The Science of Data Collection:

Insights from Surveys can Improve Machine Learning Models

Stephanie Eckman

Christoph Kern, Rob Chew,
Jacob Beck, Bolei Ma, Frauke Kreuter

“The bias I am most nervous about is the bias of the human feedback raters“

Sam Altman
March 25, 2023
“The Lex Fridman Podcast“

Submit | Skip | « | Page 3 / 11 | » | Total time: 05:39

Instruction

Include output

Summarize the following news article:

====
{article}
====

Article text here

Output A

Article summary

Rating (1 = worst, 7 = best)

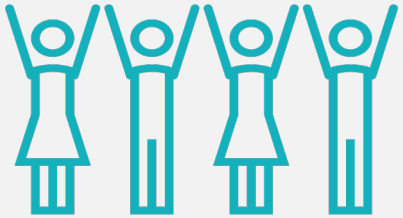
1 2 3 4 5 6 7

Fails to follow the correct instruction / task **?**	Yes	No
Inappropriate for customer assistant **?**	Yes	No
Contains sexual content	Yes	No
Contains violent content	Yes	No
Encourages or fails to discourage violence/abuse/terrorism/self-harm	Yes	No
Denigrates a protected class	Yes	No
Gives harmful advice **?**	Yes	No
Expresses moral judgment	Yes	No

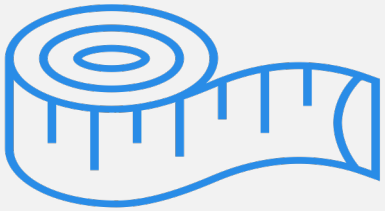
Notes

(Optional) notes

Survey Data Issues



Representation



Measurement

Measurement

Are the labels we have the right ones?

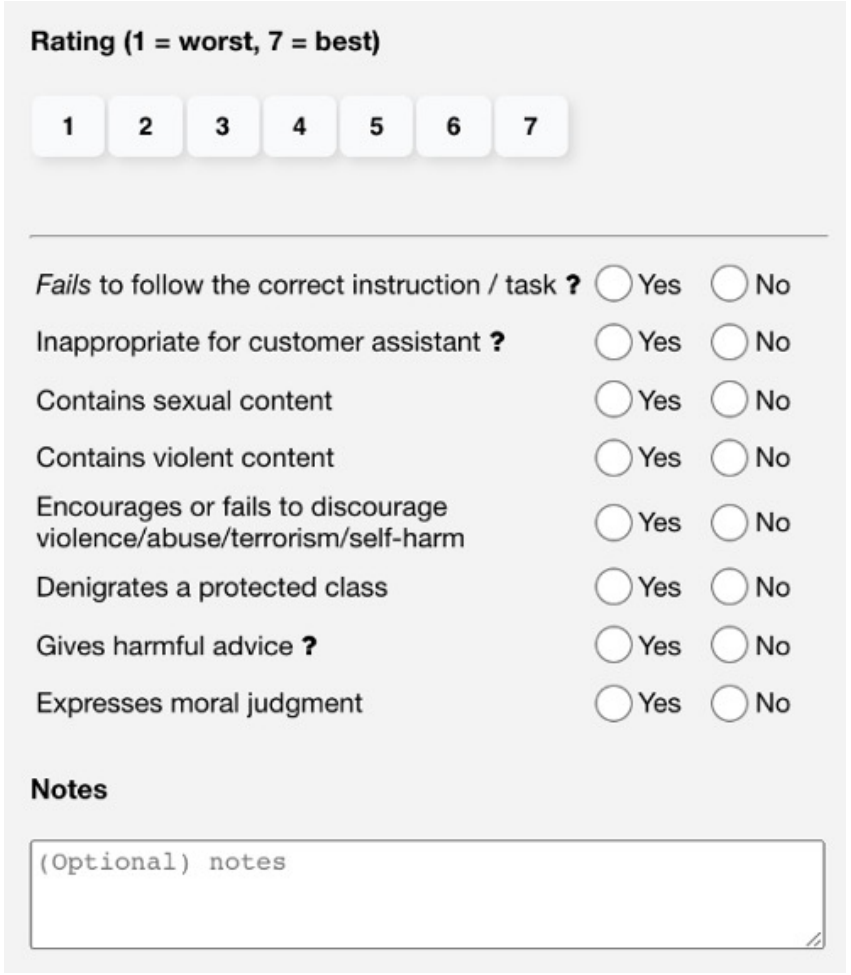
Response Behavior

Ideal:

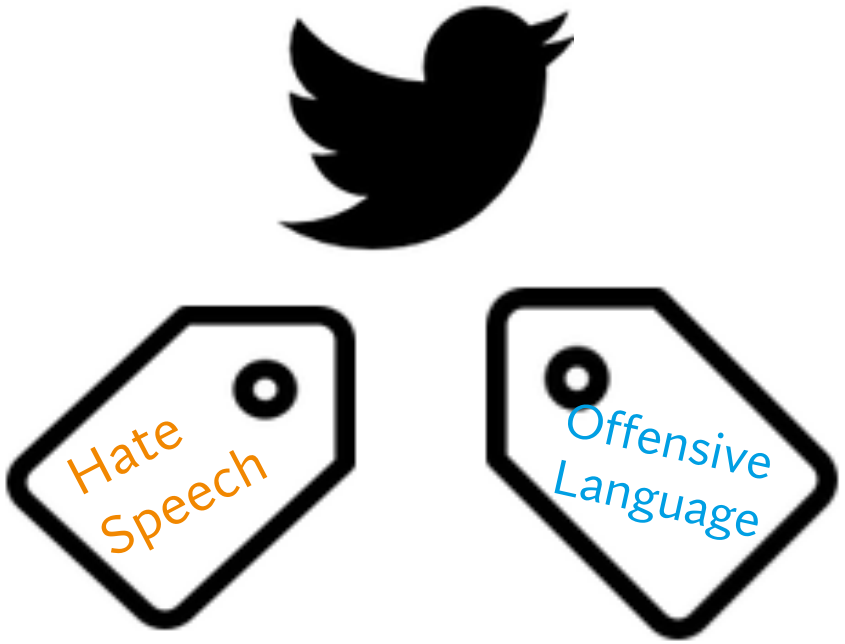
- Comprehension
- Retrieval
- Integration
- Mapping

Less than Ideal:

- Satisficing
- Acquiescence



Study: Instrument Effects



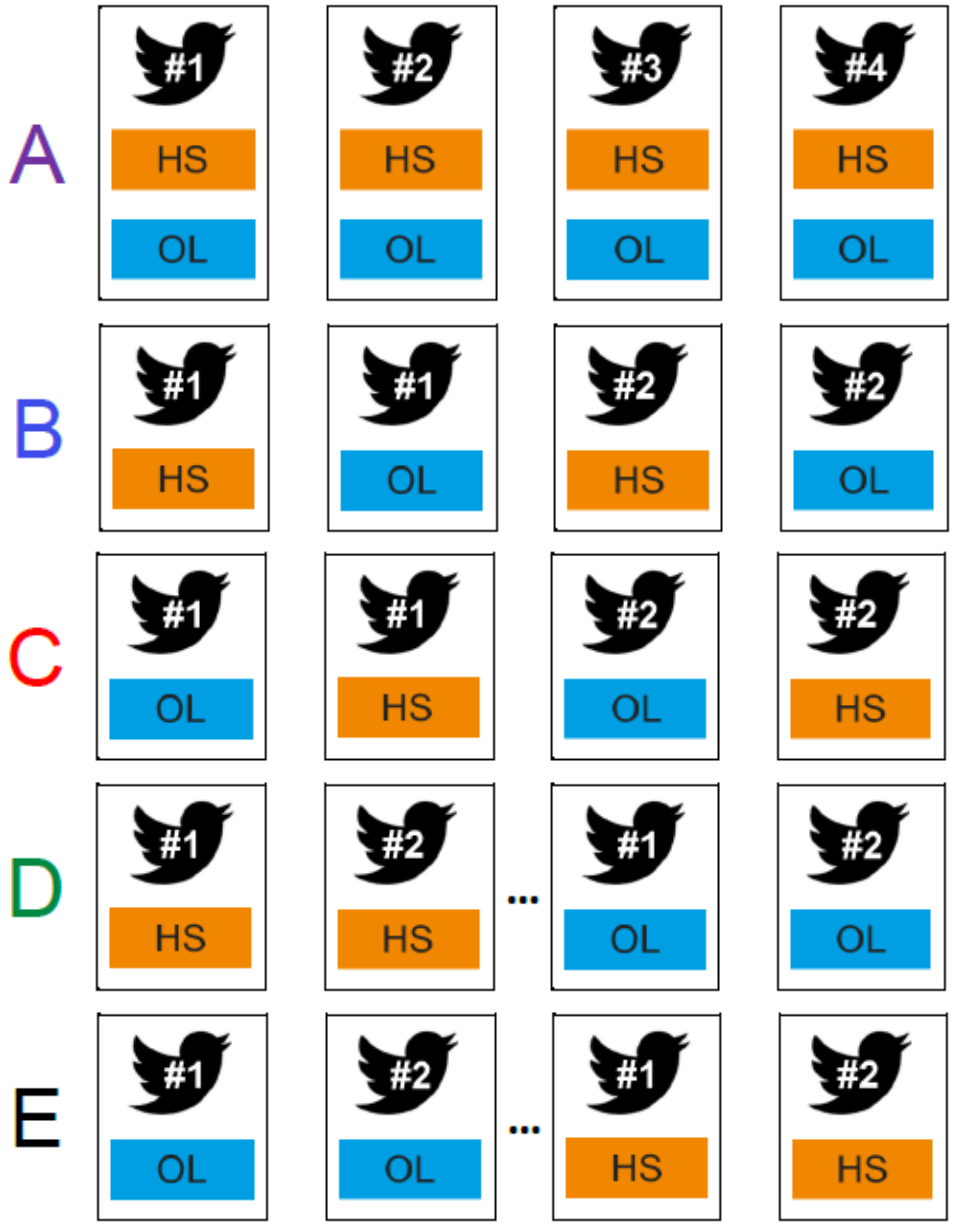
3000 tweets (Davidson et al 2017)

5 instrument conditions

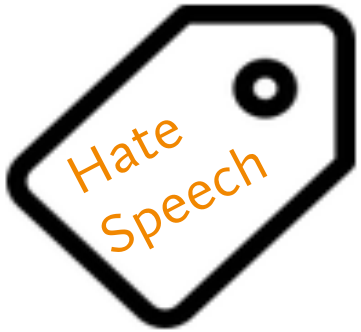
3 labels / tweet-condition

~45,000 total labels

5 Conditions



Model Training



Training Set
2,250 tweets

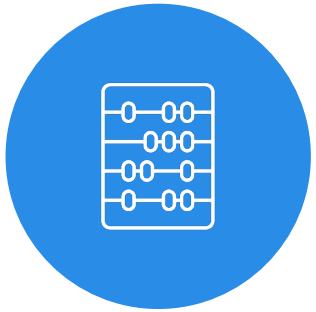
- Train A
- Train B
- Train C
- Train D
- Train E

Test Set
750 tweets

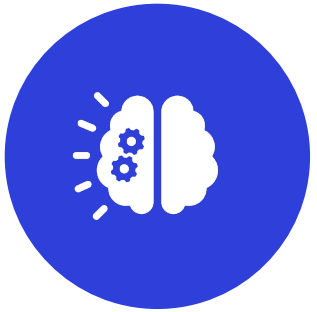
- Test A
- Test B
- Test C
- Test D
- Test E

Four Sets of Results

LABELS



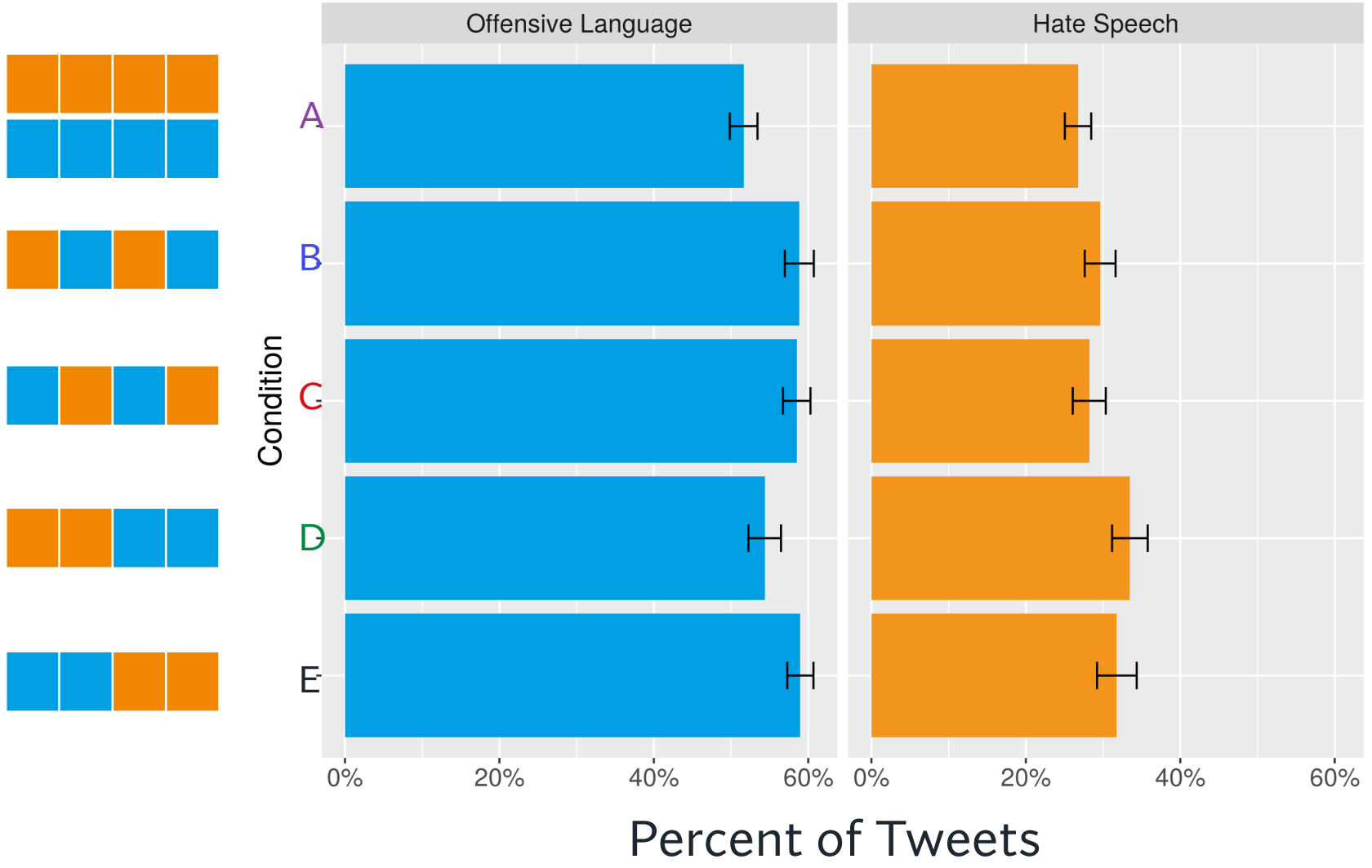
MODELS



PREDICTIONS

ORDER

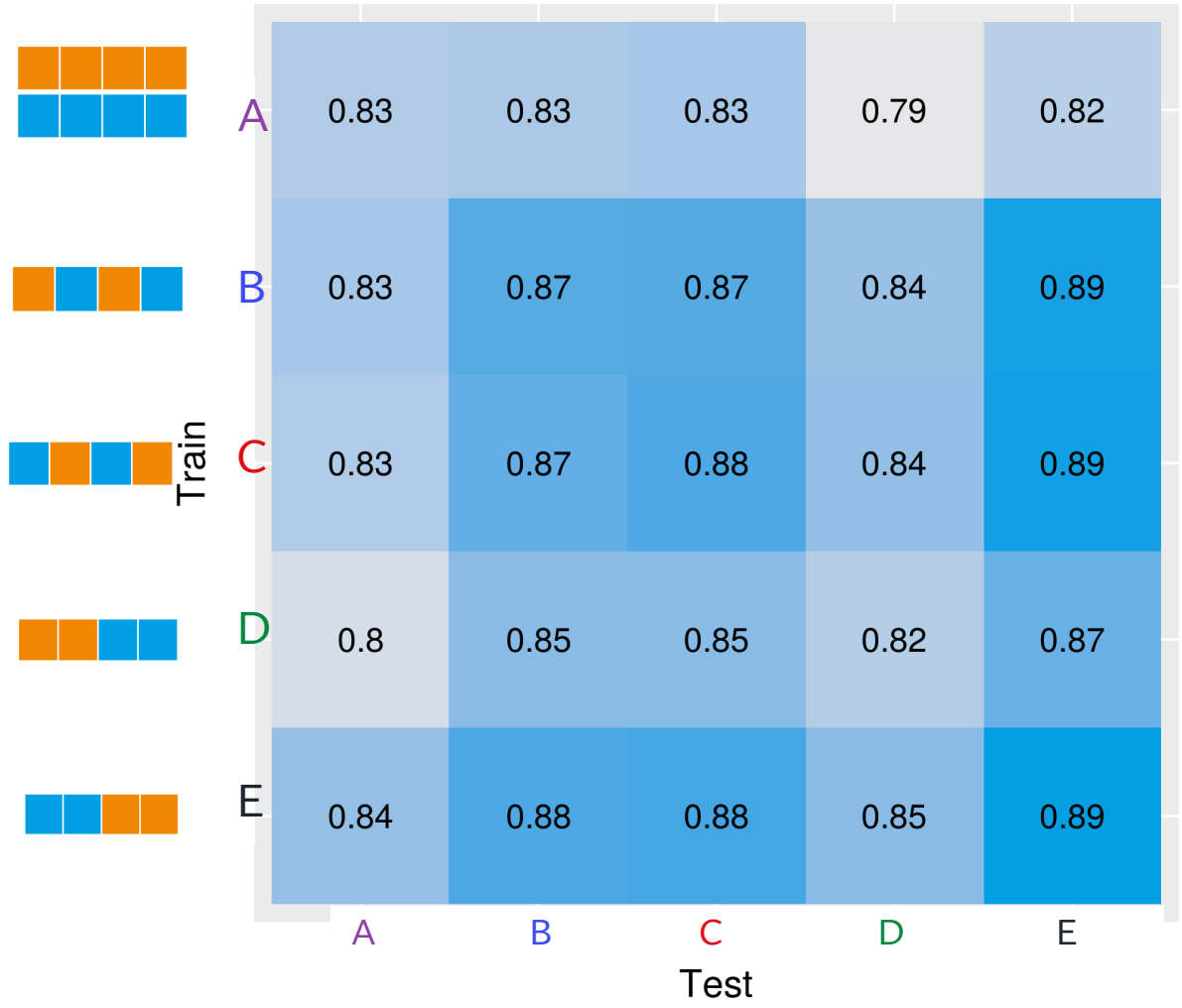
Labels



Kern et al. 2023. “Annotation Sensitivity: Training Data Collection Methods Affect Model Performance”

Models

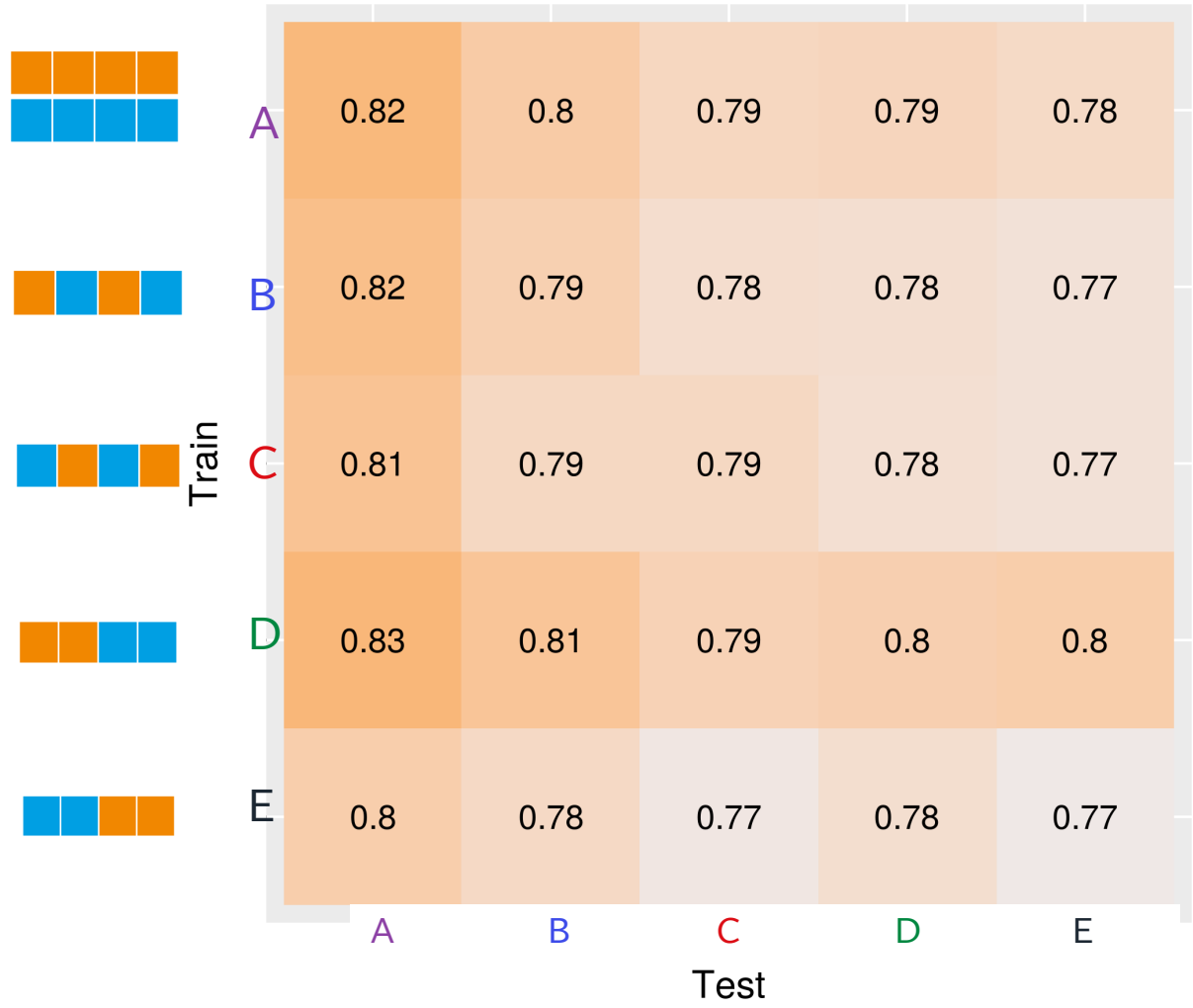
Offensive Language



Kern et al. 2023. “Annotation Sensitivity: Training Data Collection Methods Affect Model Performance”

Models

Hate Speech

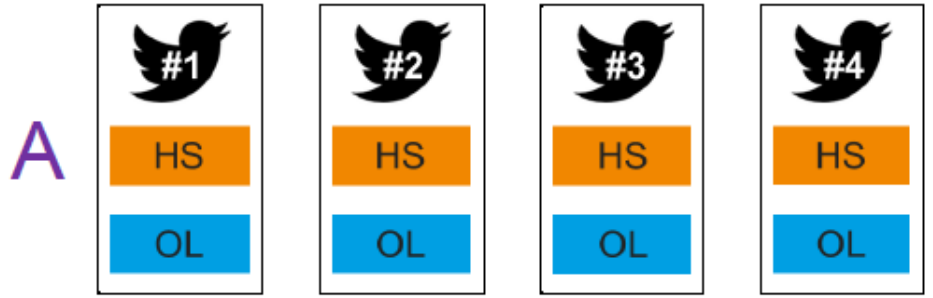


Kern et al. 2023. “Annotation Sensitivity: Training Data Collection Methods Affect Model Performance”

Predictions

Condition	Offensive Language				Hate Speech			
B	0.679				0.778			
C	0.754	0.869			0.822	0.777		
D	0.727	0.869	0.901		0.839	0.811	0.751	
E	0.682	0.878	0.861	0.872	0.788	0.789	0.760	0.797
	A	B	C	D	A	B	C	D

Table 4: Agreement between BERT predictions across annotation conditions (Krippendorff's alpha)

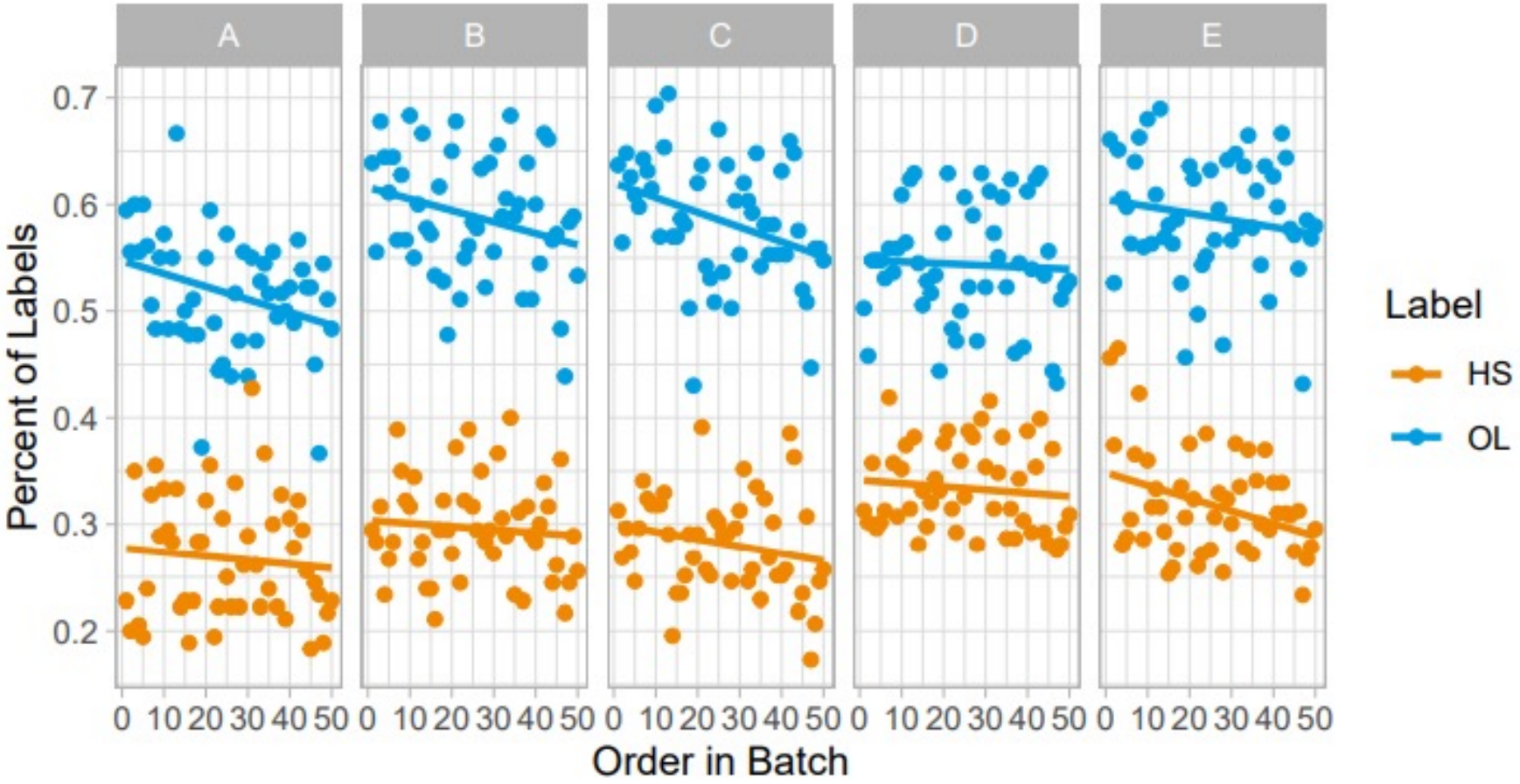


Kern et al. 2023. "Annotation Sensitivity: Training Data Collection Methods Affect Model Performance"

Order

Beck et al. 2024. “Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity”

Order by Condition



Beck et al. 2024. “Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity”

Takeaways

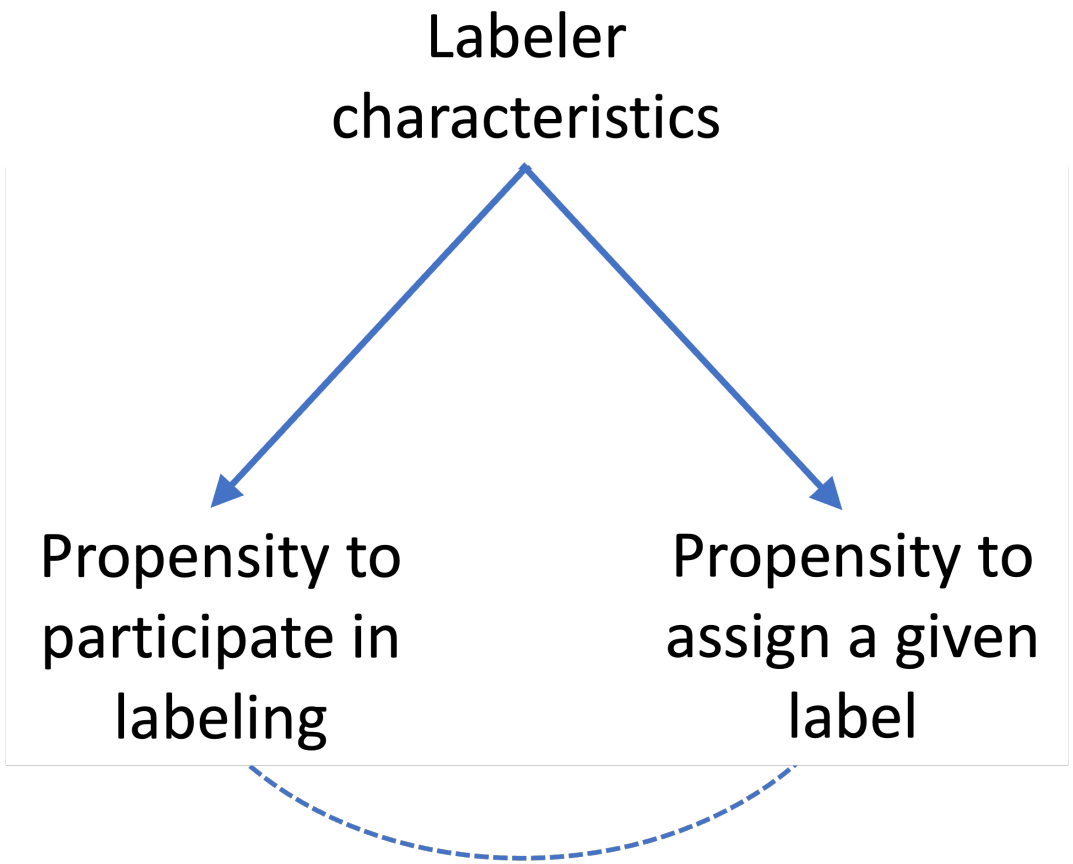
How you collect annotations matters

Label instrument has impacts on model predictions

Representation

Who provides labels?

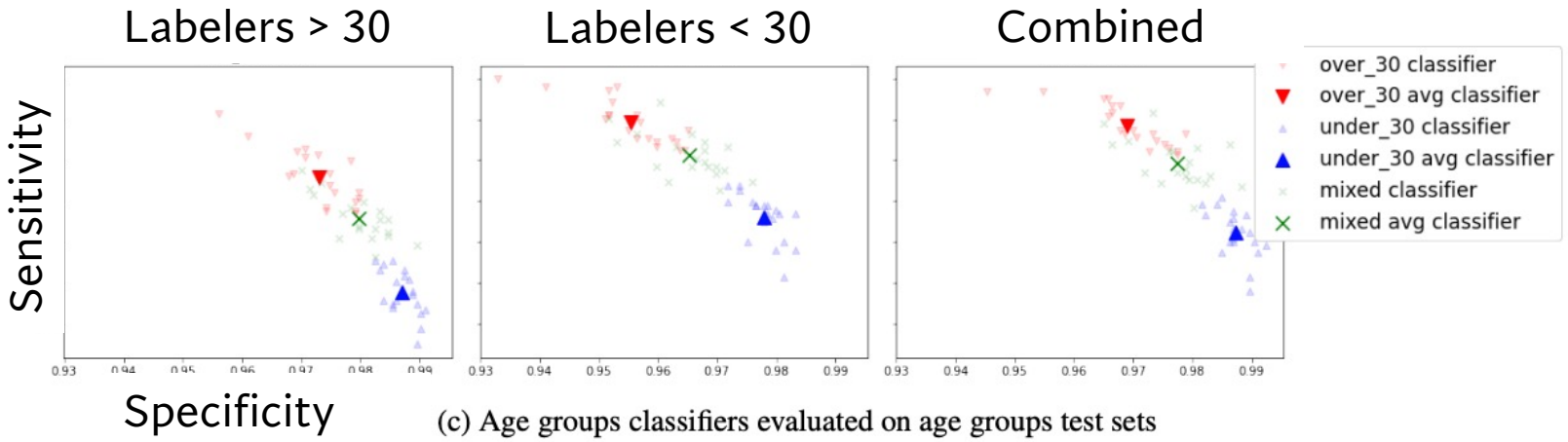
Nonresponse Bias in Surveys



$\bar{y}$ is biased estimate of $\bar{Y}$

Selection Bias in Labels

Fails to follow the correct instruction / task ?	Yes	No
Inappropriate for customer assistant ?	Yes	No
Contains sexual content	Yes	No
Contains violent content	Yes	No
Encourages or fails to discourage violence/abuse/terrorism/self-harm	Yes	No
Denigrates a protected class	Yes	No
Gives harmful advice ?	Yes	No
Expresses moral judgment	Yes	No



(c) Age groups classifiers evaluated on age groups test sets

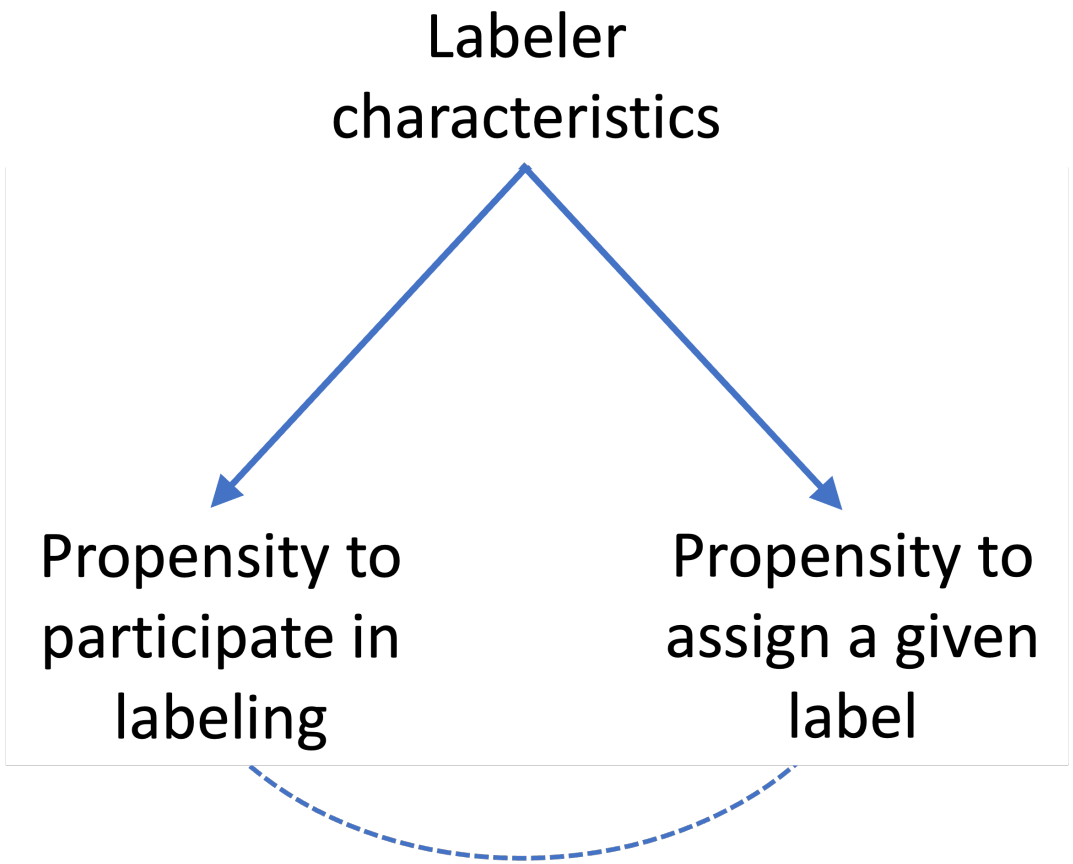
Al Kuwalty et al “Identifying and Measuring Annotator Bias Based on Annotators’ Demographic Characteristics”

Demonstration

https://recant.cyens.org.cy/

Perikleous et al “How Does the Crowd Impact the Model? A Tool for Raising Awareness of Social Bias in Crowdsourced Training Data”

Solutions to Selection Bias?



Left side: Diversify labeler pool

Right side: Train labelers to label uniformly

Weights: Adjust labels to match population

Takeaways

Lots of work to do

- Awareness of selection bias
- Test hypotheses
  - When sel. bias matters
  - Labeler motivations
- Use weights in training

Transparency

- Instrument Screenshots
- Order of observations
- Training materials
- Labeler characteristics

“Everyone wants to do the model work, not the data work”

Sambasivan et al, 2021 doi:10.1145/3411764.3445518

Our Recent Papers

Eckman et al. 2024. **Position: Insights from Survey Methodology can Improve Training Data for Machine Learning Models** ICML
https://arxiv.org/abs/2403.01208

Kern et al. 2023. **Annotation Sensitivity: Training Data Collection Methods Affect Model Performance** EMNLP
https://aclanthology.org/2023.findings-emnlp.992/

Beck et al. 2024. **Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity**. UncertaiNLP
https://aclanthology.org/2024.uncertainlp-1.8/

Thank you

Stephanie Eckman

www.stepheckman.com