Online Appendix: *Topic Classification for Political Texts with Pretrained Language Models*

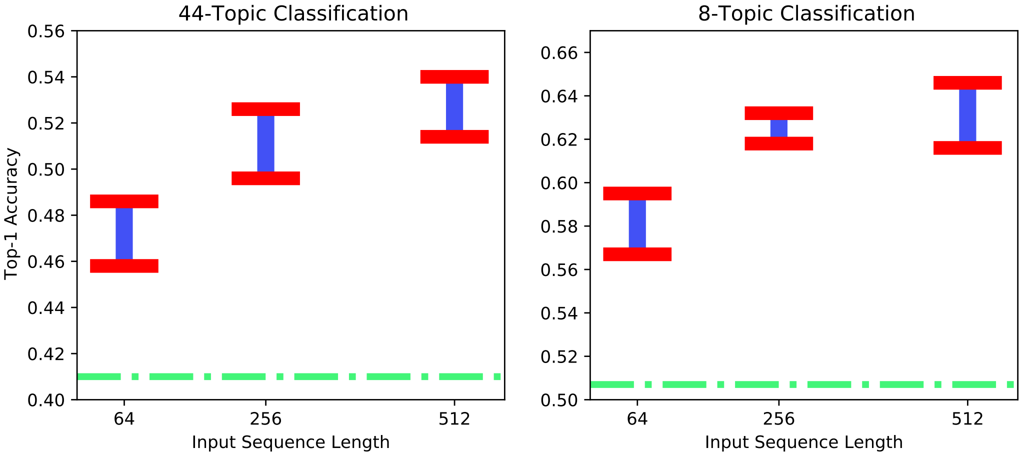
Yu Wang
University of Rochester
Email: w.y@alum.urmc.rochester.edu

Sequence Length

The max sequence length of the RoBERTa model is 512. Out of all the 4,165 samples, 533 samples (12.8%) have a sequence length greater than the max sequence length of 512 and will be truncated. The minimum sequence length is 10, and the max is 4,823, with a standard deviation of 358. The scripts for calculating the above statistics are included in the replication package.

In Figure 1, we report the top-1 accuracy as a function of input sequence length. We observe that longer sequence lengths generally lead to higher accuracies.[1] For easy comparison, we also report the accuracy of the cross-domain classifier (marked green).

Figure 1: Model performance increases as the input sequence lengths increases, for both 44-topic classification and 8-topic classification. While the finetuned language model with an input sequence length of 64 already outperforms the cross-domain classifier (marked by green dashed line), we observe further performance gains when the input sequence length is increased to 256 and then to 512.



[1] Alongside the observation that a longer sequence length yields higher accuracy, there is the question of whether the very act of truncation could "cause" miscalculation. Future work could compare the miscalculation rates of truncated documents with those of comparable lengths (i.e., around the threshold), especially when that threshold is relatively low.

Accuracy Comparison by Topic

Table 1: Cross-domain classifiers are from Osnabrügge et al. (2021). Test set is the same for both models. *N* indicates sample size. Random seed is 12. Better results are in bold.

# Classes	Topic	*N*	Cross-domain	Finetuning 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
	Nationalisation	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
	Non-economic 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**

References

Osnabrügge, M., Ash, E., & Morelli, M. (2021). Cross-Domain Topic Classification for Political Texts. *Political Analysis*.