Topic Classification for Political Texts with Pretrained Language Models

Yu Wang

University of Rochester, Rochester, NY, USA. E-mail: w.y@alum.urmc.rochester.edu

Abstract

Supervised topic classification requires labeled data. This often becomes a bottleneck as high-quality labeled data are expensive to acquire. To overcome the data scarcity problem, scholars have recently proposed to use cross-domain topic classification to take advantage of preexisting labeled datasets. Cross-domain topic classification only requires limited annotation in the target domain to verify its cross-domain accuracy. In this letter, we propose supervised topic classification with pretrained language models as an alternative. We show that language models fine-tuned with 70% of the small annotated dataset in the target corpus could outperform models trained using large cross-domain datasets by 27% and that models fine-tuned with 10% of the annotated dataset could already outperform the cross-domain classifiers. Our models are competitive in terms of training time and inference time. Researchers interested in supervised learning with limited labeled data should find our results useful. Our code and data are publicly available.[1]

Keywords: topic classification, political texts, cross domain, fine-tune, pretrained language models

1 Introduction

Supervised topic classification requires labeled data for training. This often becomes a bottleneck as high-quality labeled data are expensive to acquire. One way to overcome data scarcity is to use cross-domain topic classification (Osnabrügge, Ash, and Morelli 2021), where researchers train a model from a source domain with large labeled datasets and make inferences on the target domain where labeling is limited. This method takes advantage of two observations: rich labeled data from a source data set and high similarity between the source set and the target set. To evaluate the accuracy of the cross-domain classifier, researchers only need to annotate a small dataset in the target domain.

With the advent of language models (Devlin *et al.* 2019), however, researchers no longer have to train models from scratch as is done in Osnabrügge *et al.* (2021). Rather, researchers could take advantage of existing pretrained language models and fine-tune the already well-trained parameters on specific downstream tasks. Given that language models are known to require a relatively small number of training samples to yield good performance (Longpre, Wang, and DuBois 2020), the small annotated dataset in target domain, which is required for validating cross-domain classifiers, might be sufficient to directly train an accurate in-domain classifier.

In this letter, we present topic classification with pretrained language models as an alternative solution to the data scarcity problem. We show that language models fine-tuned with a portion (70%) of the dataset in the target domain, originally annotated for the cross-domain verification purpose, could substantially outperform cross-domain topic classifiers and that 300 training samples alone would suffice for language models to match or surpass the performance of cross-domain classifiers in Osnabrügge *et al.* (2021). We further show that fine-tuning these language models could well fit into researchers' time budgets.

Political Analysis (2023)
vol. 31: 662–668
DOI: 10.1017/pan.2023.3

Published
8 March 2023

Corresponding author
Yu Wang

Edited by
Jeff Gill

[1] The replication materials (Wang 2023) are available at the Political Analysis dataverse site.

https://doi.org/10.1017/pan.2023.3 Published online by Cambridge University Press

2 Methodology

Pretrained language models are state-of-the-art models in various natural language processing (NLP) tasks (Devlin *et al.* 2019; Lan *et al.* 2020; Liu *et al.* 2019). The heavy lifting is done during the pretraining stage, where large amounts of unlabeled text, for example, the English Wikipedia, are used to train multilayer transformer models (Vaswani *et al.* 2017) for masked language modeling and replaced token detection among other tasks (Clark *et al.* 2020).[2] Fine-tuning these pretrained language models has achieved state-of-the-art results in various NLP tasks, including classification and question answering.

Compared with other NLP models that require training with randomly initialized parameters, one advantage of pretrained language models is that they have large amounts of knowledge packed into their parameters during the pretraining stage and thus they require only a small labeled dataset for fine-tuning these parameters to achieve superb performance (Longpre *et al.* 2020). This fits well with topic classification for political texts, where labeling is expensive, and offers us an alternative to cross-domain topic classification, which trains parameters from scratch using a labeled dataset from a different but similar domain.

For this letter, we fine-tune a RoBERTa-base model (Liu *et al.* 2019) using the target dataset from Osnabrügge *et al.* (2021) for topic classification.[3] RoBERTa-base has 12 layers of transformers and 125 million parameters in total. On top of its 12 layers of transformers, we add a classification layer for 44-topic classification and 8-topic classification, respectively.[4] We use cross-entropy as the loss function. We fine-tune the RoBERTa-base model with a learning rate of 2e-5, a batch size of 16, and an input sequence length of 512 on an A100 GPU. We set the sequence input length to the maximum 512.[5] We use the validation set's accuracy to select the best epoch and the optimal checkpoint. We then use the optimal checkpoint to make inferences on the test set with a batch size of 64. For easy comparison, we use the same evaluation metrics as used in Osnabrügge *et al.* (2021).

For constructing the train, validation, and test sets, we use the 4,165 New Zealand parliamentary speeches in the target domain in Osnabrügge *et al.* (2021).[6] These 4,165 New Zealand parliamentary speeches were originally labeled to verify the effectiveness of cross-domain topic classification. In this letter, we show that these labeled speeches alone are sufficient to train a competitive topic classifier by fine-tuning a pretrained language model. In our main experiment, we randomly sample 70% from the dataset as the training set, 15% as the validation set, and the remaining 15% as the test set. In total, 2,915 samples are used for training, 625 for validation, and 625 for testing. For reproducibility, we have set a random seed for nondeterministic operations in the experiment (Zhang *et al.* 2021) and we report the averaged results of five random runs.

3 Results

3.1 Main Experiment

In the main experiment, we use 2,915 (70%) samples for training, 625 (15%) samples for validation, and 625 (15%) samples for testing and run five times with five random seeds. We report the

[2] Note that pretraining language models using generic texts and then fine-tuning them on more specific domains, such as political texts, is itself cross-domain transfer learning.

[3] Other popular pretrained language models include BERT-base-uncased, BERT-large-uncased, and RoBERTa-large. We note that BERT-large and RoBERTa-large are substantially larger and thus slower compared with RoBERTa-base. We choose RoBERTa-base because it yields competitive accuracies and is reasonably fast.

[4] There are actually 42 topics in the target corpus, so we set the number of labels to 42 for the language models. Performance difference between using 42 labels and 44 labels is minimal. To be consistent with Osnabrügge *et al.* (2021), we use "44-topic classification" throughout.

[5] We report statistics on the input sequences and the effects of sequence length on model performance in the Supplementary Material.

[6] The source data include all the English-language manifesto statements from English-speaking countries captured by the Manifesto Corpus. The target data include speeches held in the New Zealand Parliament and annotated by the Manifesto coder for New Zealand. For the complete context on the datasets, please refer to Osnabrügge *et al.* (2021).

https://doi.org/10.1017/pan.2023.3 Published online by Cambridge University Press

Table 1. Fine-tuning a RoBERTa-base model with 70% of labeled in-domain data can outperform cross-domain topic classification in both 44-topic and 8-topic tasks by a large margin. Cross-domain classifiers are from Osnabrügge *et al.* (2021). Test set is the same for both models. Mean of five random runs is reported, with standard deviation in the brackets. Better results are in bold.

Metrics	44 topics		8 topics	
	Cross-domain	Fine-tuning LM	Cross-domain	Fine-tuning LM
Top-1 accuracy/F1 micro	0.414 (0.009)	**0.527 (0.009)**	0.515 (0.006)	**0.631 (0.006)**
Top-3 accuracy	0.656 (0.008)	**0.744 (0.008)**	0.819 (0.003)	**0.904 (0.003)**
Top-5 accuracy	0.752 (0.004)	**0.828 (0.004)**	0.921 (0.008)	**0.969 (0.008)**
Balanced accuracy	0.309 (0.030)	**0.357 (0.030)**	0.454 (0.014)	**0.580 (0.014)**
F1 macro	0.294 (0.025)	**0.328 (0.025)**	0.449 (0.014)	**0.574 (0.014)**

experiment results in Table 1 with mean and standard deviation for each metric. Across all metrics and both 44-topic and 8-topic classification tasks, fine-tuning the RoBERTa model with a subset of the labeled New Zealand parliamentary speeches substantially outperforms the cross-domain topic classifier by Osnabrügge *et al.* (2021), which is trained using 115,420 annotated policy statements.

Specifically, for 44-topic classification, our top-1 accuracy stands at 52.7% and is 27.3% higher than that of the cross-domain classifier, which stands at 41.4%. For 8-topic classification, our top-1 accuracy stands at 63.1% and is 22.5% higher than that of the cross-domain classifier, which stands at 51.5%. We see large gains in other metrics as well: 10%+ gain in top-3 accuracy, 5%+ gain in top-5 accuracy, 16%+ gain in balanced accuracy, and 12%+ gain in F1 macro. The results suggest that it is feasible to train a competitive in-domain classifier with a portion of the target corpus.

3.2 Performance by Topic

In Table 2, we compare the performance of the fine-tuned RoBERTa models with that of the 44-topic cross-domain classifier (top) and the 8-topic cross-domain classifier (bottom) in terms of the accuracy of each topic in the test set using one of the five random runs from the main experiment.[7] One immediate observation is that for the 44-topic classification, the fine-tuned RoBERTa model performs better for larger topics. For this particular run, for topics with more than 10 samples in the test set, the fine-tuned RoBERTa model does better or equally well for all topics with the exception of "education" and "equality."

Our second observation is that the fine-tuned RoBERTa model's advantage over the cross-domain classifier disappears on rare topics, such as "nationalization" and "underprivileged minority groups." Because these topics are rare, the RoBERTa model did not see enough such samples during the training stage.[8] By contrast, the cross-domain classifier has seen considerably more such samples during its training stage with party manifestos. The cross-domain classifier thus has an advantage in predicting samples on rare topics correctly.[9]

Our third observation is that for 8-topic classification, the fine-tuned RoBERTa model outperforms the cross-domain classifier for seven of the eight topics. This is not surprising given that with fewer topics, the number of samples in each topic will become larger, which in turn ensures that the RoBERTa model sees enough training samples for each topic during the

[7] As a robustness check and to show inter-run variations, we report the results from another run in the Supplementary Material.

[8] For instance, among the 4,165 annotated parliamentary speeches, there are only 10 samples that fall into the "underprivileged minority groups" topic.

[9] Note that given the small sizes of some of the classes, the differences in accuracies for some classes, for example, "underprivileged minority groups," are not statistically significant in a difference in proportions test (Wang, Li, and Luo 2016).

https://doi.org/10.1017/pan.2023.3 Published online by Cambridge University Press

Table 2. Accuracy (recall) comparison by topic. Cross-domain classifiers are from Osnabrügge *et al.* (2021). Test set is the same for both models. *N* indicates sample size. Random seed is set to 11. Better results are in bold.

# Classes	Topic	*N*	Cross-domain	Fine-tuning LM
44	Political authority	140	0.550	**0.657**
	Welfare state expansion	49	0.694	**0.714**
	Democracy	44	0.318	**0.341**
	No topic	32	0.000	**0.438**
	Labour groups	31	0.387	**0.484**
	Education	26	**0.885**	0.846
	Constitutionalism	24	0.000	**0.458**
	Economic orthodoxy	21	0.238	**0.571**
	Governmental and administrative efficiency	21	0.238	0.238
	Technology and infrastructure	21	0.333	**0.524**
	Law and order	20	0.650	**0.700**
	Multiculturalism	19	0.632	**0.842**
	Equality	18	**0.389**	0.278
	Free market economy	15	0.000	**0.267**
	Economic growth	13	0.615	**0.769**
	Freedom and human rights	13	0.000	**0.231**
	Market regulation	12	0.167	**0.333**
	Traditional morality	12	0.250	**0.333**
	Military	11	0.727	**0.909**
	National way of life	10	0.300	0.300
	Political corruption	10	0.100	**0.200**
	Protectionism	10	0.200	**0.600**
	Centralization	9	0.111	**0.222**
	Environmental protection	9	0.667	**1.000**
	Agriculture and farmers	7	**0.714**	0.571
	Incentives	7	0.571	0.571
	Civic mindedness	6	0.000	0.000
	Nationalization	5	**0.400**	0.200
	Culture	3	0.000	**0.667**
	Internationalism	2	0.000	**0.500**
	Controlled economy	1	0.000	0.000
	Middle-class and professional groups	1	0.000	0.000
	Noneconomic demographic groups	1	1.000	1.000
	Peace	1	0.000	0.000
	Underprivileged minority groups	1	**1.000**	0.000
8	Political system	180	0.556	**0.622**
	Economy	105	0.600	**0.705**
	Welfare and quality of life	105	0.667	**0.810**
	Freedom and democracy	81	0.284	**0.556**
	Fabric of society	67	**0.582**	0.522
	Social groups	41	0.415	**0.537**
	No topic	32	0.000	**0.344**
	External relations	14	0.571	**0.857**

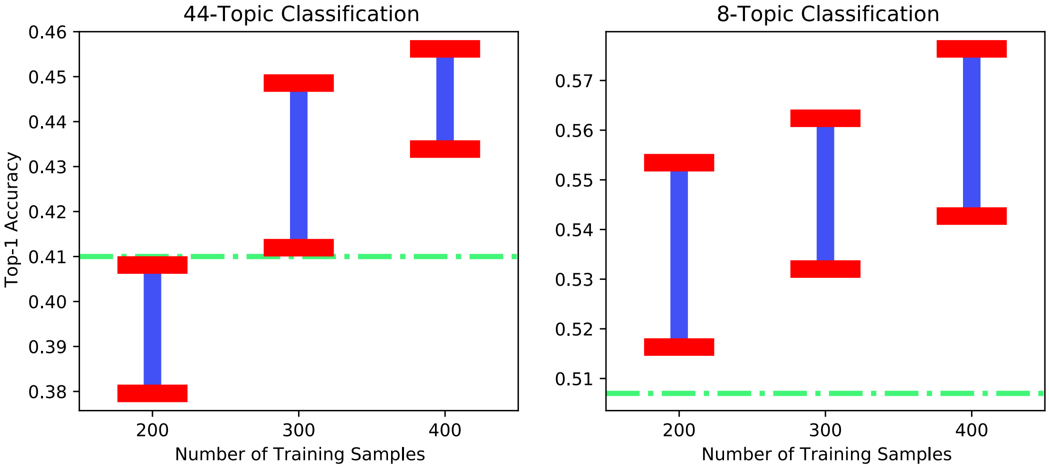

Figure 1. Model performance increases as the training size increases. With 300 training examples, the fine-tuned RoBERTa model outperforms the cross-domain classifier (the dashed green line) on the 44-topic classification task (left) and on the 8-topic classification task (right).

fine-tuning stage. In the next subsection, we explore this question from a slightly different angle: what is the minimum number of samples that we need for the fine-tuned language model to outperform the cross-domain classifiers?

3.3 Number of Training Samples

In this experiment, we study the number of training samples that the fine-tuned language model requires in order to match the performance of that in Osnabrügge *et al.* (2021). Our experiment is motivated by the observation that oftentimes researchers may not have access to an annotated target set with as many as 2,915 training samples as we did in the main experiment. Will the fine-tuned language model remain competitive with a much smaller training set? We report our results in Figure 1. We fine-tune the language model for 20 epochs with 200, 300, and 400 training samples, respectively, and split the remaining samples evenly into the validation set and the test set. We run each setting five times and report the mean of top-1 accuracy plus one standard deviation and the mean minus one standard deviation.[10] For easy comparison, we also include the corresponding performance by the cross-domain classifier as reported in Osnabrügge *et al.* (2021).

We observe that with 300 training samples, the fine-tuned language model is able to outperform the cross-domain classifier on the 44-topic classification task (left) and the 8-topic classification task (right). This suggests that depending on task difficulty, researchers with a few hundred training samples may consider a fine-tuned language model as an effective option.

3.4 Training and Inference Time

While language models are known to be slow given their large sizes, we note that their training and inference time could well fit into the time budget of most researchers. In terms of training, on a single A100 GPU with 40 GB memory, it takes 27 minutes to train the model for 20 epochs over 2,915 samples.[11] Training time will further decrease linearly as we use fewer training samples. To put that into perspective, we note that training the cross-domain classifier with cross-validation in Osnabrügge *et al.* (2021) takes 27 minutes on an iMac with 16 CPUs, and generating a single OLS regression table on large datasets could take more than 20 minutes (Stone, Wang, and Yu 2022).

[10] To ensure that the test set is the same in a run across different training sizes, we first sample 400 training samples and then downsample to 200 or 300 when necessary while keeping the dev set and test set the same.

[11] There are various ways to further reduce the training time, including freezing a few layers of the 12 transformers, reducing sequence length, using fp16, optimizing the training batch size, and distributing training across multiple GPUs. This is beyond the scope of our letter.

https://doi.org/10.1017/pan.2023.3 Published online by Cambridge University Press

It is certainly not fair to compare GPU time with other models' CPU time, but we want to note that from the researchers' point of view, the amount of time used in fine-tuning a language model is mostly comparable to other research methods.

Compared with training, inference is significantly faster. With a batch size of 64, our model makes around 145 inferences per second on a single A100 GPU, which generalizes to 10,000 inferences in a little over 1 minute. With such a quick turnaround in training and inference, our method should fit into the time budget of most researchers.

4 Conclusion

Osnabrügge *et al.* (2021) recently proposed cross-domain supervised training to take advantage of existing labeled data and to reduce data collection costs in classifying political texts. In this letter, we have proposed an alternative that builds on pretrained language models. We have shown that fine-tuning a pretrained language model requires only a small annotated dataset. As a matter of fact, we have shown that with just a small portion (10%) of the annotated dataset that was originally used to evaluate the cross-domain classifier, a fine-tuned RoBERTa-base model can outperform the cross-domain classifier. We have also noted that in topics where there are few to no in-domain training samples, the advantage of the fine-tuned language model over cross-domain classifiers largely disappears. Lastly, we have shown that the fine-tuned models are competitive in terms of training time and inference time. Future research could explore the broader application of pretrained language models, alongside cross-domain classifiers, to other research questions, such as populism prediction (Cocco and Monechi 2022), sentiment and stance analysis (Bestvater and Monroe 2022), and party position analysis (Herrmann and Döring 2021), as well as the optimization of pretrained language models in training and inference.

Acknowledgments

I thank the reviewers and the editor for their excellent comments and guidance, which substantially improved the paper and inspired new ideas for future work. It is no overstatement to say that they contributed to the paper as co-authors.

Data Availability Statement

The replication materials (Wang 2023) are available at https://doi.org/10.7910/DVN/FMT8KR.

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2023.3.

References

Bestvater, S. E., and B. L. Monroe. 2022. "Sentiment Is Not Stance: Target-Aware Opinion Classification for Political Text Analysis." *Political Analysis*.

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning. 2020. *ELECTRA: Pre-Training Text Encoders as Discriminators Rather than Generators*. ICLR.

Cocco, J. D., and B. Monechi. 2022. "How Populist Are Parties? Measuring Degrees of Populism in Party Manifestos using Supervised Machine Learning." *Political Analysis* 30 (3): 311–327.

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*, 4171–4186.

Herrmann, M., and H. Döring. 2021. "Party Positions from Wikipedia Classifications of Party Ideology." *Political Analysis* 31: 22–41.

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.

Longpre, S., Y. Wang, and C. DuBois. 2020. "How Effective Is Task-Agnostic Data Augmentation for Pretrained Transformers?" In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Osnabrügge, M., E. Ash, and M. Morelli. 2021. "Cross-Domain Topic Classification for Political Texts." *Political Analysis* 31: 59–80.

https://doi.org/10.1017/pan.2023.3 Published online by Cambridge University Press

Stone, R., Y. Wang, and S. Yu. 2022. “Chinese Power and the State-Owned Enterprise.” *International Organization* 76 (1): 229–250.

Vaswani, A., et al. 2017. “Attention Is All You Need.” In *31st Conference on Neural Information Processing Systems*.

Wang, Y. 2023. “Replication Data for: Topic Classification for Political Texts with Pretrained Language Models.” Harvard Dataverse, V1. https://doi.org/10.7910/DVN/FMT8KR

Wang, Y., Y. Li, and J. Luo. 2016. “Deciphering the 2016 U.S. Presidential Campaign in the Twitter Sphere: A Comparison of the Trumpists and Clintonists.” In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*.

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

https://doi.org/10.1017/pan.2023.3 Published online by Cambridge University Press