A. Limitations

Our work explores the possibility of data augmentation for boosting text classification performance when the downstream model is finetuned using pre-trained language models. The results show that STA consistently performs well across different bench-marking tasks using the same experimental setup, which addresses the limitation stated in the previous work (Kumar et al., 2020) calling for a unified data augmentation technique. However, similar to Kumar et al. (2020), although STA can achieve improved performance as the data size goes up to 100 examples per class in some cases (such as 100 examples per class in **EMOTION**, Table 7 and **HumAID**, Table 9), the absolute gain in performance plateaus when the training data becomes richer (such as 100 examples per class in **SST-2** and **TREC**). This suggests that it is challenging for STA to improve pre-trained classifier's model performance in more abundant data regimes.

It's also worth noting that STA currently applies to text classification exclusively using T5 and doesn't extend to other general NLP tasks without generative models. However, our approach, centered around text classification, holds the potential to expand beyond this narrow scope and encompass a wider array of NLP tasks. This flexibility arises from our use of generative models like T5 in our research. For instance, to consider the adaptation of our templates for question answering tasks, the templates used in our method can be modified to be like question-answer format. For instance, template c in Table 1 could be transformed to read, "*Given Text. Provide answer to this question: Question.*" Furthermore, our generation template can be suitably tailored to produce text based on a given question description. The capacity to adjust both classification and generation templates underpins the applicability of our approach across diverse NLP tasks. Thus, in the future, we aim to extend our approach to other downstream natural language tasks, incorporating different generation models alongside T5.

Another important consideration is the choice of templates used in STA. Ablation experiments in Section 5.2 show that our chosen set of templates yields better performance than a 'minimal subset' consisting of the two simplest templates; the question as to how to choose optimal templates for this augmentation scheme remains unanswered. Hence, in future work, we will explore better methods for constructing the prompt templates, aiming to reduce the dependency on the manual work at this step.

B. Template Example

Table 4 presents how an original training example is converted to multiple examples in STA using the prompt templates from Table 1.

C. Datasets

Table 5 lists the basic information of the four datasets used in our experiments and they are shortly described as follows.

- **SST-2** (Socher et al., 2013) is a binary sentiment classification dataset that consists of movie reviews annotated with positive and negative labels.
- **EMOTION** (Saravia et al., 2018) is a dataset for emotion classification comprising short comments from social media annotated with six emotion types, such as, sadness, joy, etc.
- **TREC** (Li and Roth, 2002) is a dataset for question topic classification comprising questions across six categories including human, location, etc.
- **HumAID** (Alam et al., 2021) is a dataset for crisis messages categorisation comprising tweets collected during 19 real-world disaster events, annotated by humanitarian categories including rescue volunteering or donation effort, sympathy and support, etc.

D. Training Details

When finetuning the generation model, we select the pre-trained T5 base checkpoint as the starting weights. For the downstream classification task, we finetune "bert-base-uncased"[6] on the original training data either with or without the augmented samples. Regarding the pre-trained models, we use the publicly-released version from the HuggingFace's transformers library (Wolf et al., 2019). For the augmentation factor (i.e., β in Section 3.2), the augmentation techniques including ours and the baselines are applied to augment 1 to 5 times of original training data. In the experiments, it is regarded as a hyper-parameter to be determined. Since our work focuses on text augmentation for classification in low-data settings, we sampled 5, 10, 20, 50 and 100 examples per class for each training dataset as per Anaby-Tavor et al. (2020). To alleviate randomness, we run all experiments 10 times so the average accuracy along with its standard deviation (std.) is reported on the full test set in the evaluation.

To select the downstream checkpoint and the augmentation factor, we select the run with the best performance on the development set for all methods. The hyper-parameters for finetuning the generation model and the downstream model are also setup based on the development set. Although using the full development set does not necessarily represent a real-life situation in low-data regime (Schick and Schütze, 2021a; Gao et al., 2021), we argue that it is valid in a research-oriented study. We choose to use the full development set since we aim to maximize the robustness of various methods' best performance given small training data available. As all augmentation methods are treated the same way, we argue this is valid to showcase the performance difference between our method and the baselines.

For all experiments presented in this work, we exclusively use *Pytorch*[7] for general code and *Huggingface*[8] for transformer implementations respectively, unless otherwise stated. In finetuning T5, we set the learning rate to 5×10^{-5} using Adam (Kingma and Ba, 2014) with linear scheduler (10% warmup steps), the training epochs to be 32 and batch size to be 16. At generation time, we use top-k ($k = 40$) and top-p ($p = 1.0$) sampling technique (Holtzman et al., 2019) for next token generation. In finetuning downstream BERT, the hyper-parameters are similar to those of T5 finetuning, although the training epoch is set to be 20. We set the training epochs to be as large as possible with the aim of finding the best model when trained on a small dataset, where the quality is based on performance on the development set. In our experiments, for a single run on all datasets, it takes around one day with a single Tesla P100 GPU (16GB) and thus estimated 10 days for 10 runs. To aid reproducibility, we will release our experimental code to the public at [9].

E. Comparing to Few-shot Baselines

Since our work explores a text augmentation approach for improving text classification in low-data regime, it is also related to few-shot learning methods that use few examples for text classification. We further conduct an experiment to compare STA to three state-of-the-art few-shot learning approaches: PET (Schick and Schütze, 2021a), LM-BFF (Gao et al., 2021), and DART (Zhang et al., 2022). For fair comparison, we set the experiment under the 10 examples per class scenario with 10 random seeds ensuring the 10 examples per class are sampled the same across the meth-

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

[7] https://pytorch.org/

[8] https://huggingface.co/

[9] removedforreview

An example from **SST-2** a sentiment classification dataset where the classes ($\mathcal{L}$): negative, positive	
Text (x)	*top-notch action powers this romantic drama.*
Label (y)	*positive*
Converted examples by classification templates ($\mathcal{C}$: c, c_{pos} and c_{neg}): source(s), target(t)	
Given sentiment: negative, positive. Classify: top-notch action powers this romantic drama.	*positive*
Text: top-notch action powers this romantic drama. Is this text about positive sentiment?	*yes*
Text: top-notch action powers this romantic drama. Is this text about negative sentiment?	*no*
Converted examples by generation templates ($\mathcal{G}$: g and g'): source(s), target(t)	
Description: positive sentiment. Text:	*top-notch action powers this romantic drama.*
Description: positive sentiment. Text: top-notch action powers this romantic drama. Another text: spielberg 's realization of	*a near-future america is masterful .*
Description: positive sentiment. Text: top-notch action powers this romantic drama. Another text: a movie in	*which laughter and self-exploitation merge into jolly soft-porn 'em powerment . '*
Description: positive sentiment. Text: top-notch action powers this romantic drama . Another text: a tightly directed	*highly professional film that 's old-fashioned in all the best possible ways .*

Table 4: The demonstration of an example conversion by the prompt templates in Table 1 where the example's text is highlighted in blue and label is highlighted in red for readability.

Dataset	**# Train**	**# Dev**	**# Test**	**# Classes**
SST-2	~6k	692	~1.8k	2
EMOTION	16k	2k	2k	6
TREC	~5k	546	500	6
HumAID	~40k	6k	~11k	8

Table 5: Datasets statistics

	SST-2	EMOTION	TREC
DART	66.5 (5.8)	26.7 (3.0)	74.0 (2.7)
LM-BFF	71.1 (9.5)	30.2 (3.8)	**77.1 (3.0)**
PET	56.7 (0.8)	28.4 (1.0)	69.1 (1.1)
STA (ours)	**81.4 (2.6)**	**57.8 (3.7)**	70.9 (6.6)

Table 6: The comparison between STA and few-shot baselines using 10 examples per class on **SST-2** and **EMOTION** and **TREC**. The results are reported as average (std.) accuracy (in %) based on 10 random experimental runs. Numbers in **bold** indicate the highest in columns.

ods. Besides, we use `bert-base-uncased`[10] as the starting weights of the downstream classifier. The results are shown in Table 6. We found that although STA loses the best score to DART and LM-BFF on the **TREC** dataset, it substantially outperforms the few-shot baselines on **SST-2** and **EMOTION**. This tells us that STA is a competitive approach for few-shot learning text classification.

F. More Results of Classification Tasks

Table 7, Table 8 and Table 9 present the results of STA comparing to baselines in low-data settings for the **EMOTION**, **TREC** and **HumAID** classification tasks respectively.

G. Demonstration

Table 10 and Table 11 demonstrate some original examples and augmented examples by different methods. In comparison, the examples generated by STA tend to be not only diverse but also highly label relevant (semantic fidelity).

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

Augmentation Method	**5**	**10**	**20**	**50**	**100**
Baseline (No Aug.)	26.7 (8.5)	28.5 (6.3)	32.4 (3.9)	59.0 (2.6)	74.7 (1.7)
EDA	30.1 (6.2)	33.1 (4.3)	47.5 (5.0)	66.7 (2.7)	77.4 (1.8)
BT	32.0 (3.0)	37.4 (3.0)	48.5 (5.1)	65.5 (2.0)	75.6 (1.6)
BT-Hops	31.3 (2.6)	37.1 (4.6)	49.1 (3.5)	65.0 (2.3)	75.0 (1.5)
CBERT	29.2 (6.5)	32.6 (3.9)	44.1 (5.2)	62.1 (2.0)	75.5 (2.2)
GPT-2	28.4 (8.5)	31.3 (3.5)	39.0 (4.1)	57.1 (3.1)	69.9 (1.3)
GPT-2-λ	28.6 (5.1)	30.8 (3.1)	43.3 (7.5)	71.6 (1.5)	80.7 (0.4)
BART-Span	29.9 (4.5)	35.4 (5.7)	46.4 (3.9)	70.9 (1.5)	77.8 (1.0)
STA w/o Self-Checking	34.0 (4.0)	41.4 (5.5)	53.3 (2.2)	65.1 (2.3)	74.0 (1.1)
STA w/o Auxiliary Prompts	41.8 (6.1)	56.2 (3.0)	**64.9 (3.3)**	75.1 (1.5)	81.3 (0.7)
STA (ours)	**43.8 (6.9)**	**57.8 (3.7)**	64.1 (2.1)	**75.3 (1.8)**	**81.5 (1.1)**

Table 7: STA on **EMOTION** in $5, 10, 20, 50, 100$ examples per class. The results are reported as average (std.) accuracy (in %) based on 10 random experimental runs. Numbers in **bold** indicate the highest in columns.

Augmentation Method	**5**	**10**	**20**	**50**	**100**
Baseline (No Aug.)	33.9 (10.4)	55.8 (6.2)	71.3 (6.3)	87.9 (3.1)	93.2 (0.7)
EDA	54.1 (7.7)	70.6 (5.7)	79.5 (3.4)	89.3 (1.9)	92.3 (1.1)
BT	56.0 (8.7)	67.0 (4.1)	79.4 (4.8)	89.0 (2.4)	92.7 (0.8)
BT-Hops	53.8 (8.2)	67.7 (5.1)	78.7 (5.6)	88.0 (2.3)	91.8 (0.9)
CBERT	52.2 (9.8)	67.0 (7.1)	78.0 (5.3)	89.1 (2.5)	92.6 (1.1)
GPT-2	47.6 (7.9)	67.7 (4.9)	76.9 (5.6)	87.8 (2.4)	91.6 (1.1)
GPT-2-λ	49.6 (11.0)	70.2 (5.8)	80.9 (4.4)	**89.6 (2.2)**	**93.5 (0.8)**
BART-Span	55.0 (9.9)	65.9 (6.7)	77.1 (5.5)	88.38 (3.4)	92.7 (1.6)
STA w/o Self-Checking	45.4 (3.2)	61.9 (10.2)	77.2 (5.5)	88.3 (1.2)	91.7 (0.8)
STA w/o Auxiliary Prompts	49.6 (9.0)	69.1 (8.0)	81.0 (5.9)	89.4 (3.0)	93.1 (0.9)
STA (ours)	**59.6 (7.4)**	**70.9 (6.6)**	**81.1 (3.9)**	89.1 (2.7)	93.2 (0.8)

Table 8: STA on **TREC** in $5, 10, 20, 50, 100$ examples per class. The results are reported as average (std.) accuracy (in %) based on 10 random experimental runs. Numbers in **bold** indicate the highest in columns.

Augmentation Method	5	10	20	50	100
Baseline (No Aug.)	29.1 (6.6)	37.1 (6.4)	60.7 (4.0)	80.0 (0.9)	83.4 (1.0)
EDA	49.5 (4.5)	64.4 (3.6)	74.7 (1.5)	80.7 (1.0)	83.5 (0.6)
BT	45.8 (5.7)	59.1 (5.2)	73.5 (2.1)	80.4 (1.2)	83.1 (0.7)
BT-Hops	43.4 (6.4)	57.5 (5.2)	72.4 (2.8)	80.1 (1.1)	82.8 (1.4)
CBERT	44.8 (7.6)	59.5 (4.8)	73.4 (1.7)	80.3 (0.8)	82.7 (1.2)
GPT-2	46.0 (4.7)	55.7 (5.7)	67.3 (2.6)	77.8 (1.6)	81.1 (0.6)
GPT-2-λ	50.7 (8.6)	68.1 (6.2)	78.5 (1.3)	82.1(1.1)	84.2 (0.8)
BART-Span	42.4 (7.3)	58.6(7.0)	70.04 (3.7)	79.3 (1.4)	83.33 (0.9)
STA w/o Self-Checking	56.4 (7.0)	70.2 (4.3)	76.3 (3.3)	79.4 (4.5)	81.8 (1.3)
STA w/o Auxiliary Prompts	68.7 (10.9	**77.6 (3.6)**	80.1 (1.7)	82.9 (1.6)	84.3 (0.7)
STA (ours)	**69.0 (3.9)**	75.8 (3.3)	**80.2 (1.6)**	**83.2 (0.5)**	**84.5 (1.1)**

Table 9: STA on **HumAID** in $5, 10, 20, 50, 100$ examples per class. The results are reported as average (std.) accuracy (in %) based on 10 random experimental runs. Numbers in **bold** indicate the highest in columns.

Original training examples and augmented examples for “Sadness" of **EMOTION**	
Original	i sit here feeling blank about this i feel ashamed that i so readily turn it aside i feel positively ashamed when i look out of the window and see the state of things i had just lost my uncle i would be sad but i feel as if i am devastated i was feeling kind of discouraged because nothing happened
EDA	i sit here opinion blank about this i feel that ashamed i so readily turn it aside i feel positively ashamed when i look out of the window and construe the state of things i had just lost my uncle i would be pitiful but i feel as if i am devastated i happened feeling kind of discouraged because nothing was
GPT-2-λ	ive seen so many girls walk around feeling ashamed of their bodi ive got to admit that i feel a little weird for a moment seeing her standing in front of my face when i walk into the shop ive always wondered what im doing right now im feeling ive read many blogs about her and how much she hates those who don’t admit to being kind or caring about others but instead blame them for not doing something about it ive never felt sympathetic towards people because of the way they look and act because of their skin to
STA-noself	i feel like the whole world is watching and feeling it’s failing me i want people to know i am not alone i feel ashamed when i look out of the window and see the state of things i walked away feeling disappointed because i don t know the answer i drank some cold drink or find some ice dessert such as chendol or ice kacang
STA	i feel sad seeing people who have to work harder to cope i walked away feeling disappointed because i don t know the answer i was feeling sad seeing the state of things that i never did i really want to see if it lasted i feel sad seeing the state of things but the truth is im not sure how to express it gracefully i feel like the whole world is watching and feeling it’s failing me

Table 10: The demonstration of original training examples and augmented examples for “sadness” of **EMOTION**. It is noted that the 5 augmented examples in each block are randomly selected instead of cherry-picked. This reveals some difference between the original training examples and the augmented examples by our STA and other methods (Here we use a rule-based heuristics method EDA, a generation-based method GPT-2-λ and STA-noself for comparison).

Original training examples and augmented examples for “missing or found people" of HumAID	
Original	UPDATE: Body found of man who disappeared amid Maryland flooding Open Missing People Search Database from Mati and Rafina areas #Greecefires #PrayForGreece #PrayForAthens @ThinBlueLine614 @GaetaSusan @DineshDSouza case in point, #California Liberalism has created the hell which has left 1000s missing 70 dead,... Heres the latest in the California wildfires #CampFire 1011 people are missing Death toll rises to 71 Trump blames fires on poor ... #Idai victims buried in mass grave in Sussundenga, at least 60 missing - #Mozambique #CycloneIdai #CicloneIdai
EDA	update flooding found of man who disappeared amid maryland boy open missing people search database from mati escape and rafina areas greecefires prayforgreece prayforathens created gaetasusan dineshdsouza hell in point california missing has thinblueline the case which has left s liberalism dead an countless people... heres blames latest in the california wildfires campfire people are missing death toll rises to trump more fires on poor... idai victims buried in mass grave in sussundenga at mozambique missing least cycloneidai cicloneidai
GPT-2-lambda	@KezorNews - Search remains in #Morocco after @deweathersamp; there has been no confirmed death in #Kerala #Cambodia - Search & Rescue is assisting Search & Rescue officials in locating the missing 27 year old woman who disappeared in ... @JHodgeEagle Rescue Injured After Missing Two Children In Fresno County #Florence #Florence Missing On-Rescue Teams Searching For Search and Rescue Members #Florence #Florence #DisasterInformer #E RT @LATTAODAYOUT: RT @HannahDorian: Search Continues After Disappearance of Missing People in Florida
STA-noself	Search Database from Matias, Malaysia, missing after #Maria, #Kerala, #Bangladesh #KeralaKerala, #KeralaFloods, ... RT @hubarak: Yes, I can guarantee you that our country is safe from flooding during the upcoming weekend! Previous story Time Out! 2 Comments The missing persons who disappeared amid Maryland flooding are still at large. More on this in the next article. the number of missing after #CycloneIdai has reached more than 1,000, reports CNN. RT @adriane@przkniewskiZeitecki 1 person missing, police confirm #CycloneIdai. #CicloneIdai
STA	The missing persons who disappeared amid Maryland flooding are still at large. More on this in the next article. Search Triangle County for missing and missing after #Maria floods #DisasterFire Just arrived at San Diego International Airport after #Atlantic Storm. More than 200 people were missing, including 13 helicopters ... Search Database contains information on missing and found people #HurricaneMaria, hashtag #Firefighter Were told all too often that Californians are missing in Mexico City, where a massive flood was devastating. ...

Table 11: The demonstration of original training examples and augmented examples for “missing or found people” of **HumAID**. It is noted that the 5 augmented examples in each block are randomly selected instead of cherry-picked. This reveals some difference between the original training examples and the augmented examples by our STA and other methods (Here we use a rule-based heuristics method EDA, a generation-based method GPT-2-λ and STA-noself for comparison).