Evaluation of Different Approaches to Training BERT for Classification Tasks in the Engineering Domain

Sergii Poluektov, 25.10.2021, Master Thesis Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

- Motivation
- Research Questions
- BERT
- Datasets and Tasks
- Domain Adaptation
- Fine-tuning and Evaluation
- Results
- Conclusion and Future Work

Outline

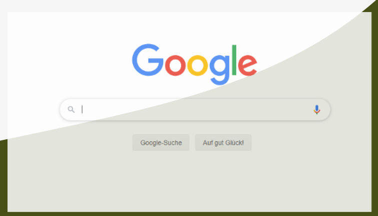
- Motivation
- Research Questions
- BERT
- Datasets and Tasks
- Domain Adaptation
- Fine-tuning and Evaluation
- Results
- Conclusion and Future Work

Motivation

Trade fairs

Trade publications

Search engines



Data Volume

1990

2020

[1]

Motivation

[1]

Outline

- Motivation
- Research Questions
- BERT
- Datasets and Tasks
- Domain Adaptation
- Fine-tuning and Evaluation
- Results
- Conclusion and Future Work

Strategies

Further Pre-training

- FinBERT [2]

Vocabulary Extension + Further Pre-training

- German LegalBERT [3]
- SciBERT [5]

Training from Scratch

- TweetBERT [4]
- SciBERT [5]
- BioBERT [6]

Dataset Extension

Generally a good idea

Research Questions

1. *How does the BERT model **further pre-trained** on texts from the engineering domain perform on the selected text classification and entity extraction tasks?*

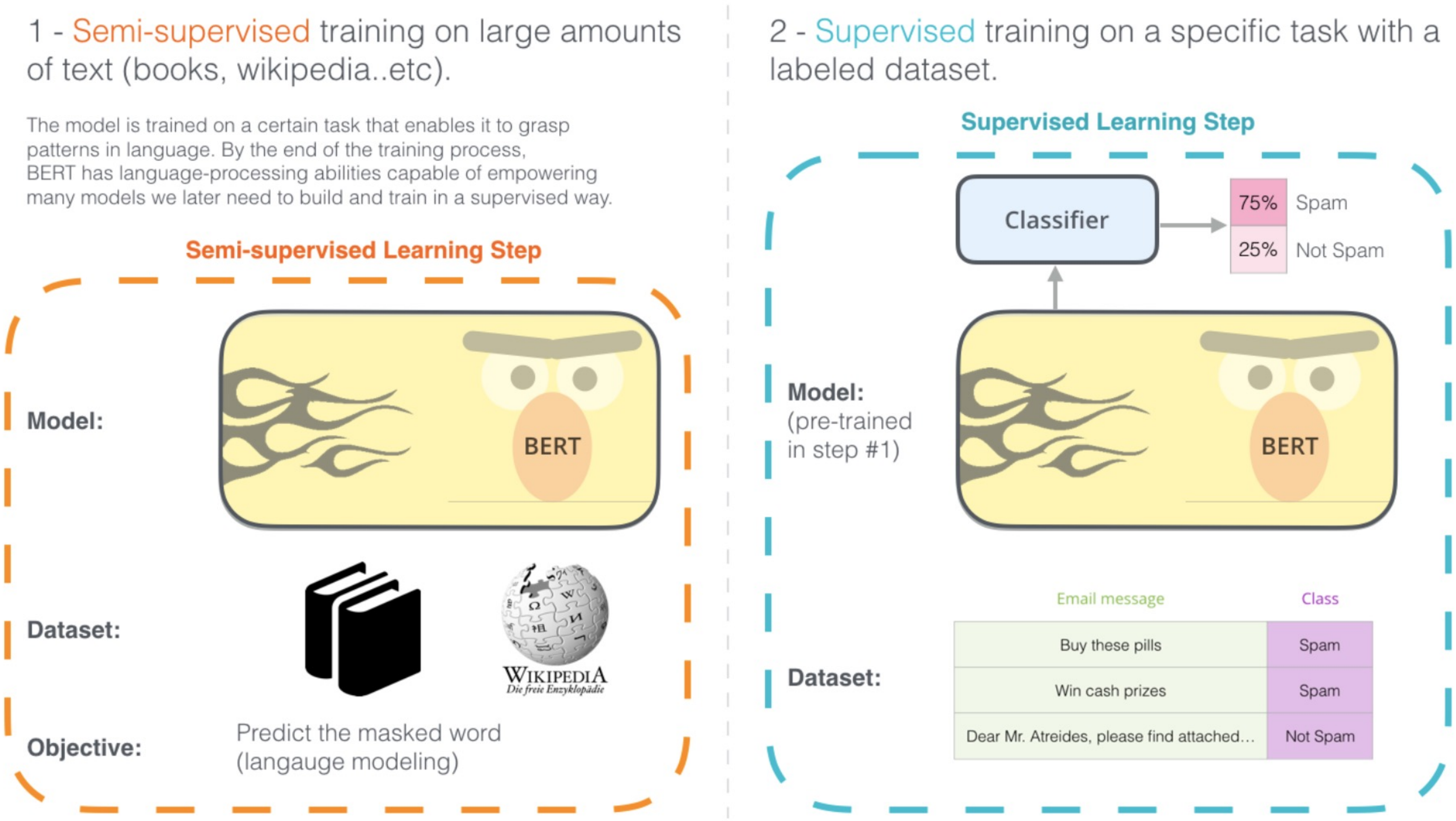
2. *How does the BERT model **with extended vocabulary** and **further pre-trained** on texts from the engineering domain perform on the selected text classification and entity extraction tasks?*

3. *How does the BERT model, **trained from scratch** on texts from the engineering domain, perform on the selected text classification and entity extraction tasks?*

4. *What effect does the **extension of labelled data** sets have on the performance of the BERT model on the selected text classification and entity extraction tasks?*

Outline

BERT



1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).
The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.
Semi-supervised Learning Step
Model:
BERT
Dataset:
WIKIPEDIA
Die freie Enzyklopädie
Objective:
Predict the masked word (langauge modeling)
2 - Supervised training on a specific task with a labeled dataset.
Supervised Learning Step
Classifier
75% Spam
25% Not Spam
Model:
(pre-trained in step #1)
BERT
Dataset:
Email message
Class
Buy these pills
Spam
Win cash prizes
Spam
Dear Mr. Atreides, please find attached…
Not Spam

[7]

Outline

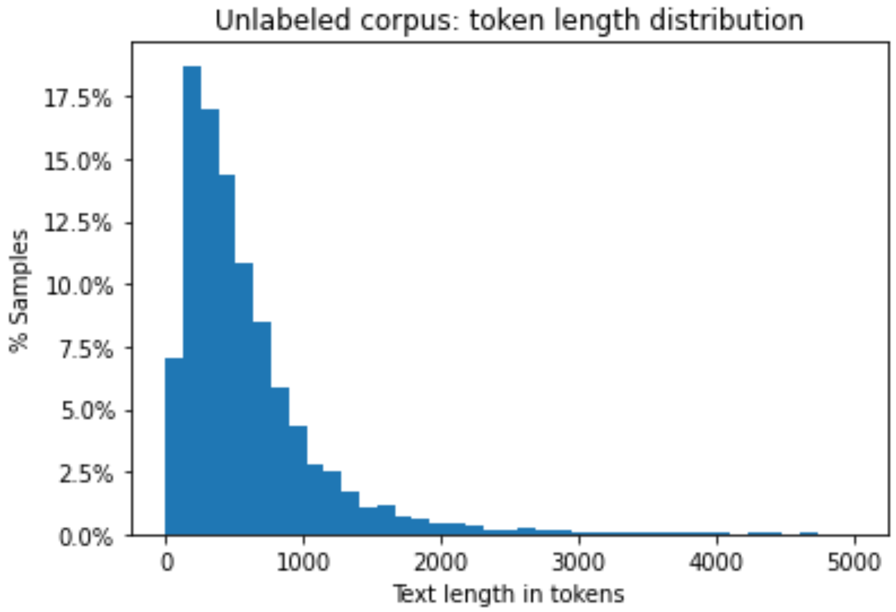
- Motivation
- Research Questions
- BERT
- Datasets and Tasks
- Domain Adaptation
- Fine-tuning and Evaluation
- Results
- Conclusion and Future Work

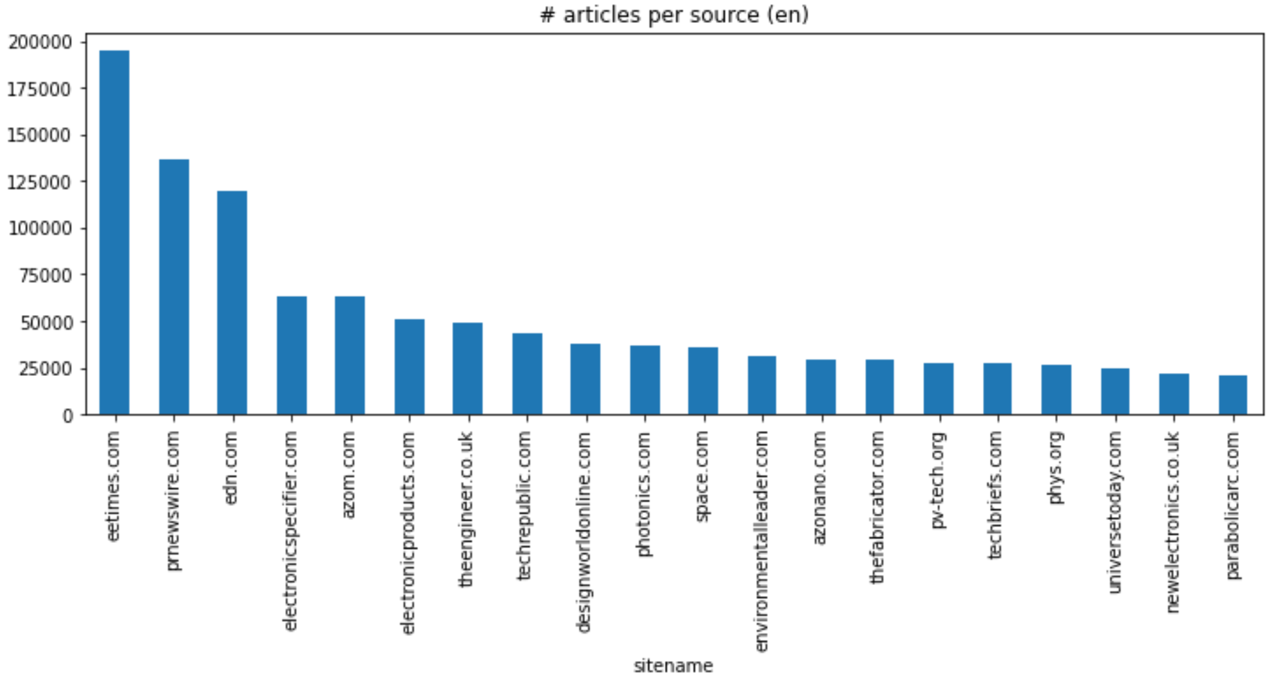
Unlabeled Engineering Articles Corpus

2 million articles

330 sources

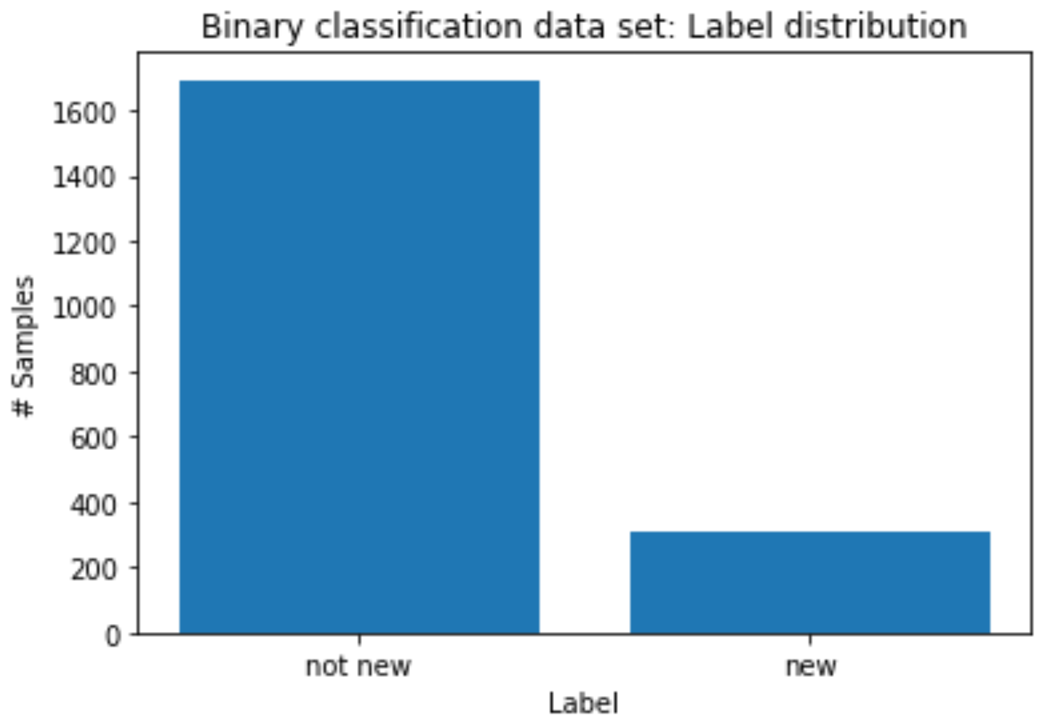
6.7 Gb





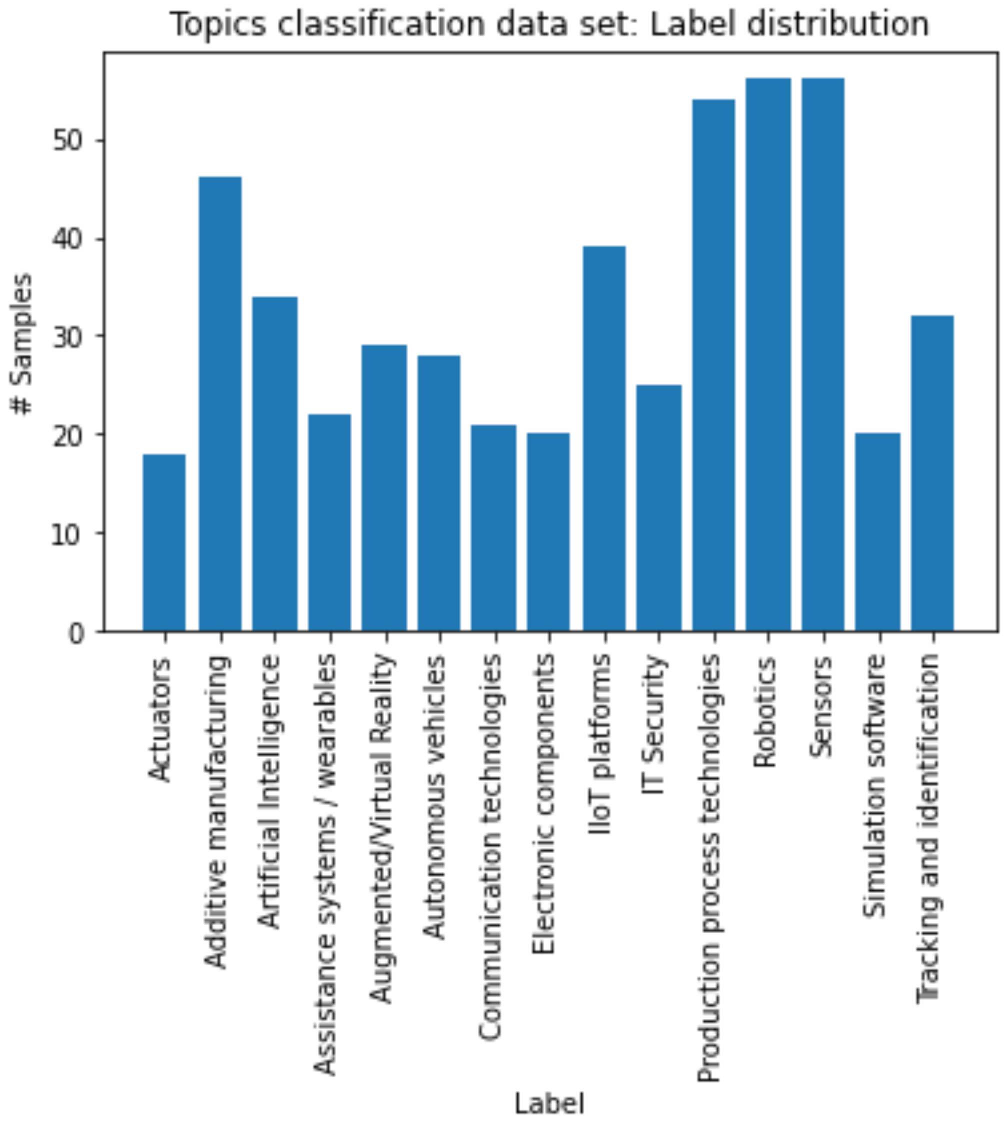
Identify articles describing new technologies

2000 articles



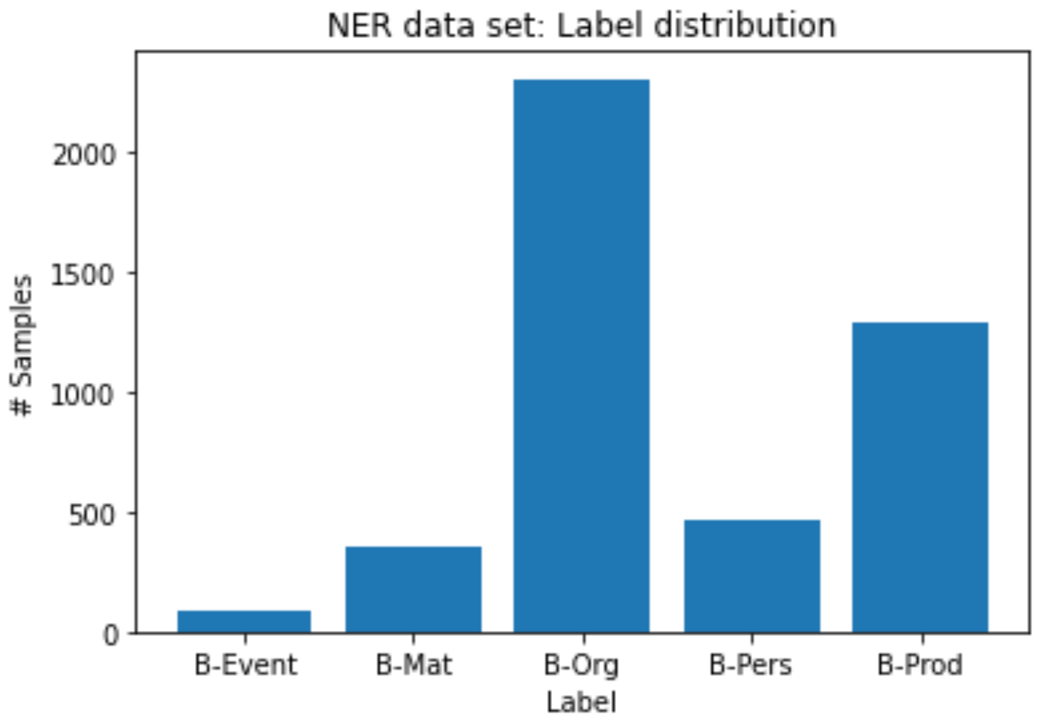
Assign articles to topics

500 articles



Extract named entities

300 articles

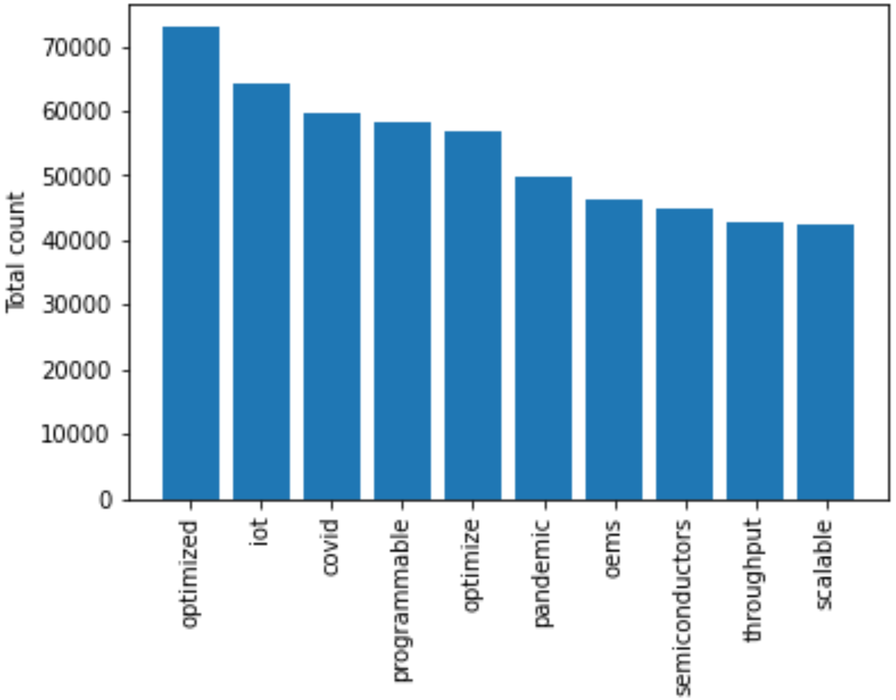


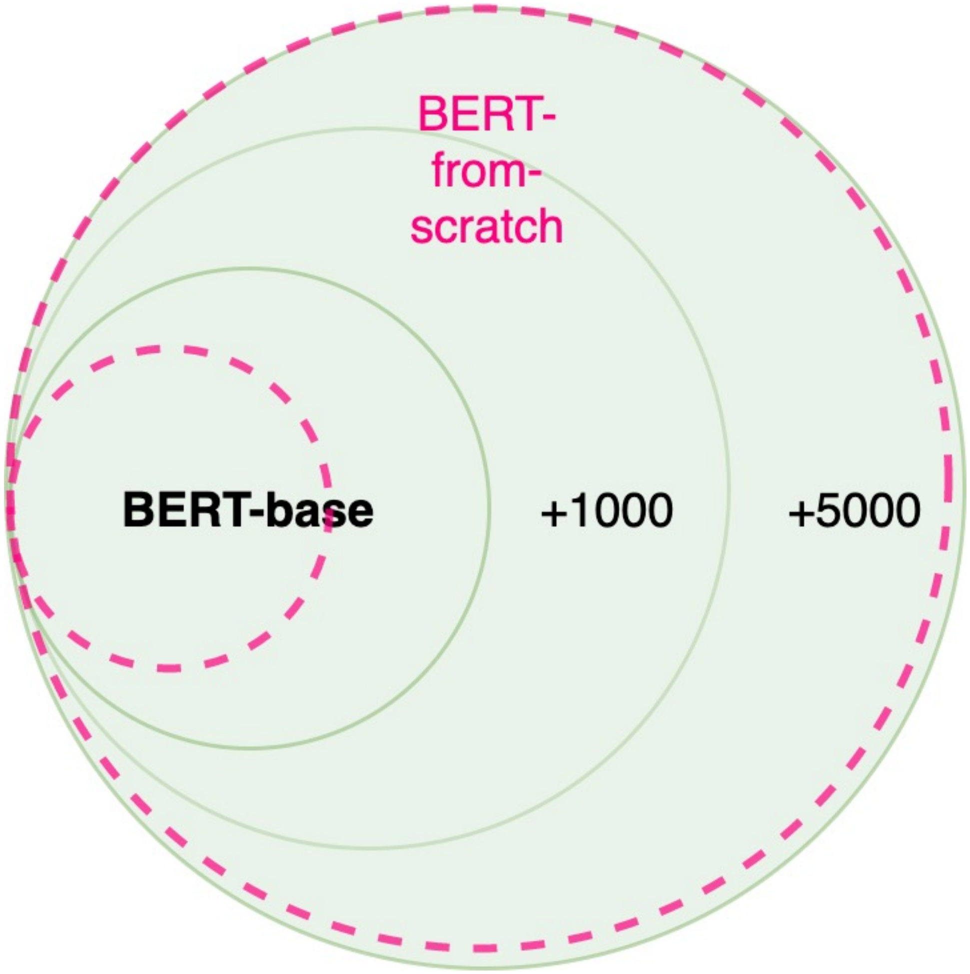
Outline

- Motivation
- Research Questions
- BERT
- Datasets and Tasks
- Domain Adaptation
- Fine-tuning and Evaluation
- Results
- Conclusion and Future Work

Domain Adaptation

- BERT-*base*
- BERT-*base-nove*
- BERT-*base-ext1000*
- BERT-*base-ext5000*
- BERT-*base-from-scratch*





Outline

Fine-tuning and Evaluation

Binary Classification

- HeadTail
- Oversampling
- Precision, Recall, F1
- 5-fold-cross-validation

Topic Classification

- HeadTail
- Accuracy
- 5-fold-cross-validation

Named Entity Recognition

- Per class precision, recall and F1 scores + weighted average.
- 5-fold-cross-validation

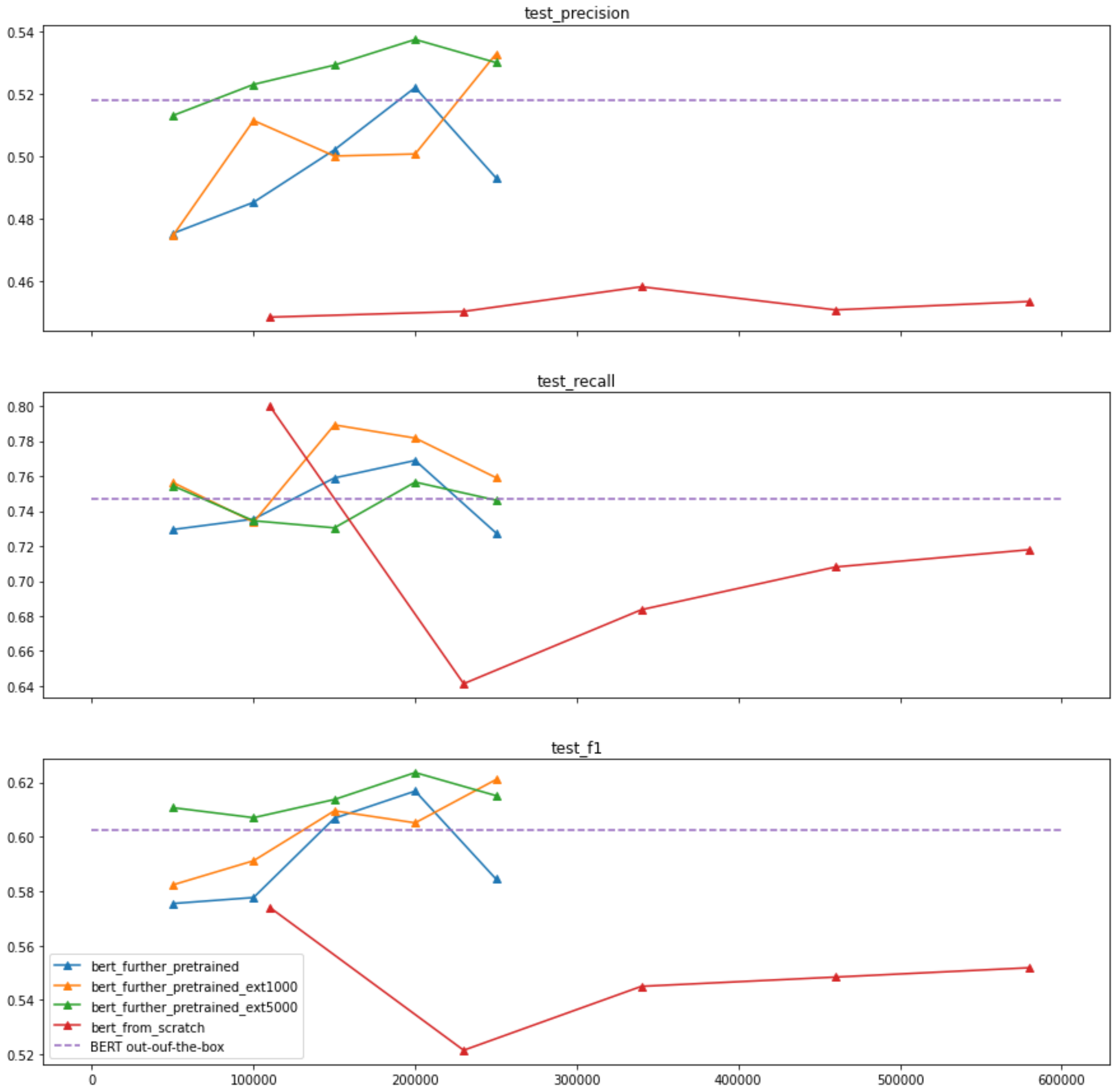
Strategy	*Head*	*HeadTail*	*Tail*
Accuracy	0.8	0.81	0.79

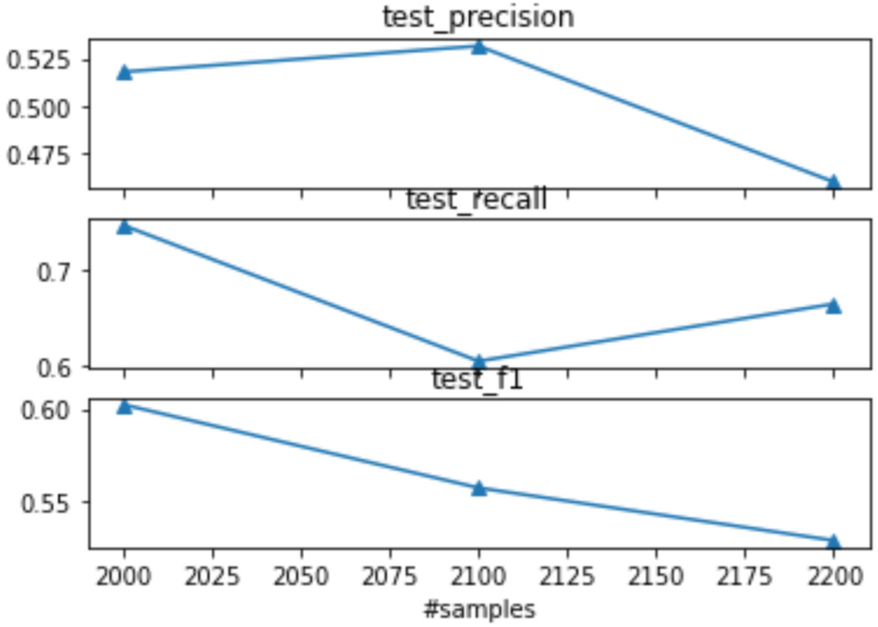
Data set	*Binary classification*	*Topic classification*	*NER*
Sizes	$2000 \rightarrow 2100 \rightarrow 2200$	$500 \rightarrow 550 \rightarrow 600$	$300 \rightarrow 340 \rightarrow 380$

Outline

- Motivation
- Research Questions
- BERT
- Datasets and Tasks
- Domain Adaptation
- Fine-tuning and Evaluation
- Results
- Conclusion and Future Work

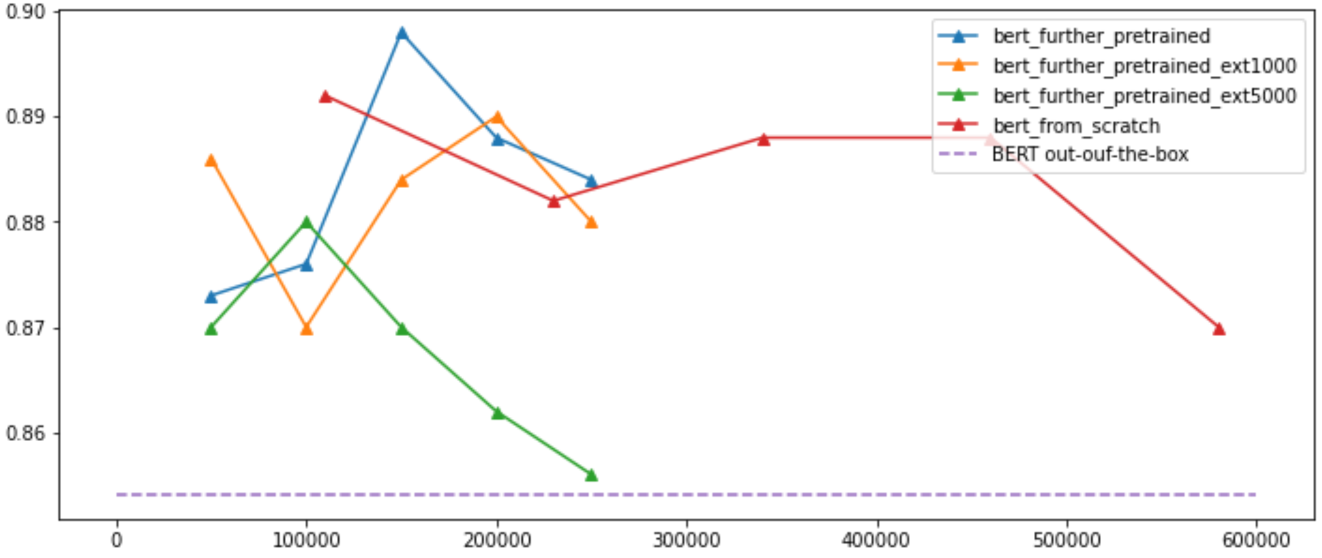
Identify articles describing new technologies

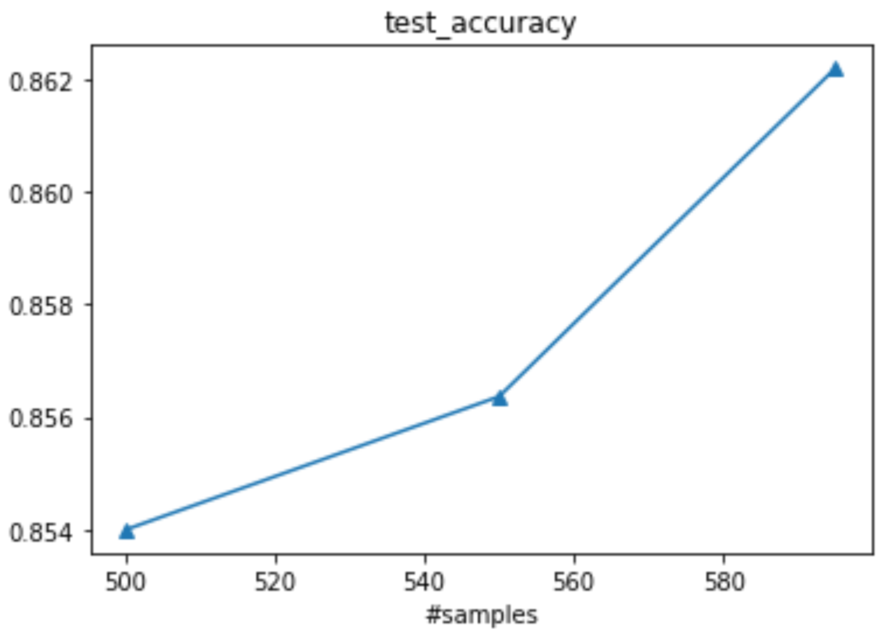




Model	**Precision**	**Recall**	**F1**
BERT-*base*	51.79	74.69	60.22
BERT-*base-nove*	52.20	76.89	61.68
BERT-*base-ext1000*	53.27	75.91	62.11
BERT-*base-ext5000*	53.74	75.65	**62.36**
BERT-*base-from-scratch*	44.83	80.00	57.40

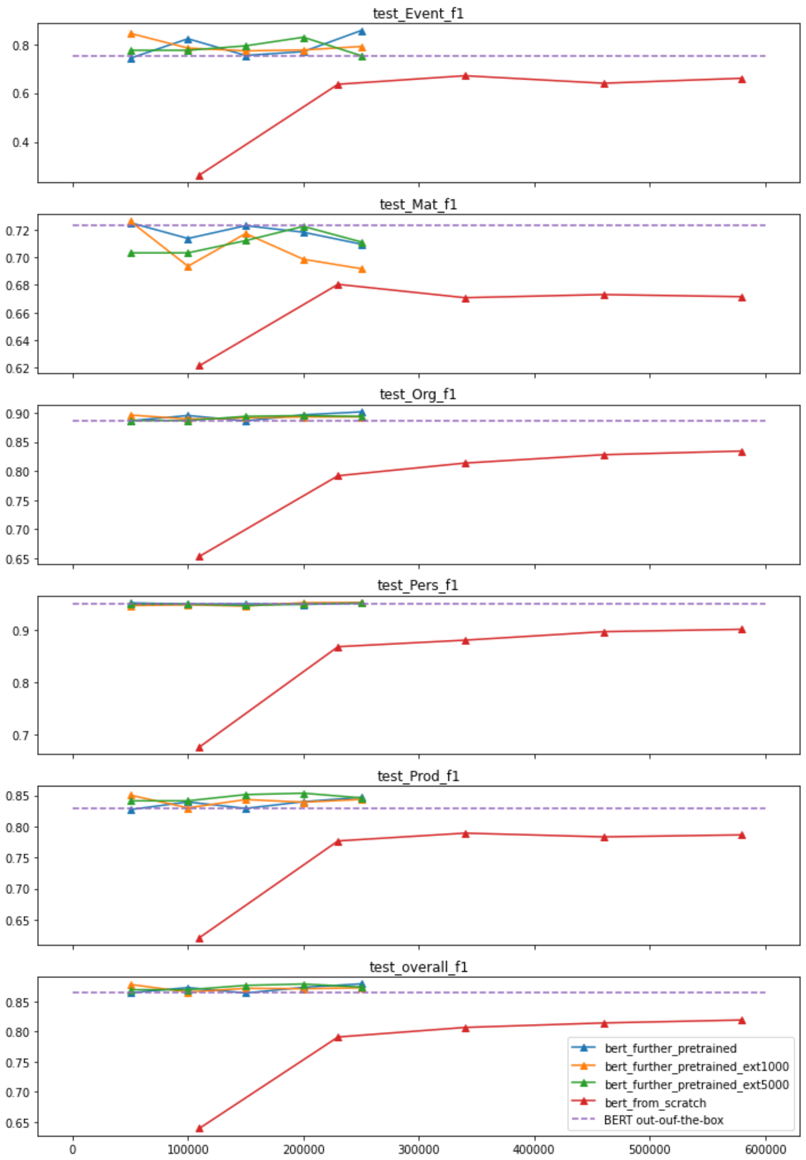
Assign articles to topics

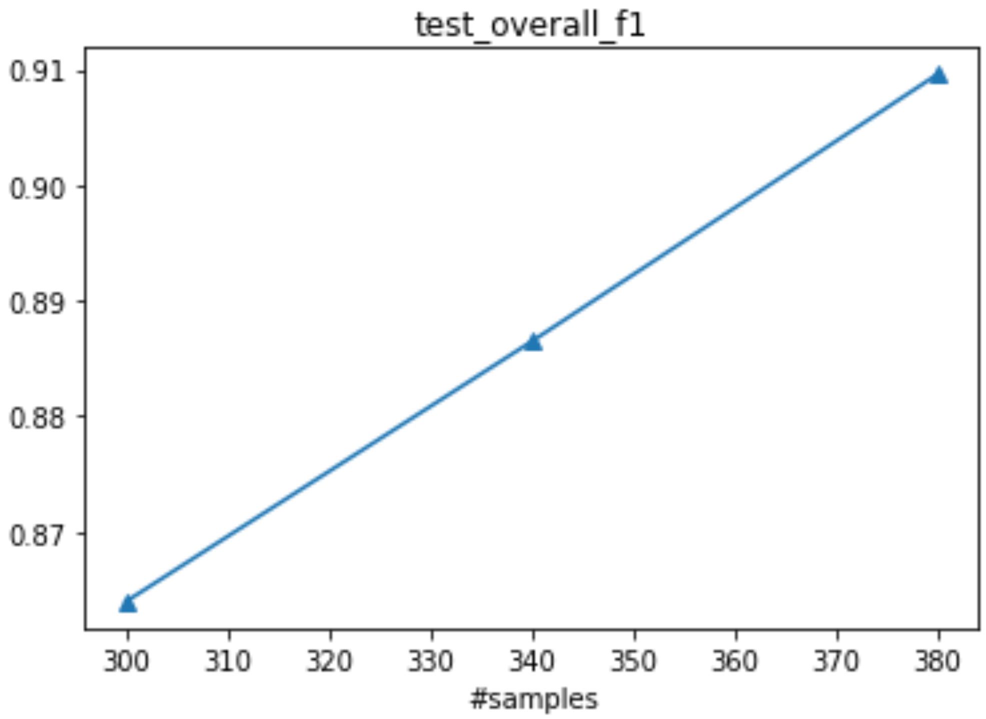




Model	Accuracy
BERT-*base*	85.40
BERT-*base-nove*	**89.80**
BERT-*base-ext1000*	89.00
BERT-*base-ext5000*	86.20
BERT-*base-from-scratch*	89.20

Extract named entities





Model	Event F1	Mat F1	Org F1	Pers F1	Prod F1	Overall F1
BERT-*base*	75.39	72.29	88.61	95.04	82.91	86.39
BERT-*base-nove*	85.57	70.95	90.17	95.18	84.69	**87.91**
BERT-*base-ext1000*	79.05	69.17	89.31	95.28	84.36	87.25
BERT-*base-ext5000*	82.84	72.23	89.52	95.00	85.36	87.89
BERT-*base-from-scratch*	66.03	67.14	83.42	90.16	78.65	81.92

Overview

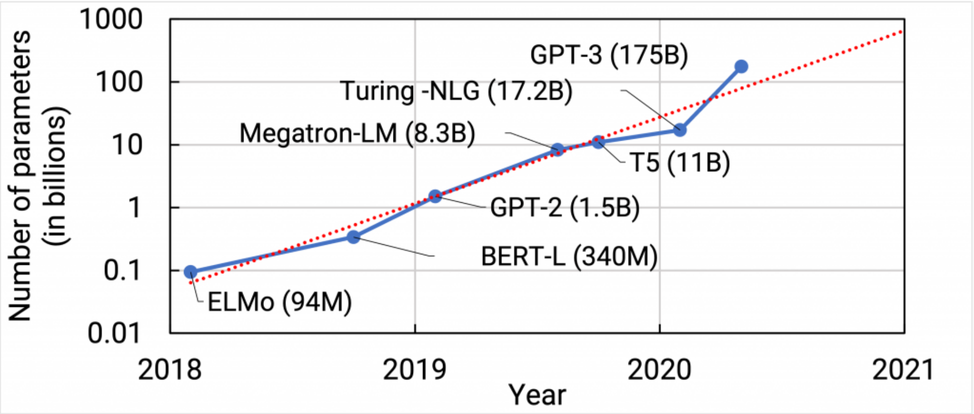
Model	Binary classification (F1)	Topic classification (Accuracy)	NER (Overall F1)
BERT-*base*	60.22	85.40	86.39
BERT-*base-nove*	61.68	**89.80**	**87.91**
BERT-*base-ext1000*	62.11	89.00	87.25
BERT-*base-ext5000*	**62.36**	86.20	87.89
BERT-*base-from-scratch*	57.40	89.20	81.92

Outline

- Motivation
- Research Questions
- BERT
- Datasets and Tasks
- Domain Adaptation
- Fine-tuning and Evaluation
- Results
- Conclusion and Future Work

Conclusion and Future Work

1. BERT-*base-nove* performed best on **two** out of three tasks.
2. BERT-*base-ext5000* performed best on **one task.** → Alternative vocabulary extension strategy.
3. Neither BERT-*base-ext1000* nor BERT-*base-ext5000* performed worse than the baseline.
4. BERT-*base-from-scratch* is undertrained → More data and evaluation on downstream tasks during training.
5. Extending labelled datasets improves the performance of the base model on **two** out of three tasks.



[8]

Sergii Poluektov

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132
Fax +49.89.289.17136

sergii.poluektov@tum.de
wwwmatthes.in.tum.de

Sources

[1] Rokin https://en.rokin.tech

[2] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao. “FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining.” In: *IJCAI*. 2020, pp. 4513– 4519.

[3] C. M. Yeung. “Effects of inserting domain vocabulary and fine-tuning BERT for German legal language”. MA thesis. University of Twente, 2019.

[4] D. Q. Nguyen, T. Vu, and A. T. Nguyen. “BERTweet: A pre-trained language model for English Tweets”. In: *arXiv preprint arXiv:2005.10200* (2020).

[5] I. Beltagy, K. Lo, and A. Cohan. “SciBERT: A Pretrained Language Model for Scientific Text”. In: *arXiv:1903.10676 [cs]* (Sept. 2019). arXiv: 1903.10676. url: http://arxiv.org/ abs/1903.10676 (visited on 04/30/2021).

[6] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. “BioBERT: a pre- trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* (Sept. 2019). arXiv: 1901.08746, btz682. issn: 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btz682. url: http://arxiv.org/abs/1901.08746 (visited on 04/30/2021).

[7] J. Alammar. *The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)*. url: http://jalammar.github.io/illustrated-bert/ (visited on 03/11/2021).

[8] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. A. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, et al. “Efficient large-scale language model training on gpu clusters”. In: *arXiv preprint arXiv:2104.04473* (2021).

Sources

[9] Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. “Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks”. In: *arXiv:2004.10964 [cs]* (May 2020). arXiv: 2004.10964. url: http://arxiv.org/abs/2004. 10964 (visited on 04/30/2021).

Performance of domain-specific models

Further Pre-training

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

[9]

Training from Scratch

Field	Task	Dataset	SOTA	BERT-Base Frozen	BERT-Base Finetune	SciBERT Frozen	SciBERT Finetune
Bio	NER	BC5CDR (Li et al., 2016)	88.85[7]	85.08	86.72	88.73	**90.01**
		JNLPBA (Collier and Kim, 2004)	**78.58**	74.05	76.09	75.77	77.28
		NCBI-disease (Dogan et al., 2014)	**89.36**	84.06	86.88	86.39	88.57
	PICO	EBM-NLP (Nye et al., 2018)	66.30	61.44	71.53	68.30	**72.28**
	DEP	GENIA (Kim et al., 2003) - LAS	**91.92**	90.22	90.33	90.36	90.43
		GENIA (Kim et al., 2003) - UAS	**92.84**	91.84	91.89	92.00	91.99
	REL	ChemProt (Kringelum et al., 2016)	76.68	68.21	79.14	75.03	**83.64**
CS	NER	SciERC (Luan et al., 2018)	64.20	63.58	65.24	65.77	**67.57**
	REL	SciERC (Luan et al., 2018)	n/a	72.74	78.71	75.25	**79.97**
	CLS	ACL-ARC (Jurgens et al., 2018)	67.9	62.04	63.91	60.74	**70.98**
Multi	CLS	Paper Field	n/a	63.64	65.37	64.38	**65.71**
		SciCite (Cohan et al., 2019)	84.0	84.31	84.85	**85.42**	**85.49**
Average				73.58	77.16	76.01	79.27

[5]