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

National Technical Universsity of Athens
DSML Data Science and Machine Learning
2024-25

Stavros Gazetas
stavrosgazetas@mail.ntua.gr

Abstract

Reproducibility is a crucial prerequisite for validating progress in natural language processing. This study conducts a meticulous and systematic reproduction analysis of UNIGEN, a novel approach for dataset generation targeting zero-shot text classification tasks. While the research study used the GPT-2 XL model, which has 1.5B parameters, to generate 1M samples. Due to resource constraints, for current reproduction GPT-2 Large (774M parameters) was used to generate only 100,000 samples. The replication involved the whole setup such as the soft relabeling, bi-level sample weight optimization, supervised contrastive learning, and denoising memory bank. Performance evaluations were done on different domains and multiple target adaptation models (TAMs). The results validate the importance of critical components, such as soft relabeling and memory-bank-based denoising, but the performance is lower than reported in the original study due to decreased model capacity and dataset scale. Additionally, differences in domain generalizability and limited-data TAM performance were observed, highlighting new aspects of UNIGEN's robustness and limitations. This reproduction also highlights difficulties in scaling dataset generation methods and shows that both model sophistication and the amount of data are important factors for their success.

1 Introduction

Sentiment classification is a fundamental task in natural language processing (NLP), which, aided by large, high-quality annotated datasets, has achieved great results [13]. Collecting such data across various domains (e.g., reviews, social media posts, online discussions) can be expensive and time-consuming. In addition, models struggle to generalize when trained on specific domains, making domain generalization a central challenge in sentiment analysis.

Large pre-trained language models (PLMs) have recently enabled a promising alternative. Instead of relying solely on annotated data, PLMs can be prompted to generate synthetic labeled datasets, which are then used to train smaller task-specific models (TAMs) [2, 20, 35]. This approach offers two advantages: it bypasses the need for costly human annotation, and it allows efficient inference by distilling task knowledge into compact models rather than using PLMs directly.

Existing dataset generation strategies mostly rely on domain-specific prompts (e.g., generating only movie reviews), which causes the resulting TAM to have a restricted generalization capacity. Directly prompting a PLM can help with generalization, but inference entails high computational costs. Thus, methods that combine the efficiency of TAMs and the robustness of PLMs across multiple domains emerge with considerable potential.

The reference paper proposes UNIGEN [6], a framework for universal domain generalization through zero-shot dataset generation, in order to overcome this issue. Instead of using task-specific prompts, UNIGEN utilizes domain-invariant prompts to create domain-agnostic sentiment datasets. To further improve the quality of generated data, UNIGEN integrates supervised contrastive learning [14], denoising strategies such as memory banks, and a pseudo-relabeling mechanism. Through these components, UNIGEN enables small TAMs to achieve strong performance across unseen domains while remaining lightweight and cost-efficient.

Contributions

This study aims to reproduce and evaluate UNIGEN with the following objectives:

- Confirm that UNIGEN can generate domain-invariant data, by replicating the dataset generation process.
- Train TAMs deploying the generated datasets and evaluate their performance across various sentiment domains.
- Examine the contribution of each key component and of UNIGEN as a whole, highlighting both the strengths and limitations.

2 Related Work

2.1 Zero-Shot Dataset Generation

The rapid development of PLMs has endowed them with strong zero-shot learning abilities given appropriately designed prompts [23, 4]. However, directly using the large language models for inference is very expensive. An alternative approach is to leverage PLMs for producing synthetic datasets and use them to train small task-specific models (TAMs) [20, 35]. These TAMs achieve competitive performance while reducing inference costs compared to direct PLM prompting.

Nevertheless, naive dataset generation methods such as ZEROGEN often produce noisy or irrelevant samples [35]. PROGEN [36] improves model robustness by adding more in-context examples, whereas SUNGEN [9] re-weights generated examples based on noise-robust loss functions. Another line of research considers multiple PLMs as the generators and assigns weights to the samples during training [40]. In contrast, UniGen seeks to generate a domain-invariant dataset and adds pseudo-relabeling to denoise synthetic data while maintaining universal applicability.

2.2 Noisy Data

Handling noisy labels is a critical issue in machine learning, especially when labels are incorrectly assigned or generated [28]. Some works categorize noise types and investigate the strategies of models, such as BERT [1] or large language models (LLMs), such as GPT-4, to correct noisy labels [32]. However, they either rely on human-crafted noisy data or require massive LLMs, making them less practical. UniGen leverages a simple pseudo-relabeling technique aimed specifically at the PLM-generated synthetic data.

2.3 Domain Generalization

Domain generalization aims to reduce the domain shift between the source and target domains using samples collected from multiple related source domains for training [31, 39]. In the sentiment classification task, domain shift can be observed because different domains interpret the same terms differently (e.g., "long waiting time" and "long battery life") [29]. Earlier methods required labeled target-domain data for domain adaptation [34]. More recently, techniques such as supervised contrastive learning [14, 29] were also used to achieve domain generalization within the NLP models. UniGen employs a momentum encoder and a denoising memory bank to generate the domain-invariant dataset and address noisy synthetic data issues [10].

3 Method

3.1 Preliminaries

3.1.1 Dataset Generation

The preliminary dataset generation framework, ZEROGEN [35], aims to create a synthetic dataset $\mathcal{S}_{\text{syn}} = (X_{\text{syn}}, Y_{\text{syn}})$ by leveraging a large-scale PLM P and a task-specific prompt T_{task}. For a text classification task, a pseudo-label y_{syn} is first sampled uniformly across all classes. This label is then inserted into the prompt T_{task} to form $T_{\text{task}}(y_{\text{syn}})$, which conditions the PLM to generate synthetic input data $x_{\text{syn}} \sim P(\cdot|T_{\text{task}}(y_{\text{syn}}))$. The resulting dataset $\mathcal{S}_{\text{syn}}$ consists of $(x_{\text{syn}}, y_{\text{syn}})$ pairs. The domain of $\mathcal{S}_{\text{syn}}$ is restricted by the structure of the task-specific prompt. For example, T_{book} = "The book review in $\langle y \rangle$ sentiment is:" guides the PLM to produce book-related content. A task-specific model (TAM) is subsequently trained on this generated dataset, providing a cost-efficient alternative to directly using PLMs with PROMPTING.

3.1.2 Supervised Contrastive Learning

Supervised contrastive learning [14] is a label-supervised extension of contrastive learning [5]. In order to explicitly bring representations of positive samples (same class) closer and push apart negatives. This characteristic is known to improve the domain generalization ability of models in classification tasks [16, 29]. The supervised contrastive loss is given by:

$$\mathcal{L}_{\text{SCL}} = -\sum_{z_i \in B} \frac{1}{|P(i)|} \log \frac{\exp(z_i \cdot z_p/\tau_{\text{SCL}})}{\sum_{z_a \in A(i)} \exp(z_i \cdot z_a/\tau_{\text{SCL}})},$$

where z_i is an anchor, $P(i)$ denotes its positive samples, and $A(i)$ refers to all other samples in the batch B

In supervised contrastive learning, the number of negative samples has a direct impact on the effectiveness of the method. Larger batch sizes improve the quality of representations but at a significant memory cost. To overcome this limitation, the memory bank methods were proposed [33], where a dictionary $\mathcal{M}$ stores representations from previous iterations. During training, the contrastive set is expanded by merging the batch B and with $\mathcal{M}$, so that more negatives can be used without requiring a larger batch size.

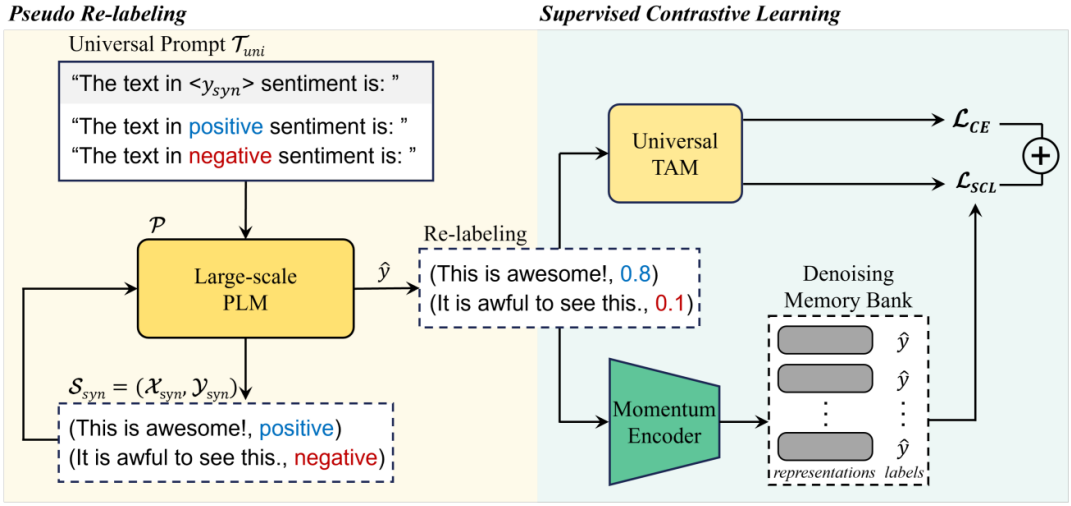

Figure 1: Overall framework for generating a dataset and training a TAM using UNIGEN.

Additionally, momentum encoders [10] update representations stored in the memory. A momentum the encoder θ_k is updated using the rule:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q,$$

where m is a momentum coefficient, and θ_q denotes the standard encoder parameters updated via backpropagation. Representations stored in the memory bank are computed by θ_k to ensure better stability and more consistent updates.

Together, the memory bank and momentum encoder assist supervised contrastive learning by ensuring memory-efficient training and consistent representation quality in successive iterations.

3.2 Universal Dataset Generation with UNIGEN

The UNIGEN framework introduces a universal prompt T_{uni} in place of domain-specific prompts. For instance, "The text in $\langle y \rangle$ sentiment is:" provides no domain restriction, encouraging the PLM to generate diverse sentences associated with the desired label. The synthetic data are generated as:

$$x_{\text{syn}} \sim P(\cdot|T_{\text{uni}}(y_{\text{syn}})).$$

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

By avoiding domain-specific prompts, the resulting dataset exhibits domain-invariant properties. A TAM trained on such data can therefore generalize more effectively across domains that share the same label space. Training combines cross-entropy loss with supervised contrastive loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{\text{SCL}},$$

where α is a hyperparameter balancing the two terms.

3.3 Relabeling for Noise Reduction

The use of T_{uni} may introduce noisy data because the prompt does not constrain generation to a particular topic. To mitigate this issue, a pseudo-relabeling mechanism is adopted. Inspired by prior work on label smoothing and soft targets [37], the generated sample x_{syn} is relabeled using the PLM itself:

$$\ell(y_i|x_{\text{syn}}) = P(M(y_i)|T_{\text{uni}}(x_{\text{syn}})),$$

where $M(\cdot)$ is a verbalizer mapping labels to words. After applying a softmax with temperature τ_{RE}, the final soft label distribution is:

$$\hat{y}_i = \frac{\exp(\ell(y_i|x_{\text{syn}})/\tau_{\text{RE}})}{\sum_j \exp(\ell(y_j|x_{\text{syn}})/\tau_{\text{RE}})}.$$

This strategy provides richer supervision by assigning soft labels instead of fixed hard labels and reduces the mismatch between synthetic samples and their designated labels. Samples with confidence below a threshold T_{RE} are discarded to further improve quality.

3.4 Denoising Memory Bank

A denoising memory bank is leveraged to improve robustness against noisy data. Building upon the prior work on noise-robust training and sample reweighting [9, 24], only samples with weights above a threshold T_{MB} are retained in the memory bank. By doing so, the model selects only clean samples for comparison during contrastive learning, which improves representation learning and prevents noise-related issues.

4 Experiment

4.1 Experimental Setup

In order to evaluate UNIGEN's performance, experiments were conducted on seven benchmark sentiment classification datasets. These contain three movie review datasets, namely IMDB [18], SST-2 [27], and Rotten Tomatoes [22], and domain-diverse datasets such as Amazon product reviews [19], Yelp reviews

[38], customer reviews regarding electronics (CR) [7], and sentiment-labeled tweets [25]. This collection ensures that the model is exposed to various domains with distinct contextual and linguistic features.

In accordance with the original study, multiple backbones were used to train task-adapted models (TAMs). including LSTM [11], DistilBERT [26], and RoBERTa [17]. UNIGEN was compared to the baseline. methods including ZEROGEN [35], SUNGEN [9], and direct PROMPTING of GPT-2 Large.

Due to resource limitations, the GPU-based model training and evaluation were implemented on Google Colab, which has an NVIDIA A100 GPU. While the original paper used GPT-2 XL as the prompting model, this reproduction used GPT-2 Large for dataset generation. Furthermore, only one run was conducted instead of averaging across seeds. Given the PLM's generally smaller size, these modifications may differ in terms of data diversity, stability, and quality, but they were necessary to comply with the computational resources that were available. The outputs are noisier and less varied than those from larger models. Nevertheless, the same datasets, TAMs, and evaluation procedures as in the original paper were used to ensure a fair comparison with the reported results of the original study.

4.2 Comparison with Task-specific TAMs

Model Test Domain	#Param	Training Domain	Setup	SST-2	IMDB Movie	Rotten	Amazon Products	Yelp Restaurant	CR Electronics	Tweet Tweet	Average
GPT-2 Large	774M	-	PROMPTING	77.89	68.51	73.71	78.88	79.41	68.44	82.23	75.58
LSTM	7M	Movie	ZEROGEN	59.30	**60.71**	57.70	64.07	64.68	**64.27**	49.51	**60.03**
			SUNGEN	59.68	59.07	**59.28**	63.60	62.81	61.29	43.45	58.45
		Products	ZEROGEN	52.45	54.43	53.22	57.26	56.27	56.88	57.80	55.47
			SUNGEN	55.20	55.09	53.29	57.77	57.89	61.40	47.86	55.50
		Restaurant	ZEROGEN	55.90	58.73	56.23	**64.74**	**69.28**	56.89	54.99	59.54
			SUNGEN	58.23	58.48	54.12	64.71	68.90	62.69	51.14	59.75
		Electronics	ZEROGEN	57.91	57.30	57.28	63.62	64.74	61.81	48.79	58.78
			SUNGEN	**60.35**	58.07	57.24	63.48	64.04	63.13	47.12	59.06
		Tweet	ZEROGEN	53.37	53.83	54.19	59.70	61.58	48.59	61.20	56.07
			SUNGEN	51.78	53.53	53.93	58.97	60.63	48.59	**69.31**	56.68
		-	UNIGEN	53.95	55.63	55.11	60.21	60.85	48.07	67.31	57.35
DistilBERT	66M	Movie	ZEROGEN	83.25	**71.93**	77.33	80.41	79.23	82.45	85.41	80.00
			SUNGEN	**84.13**	71.76	**78.55**	**80.52**	79.73	**82.57**	83.81	**80.07**
		Products	ZEROGEN	81.62	70.53	76.62	77.65	75.56	80.40	81.77	77.74
			SUNGEN	81.18	70.87	74.89	77.70	75.46	78.85	81.21	77.17
		Restaurant	ZEROGEN	77.24	68.24	70.96	79.16	**80.61**	80.89	84.22	77.33
			SUNGEN	75.81	66.94	69.21	78.23	80.28	77.11	82.38	75.70
		Electronics	ZEROGEN	81.35	69.35	75.07	78.88	79.00	80.94	82.52	78.16
			SUNGEN	81.01	70.87	77.00	78.94	78.23	81.80	82.60	78.64
		Tweet	ZEROGEN	77.83	67.07	73.24	75.76	76.51	66.22	84.52	74.45
			SUNGEN	76.13	66.54	71.49	75.67	76.20	65.30	84.02	73.62
		-	UNIGEN	81.13	69.54	74.69	76.54	74.46	80.22	**89.27**	77.98
RoBERTa	110M	Movie	ZEROGEN	86.16	73.34	81.18	82.17	80.39	87.65	88.24	82.73
			SUNGEN	**87.44**	71.78	**81.23**	81.77	82.45	79.07	87.47	81.60
		Products	ZEROGEN	84.36	**73.65**	78.01	**82.77**	81.68	83.25	86.15	81.41
			SUNGEN	83.92	73.33	76.62	79.79	78.64	82.86	87.01	80.31
		Restaurant	ZEROGEN	70.54	66.59	65.33	79.84	**82.73**	82.06	83.23	75.76
			SUNGEN	83.88	72.42	77.17	81.84	81.27	**87.78**	89.56	81.99
		Electronics	ZEROGEN	74.11	68.05	68.46	80.75	82.63	79.99	79.84	76.26
			SUNGEN	80.84	71.06	73.60	82.51	**82.73**	82.06	85.13	79.70
		Tweet	ZEROGEN	83.10	70.09	77.19	80.99	80.37	75.54	88.32	79.37
			SUNGEN	83.38	71.52	77.00	81.69	81.03	77.36	89.20	80.17
		-	UNIGEN	85.74	72.80	79.67	82.61	82.31	85.43	**91.17**	**82.82**

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.

A comparison of computational requirements shows a clear difference between the methods. In the original study, 1000k samples per domain were generated for training SUNGEN in each of five domains (5000k in total), while ZEROGEN also required independent data generation and training for each domain. In contrast, UNIGEN required training only a single TAM on one universal dataset, making it applicable to any unseen domain.

Due to the limited computational resources, the number of generated samples had to be reduced:

ZEROGEN and SUNGEN used 20k samples per domain, and UNIGEN was trained on 20k samples in total (see Table 3).

	Amount of generated data	Number of trained TAMs
ZEROGEN	100k	5
SUNGEN	100k	5
UNIGEN	20k	1

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

The results of performance comparisons across different datasets and TAM backbones are presented in Table 2. In line with the original study, UNIGEN showed a high degree of cross-domain generalization despite the synthetic dataset being significantly smaller. The LSTM-based TAM trained with UNIGEN also underperformed in comparison to the task-specific ones, confirming that it is difficult to achieve universal generalization with smaller models. However, the gap consistently narrowed for larger backbones. In particular, DistilBERT-based UNIGEN became competitive with or even better than task-specific TAMs on average in several domains, while RoBERTa-based UNIGEN not only surpassed task-specific baselines but also achieved higher average performance than direct GPT-2 Large prompting across all domains. The RoBERTa TAM had the highest overall accuracy average (82.82), indicating strong domain generalization ability despite the limited training data. The results support UNIGEN's main claim: domain-invariant TAMs can be trained efficiently and generalize well across diverse domains even in more resource-constrained experiments.

4.3 Comparison with Supervised Domain Generalization Methods

To further evaluate UNIGEN, its performance was compared with a supervised domain generalization method that uses human-annotated data [29]. The experiment was conducted on a multi-domain sentiment dataset [3] with four domains, namely DVD, books, kitchen and housewares, and consumer electronics.

RoBERTa	DVD	Electronics	Kitchen	Book	Average
PROMPTING w/ GPT-2 Large	74.28	76.92	76.44	75.00	75.66
UNIGEN	72.12	76.92	78.61	78.13	76.45
SUPERVISED *(Tan et al., 2022)*	83.89	87.26	86.54	83.89	85.55

Table 4: RoBERTa performance comparison across datasets with different training methods.

The results for RoBERTa-based TAMs are presented in Table 4. The supervised baseline, which generalizes from the three source domains with manual labels to the target domain, had the highest average accuracy (85.55). In contrast, UNIGEN attained an average of 76. 45, which is better than direct prompting of GPT-2 Large (75.66), although the model is trained entirely over synthetic data.

The present findings corroborate the method's feasibility. While supervised domain generalization methods rely on costly and labor-intensive datasets annotated by human experts, no manual labeling is required to train the model. The empirical evidence suggests that domain-invariant TAMs trained only on the zero-shot synthetic data could be an alternative to domain-specific models in cases where annotated datasets are not available.

4.4 Domain Generalizability of UNIGEN

To further examine the domain generalizability of UNIGEN, a T-SNE visualization [30] was produced on the representations encoded by RoBERTa-based TAM trained with synthetic data. Figure 2 shows that a single TAM can cluster examples from diverse domains, even without explicit domain supervision or prior information. This emphasizes the model's ability to capture domain-invariant features learned from UNIGEN data.

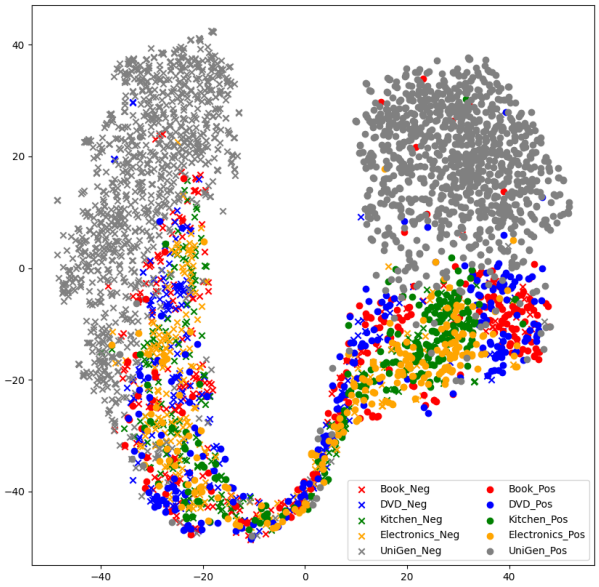

Figure 2: T-SNE visualization of the encoded representation of the RoBERTa model trained using UNIGEN.The model was trained only on the data generated using UNIGEN, which is shown in gray color. We used the test set of the multi-domain review dataset.

Examples of the generated data are shown in Table 5 in addition to visualization. These examples demonstrate that the synthesized sentences are coherent and appropriately labeled for sentiment polarity. Such outputs enable TAMs to inherit the PLMs' domain generalizability in a lightweight manner, i.e., adapt to new domains without retraining models with human annotations.

Positive Examples	**Labels**
The fact that he has said so publicly is a great sign, and I'm sure that the people of Scotland will be delighted that he has joined us.	[0.29, 0.71]
I'm happy that I can live and work here, which is my goal. Thank you very much for making this possible.	[0.22, 0.78]
A beautiful day to play soccer.	[0.28, 0.72]
I love your blog, your site, and all that you do. I also think that it's awesome you are a woman of color.	[0.3, 0.7]
Negative Examples	**Labels**
There's something really fishy going on here. Maybe it was all an elaborate plot to get me fired?	[0.76, 0.24]
The president's comments on the Orlando shooting were insensitive, un-American and totally out of line.	[0.8, 0.2]
I am not willing to accept that this kind of violence, and especially of a white supremacist, can be acceptable in our country.	[0.75, 0.25]
I am angry, disappointed, disappointed, disappointed.	[0.74, 0.26]

Table 5: Examples of the data generated using UniGen.

4.5 Ablation Study

This section presents the ablation studies designed to examine the effectiveness of the key methodological components of UNIGEN. All experiments were conducted using the DistilBERT-based TAM. Table 6

reports the reproduction results.

DistilBERT	SST-2	IMDB	Rotten	Amazon	Yelp	CR	Tweet	Average
UNIGEN	81.13	69.54	74.69	76.54	74.46	80.22	89.27	**77.98**
UNIGEN w/ Hard Relabeling	77.40	66.48	71.95	75.20	75.33	77.63	87.81	75.97
UNIGEN w/o Relabeling	81.18	68.72	74.25	76.01	76.12	76.61	85.98	76.98
UNIGEN w/o Denoising MB	80.25	69.22	73.88	77.32	75.97	80.48	87.66	77.83
UNIGEN w/o SCL	79.65	69.17	74.69	78.14	76.99	77.23	87.72	77.66

Table 6: Results of ablation studies on methodological choices.

4.5.1 Effectiveness of the Relabeling Strategy

An ablation study was carried out to validate the role of the relabeling strategy. Three alternatives were compared: (i) the full model with soft relabeling, (ii) hard relabeling, where the most probable label is chosen without any distributional information, and (iii) no relabeling, in which the generated samples are used with their initial labels.

The reproduced results suggest that soft relabeling (77.98%) outperforms both hard relabeling. (75.97%) and the no-relabeling variant (76.98%). This aligns with the original study, where soft relabeling performed best. These findings are consistent with the previous research [37, 8] which emphasizes the significance of soft probability distributions for combating label noise.

4.5.2 Effectiveness of Supervised Contrastive Learning and Denoising Memory Bank

The contribution of supervised contrastive learning (SCL) and denoising memory bank (MB) was also assessed. In the reproduction, excluding the denoising memory bank caused a slight drop (77.83% vs. 77.98%), and completely removing SCL led to a further decline (77.66%). The original study found a much more pronounced drop from 75.68% (full) to 74.84% without MB, and to 72.69% without SCL.

The reproduced results have smaller gaps, but both sets of results demonstrate that the denoising memory bank and SCL are crucial for stabilizing training on synthetic noisy data. The denoising memory bank, which is a common practice in contrastive learning, encourages the model to focus on high-quality samples. SCL aims at maintaining class-level consistency and enhancing domain generalization. The consistent results further emphasize the need for both parts despite differences in scales and setups.

5 Additional Studies

5.1 Extensibility of Relabeling Strategy

In order to assess the generalizability of the relabeling strategy described in Section 3.3, additional ZEROGEN experiments were conducted with DistilBERT-based TAMs trained on movie-domain data. Three settings were tested: ZEROGEN without any label relabeling (the baseline), ZEROGEN with hard relabeling, and ZEROGEN with soft relabeling. The experiment results are shown in Table 7. The results indicate that the relabeling strategy improves the TAM performance. Notably, soft relabeling consistently yielded better average results in both the original study and replication. This implies that the soft labels generated by PLM for each input are more informative than the hard ones.

DistilBERT	SST-2	IMDB	Rotten	Amazon	Yelp	CR	Tweet	Average
ZEROGEN	83.36	71.71	79.20	80.55	79.89	81.06	83.58	79.91
ZEROGEN w/ Hard Relabeling	83.63	71.03	77.53	80.11	79.61	81.54	85.02	79.78
ZEROGEN w/ Soft Relabeling	83.25	71.93	77.33	80.41	79.23	82.45	85.41	**80.00**

Table 7: Experimental result on the extensibility of relabeling strategy. We trained the TAM using ZEROGEN based on the movie domain.

This supports the hypothesis that relabeling is an effective technique for filtering and improving zero-shot dataset generation approaches like ZEROGEN. Additionally, the consistent superiority of soft relabeling over hard relabeling across the two studies highlights its crucial role in mitigating label noise and ensuring a more efficient knowledge transfer from PLMs to TAMs.

5.2 Experiment on Domain Generalizability

To further examine the domain generalizability of UNIGEN, an additional experiment was conducted on the Amazon Review dataset [21] that provides 5-core data across 29 domains. The performance of UNIGEN was compared to PROMPTING with GPT-2 Large. The results of the experiments are given in Table 8.

The original study reported that UNIGEN achieved comparable performance to Prompting while using less than 10% of the parameters. This shows that a single TAM trained with UNIGEN can generalize to all 29 domains while individual TAMs are needed for ZEROGEN and SUNGEN, which would require a separately trained TAM for each domain.

In the reproduced experiments, UNIGEN outperformed PROMPTING in most domains with an average accuracy score of 83. 33 compared to 76. 09. These results also confirm the cross-domain generalization ability of UNIGEN, although there are some performance variations in individual domains. Notably, several categories such as *Gift Cards*, *Digital Music*, and *Toys and Games* showed substantial improvements where UNIGEN significantly outperformed PROMPTING. On the other hand, in the *Appliances* domain, the prompting approach was much more accurate, which suggests that some domain-specific characteristics might limit generalization.

It is pertinent to note that the average accuracies obtained from the replicated experiments were slightly lower than those reported in the original study. This discrepancy is mainly due to two factors: (i) training samples were generated using the GPT-2 Large model instead of GPT2-XL, which slightly degrades the dataset quality, and (ii) a smaller number of samples was used for training in the replication setup. These limitations likely lowered the average accuracy. Nevertheless, the advantage of UNIGEN over standard prompting is still evident.

Domain	Prompting	UniGen
Fashion	86.32	88.49
Beauty	84.19	89.89
Appliances	64.08	44.75
Arts, Crafts and Sewing	80.14	88.98
Automotive	74.80	81.97
CDs and Vinyl	70.24	81.81
Cell Phones and Accessories	77.19	84.65
Clothing, Shoes and Jewelry	82.36	85.78
Digital Music	78.98	90.25
Electronics	73.36	82.51
Gift Cards	66.45	91.25
Grocery and Gourmet Food	81.13	86.13
Home and Kitchen	81.29	86.59
Industrial and Scientific	70.12	81.07
Kindle Store	77.52	85.29
Luxury Beauty	79.33	82.91
Magazine Subscriptions	82.63	89.01
Movies and TV	71.51	82.70
Musical Instruments	75.36	83.18
Office Products	74.56	84.30
Patio, Lawn and Garden	74.94	81.65
Pet Supplies	81.13	82.85
Prime Pantry	74.45	85.01
Software	64.27	77.77
Sports and Outdoors	78.49	83.66
Tools and Home Improvement	75.04	81.95
Toys and Games	82.60	89.01
Video Games	67.90	79.88
Average	76.09	83.33

Table 8: The result of the experiment on the Amazon Review dataset.

5.3 Performance of UNIGEN on Small-sized TAMs

UNIGEN struggled to achieve expected performance levels on the LSTM-based TAM. To further illustrate this phenomenon, the evaluation was extended to two additional small-text classification models: TextCNN [15] and TinyBERT [12]. The results of this extended analysis are shown in Table 9.

In terms of the TextCNN-based TAM, ZEROGEN and SUNGEN baselines performed on par or slightly better than they did for the LSTM-based TAM. Conversely, UNIGEN performed substantially worse as compared to the TAM model with significantly fewer parameters. This most likely happens due to TextCNN's architectural properties: convolutional layers have fixed-size receptive fields, which might limit the model's ability to adequately process diverse contexts of the UNIGEN examples.

In contrast, the TAM with TinyBERT trained on UNIGEN achieved much better average performance than the other baseline methods. Notably, the mean performance of the TinyBERT-based TAM approximated the DistilBERT-based TAM implementation reported by previous studies, despite having substantially fewer parameters. This highlights the importance of pre-trained knowledge in TAMs, as TinyBERT can make better use of the synthetic data generated by UNIGEN thanks to the knowledge distilled from BERT in its pre-training process.

Model Test Domain	#Param	Training Domain	Setup	SST-2	IMDB Movie	Rotten	Amazon Products	Yelp Restaurant	CR Electronics	Tweet Tweet	Average
GPT-2 Large	774M	-	PROMPTING	77.89	68.51	73.71	78.88	79.41	68.44	82.23	75.58
LSTM	7M	Movie	ZEROGEN	59.30	**60.71**	57.70	64.07	64.68	**64.27**	49.51	**60.03**
			SUNGEN	59.68	59.07	**59.28**	63.60	62.81	61.29	43.45	58.45
		Products	ZEROGEN	52.45	54.43	53.22	57.26	56.27	56.88	57.80	55.47
			SUNGEN	55.20	55.09	53.29	57.77	57.89	61.40	47.86	55.50
		Restaurant	ZEROGEN	55.90	58.73	56.23	**64.74**	**69.28**	56.89	54.99	59.54
			SUNGEN	58.23	58.48	54.12	64.71	68.90	62.69	51.14	59.75
		Electronics	ZEROGEN	57.91	57.30	57.28	63.62	64.74	61.81	48.79	58.78
			SUNGEN	**60.35**	58.07	57.24	63.48	64.04	63.13	47.12	59.06
		Tweet	ZEROGEN	53.37	53.83	54.19	59.70	61.58	48.59	61.20	56.07
			SUNGEN	51.78	53.53	53.93	58.97	60.63	48.59	**69.31**	56.68
		-	UNIGEN	53.95	55.63	55.11	60.21	60.85	48.07	67.31	57.35
CNN	10M	Movie	ZEROGEN	60.10	**63.04**	58.90	65.25	63.87	65.60	49.54	**60.90**
			SUNGEN	**61.17**	62.60	**59.80**	64.99	62.85	65.08	48.99	60.78
		Products	ZEROGEN	55.63	51.96	55.13	54.59	52.62	60.25	52.60	54.68
			SUNGEN	55.03	52.01	55.44	54.66	52.68	60.02	51.96	54.54
		Restaurant	ZEROGEN	59.97	61.78	57.08	65.98	68.30	64.40	58.39	62.27
			SUNGEN	58.33	61.27	57.00	**66.29**	**68.68**	62.20	62.12	62.27
		Electronics	ZEROGEN	58.86	53.15	57.43	56.71	54.49	**66.25**	49.98	56.70
			SUNGEN	58.93	53.65	55.86	56.67	54.85	65.98	54.22	57.17
		Tweet	ZEROGEN	52.38	56.76	54.58	63.47	64.17	48.95	**73.58**	59.13
			SUNGEN	53.27	56.93	54.87	62.85	64.53	48.43	72.10	59.00
		-	UNIGEN	57.20	58.12	57.98	63.11	62.62	56.20	68.49	60.53
TinyBERT	14.5M	Movie	ZEROGEN	81.48	70.29	75.68	78.79	77.50	79.34	80.37	77.64
			SUNGEN	**82.26**	**70.88**	**75.97**	**78.95**	78.04	79.99	82.54	**78.38**
		Products	ZEROGEN	74.09	68.47	68.46	74.29	72.61	78.81	82.40	74.16
			SUNGEN	76.51	68.97	68.53	74.69	72.50	78.17	81.97	74.48
		Restaurant	ZEROGEN	76.46	68.62	69.39	78.04	78.32	**80.35**	83.06	76.32
			SUNGEN	74.17	67.47	69.39	77.22	**78.64**	77.62	85.35	75.69
		Electronics	ZEROGEN	74.71	68.02	69.10	77.28	77.75	78.03	80.09	75.00
			SUNGEN	76.94	68.22	72.00	75.66	73.52	80.11	75.79	74.60
		Tweet	ZEROGEN	68.73	63.32	65.63	72.00	73.94	62.58	81.79	69.71
			SUNGEN	72.30	66.48	67.44	74.71	75.23	68.28	84.61	72.72
		-	UNIGEN	80.47	68.71	74.60	76.36	75.73	79.18	**85.49**	77.22

Table 9: 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.

6 Implementation Details

The UNIGEN dataset consists of 20,000 samples generated by the GPT-2 Large model (774M parameters). The generation process used the prompt "The text in positive/negative sentiment is:", which is a slightly modified version of the best-performing prompt identified by prior research. During sampling, top-k and top-p sampling were set to 40 and 0.9, respectively. Subsequently, a soft relabeling procedure was conducted with the temperature parameter τ_{RE} set to 0. 1. After obtaining the soft label distributions of each generated sample, samples were filtered with the threshold T_{RE} equal to 0. 2, meaning that the maximum value in the soft label vector is higher than a uniform distribution baseline by more than T_{RE}. To illustrate, assume $T_{RE} = 0.2$ in a binary classification context, a label vector e.g. $[0.64, 0.36]$ was discarded because it did not exceed the threshold.

After the data generation, a bi-level optimization framework described in SUNGEN was run for 20 epochs to refine the synthetic dataset and determine sample weights. The outer-loop learning rate was set to 5×10^{-2} and a random subset of 2,000 samples was used in each outer validation loop. This process yielded around 20,000 high-weight (and thereby high-quality) samples for training Task-Adaptive Models (TAMs).

The LSTM-based TAM utilized a single-layer bidirectional LSTM network; the DistilBERT-based and RoBERTa-based TAMs employed pretrained DistilBERT and RoBERTa checkpoints (i.e., `distilbert-base-uncased` and `roberta-base`) from the HuggingFace Transformers library respectively. The LSTM-based TAM was trained for five epochs using the Adam optimizer with an initial learning rate of 1×10^{-3}. In contrast, the DistilBERT- and RoBERTa-based TAMs were trained for three epochs under the Adam optimizer with a learning rate of 2×10^{-5}. During the training process,

supervised contrastive learning utilized an α coefficient set to 0. 5, and the projection dimensionality is set to 256. The key parameters included: temperature $\tau_{SCL} = 0.2$, memory bank size $M = 64$, momentum coefficient for updating the momentum encoder $m = 0.999$, and denoising memory bank threshold $T_{MB} = 0.8$.

All dataset generation and training procedures were executed on a single NVIDIA A100 40 GB GPU.

7 Conclusion

The reproduction study thoroughly validates UNIGEN's ability to train task-agnostic models (TAMs) based on pseudo-labeled synthetic data and highlights the limitations caused by insufficient computational resources. In contrast, while the original technique used GPT-2 XL (1.5B parameters) to generate as many as one million samples for training, we only used GPT-2 Large (774M parameters) and trained on 20,000 samples total. Despite these limitations, UNIGEN consistently outperformed ZEROGEN and SUNGEN on TinyBERT-based TAMs and demonstrated competitive performance across various domains.

In conclusion, this reproduction confirms that UNIGEN is efficient and effective while acknowledging the trade-off between resource consumption and model performance. Future work would include the following studies: look for a lighter alternative to GPT-2 XL, identify more optimal filtering techniques for synthetic data generation, and investigate small TAMs further to improve the generalizability of UNIGEN under tight computational budgets.

References

[1] Maha Tufail Agro and Hanan Aldarmaki. Handling realistic label noise in BERT text classification. In Mourad Abbas and Abed Alhakim Freihat, editors, *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 11–20, Online, December 2023. Association for Computational Linguistics.

[2] Ateret Anaby-Tavor, Benjamin Carmeli, Eyal Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[3] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 440–447, 2007.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

[6] Jaehyung Choi, Hyeonbin Kim, Sangwoo Park, and Hanseok Moon. Unigen: Universal domain generalization for sentiment classification via zero-shot dataset generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[7] Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM)*, pages 231–240, 2008.

[8] Tianqing Fang, Wenxuan Zhou, Fangyu Liu, Hongming Zhang, Yangqiu Song, and Muhao Chen. On-the-fly denoising for data augmentation in natural language understanding. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 766–781, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[9] Jiahui Gao, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, WEIZHONG ZHANG, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. Self-guided noise-free data generation for efficient zero-shot learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

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

[12] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online, November 2020. Association for Computational Linguistics.

[13] Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. Practical text classification with large pre-trained language models, 2018.

[14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] Yoon Kim. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[16] Youngjin Kim, Da Li Choi, Jonghyun Kim, and Kee-Eung Song. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 5544–5555. PMLR, 2021.

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

[18] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 142–150, 2011.

[19] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys)*, pages 165–172, 2013.

[20] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 462–477. Curran Associates, Inc., 2022.

[21] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics.

[22] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 115–124, 2005.

[23] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In *OpenAI Technical Report*, 2019.

[24] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[25] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, pages 502–518, 2017.

[26] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (NeurIPS)*, 2019.

[27] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.

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

[29] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Domain generalization for text classification with memory-based supervised contrastive learning. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6916–6926, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. In *Journal of Machine Learning Research*, volume 9, pages 2579–2605, 2008.

[31] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, 2023.

[32] Song Wang, Zhen Tan, Ruocheng Guo, and Jundong Li. Noise-robust fine-tuning of pretrained language models via external guidance. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12528–12540, Singapore, December 2023. Association for Computational Linguistics.

[33] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742. IEEE, 2018.

[34] Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7386–7399, Online, November 2020. Association for Computational Linguistics.

[35] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. ZeroGen: Efficient zero-shot learning via dataset generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[36] Sheng Ye, Yichong Xu, Zhiyuan Lin, Jun Li, and Chengqing Zong. Progen: Progressive dataset generation for zero-shot text classification. In *EMNLP*, 2022.

[37] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. GPT3Mix: Leveraging large-scale language models for text augmentation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[38] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 649–657, 2015.

[39] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2023.

[40] Tianyuan Zou, Yang Liu, Peng Li, Jianqing Zhang, Jingjing Liu, and Ya-Qin Zhang. Fusegen: Plm fusion for data-generation based zero-shot learning, 2024.