Don’t Stop Pretraining:

Adapt Language Models to Domain and Tasks

Authors Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, Noah A. Smith

Published in ACL 2020, Annual Meeting of the Association for Computational Linguistics (ACL)

Team 4

KAIST
Sangwook Lee

KAIST
Wenchao Dong

KAIST
Taeyeong Lee

KAIST
William Gyuho Suh

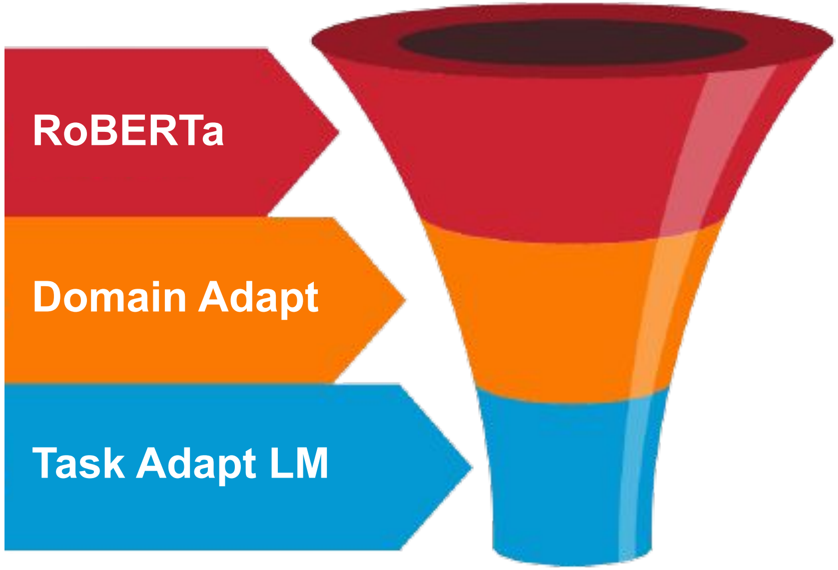
Introduction

Pre-trained Language Models

BERT (RoBERTa)

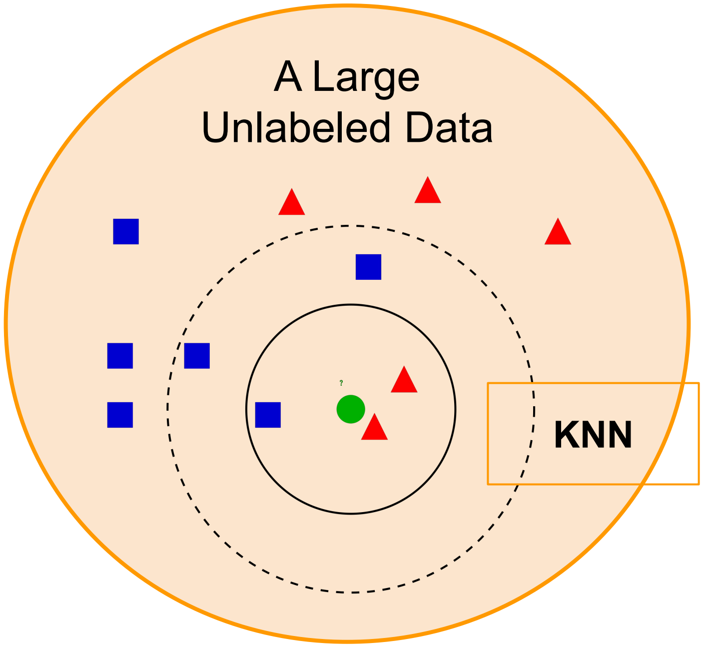
Adapt LM to specific domains

Replication Approach



1. **Baseline Model (RoBERTa)**
   - Large pre-training corpus
2. **Domain Adaptive Pre-Training (DAPT)**
   - Domain specific corpus (4 domains)
3. **Task Adaptive Pre-Training (TAPT)**
   - Task specific datasets (2 tasks/domain)

Replication Approach



Augmenting Training Data for TAPT

4. **Human Curated-TAPT**
   - Human finds task-relevant data
5. **Automated Data Selection for TAPT**
   - Automatically selects *k* candidates via nearest neighbors selection (kNN-TAPT)

Datasets

Baseline: **RoBERTa**

DAPT: **4 domain-specific** datasets, TAPT: **8 task-specific** datasets

Domain	Pretraining Corpus	Task	Label Type
BIOMED	2.68M full-text papers from S2ORC(Lo et al., 2019)	CHEMPROT(Kringelum et al., 2016) RCT(Dernoncourt and Lee, 2017)	relation classification abstract sent. roles
CS	2.22M full-text papers from S2ORC(Lo et al., 2019)	ACL-ARC(Jurgens et al., 2018) SCIERC(Luan et al., 2018)	citation intent relation classification
NEWS	11.90M atricles from REALNEWS(Zellers et al., 2020)	HYPERPARTISAN(Kiesel et al., 2019) AGNEWS(Zhang et al., 2015)	partisanship topic
REVIEWS	24.75M AMAZON reviews(He and McAuley, 2016)	HELPFULLNESS(McAuley et al., 2015) IMDB(Maas et al., 2011)	review helpfulness review sentiment

Table 1: List of 4 domain-specific unlabeled datasets sources and their corresponding task-specific datasets

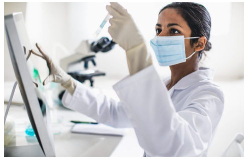
Experiment Process and Results with 4 replication experiments and 2 improvements

Overview

- Replications
  - Domain Adaptive Pretraining
  - Task Adaptive Pretraining
  - Human-Curated TAPT
  - Automated Data Selection for TAPT
- Improvements
  - Youtube Misinformation
  - GLUE Benchmark

Domain Adaptive Pretraining

- Pretrained on each of the **four domains.**

Biomedical	Computer	News	Reviews
			

- Selected two DAPT models with the **lowest correlation** between two of them.

Biomedical ⟷ Reviews   Computer Science ⟷ News

- Measured the effect of **irrelevant domain pre-training**. For example, pretraining on Biomed domain-specific dataset and Reviews task-specific dataset.

Results - Domain Adaptive Pretraining

- **DAPT** outperforms **¬DAPT** to varying degrees.
- It means that it is important to **continue pretraining** on **domain-relevant** data.

Dom.	Task	ROBA.	DAPT	¬DAPT
BM	CHEMPROT	$81.4_{1.0}$	$\mathbf{84.3}_{0.6}$	$79.5_{0.8}$
	RCT	$79.6_{0.6}$	$\mathbf{82.5}_{0.6}$	$77.3_{0.5}$
CS	ACL-ARC	$63.8_{4.3}$	$\mathbf{75.1}_{2.1}$	$63.0_{1.9}$
	SCIERC	$78.0_{3.7}$	$80.1_{0.7}$	$\mathbf{81.0}_{1.0}$
NEWS	HYP.	$\mathbf{92.3}_{2.4}$	$86.1_{10.1}$	$70.9_{3.0}$
	AGNEWS	$\mathbf{93.6}_{0.2}$	$\mathbf{93.6}_{0.2}$	$93.4_{0.2}$
REV.	HELPFUL.	$64.7_{0.4}$	$\mathbf{68.8}_{2.3}$	$65.9_{2.2}$
	IMDB	$94.5_{0.2}$	$\mathbf{95.0}_{0.1}$	$94.0_{0.2}$

Task Adaptive Pretraining

- We compared the performance between **DAPT**, **TAPT**, and **DAPT+TAPT**.
- Experimented in **4 Domains** and **8 corresponding Tasks**.

Domain	Task	RoBERTa	Additional Pretraining Phases		
			DAPT	TAPT	DAPT+TAPT
BM	CHEMPROT				
	RCT				
CS	ACL-ARC				
	SCIERC				
NEWS	HYP.				
	AGNEWS				
REV.	HELPFUL.				
	IMDB				

Results - Task Adaptive Pretraining

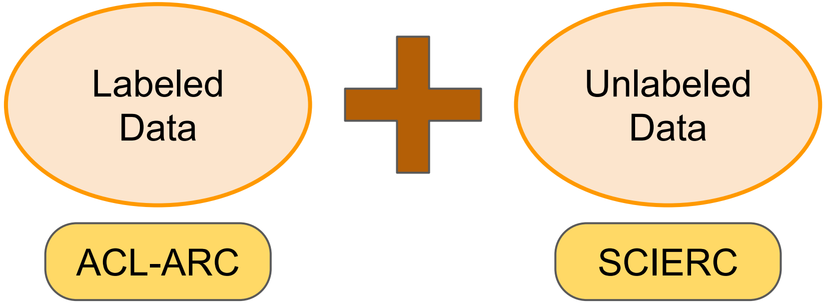
- **DAPT+TAPT** improves the **baseline model** for all tasks except one task.
- **DAPT+TAPT** also outperformed **TAPT** on most tasks.

Domain	Task	RoBERTa	Additional Pretraining Phases		
			DAPT	**TAPT**	**DAPT+TAPT**
BM	CHEMPROT	$81.4_{1.0}$	$\mathbf{84.3}_{0.6}$	$82.3_{0.6}$	$84.1_{1.0}$
	RCT	$79.6_{0.8}$	$76.2_{0.6}$	$80.3_{0.6}$	$\mathbf{82.9}_{0.1}$
CS	ACL-ARC	$63.8_{4.3}$	$\mathbf{75.1}_{2.1}$	$69.4_{1.5}$	$74.1_{3.8}$
	SCIERC	$78.0_{3.7}$	$80.1_{0.7}$	$79.2_{0.8}$	$\mathbf{80.5}_{1.0}$
NEWS	HYP.	$\mathbf{92.3}_{2.4}$	$86.1_{10.1}$	$87.5_{4.8}$	$82.9_{11.2}$
	AGNEWS	$93.6_{0.2}$	$93.5_{0.2}$	$94.1_{0.1}$	$\mathbf{94.2}_{0.1}$
REV.	HELPFUL.	$64.7_{0.4}$	$\mathbf{69.9}_{1.7}$	$69.0_{1.8}$	$68.6_{1.4}$
	IMDB	$94.5_{0.2}$	$95.0_{0.1}$	$95.0_{0.3}$	$\mathbf{95.3}_{0.2}$

Cross Task Transfer

- Checked the result of pre-training LM using both **same domain TAPT**s.
- In the CS domain, for example, we pre-trained the Language Model on ACL-ARC **unlabeled data** and fine-tuned it using SCIERC **labeled data**.

Computer Science



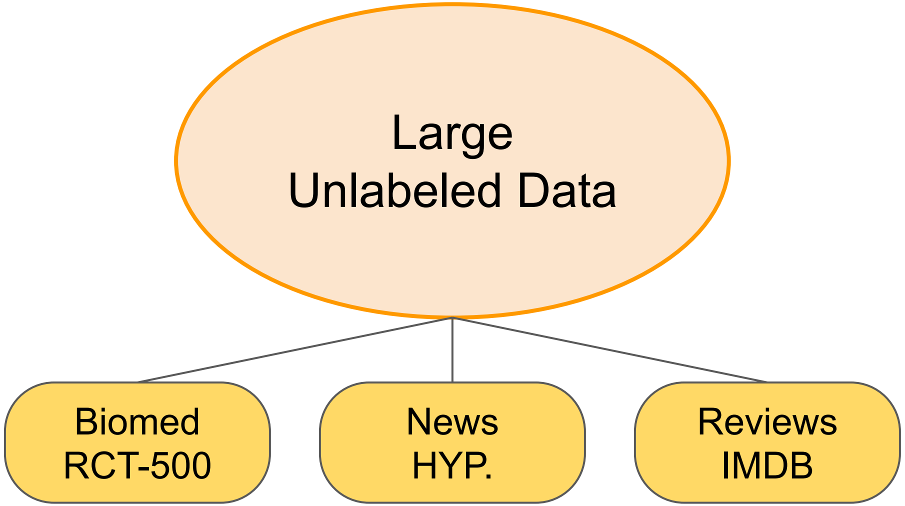
Results - Cross Task Transfer

- **Only one** experiment resulted in a slight improvement in performance.
- The remaining **7 Transfer-TAPT**s did not perform as well as the **TAPT**s.
- It suggest that in most cases, the distribution of data for different tasks belonging to the same domain is different.

BIOMED	**RCT**	**CHEMPROT**
TAPT	$80.3_{0.6}$	$82.3_{0.6}$
Transfer-TAPT	$79.5_{0.4} (\downarrow 0.8)$	$81.9_{1.2} (\downarrow 0.4)$
NEWS	**HYPERPARTISAN**	**AGNEWS**
TAPT	$87.5_{4.8}$	$94.1_{0.1}$
Transfer-TAPT	$81.6_{6.5} (\downarrow 5.9)$	$93.6_{0.3} (\downarrow 0.5)$

CS	**ACL-ARC**	**SCIERC**
TAPT	$69.4_{1.5}$	$79.2_{0.8}$
Transfer-TAPT	$70.3_{1.8} (\uparrow 0.9)$	$78.9_{1.7} (\downarrow 0.3)$
REVIEWS	**HELPFULNESS**	**IMDB**
TAPT	$69.0_{1.8}$	$95.0_{0.3}$
Transfer-TAPT	$65.2_{1.8} (\downarrow 3.8)$	$94.5_{0.1} (\downarrow 0.5)$

Human Curated TAPT



- Use human curated data for Curated-TAPT unlabeled data.
- Experiments are conducted on RCT-500, HYP., IMDB that includes a **large unlabeled data**.
- We compared Curated-TAPT with TAPT and DAPT+TAPT respectively.

Results - Human Curated TAPT

- **Curated-TAPT** produces better result compare to **TAPT** and **DAPT+TAPT**.
- On the RCT-500 and IMDB tasks, curating a data from the task distribution has a great influence on performance improvement.

Pretraining	BIOMED RCT-500	NEWS HYP.	REVIEWS IMDB
TAPT	$80.3_{0.6}$	$\mathbf{87.5}_{4.8}$	$95.0_{0.3}$
DAPT+TAPT	$82.9_{0.1}$	$82.9_{11.2}$	$95.3_{0.2}$
Curated-TAPT	$83.1_{0.3}$	$82.3_{15.4}$	$\mathbf{95.6}_{0.1}$
DAPT+Curated-TAPT	$\mathbf{83.5}_{0.5}$	$82.4_{8.9}$	$95.4_{0.2}$

Automatic Data selection for TAPT

- Human Curated TAPT is effective, but burdensome
- **VAMPIRE** - variational methods for pretraining in resource-limited environment
- N-Gram model with unlabeled text
- Consist of three part
  - VAE - pre-train network with word frequency representation of unlabeled corpus
  - VAMPIRE embedding : VAE + labeled text (world)
  - Classification : unlabeled -> label

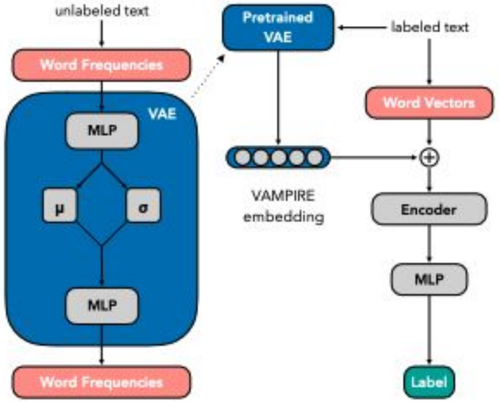
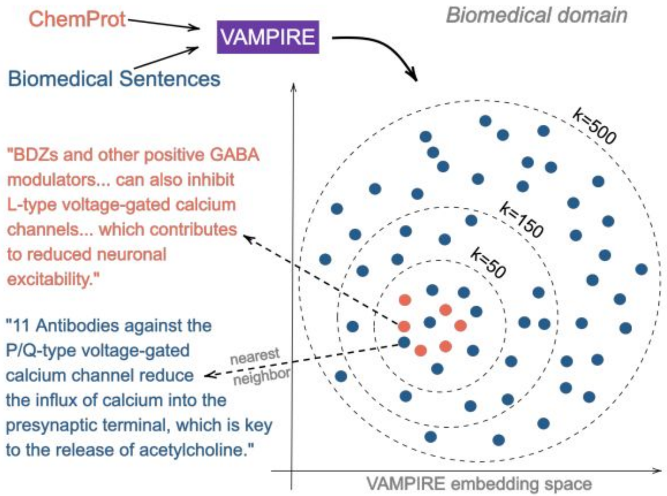


Figure 1: VAMPIRE involves pretraining a deep variational autoencoder (VAE; displayed on left) on unlabeled text. The VAE, which consists entirely of feed-forward networks, learns to reconstruct a word frequency representation of the unlabeled text with a logistic normal prior, parameterized by μ and σ. Downstream, the pretrained VAE's internal states are frozen and concatenated to task-specific word vectors to improve classification in the low-resource setting.

Strategy - Automatic Data selection for TAPT

- Label text → RCT-500
- Unlabeled corpus → BioMed (100MB)
- Restriction
  - Hard Copyright for Domain corpus in latest ver. (X CS , Small BioMed)
  - A huge time for pre-training VAMPIRE
  - Limit Computing power (X Chemprot)



Strategy - Automatic Data selection for TAPT

- Result of MRPC augment dataset with HC corpora news data as unlabeled

```
Source: At best, Davydenko's supporters were naively ignorant of tennis etiquette; at worst, they cheated - yet went without penality from umpire Lars Graf.
Neighbor_0: After being beaten  Hunter left FAMU and gave up an  82 000 scholarship
Neighbor_1: As Lee took the basketball to the rim  he glanced at Faried
Neighbor_2: And there you have the subtext that has made the 2012 redistricting so emotional that Lopez went home last week and  still in his suit  sat on his$
Neighbor_3: A ninth grade girl smiled shyly when asked about her school
Neighbor_4: A few days later  Rutgers student Scott Xu testified  Ravi told his ultimate Frisbee teammates that he had set up the webcam again
Neighbor_5: After the verdict was read  Crockam was led away in handcuffs
Neighbor_6: Adomaitis woke up at 3 a m   got his teammates up to run laps because he thought it was 6 in the morning
Neighbor_7: After winning the event  Suhr took three  unsuccessful attempts at 16 4 %  which would have broken her own U S  record
Neighbor_8: Akinyele turned the floor over to Thomas  who cleared his throat and stood up straight  He scanned the young faces looking back at him
Neighbor_9: Asked by a prosecutor why she went along with it  Young put her hands together  pressed them to her chin and bowed her head as if in prayer  As sh$
Neighbor_10: And sometimes a coach is tested with tears to earn the legitimacy of that title
Neighbor_11: After several minutes  the dance ended  I bowed to her  pressing my hands together in a universally understood gesture of thanks  The crowd appla$
Neighbor_12: As LaBove held him next to the wreckage  Kinison struggled out loud with the thought of dying  then came to grips with it   And I realized he was$
Neighbor_13: After her testimony  Hudson  clutching tissues  walked slowly directly in front of the jury as she crossed the courtroom  She then took a seat in$
Neighbor_14: And there was a brief   tantalizingly  sadistically brief   shot of Batman and Bane finally going at it  mano a mano
Neighbor_15: After the prosecution finished its opening statements  Swor took to the podium and said   Wow  great show
Neighbor_16: After the final arguments had ended and after the courtroom had emptied  Senser  45  of Edina  was asked about her thoughts now that seven days o$
Neighbor_17: Andrea said she and the other students were threatened with disciplinary action for walking out   but we did it anyway
Neighbor_18: Anoka High counselor Barry Terrass said he wore the red T shirt Friday because a student asked him to
Neighbor_19: About a half dozen students watched a piece of the encounter  in which Clementi and M B  were seen kissing
Neighbor_20: Although still weak at times  Kendall finished fifth  As she walked down the 16th fairway during the final round  she felt her eyes fill with tea$
Neighbor_21:  Actions speak louder than words   Russell told jurors as she began her final argument   That is a phrase I would like you to keep in mind
Neighbor_22: Altman questioned M B  s appearance at the time of the meetings  trying to strengthen the defense s contention that Ravi watched the two men only$
Neighbor_23: Asked how much he thought about defense while in high school  Faust laughed
Neighbor_24: Applause broke out among several dozen onlookers on the street when Goodwin reached the balcony around 5 20 and waved before his arrest
```

Result - Automatic Data selection for TAPT

- **Aug + TAPT** improves the **baseline model & only TAPT**
- Even unlabeled corpus is small than the origin paper, the results are similar
- As K increases higher(50→500), the performance improve better **except 500NN-TAPT**

Pretraining	BioMed, RCT-500			
	Paper		Replication	
ROBERTA	79.3	0.6	79.6	0.6
TAPT	79.8	1.4	80.3	0.6
50NN-TAPT	80.8	0.4	80.7	0.7
150NN-TAPT	81.2	0.6	81.1	0.2
500NN-TAPT	81.7	0.8	80.3	0.4
DAPT	**82.5**	**0.5**	**82.5**	**0.6**

Result - Automatic Data selection for TAPT

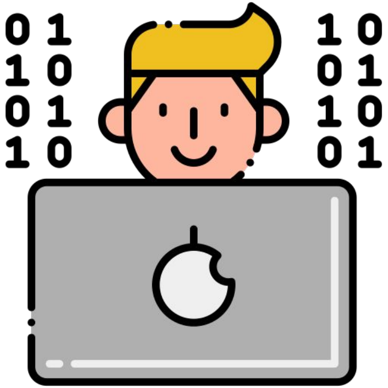
- Too much Repetition in Small Unlabel Corpus
- Effect way to augment for small task data if we has appropriate corpus

→ **Use in Improvement part**

```
The company added that it would reevaluate its commitment to the remaining New York show, which took place last month.
32  New York
6 New York
2  New York
 05 04  09 35 PDT New York  AP
 05 02  21 00 PDT New York  AP
 05 01  12 58 PDT NEW YORK   AP
 05 04  08 36 PDT New York  AP
 05 06  15 21 PDT New York  AP
 05 04  07 57 PDT NEW YORK   AP
 05 06  12 38 PDT New York  AP
 05 02  14 30 PDT NEW YORK   AP
 05 04  08 32 PDT New York  AP
 05 06  10 35 PDT New York  AP
 05 11  15 28 PDT New York  AP
 05 08  12 55 PDT NEW YORK   AP
 05 02  19 19 PDT New York  AP
 05 11  12 30 PDT NEW YORK   AP
 05 08  14 16 PDT NEW YORK   AP
 05 11  19 05 PDT New York  AP
 05 10  08 18 PDT New York  AP
 05 10  10 34 PDT NEW YORK   AP
 05 04  13 15 PDT New York  AP
 05 07  07 34 PDT New York  AP
 05 03  08 19 PDT NEW YORK   AP
 05 04  15 29 PDT NEW YORK   AP
 05 03  07 59 PDT New York  AP
```

repetitions

Why is it important?



I want to make a NLP model for classifying sociology papers

SOTA Pretrained Models

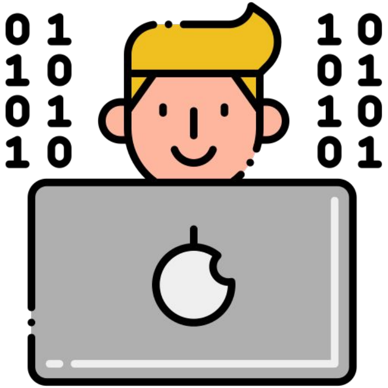
Fine-tuning

Sociology Papers

Sociology Paper Classifier

Acc. **90%!!!!**

Why is it important?



I want to make a NLP model for classifying sociology papers

data related to the task and domain

Continued pretraining

SOTA Pretrained Models

Fine-tuning

Sociology Papers

Sociology Paper Classifier

Acc. 92%!!!!

Improvement Approaches and Results

Overview

- Conduct **statistical tests for identifying** the effects of adaptive pretraining
- Task Adaptive Pretraining for a task to **classify YouTube comments**
- Domain and Task Adaptive Pretraining for the **GLUE benchmark**

Approach - Statistical test

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	$95.0_{0.2}$	**$95.4_{0.2}$**	$94.1_{0.4}$

Table 3: Comparison of RoBERTa (RoBA.) and DAPT to adaptation to an *irrelevant* domain (¬ DAPT). Reported results are test macro-F_1, except for ChemProt and RCT, for which we report micro-F_1, following Beltagy et al. (2019). We report averages across five random seeds, with standard deviations as subscripts. † indicates high-resource settings. Best task performance is boldfaced. See §3.3 for our choice of irrelevant domains.

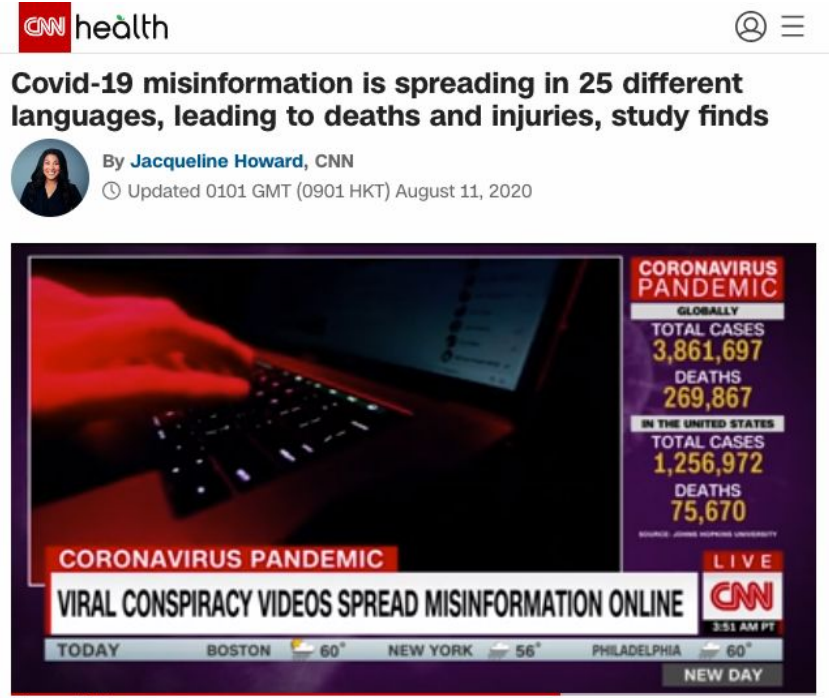
- In the original paper, they report averages across five random seeds, with standard deviations.
- However, if the standard deviation is **large**, we cannot affirm whether there is a statistically significant improvement.
- So, we trained and tested the models 30 times per each experiment and conducted **Two sample T-Test**.

Improvement Approaches and Results

Overview

- Conduct **statistical tests for identifying** the effects of adaptive pretraining
- Task Adaptive Pretraining for a task to **classify YouTube comments**
- Domain and Task Adaptive Pretraining for the **GLUE benchmark**

Motivation - YouTube comments

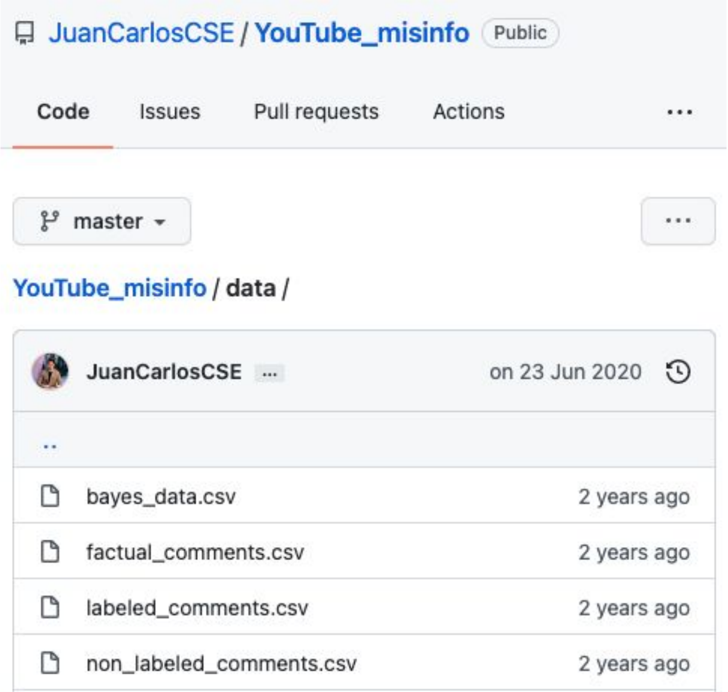


Source: CNN

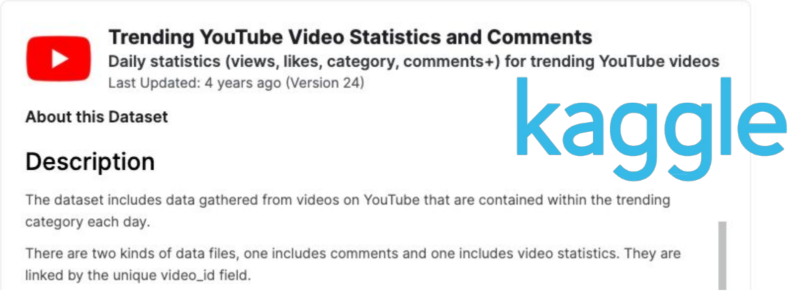
During the COVID periods, viral conspiracy videos spread misinformation through YouTube and Facebook

→ We want to check whether adaptive pretraining is helpful to build a classifier of COVID conspiracy comments on YouTube

Datasets - YouTube comments



1. **Baseline Model (RoBERTa)**
2. **Task Adaptive Pre-Training (TAPT)**
   - labeled comments dataset
3. **Human Curated-TAPT**
   - unlabeled comments dataset
4. **Automatic Data Selection for TAPT**
   - Task: labeled comments
   - Domain: YouTube comments



Results - YouTube comments

	ROBERTA	TAPT	Curated-TAPT	10NN-TAPT	25NN-TAPT
avg	74.26	75.59	75.88	74.03	75.97
std	1.82	1.88	2.13	2.18	1.82

- TAPT and Curated-TAPT improves the model performance compared to the pure ROBERTA model.
  - vs. TAPT : $p = 0.007$
  - vs. Curated-TAPT: $p = 0.002$
- No significant difference between TAPT and Curated-TAPT ($p = 0.57$)

- 10NN-TAPT is not helpful for the model performance.
- 25NN-TAPT has better performance than 10NN-TAPT ($p = 0.0004$)

Interpretation - YouTube comments

	ROBERTA	TAPT	Curated-TAPT	10NN-TAPT	25NN-TAPT
avg	74.26	75.59	75.88	74.03	75.97
std	1.82	1.88	2.13	2.18	1.82
data size	0 B	660 KB	5.9 MB	5.6 MB	12.4 MB

9x. large

Pretraining	Steps	Docs.	Storage	F_1
RoBERTa	-	-	-	$79.3_{0.6}$
TAPT	0.2K	500	80KB	$79.8_{1.4}$
50NN-TAPT	1.1K	24		[illegible]$_{0.6}$
150NN-TAPT	3.2K	66		[illegible]$_{0.8}$
500NN-TAPT	9.0K	185		[illegible]$_{0.4}$
Curated-TAPT	8.8K	180K	27MB	**83.4**$_{0.3}$
DAPT	12.5K	25M	47GB	$82.5_{0.5}$
DAPT + TAPT	12.6K	25M	47GB	$83.0_{0.3}$

337x. large

Pretraining	BioMed RCT-500
TAPT	$79.8_{1.4}$
DAPT + TAPT	$83.0_{0.3}$
Curated-TAPT	$83.4_{0.3}$
DAPT + Curated-TAPT	**83.8**$_{0.5}$

- TAPT and Curated-TAPT improves the model performance compared to the pure ROBERTA model.
- No significant difference between TAPT and Curated-TAPT
  - There is not much difference in the size of the dataset.
- 10NN-TAPT is not helpful for the model performance.
  - Domain dataset for YouTube comments is much smaller.

Interpretation - YouTube comments

	ROBERTA	TAPT	Curated-TAPT	10NN-TAPT	25NN-TAPT
avg	74.26	75.59	75.88	74.03	75.97
std	1.82	1.88	2.13	2.18	1.82
data size	0 B	660 KB	5.9 MB	5.6 MB	12.4 MB

Pretraining	Steps	Docs.	Storage	F_1
ROBERTA	-	-	-	$79.3_{0.6}$
TAPT	0.2K	500	80KB	$79.8_{1.4}$
50NN-TAPT	1.1K	24K	3MB	$80.8_{0.6}$
150NN-TAPT	3.2K	66K	8MB	$81.2_{0.8}$
500NN-TAPT	9.0K	185K	24MB	$81.7_{0.4}$
Curated-TAPT	8.8K	180K	27MB	$\mathbf{83.4}_{0.3}$
DAPT	12.5K	25M	47GB	$82.5_{0.5}$
DAPT + TAPT	12.6K	25M	47GB	$83.0_{0.3}$

- TAPT and Curated-TAPT improves the model performance compared to the pure ROBERTA model.
- No significant difference between TAPT and Curated-TAPT
  - There is little difference in data size.
- 10NN-TAPT is not helpful for the model performance.
  - **Domain dataset for YouTube comments is much smaller.**

122 MB

Interpretation - YouTube comments

Suggestions for Task adaptive pretraining

- If you try **human-curated TAPT**, you need to use **sufficiently large dataset**.
    - Though you use human-curated TAPT, if the dataset is not large, you may not see a significant improvement.
- When you try to **extract the data for TAPT** from a domain dataset, **use as large domain dataset as possible.**
    - If the domain dataset is small, it can rather reduce the performance.

Improvement Approaches and Results

Overview

- Conduct **statistical tests for identifying** the effects of adaptive pretraining
- Task Adaptive Pretraining for a task to **classify YouTube comments**
- Domain and Task Adaptive Pretraining for the **GLUE benchmark**

Motivation - GLUE Benchmark

Is adaptive pretraining effective in **GLUE** benchmark?

GLUE Tasks

Corpus	|Train|	|Test|	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

Datasets - GLUE Benchmark

data	train	dev	test	domain	input	task	metrics
Stanford Sentiment Treebank (SST-2)	67k	872	1.8k	movie reviews	single-sentence	- sentiment - binary classification (positive / negative)	acc.
Microsoft Research Paraphrase Corpus (MRPC)	3.7k	408	1.7k	news	two sentences	paraphrase	acc./F1
The Recognizing Textual Entailment (RTE)	2.5k	276	3.0k	news wikipedia	two sentences	binary classification (entailment / not_entailment)	acc.

DAPT models

Available `DAPT` models:

```
allenai/cs_roberta_base
allenai/biomed_roberta_base
allenai/reviews_roberta_base
allenai/news_roberta_base
```

Datasets: **sst** | like 2

Tasks: sentiment-classification sentiment-scoring

Task Categories: text-classification text-scoring Languages: en

Multilinguality: monolingual

Size Categories: 10K<n<100K 100K<n<1M

Language Creators: found Annotations Creators: crowdsourced

Microsoft Research Paraphrase Corpus

Important! Selecting a language below will dynamically change the complete page content to that language.

Language: English Download

This download consists of data only: a text file containing 5800 pairs of sentences which have been extracted from news sources on the web, along with human annotations indicating whether each pair captures a paraphrase/semantic equivalence relationship. Last published: March 3, 2005.

Results - GLUE Benchmark

	mrpc(news)		rte(news, wikipedia)		sst(reviews)	
	avg	std	avg	std	avg	std
base	84.50	1.26	58.63	7.60	93.92	0.31
news	83.73	0.86	53.65	1.72	93.85	0.34
cs	83.82	1.46	53.29	0.75	91.58	0.37
review	82.23	0.84	53.00	2.58	**93.67**	**0.29**
biomed	**84.87**	**0.70**	**60.58**	**5.02**	91.74	0.42

	mrpc(news)		sst(reviews)	
	base	TAPT	base	TAPT
avg	86.22	77.20	93.96	93.90
std	2.30	1.02	0.36	0.26

- News DAPT models are less effective on MRPC and RTE.
- Review DAPT models are effective on SST.
- The model pretrained by MRPC datasets **reduce the task performance**.
- The model pretrained by SST-2 datasets doesn't affect the performance. ($p = 0.506$)

Interpretation - GLUE benchmark

	mrpc(news)		rte(news, wikipedia)		sst(reviews)	
	avg	std	avg	std	avg	std
base	84.50	1.26	58.63	7.60	93.92	0.31
news	83.73	0.86	53.65	1.72	93.85	0.34
cs	83.82	1.46	53.29	0.75	91.58	0.37
review	82.23	0.84	53.00	2.58	**93.67**	**0.29**
biomed	**84.87**	**0.70**	**60.58**	**5.02**	91.74	0.42

Sentence #1 —— Sentence #2

	mrpc(news)			sst(reviews)	
	base	TAPT	New-TAPT	base	TAPT
avg	86.22	77.20	89.23	93.96	93.90
std	2.30	1.02	0.58	0.36	0.26

Microsoft Research Paraphrase Corpus (MRPC)	3.7k	408	1.7k	news	two sentences	paraphrase

- DAPT models are not always effective for the task in the same domain.
  - The distribution of task data may be far from news datasets used for pretraining our DAPT model.
- The model pretrained by MRPC datasets reduce the task performance.
  - MRPC is different with previous tasks: two sentence input, paraphrase
  - The model pretrained with the first sentences improve the performance (**p = 0.00**)

Interpretation - GLUE benchmark

Suggestions for Task adaptive pretraining

- When you use a DAPT model, **check the dataset** used for pretraining.
  - Even in the same field, it might not help improving the model.
- For a task with **two sentence inputs**, pretrain the model with the task dataset made of only **first sentence**.

Don’t Stop Pretraining:

Adapt Language Models to Domain and Tasks

References

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. 2016. Chemprot-3.0: a global chemical biology diseases mapping. *Database*, 2016.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. *arXiv preprint arXiv:1906.07348*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *arXiv preprint arXiv:1808.09602*.

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 the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news. *Neurips*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.