Appendix

Treating machine learners more like crowd workers:

Increasing validity and robustness of language models with instructions

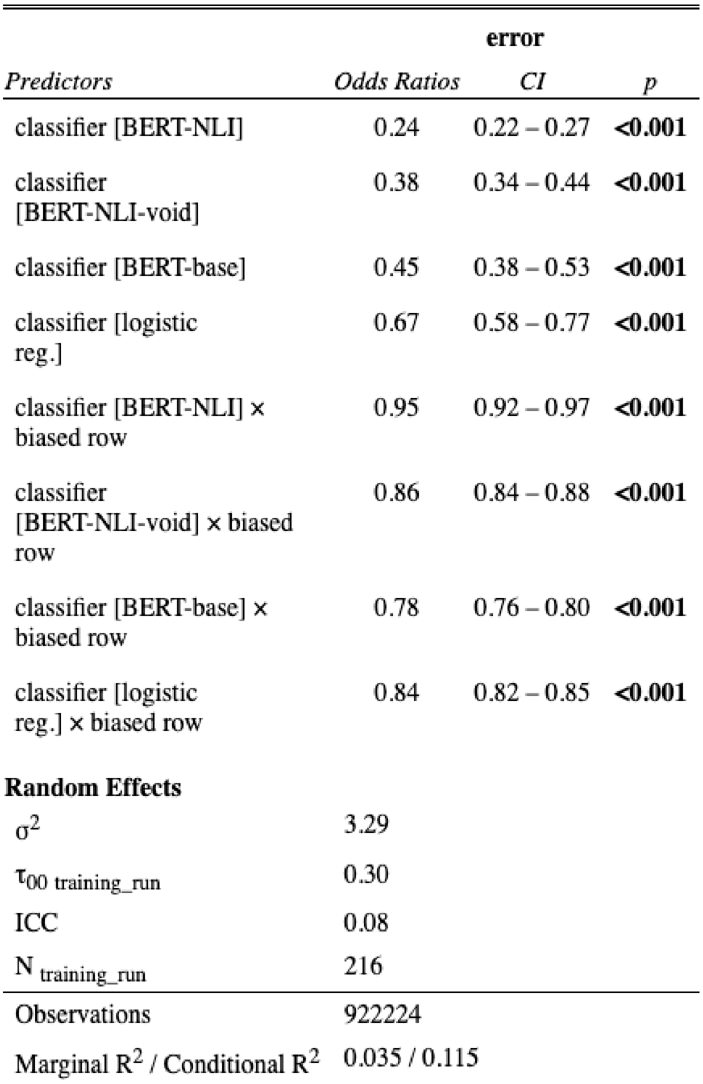
01.11.2023

Anonymous

Appendix A: Analysis Details

The following tables display the results from the binomial mixed-effects regression analysis. For the exact code that produced these tables, see our GitHub repository.[1]

Table 1: Regression table



	error		
Predictors	*Odds Ratios*	*CI*	*p*
classifier [BERT-NLI]	0.24	0.22 – 0.27	**<0.001**
classifier [BERT-NLI-void]	0.38	0.34 – 0.44	**<0.001**
classifier [BERT-base]	0.45	0.38 – 0.53	**<0.001**
classifier [logistic reg.]	0.67	0.58 – 0.77	**<0.001**
classifier [BERT-NLI] × biased row	0.95	0.92 – 0.97	**<0.001**
classifier [BERT-NLI-void] × biased row	0.86	0.84 – 0.88	**<0.001**
classifier [BERT-base] × biased row	0.78	0.76 – 0.80	**<0.001**
classifier [logistic reg.] × biased row	0.84	0.82 – 0.85	**<0.001**
Random Effects			
σ^2	3.29		
$\tau_{00\ \mathrm{training\_run}}$	0.30		
ICC	0.08		
$N_{\mathrm{training\_run}}$	216		
Observations	922224		
Marginal R^2 / Conditional R^2	0.035 / 0.115		

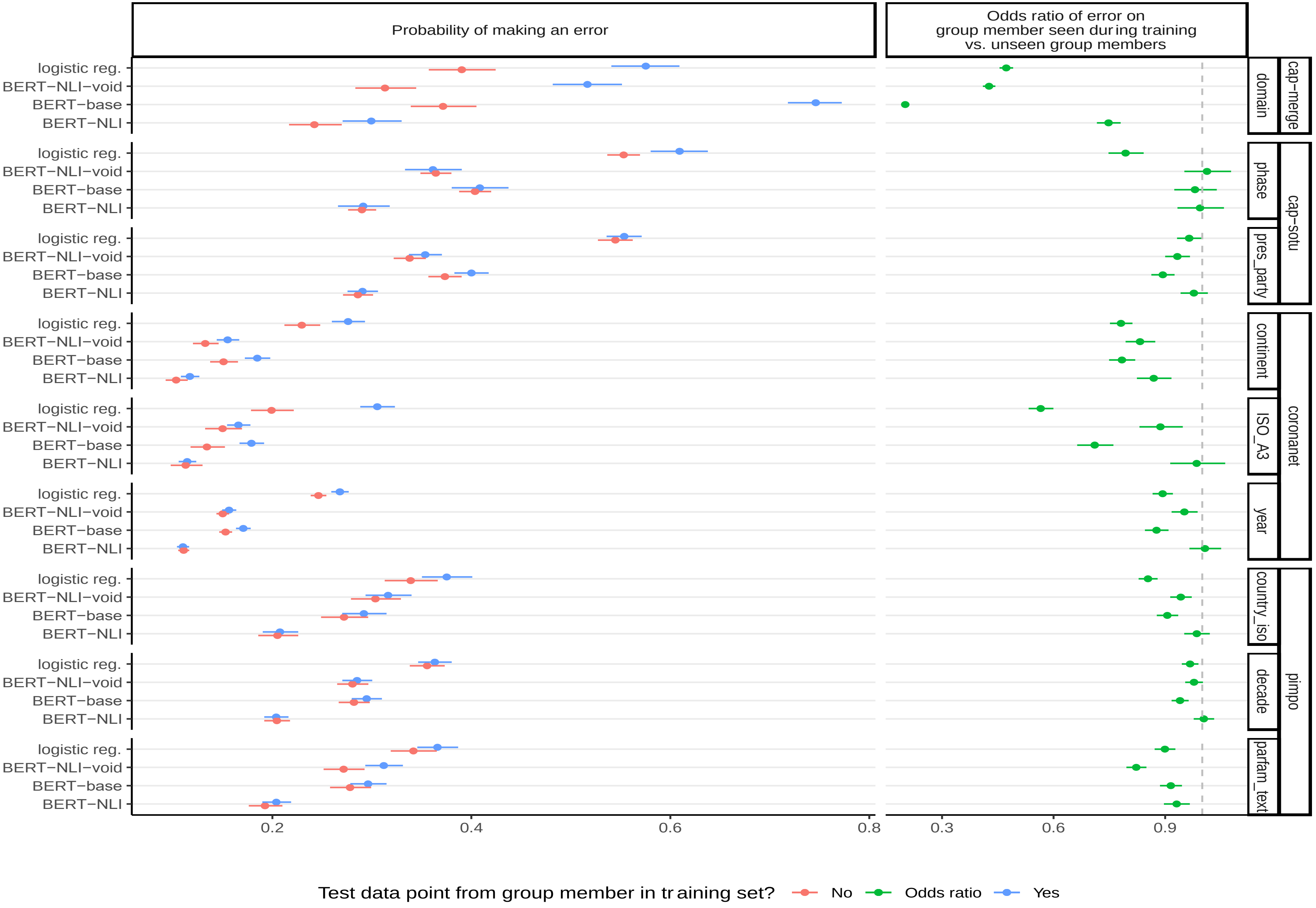
[1] https://anonymous.4open.science/r/language-model-bias-validity-D6E4/

Table 2: Regression disaggregated by groups

	pimpo - country_iso	**pimpo - decade**	**pimpo - parfam_text**	**coronanet - continent**	**coronanet - year**	**coronanet - ISO_A3**	**cap-merge - domain**	**cap-sotu - phase**	**cap-sotu - pres_party**
Predictors	*Odds Ratios*	*Odds Ratios*	*Odds Ratios*	*Odds Ratios*	*Odds Ratios*	*Odds Ratios*	*Odds Ratios*	*Odds Ratios*	*Odds Ratios*
classifier [BERT-NLI]	0.26 ***	0.26 ***	0.26 ***	0.13 ***	0.12 ***	0.13 ***	0.43 ***	0.41 ***	0.41 ***
classifier [BERT-NLI-void]	0.46 ***	0.40 ***	0.45 ***	0.18 ***	0.19 ***	0.20 ***	1.07	0.57 ***	0.55 ***
classifier [BERT-base]	0.41 ***	0.42 ***	0.42 ***	0.23 ***	0.21 ***	0.22 ***	2.94 ***	0.69 ***	0.67 ***
classifier [logistic reg.]	0.60 ***	0.57 ***	0.58 ***	0.38 ***	0.37 ***	0.44 ***	1.35 ***	1.56 ***	1.24 ***
classifier [BERT-NLI] × biased row	0.99	1.00	0.93	0.87 **	1.01	0.98	0.75 ***	0.99	0.98
classifier [BERT-NLI-void] × biased row	0.94	0.98	0.82 ***	0.83 ***	0.95	0.89	0.43 ***	1.01	0.93
classifier [BERT-base] × biased row	0.91 **	0.94 *	0.92 **	0.78 ***	0.88 ***	0.71 ***	0.20 ***	0.98	0.89 **
classifier [logistic reg.] × biased row	0.85 ***	0.97	0.90 ***	0.78 ***	0.89 ***	0.57 ***	0.47 ***	0.79 ***	0.96
Random Effects									
σ^2	3.29	3.29	3.29	3.29	3.29	3.29	3.29	3.29	3.29
τ_{00}	$0.02_{\text{training\_run}}$	$0.01_{\text{training\_run}}$	$0.01_{\text{training\_run}}$	$0.01_{\text{training\_run}}$	$0.00_{\text{training\_run}}$	$0.01_{\text{training\_run}}$	$0.03_{\text{training\_run}}$	$0.00_{\text{training\_run}}$	$0.00_{\text{training\_run}}$
ICC	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
N	$24_{\text{training\_run}}$	$24_{\text{training\_run}}$	$24_{\text{training\_run}}$	$24_{\text{training\_run}}$	$24_{\text{training\_run}}$	$24_{\text{training\_run}}$	$24_{\text{training\_run}}$	$24_{\text{training\_run}}$	$24_{\text{training\_run}}$
Observations	163008	163008	163008	91968	91968	91968	46272	55512	55512
Marginal R^2 / Conditional R^2	0.026 / 0.031	0.023 / 0.025	0.026 / 0.029	0.043 / 0.046	0.039 / 0.039	0.054 / 0.056	0.125 / 0.132	0.050 / 0.051	0.048 / 0.049

** p<0.05   ** p<0.01   *** p<0.001*

Figure 1: Error probability and odds ratios by group



Appendix B: Instructions

The instructions used for BERT-NLI for each dataset are below. The words to the left (e.g. “neutral”) indicate the label which the instruction (hypothesis) belongs to. Note that we use the formulation “The quote…” for the instruction, because we wrap every input text into the string ‘The quote: “{text}”’. This enables us to make the instruction (hypothesis) fit more naturally to the input text.

PimPo

- "neutral": "The quote is neutral towards immigration/integration or describes the status quo of immigration/integration."
- "sceptical": "The quote is sceptical of immigration/integration."
- "supportive": "The quote is supportive of immigration/integration."
- "no_topic": "The quote is not about immigration/integration."

CoronaNet

- "Health Resources": "The quote is related to health resources, materials, infrastructure, personnel, mask purchases"
- "Restriction and Regulation of Businesses": "The quote is related to restricting or regulating businesses"
- "Restrictions of Mass Gatherings": "The quote is related to restrictions of mass gatherings"
- "Public Awareness Measures": "The quote is related to public awareness measures"

CAP-SotU:

- "Defense": "The quote is related to defense, or military"
- "Government Operations": "The quote is related to government operations, or administration"
- 'Health': "The quote is related to health"
- 'International Affairs': "The quote is related to international affairs, or foreign aid"
- "Macroeconomics": "The quote is related to macroeconomics"

CAP-2

- "Civil Rights": "The quote is related to civil rights, or minorities, or civil liberties"
- 'Domestic Commerce': "The quote is related to banking, or finance, or commerce"
- "Government Operations": "The quote is related to government operations, or administration"
- 'Labor': "The quote is related to employment, or labour"
- "Law and Crime": "The quote is related to law, crime, or family issues"

The instructions for BERT-NLI-void are always structured in the same way for each dataset: “The quote is about category A.”, “The quote is about category B.” etc. We randomly determined the order of the instructions.

Appendix C: Hyperparameters

All Transformers are always based on DeBERTav3. The exact model used is publicly available on the Hugging Face Hub.[2] The hyperparameters are based on the extensive hyperparameter search discussed in the appendix of Laurer et al. 2023.[3]

BERT-base:
{'lr_scheduler_type': 'constant', 'learning_rate': 2e-5, 'num_train_epochs': 40, 'seed': SEED_RUN, 'per_device_train_batch_size': 32, 'warmup_ratio': 0.06, 'weight_decay': 0.01, 'per_device_eval_batch_size': 256}

BERT-NLI-void:
{'lr_scheduler_type': 'linear', 'learning_rate': 2e-5, 'num_train_epochs': 30, 'seed': SEED_RUN, 'per_device_train_batch_size': 32, 'warmup_ratio': 0.20, 'weight_decay': 0.01, 'per_device_eval_batch_size': 256}

BERT-NLI:
{'lr_scheduler_type': 'linear', 'learning_rate': 2e-5, 'num_train_epochs': 7, 'seed': SEED_RUN, 'per_device_train_batch_size': 32, 'warmup_ratio': 0.20, 'weight_decay': 0.01, 'per_device_eval_batch_size': 256}

Logistic regression classifier
Vectorizer: {"analyzer": "word", "ngram_range": (1, 2), "min_df": 0.01, "max_df": 0.8}
Classifier: {"random_state": SEED_RUN, "C": 50.0, "max_iter": 200}

[2] https://huggingface.co/MoritzLaurer/deberta-v3-base-zeroshot-v1

[3] https://www.cambridge.org/core/journals/political-analysis/article/less-annotating-more-classifying-addressing-the-data-scarcity-issue-of-supervised-machine-learning-with-deep-transfer-learning-and-bertnli/05BB05555241762889825B080E097C27