	71.53
			SUNGEN	**80.78**	**69.86**	**73.47**	72.36	72.42	73.75	70.81	73.35
		Products	ZEROGEN	69.22	62.79	63.44	72.57	69.70	73.22	71.21	68.88
			SUNGEN	71.74	64.38	64.51	**75.81**	73.76	74.17	72.86	71.03
		Restaurant	ZEROGEN	75.79	64.62	65.53	71.33	77.10	73.52	70.84	71.25
			SUNGEN	77.45	67.41	68.01	74.41	**79.16**	75.86	72.11	73.49
		Electronics	ZEROGEN	71.22	64.37	63.06	69.51	70.75	75.71	69.49	69.16
			SUNGEN	73.10	65.81	66.71	71.33	74.86	78.43	73.88	72.02
		Tweet	ZEROGEN	70.76	63.40	64.43	68.74	70.44	73.72	78.14	69.95
			SUNGEN	73.94	64.87	66.31	71.39	72.21	78.16	**81.23**	72.59
		-	UNIGEN	76.74	66.88	69.63	73.29	72.10	**78.64**	80.52	**73.97**

Table 11: Result of ablation study that examines the performance of UNIGEN and baselines on small-sized TAMs. The performance of TAM, which is superior to that of PROMPTING, is underlined, and the best result in each test dataset within the group for each TAM is presented in boldface.