Political Analysis (2024), **32**, 379–383
doi:10.1017/pan.2023.36

PA

LETTER

On Finetuning Large Language Models

Yu Wang

Fudan Institute for Advanced Study in Social Sciences, Fudan University, Shanghai, China

Email: yuwang.aiml@gmail.com

(Received 2 June 2023; revised 26 July 2023; accepted 24 August 2023; published online 28 November 2023)

Abstract

A recent paper by Häffner *et al.* (2023, *Political Analysis* 31, 481–499) introduces an interpretable deep learning approach for domain-specific dictionary creation, where it is claimed that the dictionary-based approach outperforms finetuned language models in predictive accuracy while retaining interpretability. We show that the dictionary-based approach's reported superiority over large language models, BERT specifically, is due to the fact that most of the parameters in the language models are excluded from finetuning. In this letter, we first discuss the architecture of BERT models, then explain the limitations of finetuning only the top classification layer, and lastly we report results where finetuned language models outperform the newly proposed dictionary-based approach by 27% in terms of R^2 and 46% in terms of mean squared error once we allow these parameters to learn during finetuning. Researchers interested in large language models, text classification, and text regression should find our results useful. Our code and data are publicly available.

Keywords: finetuning; large language models; text as data

Edited by: John Doe

1. Introduction

Large language models have been gaining popularity among political scientists (Bestvater and Monroe 2023; Wang 2023b). These models are known for being easy to use for end-to-end training and accurate in making predictions. The downsides of these large language models, however, are that they are slow to run and hard to interpret. In an effort to overcome these shortcomings, Häffner *et al.* (2023) introduce an interpretable deep learning approach to domain-specific dictionary creation. Such an approach, coupled with Random Forest or XGBoost (Wang 2019), creates accurate and interpretable models. Häffner *et al.* (2023) claim that these new models outperform finetuned ConfliBERT models.[1] In this letter, we show that the apparent superiority of the newly proposed dictionary-based method stems from the fact that most of the parameters in the BERT model are excluded from training at the finetuning stage. We first illustrate the BERT model's architecture and explain which components are actually being finetuned, then we demonstrate how we can maximize learning by making all parameters trainable, and lastly we report the new results with fully finetuned BERT models that outperform the dictionary-based approach by a large margin.

2. Finetuning BERT Models

2.1. The BERT Architecture

BERT models are first introduced in Devlin *et al.* (2019). These are large language models that are initially pretrained with billions of words on tasks, such as masked language modeling and next sentence

[1] ConfliBERT (Hu *et al.* 2022) is a particular version of BERT (Devlin *et al.* 2019) where the authors pretrain the BERT models on texts in the domain of conflicts and political violence. In this paper, we use BERT and ConfliBERT interchangeably.

© The Author(s), 2023. Published by Cambridge University Press on behalf of The Society for Political Methodology.

https://doi.org/10.1017/pan.2023.36 Published online by Cambridge University Press

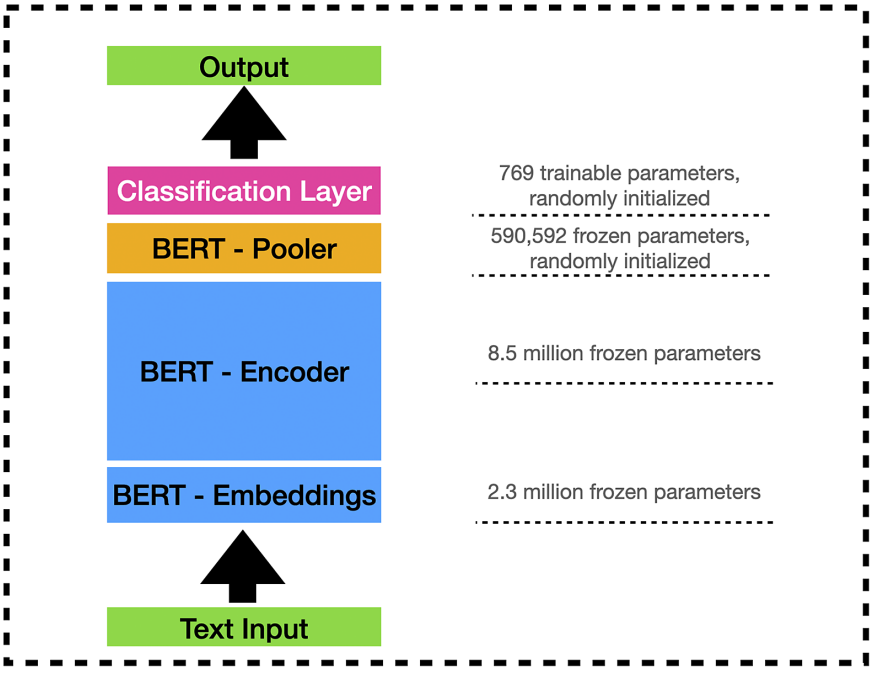


Figure 1. Illustration of the BERT model's architecture. Numbers are based on Häffner *et al.* (2023) and Hu *et al.* (2022). Code for calculating the number of parameters is included in the replication package. Diagram is not drawn to proportion.

prediction. They are then finetuned on specific downstream tasks, such as topic classification and conflict prediction.[2] Under this pretrain-and-finetune paradigm, BERT models have achieved the state-of-the-art results in various natural language processing tasks. There have also been many variations of BERT models based on more training data (Liu *et al.* 2019), new data domains (Hu *et al.* 2022), and new pretraining tasks (Lan *et al.* 2020).

In Häffner *et al.* (2023), for the task of predicting the natural logarithm of fatalities on a country month level using the CrisisWatch text, the BERT model attains an R^2 of 0.60, which is lower than the 0.64 achieved by the recently introduced dictionary-based approach.[3] We demonstrate that BERT's apparent inferior performance is due to the fact that Häffner *et al.* (2023) freeze all the parameters in the BERT model except the classification layer. In Figure 1, we illustrate the architecture of a BERT model. A BERT model consists of four components: embedding layers (2.3 million parameters), encoder (8.5 million parameters), the pooler layer (0.6 million parameters), and the classification layer (769 parameters). The majority of the parameters lie in the encoder and the embeddings (blue boxes).[4]

There are two observations with regard to training exclusively over the classification layer (pink box).[5] One is that this layer accounts for only a tiny portion (less than 0.1%) of the BERT model. As a result, this considerably limits the model's expressive power. Another is that the pooler layer (yellow

[2]For best practices for finetuning language models, please refer to Devlin *et al.* (2019), Mosbach, Andriushchenko, and Klakow (2021), Dodge *et al.* (2020), and Zhang *et al.* (2021).

[3]The embedding layer, the encoder layer(s), and the pooler layer mostly serve as a feature representation for the input text. The classification layer (pink box in Figure 1) of BERT serves as a linear regression model with mean squared error as the loss function.

[4]The exact numbers could vary among models. The numbers that we report in this letter are based on the specific model used in Häffner *et al.* (2023).

[5]There is a large literature on the benefits of freezing layers in language models, mostly in the context of multi-task learning where frozen parameters can be shared among different tasks and in the context of catastrophic forgetting where low learning rates (or zero) could help preserve learned information from previous tasks. Interested readers could refer to Houlsby *et al.*

https://doi.org/10.1017/pan.2023.36 Published online by Cambridge University Press

Table 1. Metrics comparison between finetuning only the classification layer and finetuning the entire model. Due to space limit, we only report the first 5,000 steps. Interested readers could find all the training steps in the replication package.

	Finetuning classification layer		Finetuning entire model	
Step	Training loss	Validation loss	Training loss	Validation loss
500	1.853	2.123	1.359	1.229
1,000	1.672	1.971	0.899	1.235
1,500	1.634	1.951	0.835	1.186
2,000	1.632	1.898	0.642	1.108
2,500	1.642	1.944	0.666	0.941
3,000	1.643	1.786	0.630	1.204
3,500	1.699	1.862	0.571	1.080
4,000	1.555	1.794	0.511	1.023
4,500	1.558	1.831	0.432	1.013
5,000	1.652	1.834	0.477	0.927

box), which has 0.6 million parameters, is not pretrained, but randomly initialized.[6] By freezing the pooler layer, we end up with 0.6 million randomly initialized parameters that are neither pretrained nor finetuned. Both observations contribute to limiting the performance of the finetuned BERT model.

2.2. Finetuning over All Parameters

In this subsection, we study the effects of making all the parameters trainable. We use the same setting as in Häffner *et al.* (2023), except that we change the learning rate from 2e-3 to 2e-5 (Devlin *et al.* 2019). We need to lower the learning rate during finetuning to reflect the fact that most of the parameters are pretrained (Howard and Ruder 2018).[7] We also reduce the number of training epochs from 20 to 10, as it becomes clear that we do not need that many epochs. This is mostly to save computational costs. In Table 1, we compare the learning trajectories of finetuning only the classification layer as is done in Häffner *et al.* (2023) (left) and finetuning the entire BERT model (right). We observe that compared with finetuning only the classification layer, by finetuning the entire model, we are able to learn at a much faster pace and achieve a training loss and validation loss, both calculated in terms of mean squared error (MSE), that are considerably lower. As a matter of fact, by finetuning the classification layer only, we reach the lowest validation loss at 1.776 after training for 8,000 steps. By contrast, we can attain a validation loss of 1.229 after training for the first 500 steps once we allow the entire model to be finetuned.

2.3. Experimental Results

In this subsection, we report two groups of experimental results: one group, *ConfliBERT Unrestricted*, where we make all parameters trainable, and another group, *ConfliBERT Max Length*, where we not only make all parameters trainable but also increase the maximum sequence length from 256 (Häffner *et al.*

(2019) and Ding *et al.* (2023) for its application in multi-task learning and to Howard and Ruder (2018) for its application in the context of catastrophic forgetting.

[6] Please see Section 5 of the Supplementary Material for a more detailed analysis.

[7] Failure to do so could result in poor model performance. Please see Footnote 1 in the Supplementary Material for a concrete example, which helps illustrate the importance of selecting proper learning rates when finetuning large language models.

https://doi.org/10.1017/pan.2023.36 Published online by Cambridge University Press

Table 2. By making all parameters learnable, finetuned BERT models outperform dictionary-based models by a large margin. Results in columns 1–3 are from Table 2 in Häffner *et al.* (2023). By increasing the maximum sequence length to 512, we are able to further improve the performance of those finetuned models. Best results in bold.

Model	OCoDi Random forest (1)	OCoDi XGBoost (2)	ConfliBERT Restricted (3)	ConfliBERT Unrestricted (4)	ConfliBERT Max length (5)
MSE	1.59	1.60	1.75	0.99	**0.87**
R^2	0.64	0.63	0.6	0.77	**0.80**

2023) to 512 (Wang 2023b).[8] In Table 2, we compare the performance of these different models. OCoDi-Random Forest is the dictionary-based model that leverages random forest, where OCoDi stands for Objective Conflict Dictionary. OCoDi-XGBoost is the dictionary-based model that leverages XGBoost. *ConfliBERT Restricted* is the model from Häffner *et al.* (2023) where most parameters are excluded from finetuning.

We observe that while OCoDi-Random Forest (column 1) and OCoDi-XGBoost (column 2) both achieve lower MSE and higher R^2 than ConfliBERT Restricted (column 3) where we finetune only the classification layer of ConfliBERT, once we finetune the entire ConfliBERT (column 4), we are able to achieve much lower MSE and higher R^2 than the dictionary-based approaches. Further, by increasing the maximum sequence length from 256 to 512, we observe that the MSE on the test set deceases from 0.99 to 0.87 and that the R^2 on the test set increases from 0.77 to 0.80 (column 5). Comparing *OCoDi-XGBoost* and *ConfliBERT Max Length*, we observe that ConfliBERT is able to achieve an MSE that is 46% lower than OCoDi-XGBoost and an R^2 that is 27% higher.

In terms of computational costs, ConfliBERT Restricted takes 37 minutes to run on an A100 GPU, ConfliBERT Unrestricted (column 4) 50 minutes, and ConfliBERT Max Length (column 5) 98 minutes. While finetuning over all parameters takes some more time, we believe both models are well within the time budget for most researchers. Moreover, there are various ways to further speed up the training process, including, for example, using larger input batch sizes.[9]

3. Conclusion

Häffner *et al.* (2023) have made a significant contribution to the study of interpretable machine learning by developing a method for domain-specific dictionary creation and demonstrating its effectiveness and interpretability in conflict prediction. In this letter, we have discussed the architecture of BERT models and explained the limitations of finetuning only the classification layer. While the dictionary-based approaches are definitely easier to interpret, we have shown that when fully finetuned, BERT models still outperform these dictionary-based approaches by a sizeable margin. Researchers interested in large language models, text classification, and text regression should find our results useful.

Acknowledgments. We thank the reviewers and the editor for their excellent comments and guidance, which substantially improved the paper.

Data Availability Statement. The replication materials are available in Wang (2023a) at https://doi.org/10.7910/DVN/7PCLRI.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2023.36.

[8] We note that 533 out of 4,165 samples (13%) contain more than 256 tokens. By setting the max sequence length to 256, we are effectively truncating these long samples to 256 tokens, thus reducing the amount of information that we give to the model. Setting the max sequence length to 512, which is the longest input length possible for BERT, helps alleviate this problem.

[9] For more details about the computational costs, please refer to Section 6 of the Supplementary Material.

https://doi.org/10.1017/pan.2023.36 Published online by Cambridge University Press

References

Bestvater, S. E., and B. L. Monroe. 2023. "Sentiment Is Not Stance: Target-Aware Opinion Classification for Political Text Analysis." *Political Analysis* 31 (2): 235–256. https://doi.org/10.1017/pan.2022.10

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of NAACL-HLT*, edited by J. Burstein, C. Doran, and T. Solorio, 4171–4186. Minneapolis: Association for Computational Linguistics.

Ding, N., et al. 2023. "Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models." *Nature Machine Intelligence* 5: 220–235.

Dodge, J., G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. 2020. "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping." Preprint, arXiv:2002.06305.

Häffner, S., M. Hofer, M. Nagl, and J. Walterskirchen. 2023. "Introducing an Interpretable Deep Learning Approach to Domain-Specific Dictionary Creation: A Use Case for Conflict Prediction." *Political Analysis* 31 (4): 481–499. https://doi.org/10.1017/pan.2023.7

Houlsby, N., et al. 2019. "Parameter-Efficient Transfer Learning for NLP." *Proceedings of the 36th International Conference on Machine Learning*, 2790–2799. PMLR.

Howard, J., and S. Ruder. 2018. "Universal Language Model Fine-Tuning for Text Classification." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 328–339. Melbourne: Association for Computational Linguistics.

Hu, Y., et al. 2022. "ConfliBERT: A Pre-Trained Language Model for Political Conflict and Violence." *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, 5469–5482, Seattle: Association for Computational Linguistics.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. *ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations*. ICLR.

Liu, Y., et al. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." Preprint, arXiv:1907.11692.

Mosbach, M., M. Andriushchenko, and D. Klakow. 2021. "On the Stability of Fine-Tuning BERT: Misconceptions, Explanations, and Strong Baselines." ICLR.

Wang, Y. 2019. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data: A Comment." *Political Analysis* 21 (1): 107–110.

Wang, Y. 2023a. "Replication Data for: On Finetuning Large Language Models." Harvard Dataverse. https://doi.org/10.7910/DVN/7PCLRI

Wang, Y. 2023b. "Topic Classification for Political Texts with Pretrained Language Models." *Political Analysis* 31 (4): 662–668.

Zhang, T., F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi. 2021. "Revisiting Few-Sample BERT Fine-Tuning." ICLR.

Cite this article: Wang, Y. (2024). On Finetuning Large Language Models. *Political Analysis*, 379–383. https://doi.org/10.1017/pan.2023.36

https://doi.org/10.1017/pan.2023.36 Published online by Cambridge University Press