DMLR@ICLR Extended Abstract

Anonymous submission

Reviewed on OpenReview:

Editor:

Introduction When developing Machine Learning (ML) solutions, many efforts and resources go into algorithm optimization to maximize performance metrics and reduce the resources employed. However, at some point the real-life performance of ML applications will be limited by the quality of the underlying training data. More importantly, unwanted biases and flaws within the annotated training data can also creep into the resulting models and lead to an overreliance on erroneous data. A data-centric approach can help gaining a better understanding of determinants of bias and data quality in ML. Thorough experimental research, that carefully evaluates current practices in the light of their effect on training data and models, is key to develop new best practices for annotation. To foster the improvement of annotation practices, we follow a research agenda that assesses the quality of ML training data and its drivers. Inspired by the realization that annotation tasks are similar to web surveys, we derive hypotheses from research in survey methodology and social psychology. More specifically, surveys and annotation tasks both provide the human with a fixed stimulus and ask to select one or more fixed response categories. Informed by a rich interdisciplinary body of literature we conduct experimental research to gain an understanding of mechanisms that impact the quality of annotated training data.

Data We collect annotations of tweets to assess how sensitive hate speech annotations are to variations in task design and annotator sample. Based on a pre-annotated tweet corpus provided by Davidson et al. (2017), we collect annotations in two survey experiments. In Experiment 1, 1000 annotators from the Prolific panel annotated the same 20 tweets, randomly assigned to one of six experimental conditions Beck et al. (2022). The experimental conditions varied by task structure and availability of a "Don't know" option (see Figure 1). In Experiment 2, 3000 tweets were annotated up to 3 times in each of five experimental conditions (Experiment 2) by a total of 900 annotators. While we dropped the experimental conditions containing "Don't know", we added two new ones (see Figure 3). Tweets were split up into batches of 50 tweets, and the order remained constant within one batch. Each annotator was randomly assigned to annotate one batch in one experimental condition. In line with Davidson et al. (2017) we use the classes Hate Speech (HS), Offensive Language (OL) and Neither (NE). Both annotation tasks concluded with a variety of demographic and task perception items. [1]

1. See Appendix for more detailed description of the experimental conditions. The two data sets are hosted at `https://huggingface.co/soda-lmu`. All figures and tables referenced in this document can be found in the Appendix.

Results

Task Structure Effects Results from Experiment 1 explicitly show that even small changes in the configuration of an annotation task impact how the human annotator completes the task. Figure 2 shows the annotation distributions by experimental condition. Querying both labels (OL and HS) on one screen (left column) changed the distribution of annotations by up to eight percentage points compared to splitting the same task up in two screens (center and right column). The order of collecting the annotations also mattered: Flipping the sequence of HS and OL annotation retrieval led to a significant shift of HS annotations by seven percentage points (center vs right column). The "Don't know" option (in the bottom row of the figure) was rarely used across all conditions.

Downstream Effects The tailored data collection design in Experiment 2 allowed us to train models on the collected annotations. We trained BERT models using data from each of the experimental conditions and evaluated the models on annotations from the other conditions. Each cell contains the ROC-AUC[2] of a model trained on the data from the condition on the y-axis and evaluated on the data from the condition on the x-axis. We do not observe the main diagonal cells to show the highest ROC-AUC (e.g. a model trained on data from Cond. B also performs best when tested on Cond. B). Much rather, certain columns (e.g. A and D for the test data of OL models) indicate structurally lower model performance. The bias introduced by the structure of the annotation task affects not only the data itself but also transfers downstream into the models (Figures 4 and 5).

Order Effects As human annotators are prone to cognitive biases, the order in which annotation objects are presented likely influences the label assigned. Our analyses show that a tweet's order in its batch negatively correlates with its probability to be annotated as either HS and OL in all five experimental conditions (Figure 6). Regression analysis confirmed a statistically significant, yet small, negative effect for both outcome variables (Table 1).

Demographic Effects In addition to *how* an object is annotated, it is highly relevant *who* conducts this task. Past research observed that a variety of annotator characteristics correlate with annotations (Beck, 2023). Collecting annotator characteristics alongside the annotations in Experiment 1 allowed us to run empirical analyses. Figure 7 shows a significant distribution shift in HS annotations by the annotator's first language. First language english speakers selected significantly more tweets to be hateful compared to non-native speakers. Most likely certain jargon, slurs or irony are much harder to pick up for non-native speakers of a language. In addition, ongoing work provided first evidence that african-american annotators were less likely to flag anti-asian hate speech and vice versa.

Conclusion and Future Work Our work highlights a variety of components within annotation that impact the resulting training data. The results stress the sensitivity of annotations and models to slight changes in the data collection process. Future work could shift from text classification to other annotation applications such as image or audio annotation. Furthermore, follow-up research should examine and compare mechanisms within different annotator profiles (e.g., experts vs. laypeople).

2. "Receiver Operating Characteristic - Area under Curve"

Reproducibility Statement

The following resources foster the reproducibility of our work :

1. The data used for Beck et al. (2022) and Kern et al. (2023) via `https://huggingface.co/soda-lmu`
2. The model training code used in Kern et al. (2023) via `https://osf.io/mn9ux/?view_only=75c84803b70947cb9831bd897cf8f01e`
3. Thorough descriptions of our data collection processes and analyses in the manuscripts.

References

Jacob Beck. Quality aspects of annotated data: A research synthesis. *AStA Wirtschafts-und Sozialstatistisches Archiv*, pages 1–23, 2023.

Jacob Beck, Stephanie Eckman, Rob Chew, and Frauke Kreuter. Improving labeling through social science insights: Results and research agenda. In Jessie Y. C. Chen, Gino Fragomeni, Helmut Degen, and Stavroula Ntoa, editors, *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, pages 245–261, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-21707-4.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515, 2017. doi: 10.1609/icwsm.v11i1.14955. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/14955`.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. Annotation sensitivity: Training data collection methods affect model performance. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore, December 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.findings-emnlp.992`.

Appendix

	All label options in one screen	First hate speech, then offensive language	First offensive language, then hate speech
Without „Don‘t know“ Option	Condition 1 n = 164 annotators	Condition 3 n = 183 annotators	Condition 5 n = 160 annotators
With „Don‘t know“ option	Condition 2 n = 178 annotators	Condition 4 n = 158 annotators	Condition 6 n = 164 annotators

Figure 1: Experimental conditions in Experiment 1

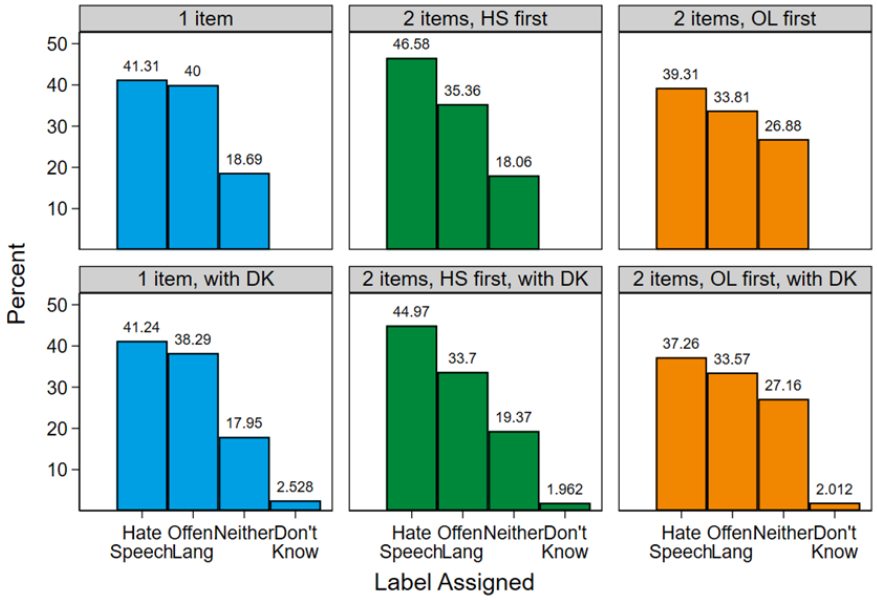

Figure 2: Annotation distribution by condition in Experiment 1

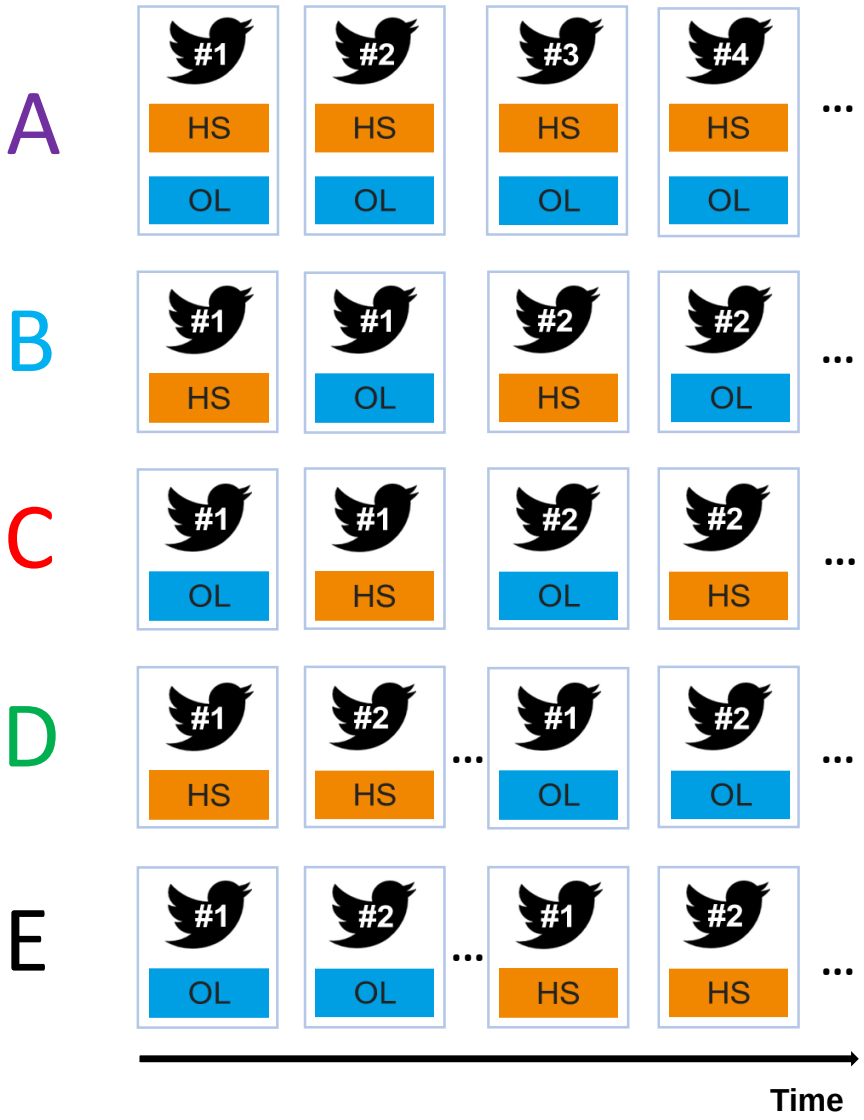

Figure 3: Experimental conditions in Experiment 2

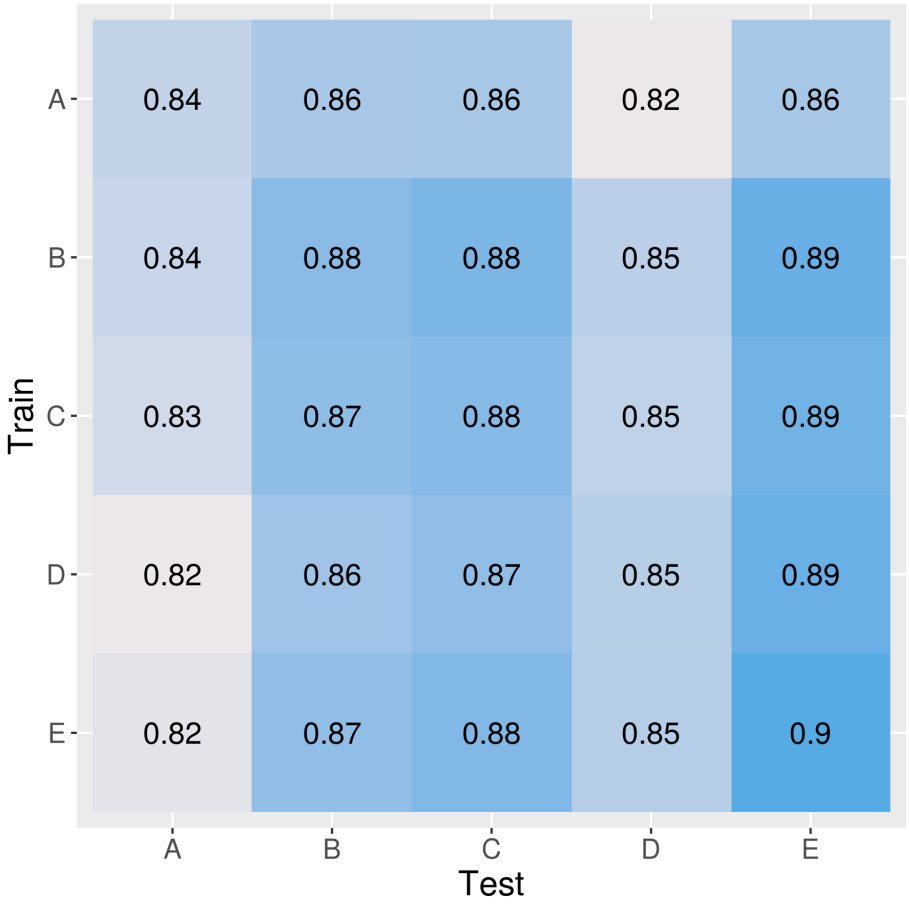

Figure 4: ROC-AUC of BERT-Models trained and tested on annotations from each of five conditions. Outcome: Offensive Language

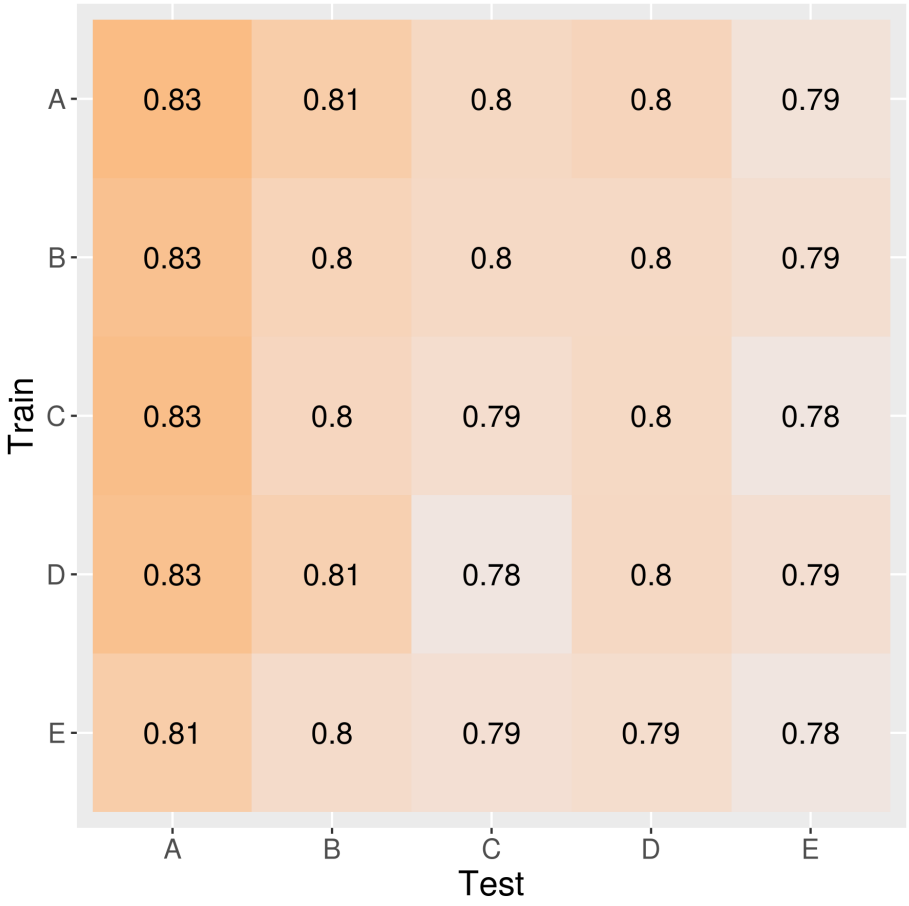

Figure 5: ROC-AUC of BERT-Models trained and tested on annotations from each of five conditions. Outcome: Hate Speech

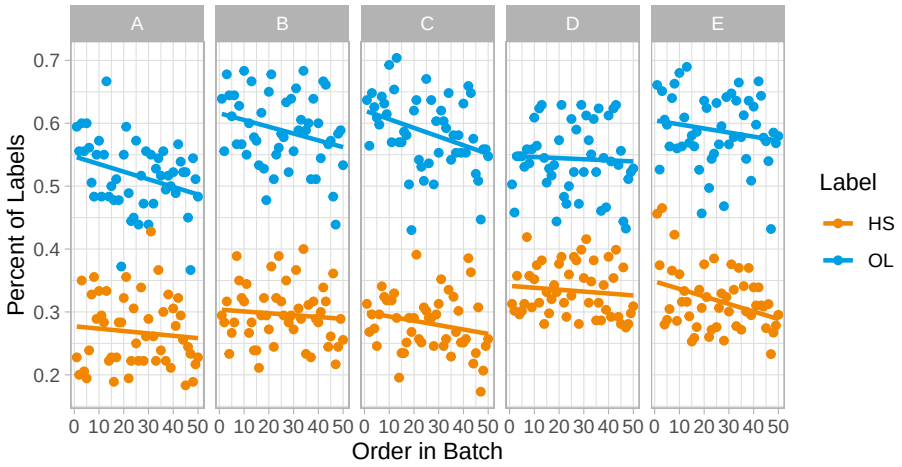

Figure 6: Relationship between batch order and annotation probability across experimental conditions

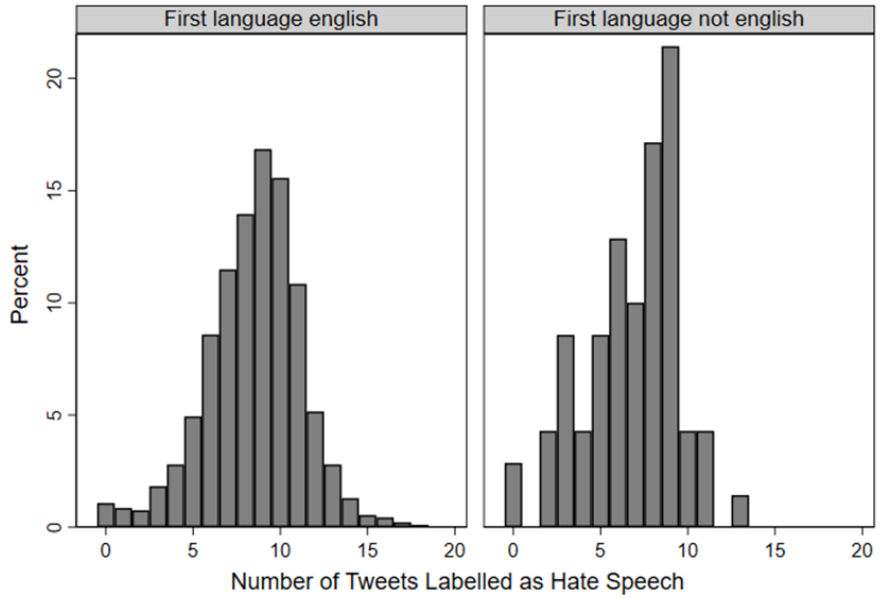

Figure 7: Distribution of number of HS annotations (of 20 total) by annotator First Language

	HS	OL
Order	-0.00057***	-0.00090***
	(0.00015)	(0.00016)
N	44,550	44,550

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Estimated intercept not shown

Table 1: Order Effects in Hate Speech and Offensive Language Annotations