Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks

The paper discusses about the effects of continued Pretraining(the process where a neural language model (LM), with large number of parameters, is trained on large unlabeled data and for any further downstream tasks, the weight after the pretraining are fine tuned) of an existing model and extends the idea about involving two techniques i.e Domain Adaptive Pretraining(***DAPT***) & *Task Adaptive Pretraining (**TAPT**)* , which can be used to improve performance gains in both low and high resource setting for a domain specific task.

The model selected here is RoBerta - a transformer based Language Model (LM) trained with a masked language modeling objective on unlabeled data(derived from various sources) of over 160 Gb.

For the analysis, the paper uses four domains with both high and low resource datasets for a classification task: (i)Biomedical(**BM)** Domain (ii)CS(**CS)** domain, (iii)News(**NEWS)** domain and Reviews(**REVIEWS**) domain.

Domain Adaptive Pretraining(*DAPT*)

The main idea is quite direct i.e pretraining ROBERTA on a large corpus of unlabeled domain related text. For this task , the **domain similarity** plays an important role in the analysis as domain relevance to a certain domain , enhances the claim that the better/improved result is not simply a result of exposure to more data. So for checking the similarity , top 10000 most frequent words(excluding stop words) from the held out samples out of the large unlabeled corpus(50K for other sources ,including RoBerta and 150k for Reviews) and checked the vocabulary overlap.
It was observed that RoBerta’s pretraining domain has strong vocabulary overlap with NEWS and REVIEWS , thus more related as compared to CS and BIOMEDICAL.
Experiments → For each domain , two classification based tasks were selected and were classified on the standard classification architecture(discussed in the paper of Jacob Devlin) with a final layer of feed forward Neural Network for the task of prediction for 12.5k time steps . The tasks involve both high resource and low resource(less than 5K labeled training examples, and no additional unlabeled data) settings .
There was another setting in which the experiment was conducted . In it , a Language Model out of the domain of interest was trained upon the Domain specific data to be used for DAPT . This was performed to showcase the evidence that the surge in performance isn’t attributed to exposure to more data. This was called ¬DAPT.
Results/Conclusions → The adaptation to DAPT resulted as an enhancement to performance in all the tasks. The enhancement was more in domains with more domain dissimilarity as compared to corpus used to train the RoBerta model(CS and BM exhibited more improvement as compared to NEWS and REVIEWS), thus implying the more dissimilar the domain, the higher the potential for improvement using DAPT.
The results for ¬DAPT was a worse performing setting in which the results extracted were worse in comparison to RoBerta base . Thus , consolidates the claims on how relevant the domain is.

Even though tasks for different domains might be differentiable but all this difference is not mutually exclusive and **DOMAIN OVERLAP** does occur.

Task Adaptive Pretraining(*TAPT*)

Here the LM is pretrained on the unlabeled training set for a given task,thus the data distribution in a task is realised.

Experiment → The training process is still the standard process that we used in *DAPT* (one involving the latest layer of Feed Forward Neural Network) and training it for 100 epochs.The corpus size is small in comparison to the one required for DAPT.The number of settings in which the TAPT approach was tried can be categorised as :- (i) **Combined DAPT and TAPT**[a computationally expensive approach where we first train RoBerta under DAPT and then TAPT] (ii) **Cross Transfer**[here we pretrain on one task and finetune on the other task of the same domain] & (iii) **Data Augmented TAPT**[here we consider even larger size of unlabeled data which is either hand-curated or retrieved from sources].

Results → The performance enhanced by using TAPT across all domains and in some tasks , gets the better of a much more computationally expensive , DAPT approach. The setting of Combined DAPT and TAPT further enhanced the performance as compared to TAPT , but this approach was "order specific" as Cross Transfer setting's results further demonstrated that data distributions of tasks within a given domain do differ so reversing the order will lead to catastrophic result as predicted in Yogatama's Paper.

Data Augmentation → This setting involved two approaches :- (i) Human Curated TAPT (ii)Automated Data Selection for TAPT . Human Curated TAPT involves sampling some labeled data and treating rest as unlabeled from existing dataset & extracting human annotated unlabeled data. Whereas Automated Data Selection uses **VAMPIRE**, bag-of-words language model pretrained on samples of similar domain to obtain embeddings of the text from both the task and domain sample, and then select either top-k similar embeddings or select randomly.

Results show that Human Curated TAPT outperforms every other setting and comes close even with a very small percentage of labeled data while Automated Data selection approach outperforms TAPT approach and the performance steadily increases as we increase k . But the random selection Automated Data Selection setting generally performs worse than TAPT.

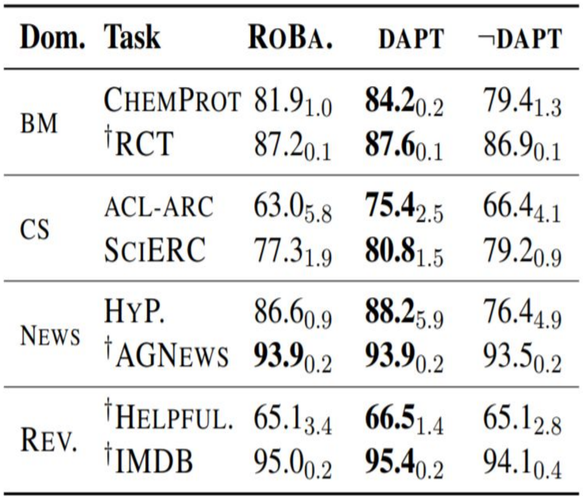
Related Work

1. Transfer Learning from Domain Adaptation → This paper by Alsentzer demonstrated how continued pretraining in domain acts beneficial for certain tasks.
2. The approach of TAPT as described in the paper was not a novel approach but was an idea used by Howard and Sebastian Ruder in their paper "Universal language model fine-tuning for text classification "
3. The Data Selection involved Selecting data for transfer learning . , Aharoni and
4. Goldberg (2020) proposed data selection methodologies for NMT based on cosine similarity in embedding space, using DISTILBERT

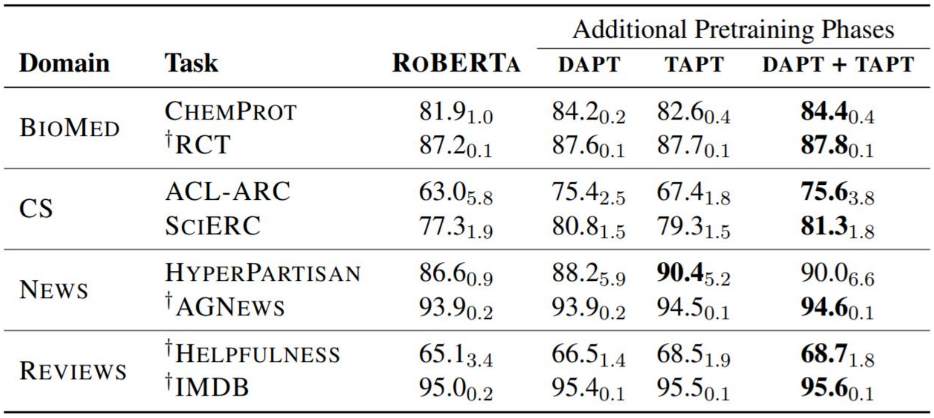
So the paper concludes that even a model of hundreds of millions of parameters does struggle to encode the complexity of a single textual domain and thus pretraining the model towards a specific task or small corpus can prove beneficial.

Evidences

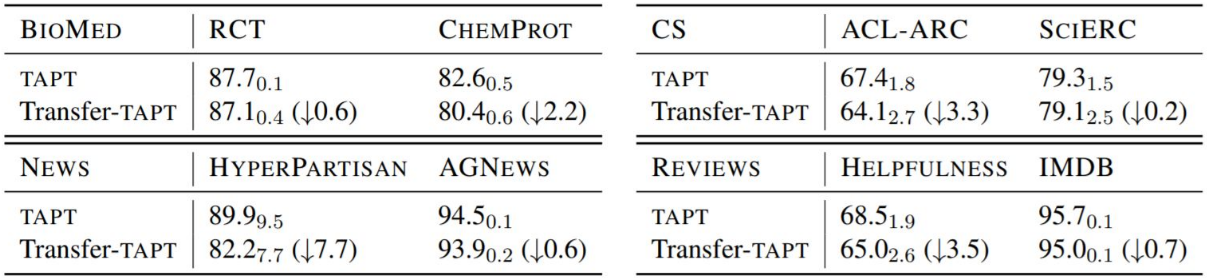
Following are the results' measures as measured by micro-F1 parameters on each of the settings . The naming remains same as mentioned in the above study.



Dom.	Task	RoBa.	DAPT	¬DAPT
BM	ChemProt	81.9$_{1.0}$	**84.2**$_{0.2}$	79.4$_{1.3}$
	†RCT	87.2$_{0.1}$	**87.6**$_{0.1}$	86.9$_{0.1}$
CS	ACL-ARC	63.0$_{5.8}$	**75.4**$_{2.5}$	66.4$_{4.1}$
	SciERC	77.3$_{1.9}$	**80.8**$_{1.5}$	79.2$_{0.9}$
News	HyP.	86.6$_{0.9}$	**88.2**$_{5.9}$	76.4$_{4.9}$
	†AGNews	**93.9**$_{0.2}$	**93.9**$_{0.2}$	93.5$_{0.2}$
Rev.	†Helpful.	65.1$_{3.4}$	**66.5**$_{1.4}$	65.1$_{2.8}$
	†IMDB	95.0$_{0.2}$	**95.4**$_{0.2}$	94.1$_{0.4}$



			Additional Pretraining Phases		
Domain	**Task**	**RoBERTa**	**DAPT**	**TAPT**	**DAPT + TAPT**
BioMed	ChemProt	81.9$_{1.0}$	84.2$_{0.2}$	82.6$_{0.4}$	**84.4**$_{0.4}$
	†RCT	87.2$_{0.1}$	87.6$_{0.1}$	87.7$_{0.1}$	**87.8**$_{0.1}$
CS	ACL-ARC	63.0$_{5.8}$	75.4$_{2.5}$	67.4$_{1.8}$	**75.6**$_{3.8}$
	SciERC	77.3$_{1.9}$	80.8$_{1.5}$	79.3$_{1.5}$	**81.3**$_{1.8}$
News	HyperPartisan	86.6$_{0.9}$	88.2$_{5.9}$	**90.4**$_{5.2}$	90.0$_{6.6}$
	†AGNews	93.9$_{0.2}$	93.9$_{0.2}$	94.5$_{0.1}$	**94.6**$_{0.1}$
Reviews	†Helpfulness	65.1$_{3.4}$	66.5$_{1.4}$	68.5$_{1.9}$	**68.7**$_{1.8}$
	†IMDB	95.0$_{0.2}$	95.4$_{0.1}$	95.5$_{0.1}$	**95.6**$_{0.1}$



BioMed	RCT	ChemProt
TAPT	87.7$_{0.1}$	82.6$_{0.5}$
Transfer-TAPT	87.1$_{0.4}$ (↓0.6)	80.4$_{0.6}$ (↓2.2)

CS	ACL-ARC	SciERC
TAPT	67.4$_{1.8}$	79.3$_{1.5}$
Transfer-TAPT	64.1$_{2.7}$ (↓3.3)	79.1$_{2.5}$ (↓0.2)

News	HyperPartisan	AGNews
TAPT	89.9$_{9.5}$	94.5$_{0.1}$
Transfer-TAPT	82.2$_{7.7}$ (↓7.7)	93.9$_{0.2}$ (↓0.6)

Reviews	Helpfulness	IMDB
TAPT	68.5$_{1.9}$	95.7$_{0.1}$
Transfer-TAPT	65.0$_{2.6}$ (↓3.5)	95.0$_{0.1}$ (↓0.7)

Pretraining	BioMed		CS
	ChemProt	RCT-500	ACL-ARC
RoBERTa	$81.9_{1.0}$	$79.3_{0.6}$	$63.0_{5.8}$
TAPT	$82.6_{0.4}$	$79.8_{1.4}$	$67.4_{1.8}$
RAND-TAPT	$81.9_{0.6}$	$80.6_{0.4}$	$69.7_{3.4}$
50NN-TAPT	$83.3_{0.7}$	$80.8_{0.6}$	$70.7_{2.8}$
150NN-TAPT	$83.2_{0.6}$	$81.2_{0.8}$	$73.3_{2.7}$
500NN-TAPT	$83.3_{0.7}$	$81.7_{0.4}$	**75.5**$_{1.9}$
DAPT	**84.2**$_{0.2}$	**82.5**$_{0.5}$	$75.4_{2.5}$

Pretraining	BioMed	News	Reviews
	RCT-500	HyP.	IMDB †
TAPT	$79.8_{1.4}$	$90.4_{5.2}$	$95.5_{0.1}$
DAPT + TAPT	$83.0_{0.3}$	$90.0_{6.6}$	$95.6_{0.1}$
Curated-TAPT	$83.4_{0.3}$	$89.9_{9.5}$	$95.7_{0.1}$
DAPT + Curated-TAPT	**83.8**$_{0.5}$	**92.1**$_{3.6}$	**95.8**$_{0.1}$