Appendix A. Example class distributions of the imbalanced data and hyperparameters used

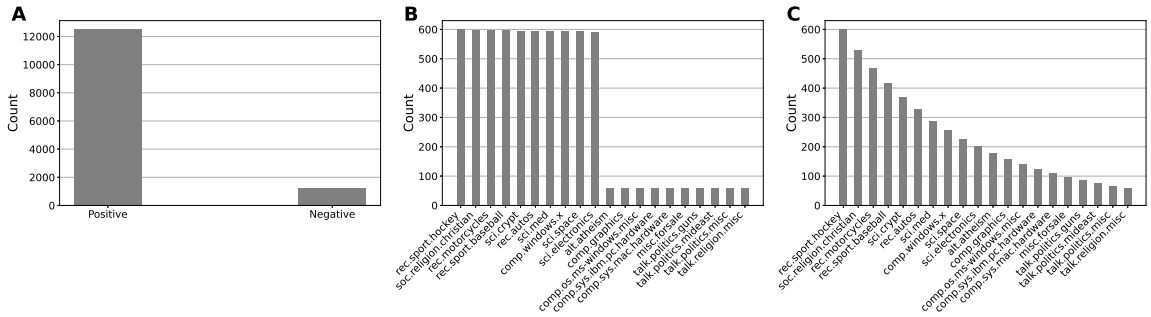

Figure 6: Number of training samples per class in artificially created imbalanced data. (A) The IMDB with step-imbalance class distribution. (B) The 20 Newsgroups with step-imbalance class distribution. (C) The 20 Newsgroups with long-tailed class distribution.

Table 2: Hyperparameters used for individual datasets

		IMDB	20 Newsgroups	ADME
Batch Size		8	8	16
Maximum Sequence Length		512	512	128
Vanilla Finetuning	Learning Rate	1e-5	1e-5	1e-5
	Epoch	3	5	3
Pre-finetuning with DA	# of Top Layers	1	1	1
	Learning Rate (Pre-finetune/Finetune)	1e-4/1e-5	1e-4/1e-5	1e-4/1e-5
	Epoch (Pre-finetune/Finetune)	1/2	1/5	1/3

Appendix B. Results of the 20 Newsgroups Dataset

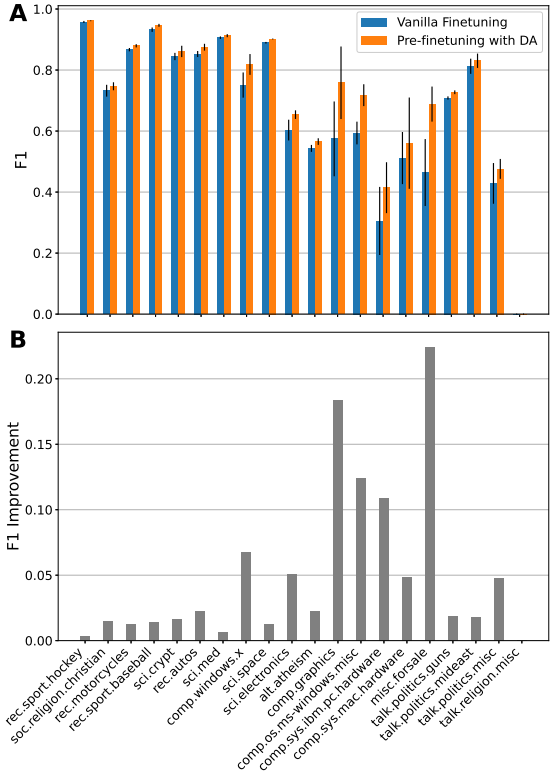

Figure 7: Per-class F1 comparison between the vanilla finetuning and our method on the 20 Newsgroups benchmark with *long-tailed* data. A: Comparison of the F1 between the vanilla finetuning and our method. The error bar: SEM. B: The absolute improvements, where substantial improvements were observed in minority classes.

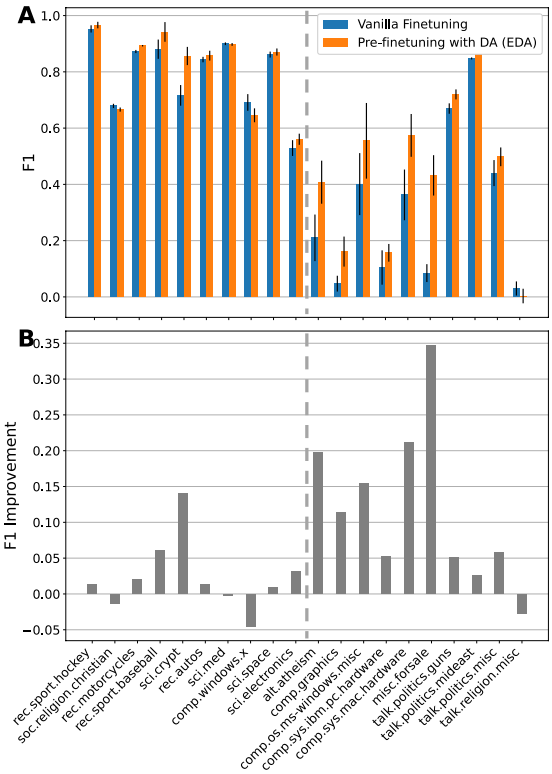

Figure 8: Per-class F1 comparison between the vanilla finetuning and our pre-finetuning with data augmentation by *EDA* on the 20 Newsgroups data with *step-imbalance*. The classes on the left of the dash line are the majority classes, and the ones on the right are the minority classes. A: Comparison of the F1 between the vanilla finetuning and our method. The error bar: SEM. B: The absolute improvements, where substantial improvements were observed in minority classes, which have a similar trend as the back translation.

Appendix C. Results of Imbalanced Testing Dataset

Table 3: F1 and per-class F1 (SEM) of IMDB benchmark with the step-imbalance testing data

	F1	Per-class F1	
		Negative	Positive
Vanilla Finetuning	0.9637 (0.0003)	0.7869 (0.0028)	0.9801 (0.0001)
Pre-finetuning with DA	0.9657 (0.0005)	0.7982 (0.0029)	0.9812 (0.0002)

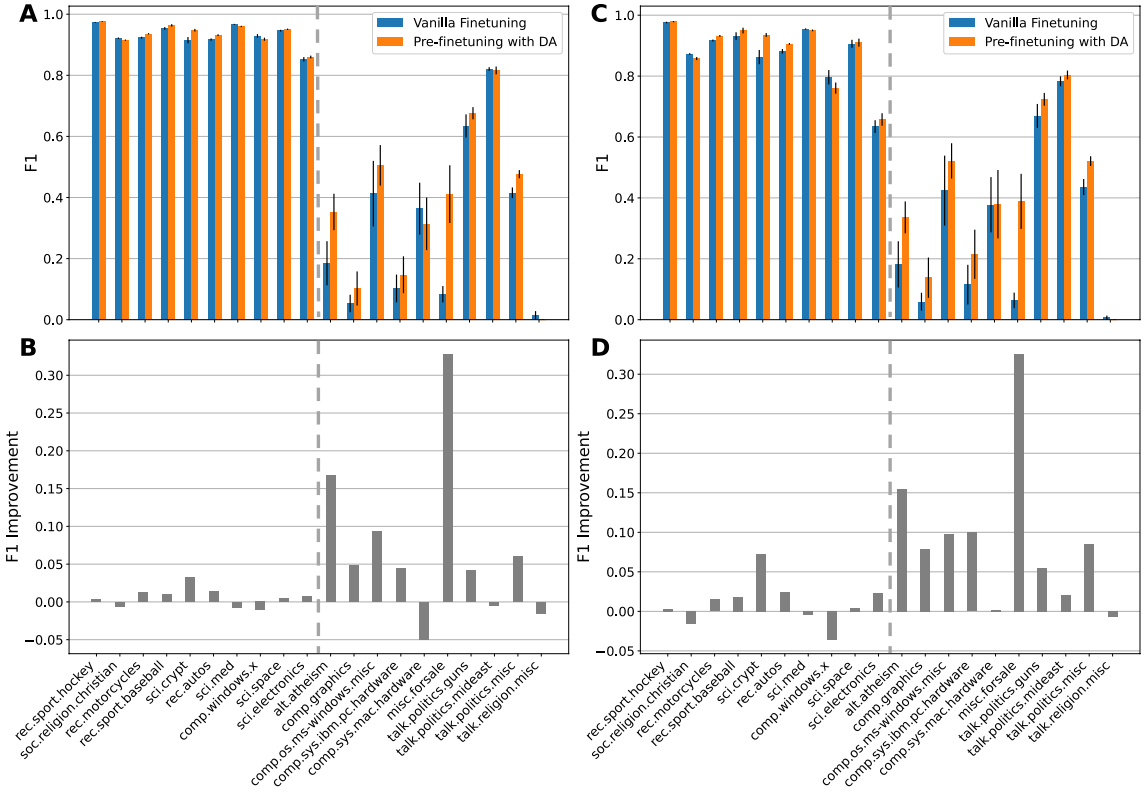

Figure 9: Per-class F1 comparison between the vanilla finetuning and our method on the 20 Newsgroups data with *step-imbalance*. A, B: The per-class F1 and improvements with *step-imbalance* testing dataset, respectively. C, D: The per-class F1 and improvements with *long-tailed* testing dataset, respectively. The classes on the left of the dash line are the majority classes, and the ones on the right are the minority classes. The per-class F1 of minority classes had more improvements, a trend similar to that in the balanced testing dataset.

Appendix D. Results for the 5-fold Cross Validation

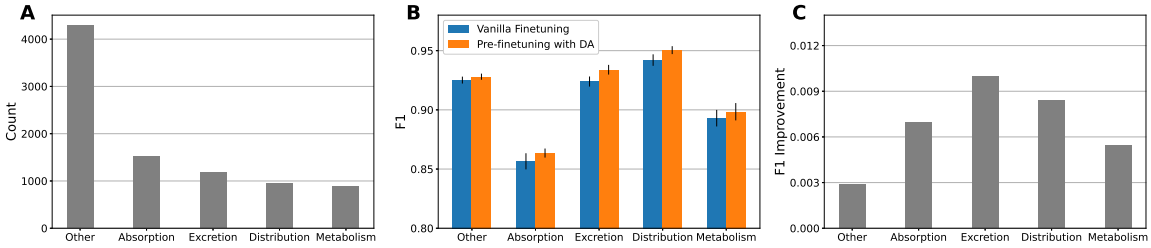

Figure 10: Per-class F1 scores and improvements on ADME dataset with stratified 5-fold cross validation. A: The class distribution of the training set. B: The per-class F1-score comparison between the two methods (The vanilla method and our pre-finetuning with DA). C: Improvement in per-class F1 of our method over the vanilla method. The F1-score increase is observed in all classes with our method. The improvement becomes larger in the less frequent classes. Error bars: SEM.