Assessing the downstream effects of training data annotation methods on supervised machine learning models

Keywords: Annotation, Bias, Twitter, Machine Learning, Epistemic Error

Extended Abstract

Machine learning (ML) training datasets often rely on human-annotated data collected via online annotation instruments. These instruments have many similarities to web surveys, such as the provision of a stimulus and fixed response options. Survey methodologists know that item and response option wording and ordering, as well as annotator effects, impact survey data (Grove et al. 2004, Biemer et al. 2017). Previous research (Beck et al. 2022) showed that these effects also occur when collecting annotations for ML model training and that small changes in the annotation instrument impacted the collected annotations. This study builds on those results, exploring how instrument structure and annotator composition impact models trained on the resulting annotations.

Using Twitter data on hate speech, we collect annotations with five experimental versions of an annotation instrument, randomly assigning annotators five different versions. In condition A they were asked to indicate whether this tweet contains hate speech or offensive language (or both). In condition B, the question was split into two different tasks first evaluating hate speech then offensive language, with condition C reversing that order. Condition D asked first the hate speech classification for all tweets, to then evaluate them also for offensive language, and condition E reversed that order.

Our data includes 3,000 Tweets labeled by 1,897 annotators, resulting in a total of over 90,000 annotations. We train and fine-tune state of the art ML models such as LSTM and BERT for hate speech classification on five training data sets that consist of annotations collected with five different instrument versions and evaluate on the corresponding five test sets.

By comparing model performance across the instruments, we aim to understand 1) whether the way annotations are collected impacts the predictions and errors of the trained models; and, 2) which instrument version leads to robust and efficient models, judged by the performance scores across instrument versions and model learning curves. In addition, we expand upon on earlier findings that annotators' demographic characteristics impact the annotations they make.

We find considerable performance differences across test sets and, to some extent, also across models trained with tweets that were labeled using different instrument versions. Our results emphasize the importance of careful annotation instrument design. Hate speech detection models are likely to hit a performance ceiling without increasing data quality; By paying additional attention to the training data collection process, researchers can better understand how their models perform and assess potential misalignment with the underlying concept of interest they are trying to predict.

References

Beck, J., Eckman, S., Chew, R., and Kreuter, F. (2022). Improving labeling through social science insights: Results and research agenda. In Chen, J. Y. C., Fragomeni, G., Degen, H., and Ntoa, S., editors, *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, pages 245–261. Springer Nature Switzerland.

Biemer, P., de Leeuw, E.D., Eckman, S., et al. (2-17): *Total Survey Error in Practice*. Wiley, Hoboken

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019): *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT (1) 2019: 4171-4186

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., Tourangeau, R. (2004): *Survey Methodology*. Wiley, Hoboken

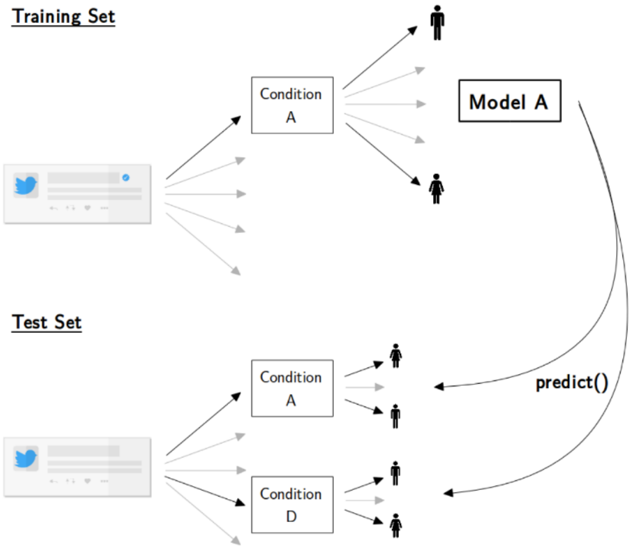
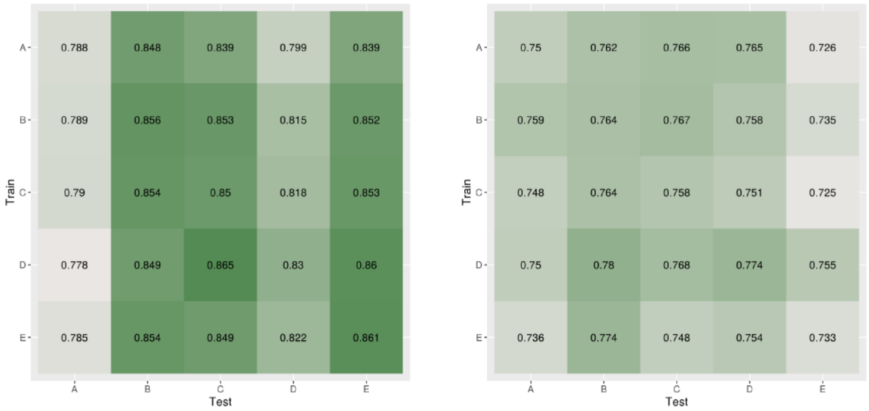


Figure: Experimental Conditions



(a) Offensive Language

(b) Hate Speech

Figure: Performance (ROC-AUC) of LSTM models across labeling conditions