Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity

Jacob Beck[♣] **Stephanie Eckman**[♢] **Bolei Ma**[♣]
Rob Chew[♡] **Frauke Kreuter**[♣,♢]

[♣]LMU Munich & MCML [♢]University of Maryland, College Park
[♡]RTI International

{jacob.beck, bolei.ma, frauke.kreuter}@lmu.de
steph@umd.edu rchew@rti.org

Abstract

The data-centric revolution in AI has revealed the importance of high-quality training data for developing successful AI models. However, annotations are sensitive to annotator characteristics, training materials, and to the design and wording of the data collection instrument. This paper explores the impact of observation order on annotations. We find that annotators' judgments change based on the order in which they see observations. We use ideas from social psychology to motivate hypotheses about why this order effect occurs. We believe that insights from social science can help AI researchers improve data and model quality.

1 Introduction

When annotating training data for AI models, the primary focus is often on the quantity of labeled data rather its quality or how it is collected. Introductory machine learning courses often portray training data as generated by noise-free independent draws from an underlying distribution. However, data annotation is a human rather than a statistical process and this human label variation has often been neglected (Plank, 2022); it is unclear if observations and their annotations are truly independent from the perspective of the annotator.

This paper tests the hypothesis that the order of observations presented during annotation impacts the labels assigned. We experiment with hate speech and offensive language annotations of tweets. Our findings demonstrate that observation order impacts annotations. Moreover, the order effect differs across the five conditions of the annotation instrument we tested, highlighting the nuanced influence of observation order on the annotation process.

2 Relevant Research

The hypothesis that annotators annotate observations differently based on the order in which they

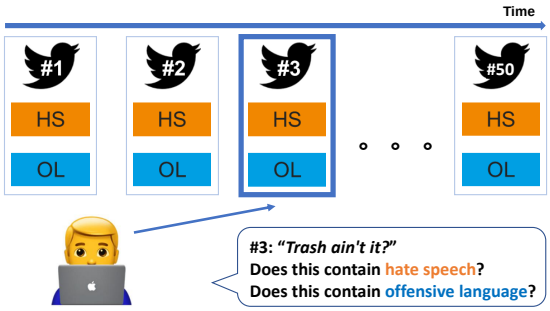

Figure 1: An example view of tweet annotations in a sequentially ordered batch of 50 tweets

are presented rests on several streams of literature. Our previous papers on *annotation sensitivity* suggest that the design of the data collection instrument impacts the annotations collected (Beck et al., 2022) and the machine learning models trained on those annotations (Kern et al., 2023). Spreading two annotations of a single tweet across two screens led to changes in Hate Speech and Offensive Language annotation rates of five to seven percentage points compared to conducting both annotations on the same screen (Beck et al., 2022). The impact of these small manipulations of the data collection instrument prompted us to think that the order of annotation observations may also impact the annotations collected.

Research from social psychology and survey methodology suggests two additional factors that may affect annotation behavior: *context* and *burden*. Context effects (also called anchoring or priming effects) concern how human perception is influenced by information perceived previously (Tversky and Kahneman, 1974; Strack, 1992). Context effects can make two objects appear more similar or more different than they otherwise would. For example, a very tall person can make others seem shorter: a contrast effect. An unethical politician can make other politicians seem less ethical: an assimilation effect (Bless and Schwarz, 2010).

Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024), pages 81–86

In surveys, context effects can lead to question order effects: earlier questions impact how later questions are understood. For example, exchanging the order of two questions about abortion meaningfully changed respondents' reported opinions. Reordering the questions in a scale that measures anxiety resulted in increased anxiety scores (Chapter 2, Schuman and Presser, 1996). A similar effect may occur during annotation: early observations may change how annotators perceive later observations.

In surveys, respondents often alter their response behavior across the length of a survey, which can also introduce order effects. Most investigations suggest that response quality decreases with length rather than increasing. Respondents are more likely to satisfice (choose an answer that is good enough rather than evaluating all response options (Krosnick et al., 1996)) as survey length increases (Galesic and Bosnjak, 2009). Annotators may also engage in satisficing behavior, annotating later observations with less care. Alternatively, annotators may gain expertise as they annotate and assign more accurate labels to later observations (Lee et al., 2022).

These research findings lead us to hypothesize that annotations may show order effects, that is, that observations which appear earlier receive different annotations than they would if they appeared later. This paper analyzes annotations of tweets in the context of hate speech and offensive language to test whether the order of the tweets impacts the annotations assigned. This preliminary research will inform future studies and contribute to the development of annotation best practices for the NLP community.

3 Data

We use our previously collected dataset (Kern et al., 2023) that contains annotations of 3000 tweets as containing Hate Speech (HS) and Offensive Language (OL). Tweets were selected from the Davidson et al. (2017) corpus and randomly grouped into batches of 50 tweets. Each batch was annotated 15 times: three times in five experimental conditions. This data set supports the estimation of order effects because the tweets were annotated in random order and that order was recorded in the data set.[1]

Figure 2 illustrates the five conditions. Condition A collected both labels for a tweet on one screen, offering options for HS, OL, or neither. Conditions B and C divided the annotation for a tweet over two screens. For Condition B, the first screen prompted annotators to indicate whether the tweet contained HS, and the subsequent screen addressed OL. Condition C mirrored Condition B but reversed the order of questions for each tweet. In Condition D, annotators first identified HS for all assigned tweets and then annotated OL for the same tweets in the same order. Condition E followed a similar approach but began with the OL annotation task for all tweets, followed by the HS annotation task.

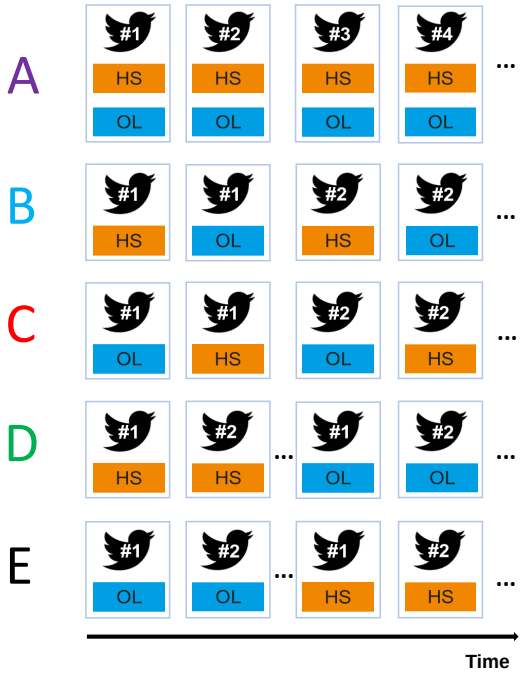

Figure 2: Five Experimental Conditions

The annotators were 908 members of the Prolific panel living in the US.[2] Each participant annotated one batch of 50 tweets in one condition (see Figure 1). Within each batch, the tweets were randomly ordered. However, the order of the tweets was fixed across the 15 annotators. The data set has 44,550 annotations of 3,000 tweets.[3] See Kern et al. (2023) for details of the annotation process.

4 Methods

We first look graphically at order effects, plotting the percent of tweets labeled as HS and OL against

[1]Data are available at https://huggingface.co/datasets/soda-lmu/tweet-annotation-sensitivity-2

[2]https://www.prolific.com/

[3]Some annotators stopped before annotating all 50 assigned tweets; annotations by two annotators of 50 tweets each were corrupted and omitted from the analysis. And the N/A annotations were omitted from the analysis.

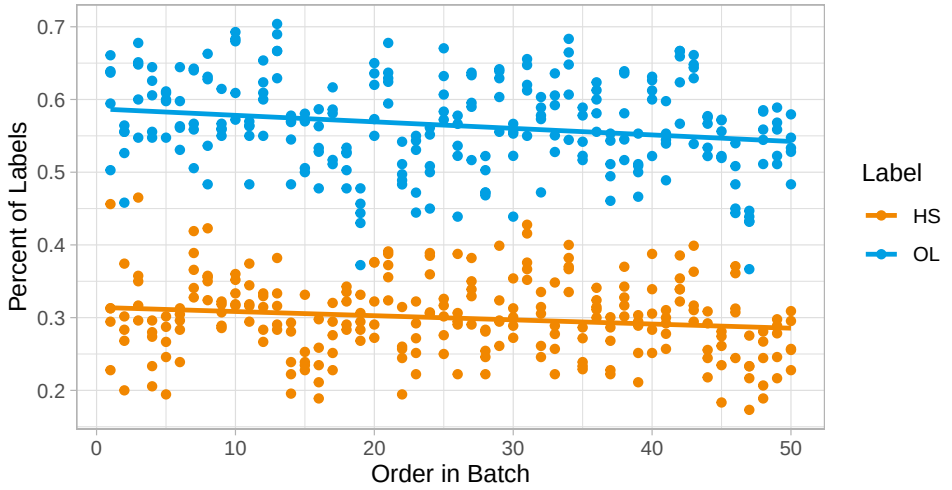

Figure 3: Percentage of Hate Speech, Offensive Language Annotations versus Tweet Order

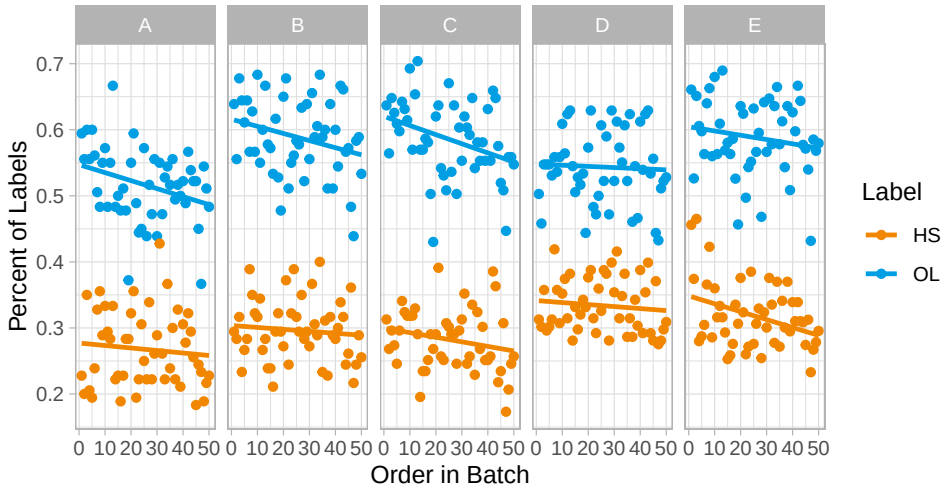

Figure 4: Percentage of Hate Speech, Offensive Language Annotations versus Tweet Order, by Condition

tweet order in the annotation batch. Then we estimate linear probability models to test for order effects. We also test which conditions of the annotation instrument are more vulnerable to order effects using linear probability models.

5 Results

Figure 3 shows the percentage of HS and OL labels by batch order. If order had no effect, the line through the points would be horizontal, because tweets were randomly assigned to batches and randomly ordered within batches. Instead, we see that the percentage of labels that are hate speech or offensive language decreases with batch order.

We ran linear probability models to test these order effects. The dependent variable in each model is an indicator of whether a tweet was annotated as hate speech or offensive language. The independent variable is the order of a tweet within a batch (1 through 50). Table 1 shows the slope coefficients of the HS and OL regression models in Figure 3. The negative and significant coefficients in both models indicate that tweets later in a batch are less likely to be labeled as HS and OL. Because tweets were randomly ordered within batches, we

	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

	HS	OL
Condition A	0.2772***	0.547***
	(0.0098)	(0.011)
Condition B	0.3037***	0.616***
	(0.0098)	(0.011)
Condition C	0.2998***	0.620***
	(0.0098)	(0.011)
Condition D	0.3415***	0.548***
	(0.0098)	(0.011)
Condition E	0.35***	0.605***
	(0.01)	(0.011)
Cond. A × Order	-0.00038	-0.00121***
	(0.00033)	(0.00036)
Cond. B × Order	-0.00030	-0.00108**
	(0.00033)	(0.00036)
Cond. C × Order	-0.00069*	-0.00138***
	(0.00033)	(0.00036)
Cond. D × Order	-0.00030	-0.00017
	(0.00034)	(0.00036)
Cond. E × Order	-0.00119***	-0.00066
	(0.00034)	(0.00037)
N	44,550	44,550

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

Intercept not included in models

Table 2: Order Effects by Condition

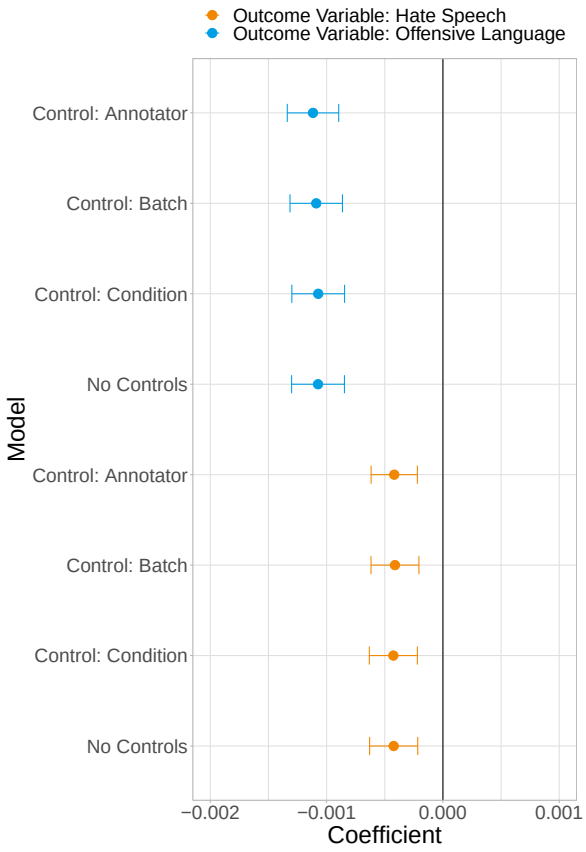

Figure 5: Estimated Coefficient on Tweet Order in linear Regression Models

interpret these coefficients as order effects in tweet annotation.

In addition, we ran models that controlled (separately) for annotator, condition, and batch fixed effects for both outcome variables. Figure 5 indicates that the coefficient on the order variable was substantively unchanged and significant in each model.

Figure 4 and Table 2 analyze order effects separately for the five instrument conditions. We see strong main effects for condition: the annotation collection instrument influences the annotations collected, as reported in Kern et al. (2023). When collecting hate speech annotations, order effects are negative and significant in Conditions C and E. However, the order effects in Conditions C and E are not significantly different from each other. Here it seems important that in both conditions the OL annotation preceded the HS annotation. This could be a potential explanation for the significant condition-specific order effects for HS annotation in Conditions C and E. Contrary to this theory, when collecting OL annotations, order effects are negative and significant in Conditions A, B, and C. Again, the condition-specific order effects are not different from each other.

6 Discussion & Conclusion

The order in which observations are presented to annotators influences the annotations they assign. The later a tweet appeared in a batch, the less frequently it was annotated as hate speech or offensive language. The estimated effects are small, however. The fiftieth tweet in a batch is approximately 2.8 percentage points less likely to be annotated as hateful and 4.5 percentage points less likely to be annotated as offensive language than the first tweet in a batch. However, annotators often label more than 50 observations, which could lead to stronger order effects. We see no evidence of a positive order effect, overall or in any of the five experimental conditions: later tweets are not more likely to be annotated as hate speech or offensive language.

Statistically significant order effects are present in five of the ten conditions we tested (five annotation conditions (Figure 2) times two labels). While Condition E showed a highly significant order ef-

fect for HS annotations, its converse, Condition D, does not have significant order effects for either annotation. The condition-specific results do not seem in line with the order effect mechanisms described in the Relevant Research section. While it is unclear to what degree these small but significant order effects can be accounted to context effects in place our empirical findings might be interpreted as the first evidence for annotation burden effects. Annotator fatigue or boredom could have been the main drivers of diminishing annotation probability, which would explain why no interpretable condition-specific order effects were observed; the burden of two annotations for 50 tweets was constant across all conditions.

The order effects we find in this study suggest that researchers collecting annotations for model building should give more thought the order of observations presented to annotators. Often, the order in which observations are annotated is driven by the needs of the model, as in Active Learning (Wang and Plank, 2023). We suggest that label collectors should also consider the impact of observation order on annotators. Until we better understand the causes of order effects, we recommend random ordering of observations. We also recommend thorough documentation of annotation collection methods to foster replicability and reproducibility.

More research is needed to identify the underlying mechanisms of the order effects we have detected to better understand when they may appear and at what intensity. The literature reviewed above suggests several hypotheses that future research should test. Follow-up research could investigate whether order effects are stronger when the annotation task is more challenging or difficult, as suggested by the survey literature (Schuman and Presser, 1996), and measure whether annotation guidelines/tutorials anchor the annotation behavior. An in-depth qualitative analysis of single tweets or sequences of two tweets could yield valuable insights into linguistic determinants of order effects. The deeper understanding provided by future research should help us design annotation procedures to reduce order effects.

This preliminary work adds to the growing literature on annotation sensitivity. Even large language models are fine-tuned on human feedback about the most appropriate and relevant response. This research and others cited above demonstrate that social science theories about how people answer questions and make judgments are crucial to the collection of high-quality training data for NLP and other AI models.

Limitations

Several limitations challenge the validity and generalizability of our results. First, our data do not allow us to test hypotheses about the causes of the order effects. Although tweets were randomly assigned to batches and randomly ordered within batches, each tweet always appeared in that same order across annotators. The lack of randomization of order across annotators limits our ability to test hypotheses about contrast and assimilation or about learning and burden over time. It also hampers our ability to uncover the reasons behind the different order effects by conditions. It is also possible that the downward slopes in the graphs (Figures 3 and 4) might be caused by a failure in the randomization process in each of the conditions. We encourage future work on this issue.

We are also not able to assess whether order effects improve or worsen after 50 tweets. Annotators often perform many more than 50 annotations. If fatigue is a factor in the order effect we detected, annotation quality may worsen as annotators perform more annotations. In addition, we used only English tweets and only American annotators. Future work should look at other tweets and other populations, as well as other types of NLP and non-NLP tasks.

In addition, while the five experimental conditions contained the same number of tweets (50), each annotator in Condition A saw 50 screens while the others saw 100 screens. However, we did not detect meaningful differences between Condition A and the other conditions, suggesting that the number of tweets is more important to order effects than the number of screens.

Ethics Statement

This data collection was reviewed by the IRB of RTI. Annotators were paid a wage in excess of the US federal minimum wage. Our work deals with hate speech and offensive language, which could cause harm (directly or indirectly) to vulnerable social groups. We do not support the views expressed in these tweets.

Acknowledgements

This research received funding support from RTI International, MCML, and BERD@NFDI.

References

Jacob Beck, Stephanie Eckman, Rob Chew, and Frauke Kreuter. 2022. Improving labeling through social science insights: Results and research agenda. In *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, pages 245–261, Cham. Springer Nature Switzerland.

Herbert Bless and Norbert Schwarz. 2010. Mental construal and the emergence of assimilation and contrast effects: The inclusion/exclusion model. *Advances in Experimental Social Psychology*, 42:319–373.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. 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.

Mirta Galesic and Michael Bosnjak. 2009. Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. *Public Opinion Quarterly*, 73(2):349–360.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.

Jon A. Krosnick, Sowmya Narayan, and Wendy R. Smith. 1996. Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 1996(70):29–44.

Ji-Ung Lee, Jan-Christoph Klie, and Iryna Gurevych. 2022. Annotation curricula to implicitly train non-expert annotators. *Computational Linguistics*, 48(2):343–373.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Howard Schuman and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Fritz Strack. 1992. "order effects" in survey research: Activation and information functions of preceding questions. In *Context Effects in Social and Psychological Research*, pages 23–34, New York, NY. Springer New York.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Xinpeng Wang and Barbara Plank. 2023. ACTOR: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. Association for Computational Linguistics.