A new Benchmark for NLP in Social Sciences: Evaluating the usefulness of pre-trained language models for classifying open-ended survey responses

Maximilian Meidinger and Matthias Aßenmacher

Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany

The ANES 2008 data

- High-quality data for Social Science research
- Surveys on public opinion and voting behaviour
- Supplemented by a coding project for open-ended responses for the 2008 data set [1]

Data preparation

- Grouping of questions by same "*Code sets*"
- Raw "*codes*" cannot be used directly for machine learning purposes
- We understand the codes associated to each response in the data as the labels encountered in a multi-label learning problem, just as [2]
- *Idea:* Represent raw codes as q-dimensional binary vectors which results in "multi-hot encoded" label vectors $\mathbf{y} = (y_1, .., y_n) = \{0, 1\}^q$ associated to each observation
- Train/Test data split is obtained by using an iterative stratification method proposed by [3] for balancing the label distributions, which previously has not been used [2]
- *To get an impression of the prepared data:* Watch the accompanying video for a quick tour through our GitHub repository

Models

- *Simple Baseline:* Logistic regression model + `fastText` vectors [4]
- *External Baseline:* Card & Smith (2015) [2]
- *Transfer learning:* Recent SOTA NLP models, namely BERT [5], RoBERTa [6] and XLNet [7].
- Implementations from the huggingface transformers library (`base, cased` versions)

Multi-label measures

- *In agreement with Card & Smith (2015):* Sample-based F1-Score:

$$F_1^{sample} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i \cap P_i|}{|Y_i| + |P_i|}$$

- *Additional measures:* Macro- & Micro-averaged versions of the traditional F1-Score:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

- *Further: Label Ranking Average Precision* (LRAP) and *Subset Accuracy* (Exact match ratio)

Model Performance on the different benchmark data sets

	Dataset-ID	1	2	3	4	5	6	7	8	9	10
	n	238	288	4393	4672	2100	8399	2096	2094	2094	2092
	#labels	34	29	33	34	26	72	9	11	14	9
F_1^{sample}	Baseline	0.44	0.51	0.57	0.54	0.68	**0.88**	0.92	0.95	0.90	0.91
	BERT [5]	0.00	0.02	0.44	0.35	0.41	0.79	0.94	0.95	0.91	0.93
	RoBERTa [6]	0.00	0.00	0.56	0.55	0.57	0.85	0.95	0.97	**0.93**	0.94
	XLNet [7]	0.00	0.00	0.54	0.58	0.55	0.86	**0.96**	**0.98**	0.91	0.92
	Card & Smith (2015) [2]	**0.55**	**0.67**	**0.71**	**0.71**	**0.81**	0.86	0.94	0.96	**0.93**	**0.96**
F_1^{micro}	Baseline	**0.40**	**0.48**	0.53	0.51	0.61	0.84	0.89	0.93	0.85	0.90
	BERT	0.00	0.03	0.51	0.44	0.46	0.79	0.94	0.95	0.91	0.93
	RoBERTa	0.00	0.00	**0.60**	0.60	**0.62**	**0.85**	**0.96**	**0.97**	**0.94**	**0.95**
	XLNet	0.00	0.00	0.59	**0.61**	0.61	**0.85**	**0.96**	**0.97**	0.90	0.93
F_1^{macro}	Baseline	**0.23**	**0.29**	**0.33**	**0.34**	**0.47**	**0.46**	**0.62**	0.51	**0.56**	**0.71**
	BERT	0.00	0.01	0.11	0.16	0.12	0.09	0.47	0.40	0.39	0.58
	RoBERTa	0.00	0.00	0.18	0.26	0.21	0.14	0.51	0.51	0.44	0.58
	XLNet	0.00	0.00	0.20	0.27	0.21	0.16	0.58	**0.53**	0.43	0.66
LRAP	Baseline	**0.59**	**0.65**	**0.70**	**0.70**	**0.75**	**0.93**	**0.95**	**0.98**	0.92	**0.95**
	BERT	0.09	0.10	0.41	0.32	0.40	0.71	0.94	0.95	0.90	0.93
	RoBERTa	0.09	0.09	0.51	0.49	0.55	0.79	**0.95**	0.97	**0.93**	0.94
	XLNet	0.09	0.09	0.49	0.52	0.53	0.80	**0.95**	0.97	0.90	0.92
subset acc.	Baseline	0.00	**0.10**	0.20	0.17	**0.35**	**0.70**	0.80	0.89	0.76	0.80
	BERT	0.00	0.00	0.16	0.08	0.20	0.41	0.89	0.90	0.81	0.87
	RoBERTa	0.00	0.00	**0.22**	0.20	0.32	0.54	0.91	**0.94**	**0.87**	**0.89**
	XLNet	0.00	0.00	**0.22**	**0.22**	0.31	0.58	**0.92**	**0.94**	0.80	0.87

Table 1:Model performances (measured as micro- and macro-averaged F_1-scores, LRAP and Subset Accuracy) for all considered architectures. Results are displayed separately for each data set with the best performance per data set in bold. We report F_1^{sample} to ensure comparability to the results reported by Card & Smith (2015) [2].

Results

- *Very* poor performance of BERT & Co. on the most challenging data sets (IDs 1 & 2) with very low n and a high #labels.
- Performance increases with increasing n and decreasing #labels
- Strong performance of the two baselines with regard to the F_1^{sample} measure
- RoBERTa & XLNet pretty competitive when performance is measured with F_1^{micro}, clearly outperforming the "ordinary" BERT model

Conclusion

- Unified preprocessing of a hallmark Social Science data set (ANES 2008) for NLP
- Pre-trained language models do not perform as well here as on other commonly used tasks
- Still room for improvement regarding a real-life scenario from Social Science with scarce data

Additional Contributions:

- Fixed Train/Test split enables a valid comparison against our baseline for future research
- Extension of common benchmark data sets used for transfer learning models in NLP

References (Poster)

[1] J. A. Krosnick, A. Lupia, and M. K. Berent, "2008 open ended coding project," 2012.

[2] D. Card and N. A. Smith, "Automated coding of open-ended survey responses," 2015.

[3] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Machine Learning and Knowledge Discovery in Databases* (D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, eds.), (Heidelberg), pp. 145–158, Springer, 2011.

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[7] Z. Yang *et al.*, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, pp. 5753–5763, 2019.

Further Information // Contact

Working group: misoda.statistik.uni-muenchen.de
Code: github.com/mxli417/co_benchmark
M. Meidinger: mx.meidinger@gmail.com
M. Aßenmacher: matthias@stat.uni-muenchen.de