Replication and Improvements for Paper: Don't Stop Pretraining

Sangwook Lee, Wenchao Dong, Taeyeong Lee, William Gyuho Suh
School of Computing, KAIST
{sangwooklee, wenchao.dong, ooasd990, williamgyuhosuh}@kaist.ac.kr

Abstract

By continuing to pretrain the existing pretrained language model on a domain-specific corpus (DAPT), it is still possible to improve the performance of the task completion. Further experiments demonstrated that continued pretraining on a task-specific corpus(TAPT) can also significantly improve performance on a given task while using fewer computational resources.

Further, adaptive-pretraining can also be applied to the YouTube misinformation comment classification task. This study also proposes to evaluate the above models using a statistical approach as well as the GLUE benchmark, and finally, adaptive pretraining is applied to the GLUE non-classification task dataset.

1 Introduction

A variety of today's pretrained LMs can achieve good performance in a diverse range of downstream tasks, and these models are trained on a large corpus to learn a sufficiently good representation. We wonder if pretrained models can work universally well, or can we still build better-performing models by further training in specific domains.

We selected RoBERTa as the baseline model and continued to pretrain it in the four domains to obtain the corresponding DAPTs. Since DAPT is computationally expensive, we wondered whether pretraining on a more task-relevant subset of the domain would work well for completing a particular task. Further, we speculate whether there is still a performance gain from performing TAPT on the basis of DAPT.

To enhance the above research, the results were analysed again using statistical methods.

The adaptive pretraining was applied to the YouTube COVID conspiracy comments classification task, extending the scope of application of the research.

Different tasks of GLUE were used to measure the performance of TAPT, introducing criteria for new dimensions.

Finally, we combine the above ideas again. We obtained TAPT models for SST-2 and MRPC tasks by pretraining using datasets from the MRPC and SST-2 task in GLUE. This contribution is outstanding since we investigated the performance of adaptive-pretraining on non-classification task(MRPC).

2 Approach

Some studies have demonstrated the benefit of further pretraining the pretrained language model on domain-specific unlabeled data(Lee et al., 2020), but the pretrained models employed in these studies used a small corpus with low richness and considered only one domain at a time.

To address this problem, we use BoBERTa(Liu et al., 2019) as the baseline model. 160GB of uncompressed raw data including bookcorpus(Zhu et al., 2015), CCNews is used for RoBERTa pretraining corpus, and we hope to understand whether the pretraining corpus of RoBERTa is diverse enough to be able to generalize to all domains.

We take four domains for domain-adaptive pretraining, which means we continue to pretrain RoBERTa on a large unlabeled domain-specific corpus. The four domains are biomedical papers, CS publications, news, and Amazon reviews.

The above domains are defined by genres, but we can still narrow down the domains to only task-related ones. Task data is a narrowly-defined subset of the domain in which it resides, so we are interested in understanding whether further pretraining on a corpus directly related to the task can further improve performance. So we used eight classified tasks (two in each of the four previous domains) to retrain the pretrained model.

3 Paper Replication

3.1 Datasets

According to the experiments, the corpus of the four domains we selected has different degrees of overlap with the pretrained corpus of our baseline model, where the NEWS domain has more than 50% overlap with the RoBERTa corpus and the CS domain has less than 20% overlap(Gururangan et al., 2020). Next, we list the data sources for the four domains and eight tasks in Table 1.

3.2 Baseline Model

In this experiment, we use RoBERTa(Liu et al., 2019) as the baseline model because its performance is not much worse than SOTA, and it is a single language model that can be adapted to different domains, so we can choose the RoBERTa model to fine-tune its parameters to accomplish the the corresponding tasks.

3.3 Domain-Adaptive Pretraining

3.3.1 Experiment process

We continued to pretrain RoBERTa on each of the four domains, which consisted of large-scale unlabeled text in the following areas: BIOMED, NEWS, CS, REVIEWS. To ensure that the performance improvement of DAPT did not simply come from exposing the pretrained LM to more data, we selected two DAPT models with the lowest correlation between two of them and tested them against each other based on the correlation between the four domains.

We hope to demonstrate that continued pretraining on a domain-specific corpus is a further boost to task completion performance by comparing the performance of DAPT on different tasks with RoBERTa. The results for the baseline models RoBERTa and DAPT on the eight tasks are presented in Table 2.

3.3.2 Results

We observe that DAPT improves on RoBERTa in all three domains except the NEWS domain. In particular, the highest performance improvement was achieved on the ACL-ARC task in the CS domain, but on the HYPERPARTISAN task, RoBERTa still performed much better than DAPT.

We found a direct relationship between the results and the overlap between domains. DAPT boosted RoBERTa the most in the CS domain, and this domain had the least overlap with RoBERTa at 19.2%. DAPT did not boost RoBERTa in the NEWS domain, and this domain had an overlap with RoBERTa of 54.1%(Gururangan et al., 2020).

Noting that RoBERTa achieved a good performance of over 92% on both tasks in the NEWS domain and considering the large overlap between this domain and RoBERTa's domain, we infer that on tasks similar to RoBERTa's pretrained corpus, RoBERTa's performance is good enough that do not need to additionally perform domain-adaptive pretraining again. However, for those domains that differ significantly from the RoBERTa pretrained corpus, the effect of continuing to pretrain on the basis of LM is significant, and the greater the difference from RoBERTa's source domains, the more likely that DAPT will yield good performance.

We wanted to explore whether the performance improvement of DAPT came from exposing RoBERTa to more data, regardless of the domains, so we selected the two DAPTs with the least domain overlap, and presented this part of the results in the ¬DAPT column in Table 2, named irrelevant domains. For the BM domain, REVIEW LM was used and vice versa, and for the CS domain, NEWS LM was used and vice versa.

It can be observed that DAPT outperforms ¬DAPT to varying degrees, except for the SCIERC task, showing that it is crucial to continue pretraining on domain-relevant data. Continued pretraining on the NEWS domain may provide a boost to performance in one CS task, suggesting that, in some cases, it would be useful to continue pretraining on additional data.

Furthermore, we compared the results of ¬DAPT with the baseline model and found that RoBERTa outperformed ¬DAPT except for SCIERC and HELPFULNESS tasks. Therefore, we can infer that, in most cases, exposing LM to more data that is not relevant to the task is detrimental to end-task completion.

3.4 Task-Adaptive Pretraining

3.4.1 Experiment process

Although research has been conducted to demonstrate the effectiveness of TAPT(Howard and Ruder, 2018), we aimed to experimentally compare the performance differences between DAPT and TAPT, and how the two could be combined to obtain further performance improvements. TAPT refers to pretraining on a subset of the domain that is directly relevant to the task. TAPT uses a much

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

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

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	$94.5_{0.2}$	$\mathbf{95.0}_{0.1}$	$94.0_{0.2}$

Table 2: Comparison of RoBERTa(ROBA.) and DAPT as well as the adaptation to an irrelevant domain(¬DAPT). Reported results are test macro-F1. except for CHEMPROT and RCT, for which we report micro-F1, following Beltagy et al. (2019). Best task performance is boldfaced.

smaller corpus than DAPT and requires fewer computational resources.

Both DAPT and TAPT include a second stage of pretraining RoBERTa, which is the fine-tuning of LM learned representations and the completion of downstream tasks. We therefore designed experiments and tested the performance of RoBERTa, DAPT, TAPT, and DAPT+TAPT on eight tasks. The term DAPT+TAPT refers to RoBERTa being first applied to domain-adaptive pretraining and then to task-adaptive pretraining, i.e. the process consists of three stages of pretraining. We present the results in Table 3.

3.4.2 Results

It can be seen that TAPT outperforms RoBERTa on all tasks except HYPERPATISAN, which shows the advantage of task adaptation. More importantly, TAPT outperforms DAPT on RCT, HYPERPATISAN, and AGNews, which is noteworthy because TAPT requires far fewer time and computational resources than DAPT, providing an effective option for practical applications.

It can be noticed that DAPT+TAPT improves the baseline model for all tasks except HYPERPATISAN. Further, this combination outperforms DAPT to the highest level on RCT, SCIERC, AGNews, and IMDB tasks. This approach also outperformed TAPT on most tasks.

3.5 Human Curated-TAPT

3.5.1 Experiment process

The process of building a dataset generally involves obtaining a large unlabeled corpus and subsequently annotating it, so that the larger the unlabeled corpus the closer the distribution is to the training data for the task.

We obtained RCT-500 by downsampling the RCT training data to 500 examples and treating the rest of the training data as unlabeled. The HYPERPARTISAN shared task(Kiesel et al., 2019) has two tracks: low- and high-resource. We selected 5K data from the high-resource as curated-TAPT and used the low-resource training data for task fine-tuning. For IMDB, we used additional data(Maas et al., 2011), which had a similar distribution to the labeled data.

We compared Curated-TAPT with DAPT and DAPT+TAPT respectively and the detailed test results are shown in Table 4.

3.5.2 Results

We found that Curated-TAPT outperformed TAPT on the RCT-500 and IMDB tasks. And implementing Curated-TAPT after DAPT can lead to further improvements in the RCT-500 task. On the RCT-500 and IMDB tasks, Curated-TAPT outperformed DAPT+TAPT, further suggesting that we are not getting better performance from larger models on certain 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	$94.5_{0.2}$	$95.0_{0.1}$	$95.0_{0.3}$	$\mathbf{95.3}_{0.2}$

Table 3: Results on different phases of adaptive pretraining compared to the baseline RoBERTa. Reported results follow the same format as Table 2.

Pretraining	BIOMED	NEWS	REVIEWS
	RCT-500	HYP.	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	$\mathbf{83.5}_{0.5}$	$82.4_{8.9}$	$95.4_{0.2}$

Table 4: Mean test set macro-F1(for HYP. and IMDB) and micro-F1(for RCT-500) with standard deviations as subscripts

ROBA.	TAPT	50NN	150NN	500NN
$79.6_{0.6}$	$80.3_{0.6}$	$80.7_{0.7}$	$\mathbf{81.1}_{0.2}$	$80.3_{0.4}$

Table 5: Results of automatic data selection for RCT-500 task in BIOMED domain

3.6 Automated Data Selection for TAPT

3.6.1 Experiment process

Automatic data selection refers to finding the nearest sentence from a labeled sentence in an unlabeled corpus. There is a condition that task data should be possible only when included in the domain. We used VAMPIRE, a bag-of-words language model to observe neighboring sentences, to finish the data selection. VAMPIRE first learned the word frequency representation of unlabeled text, then mixed it with labeled text, constructing embedding. Finally, it annotated unlabeled data and found neighbors from task data according to the label.

We use only RCT-500 as task data and BioMed as an unlabeled corpus in our replication because there is complex copyright for domain corpus data in the present. We use the small BioMed offered by Github.

3.6.2 Results

Table 5 shows that kNN-TAPT surpasses RoBERTa and TAPT for all cases. The expectation was that performance would improve as the K increased, but when $K = 500$, we found that performance decreased instead. This is because our unlabeled corpus is too small, resulting in duplicate data selection. We, therefore, recommend data selection with a small number of neighbors when there is small unlabeled data.

4 Improvements

The adaptive pretraining was applied to the YouTube Covid conspiracy comments classification task as well as to the GLUE benchmark task. Furthermore, we took a statistical approach and used a t-test to determine whether the metric improvement was significant. In the original paper, the averages and standard deviations for five random seeds were presented. However, TAPT produced somewhat better but more volatile outcomes throughout replication. To compare the TAPT results, we fine-tuned a model 30 times and double-checked the test p-values.

4.1 YouTube Covid Conspiracy Comment Classification

4.1.1 Datasets

We adapted the YouTube dataset from previous research(Serrano et al., 2020), which is used to detect COVID-19 misinformation videos on YouTube by leveraging user comments. The videos were classified into misinformation and factual videos, and a 10% random sample of comments from the misinformation videos was manually assigned two agreements and two conspiracy labels. We divided the labeled comments into 6:2:2 (train:dev:test) and made a binary classifier to detect conspiracy comments.

To perform the automated data selection for TAPT, we selected a dataset containing over one million YouTube comments from the US and UK[1].

4.1.2 Experiment process

We made TAPT, human curated-TAPT, 10NN-, and 25NN-TAPT models by continuing pretraining RoBERTa on the YouTube conspiracy comments dataset. Due to the lack of GPU memory, we reduced the batch size from 16 to 8 while keeping other settings.

4.1.3 Results

We compared all TAPT models with RoBERTa in Table 6. TAPT and Curated-TAPT significantly improve the model performance compared to RoBERTa(both p<0.01). The difference in performance between Curated-TAPT and TAPT was not statistically significant (p>0.05).10NN-TAPT did not improve the base model's performance (p<0.05), but 25NN-TAPT enhanced it. (p<0.01) However, both have no significant improvement compared to base TAPT (p>0.05).

These results highlight the importance of dataset size for curated-TAPT and automatic data selection. In Table 4, curated-TAPT improved the task performance significantly for RCT-500, but curated-TAPT models did not. We expect that the size of the curated dataset made this difference. For the RCT task, the size of the curated dataset is 27MB, which is 337 times larger than the task dataset(80KB). However, our curated dataset size is 5.9MB, and it is only nine times larger than the dataset for TAPT(660KB). It shows that the effect of curated-TAPT depends on the size difference between task and curated dataset.

In Table 5, 50NN and 150NN-TAPT worked better than TAPT, but in our improvements, 10NN-TAPT and 25NN-TAPT did not show better results than TAPT. We guess that the size of the YouTube comments domain dataset mattered. The size of our domain dataset is only 122 MB, which is much smaller than the domain dataset size of the original paper(over 10GB). Thus, our augmenting algorithm would not have been able to extract enough diverse data similar to task data.

ROBA.	TAPT	Curated	10NN	25NN
$74.3_{1.8}$	$75.6_{1.9}$	$75.9_{2.1}$	$74.0_{2.2}$	$76.0_{1.8}$

Table 6: Performance comparison among baseline model RoBERTa, task-adaptive pretraining, human-curated TAPT and automatic data selection for TAPT

4.2 GLUE benchmark with DAPT and TAPT models

In this section, we explored the combination of adaptive pretraining and GLUE benchmark(Wang et al., 2018).

4.2.1 Datasets

We selected three tasks from the GLUE benchmark: MSRP, SST-2, and RTE because their domains include reviews and news, which are covered by our DAPT domains.

We pretrained TAPT by using the MRPC dataset and the SST-2 dataset from GLUE tasks[2]. The jiant toolkit(Phang et al., 2020) was also used to check the performance of each model. Then we compared the performance of these two TAPTs: SST-2 is for single sentence input and finished binary classification; MRPC is two sentences input with a paraphrasing task.

4.2.2 Experiment process

We tried to evaluate four DAPTs and RoBERTa with the GLUE benchmark and adapted from the original paper setting by training five times with random seeds.

Also, we pretrained the RoBERTa on SST-2 and MRPC datasets, obtaining corresponding TAPTs.

4.2.3 Results

In Table 7, When compared to RoBERTa, DAPT significantly reduced performance. Although MRPC and RTE are in the news domain, the NEWS-DAPT model still underperforms the baseline model. For SST-2 in a review domain, DAPT with news and reviews datasets works better than CS and Biomed ones. In Table 8, TAPT for SST-2 does not affect the task results(p>0.05), but TAPT for MRPC significantly improves the task performance(p<0.01).

We think that the different distribution between the task dataset and the domain dataset may reduce the performance of DAPT in the MRPC and

[1] https://www.kaggle.com/adepvenugopal/sentiment-analysis-of-youtube-comments/data

[2] https://gluebenchmark.com/

Model	MRPC (News)	RTE (News, Wikipedia)	SST-2 (Reviews)
RoBERTa	84.5	58.63	92.92
NEWS-DAPT	83.73	53.65	**93.85**
CS-DAPT	83.82	53.29	91.58
REVIEW-DAPT	82.23	53.00	93.67
BIOMED-DAPT	**84.87**	**60.58**	91.74

Table 7: Test results for RoBERTa, NEWS-DAPT, CS-DAPT, REVIEW-DAPT, and BIOMED-DAPT on the MRPC, RTE, and SST-2 tasks.

SST-2 (Reviews)		MRPC (News)	
ROBA.	TAPT	ROBA.	TAPT
$94.0_{0.4}$	$\mathbf{93.9}_{0.3}$	$86.2_{2.3}$	$87.8_{0.6}$

Table 8: Performance comparison of RoBERTa and pretrained TAPT on the MPRC,SST dataset.

RTE tasks. The news DAPT models from original papers were pretrained by RealNews datasets, crawled from December 2016 through March 2019 by Common Crawl (Zellers et al., 2020). However, the MRPC dataset was extracted from the World Wide Web over 2 years before 2005. (Dolan and Brockett, 2005) Also, RTE datasets are the combination of RTE1, 2, 3, and 5, sampled from Wikipedia and news before 2009 (Bentivogli et al., 2009). Therefore, the vocabulary distributions might differ due to time differences and adversely affect task performance. It shows that DAPT with time-variant and diverse datasets like news may not always be effective for the exact domains.

5 Discussions

5.1 Practical suggestions for adaptive pretraining

We suggest three tips about adaptive pretraining based on the insights that we got during our replications and improvements:

- When you try to create a curated task dataset using human resources, you should consider the cost and effect of the estimated data size. Curated-TAPT may cost too much if you have a task dataset that is large enough.

- When you try automatic selection for TAPT, you should use as large a domain dataset as possible. If you use a small domain dataset, the selection algorithm may create augmented datasets with a narrow range of diversity or a lot of duplication.

- When using the existing DAPT models, you should check the relevance between the task and domain datasets. As you can see from previous results, DAPT in the same domain with the specific task does not constantly improve performance. Like in the original paper, you can simply check the vocabulary overlap or try another way to check the relevance.

5.2 Future work

Relationship between domain size and k-candidate In the replication result in Table 5, we observe that too much-increased number of neighbor points diminishes the performance dislike the paper's result. Experiments show that if the value of k-candidate is larger than the size of the domain, it becomes unnecessary repetition. If we had more time, we would have been able to find the optimal k value. We would be able to find the tendency between the size of a domain and the k value, which is helpful for hyperparameter tuning.

Overlapping between domain and task data We used a lot of domain data for pretraining. However, it was insufficiently verified whether the test data belonged to the domain. In this paper, the similarity between domains was verified through Vocabulary Overlapping. It will be possible to find an appropriate domain by similarly applying this to the test data. To this end, it seems that research on the optimal bag of word models can be conducted.

Domain Similarity between different era Although we tried to improve performance by using TAPT + Data Augmentation in YouTube comments, it was found that the performance was not significantly improved compared to the baseline model. All the corpus of the paper has experimented with formal writing. Still, it might be quite different in the informal domain, such as SNS, which is frequently used as slang or composed of sentences on various topics. It is possible to compare similarities by classifying words used in past SNS and current SNS as corpus.

5.3 Limitation

The biggest problem we encountered during the experiment was the lack of computing resources. Some tasks cannot be completed due to memory capacity limitations. If we had more computing resources, we could pretrain RoBERTa on a large corpus and discover more appealing findings.

References

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

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

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

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

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*.

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.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. http://jiant.info/.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

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.