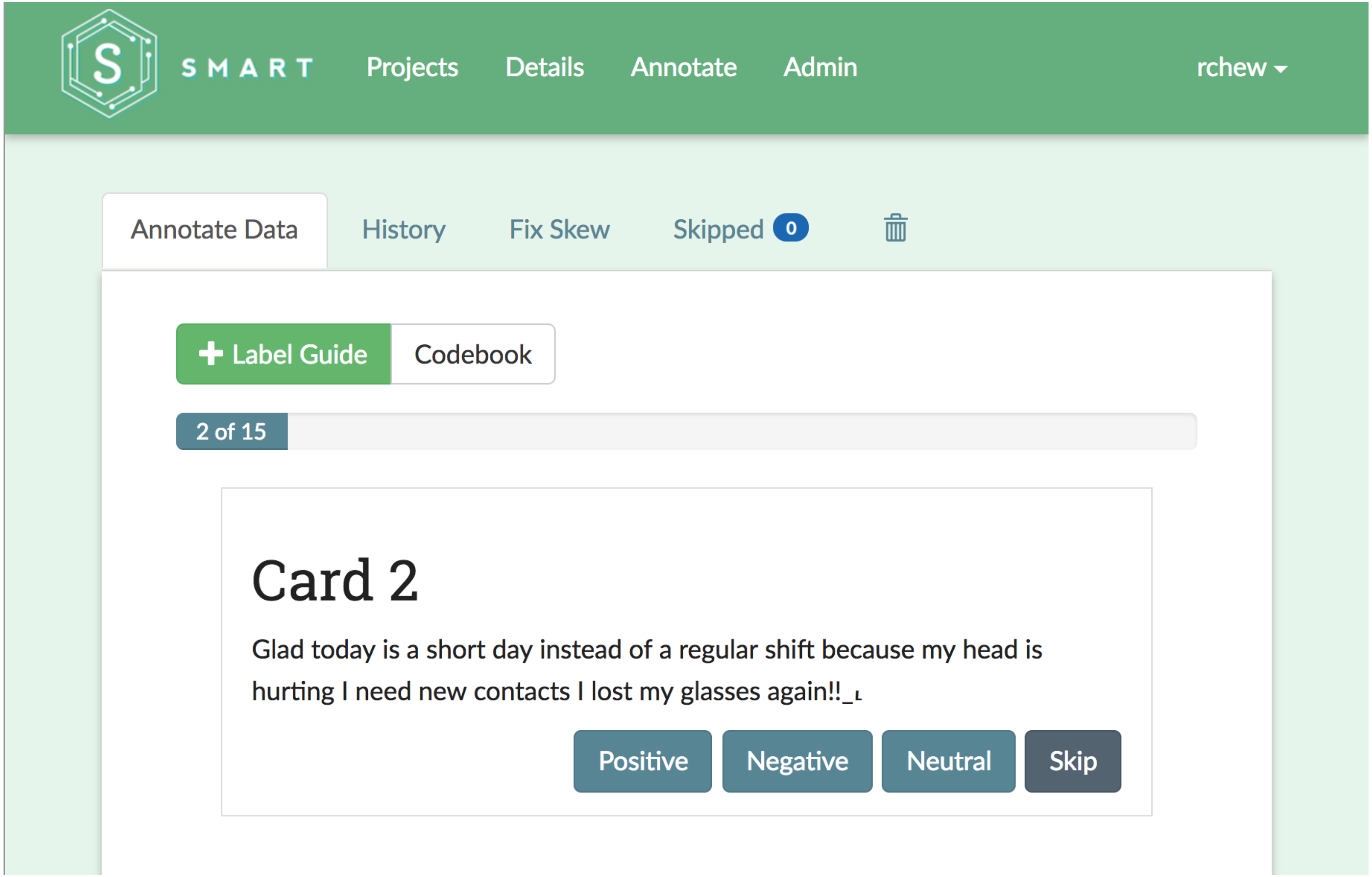
Improving Label Collection Through Social Science Insights: Preliminary Results and Research Agenda

Stephanie Eckman, RTI

Jacob Beck, LMU

Rob Chew, RTI

Frauke Kreuter, LMU, JPSM



https://rtiinternational.github.io/SMART/

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

Sambasivan et al, 2021 doi:10.1145/3411764.3445518

Relevant Literature

Machine Learning

- Annotator effects
- Annotator characteristics

Social Psychology

- Contrast and assimilation effects

Survey Methodology

- Question wording & response options
- Question order
- Interviewer effects

Data Collection

- Label 20 tweets
  - Davidson et al: “Automated Hate Speech Detection and the Problem of Offensive Language”
- Labels:
  - Hate speech
  - Offensive language
  - Neither
- 1007 annotators from Prolific

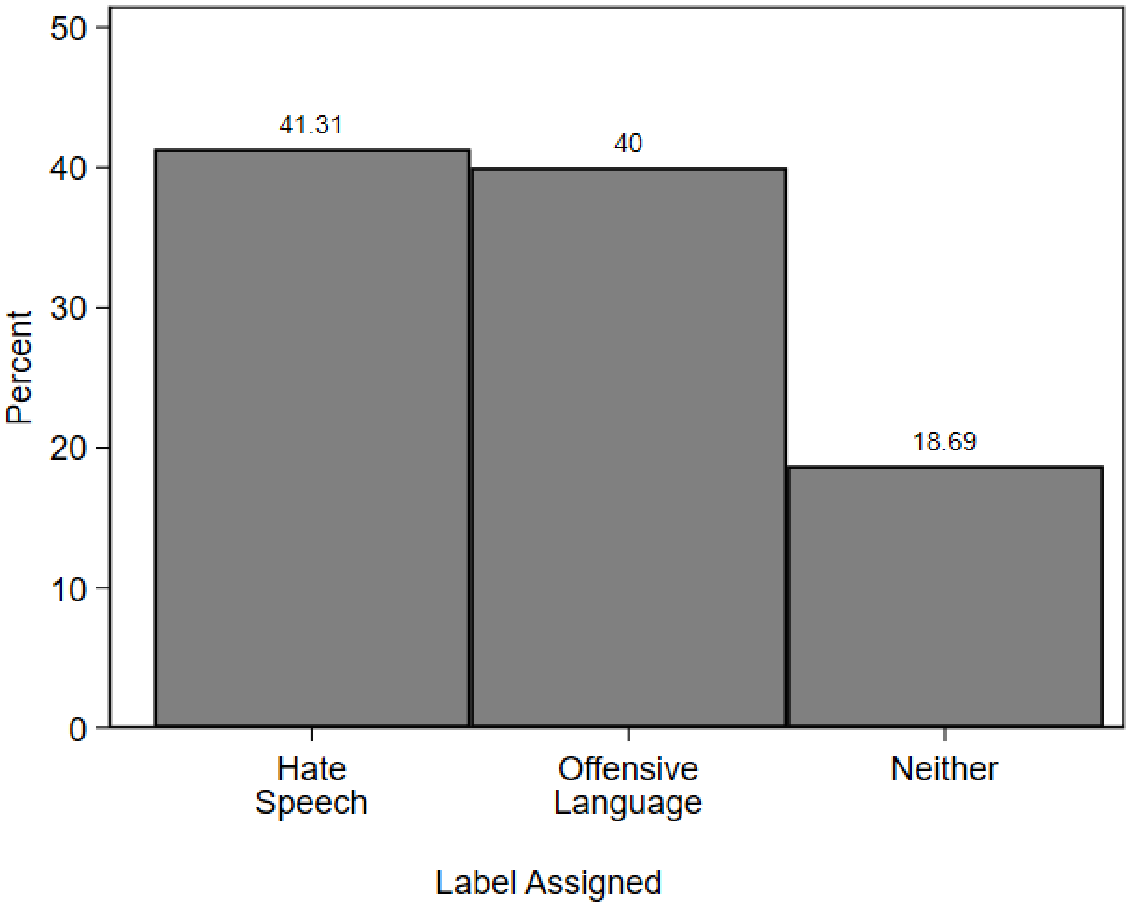
- **1,007 labels**
- **of 20 tweets**
- **Annotator characteristics**

- **Varied 2 factors:**
  - **3 wordings**
  - **2 response options**

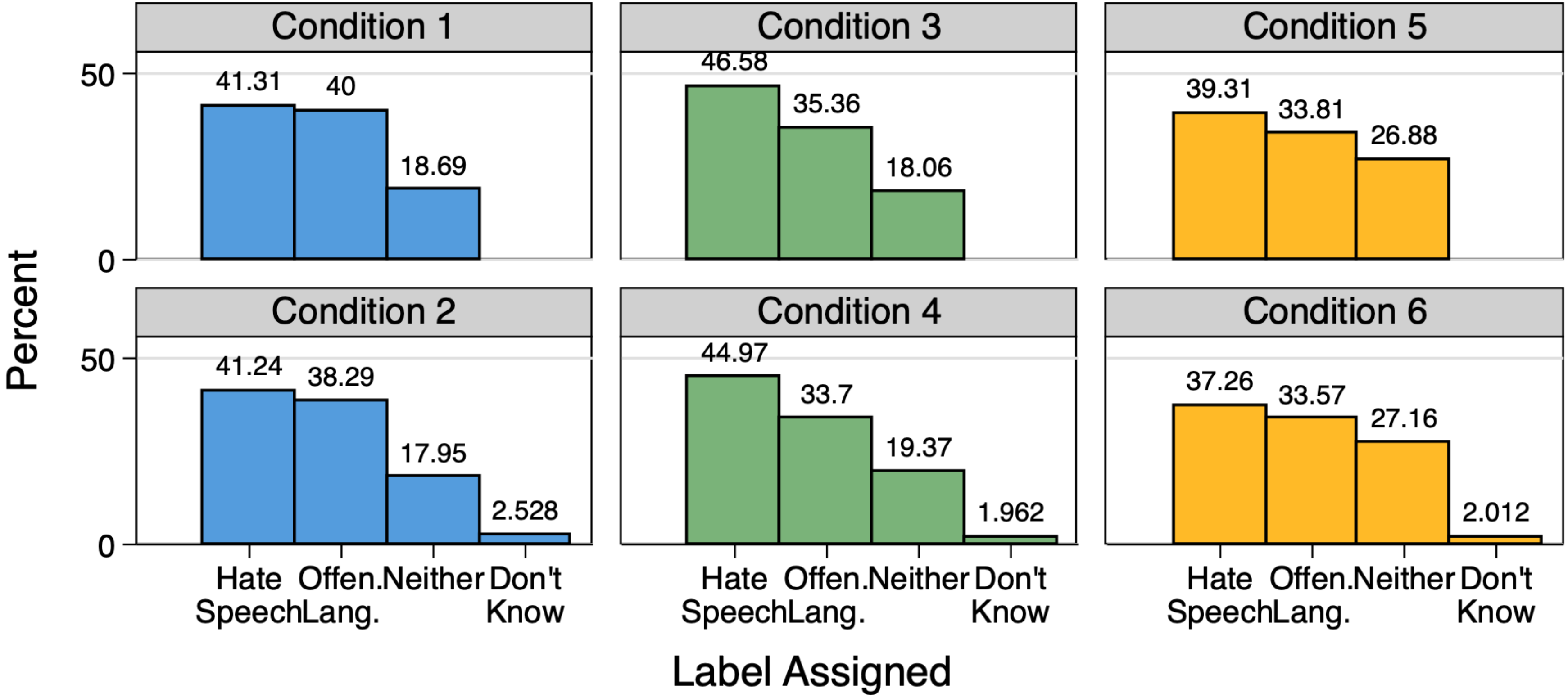
6 Task Structure Conditions

Condition 1: **1 item** Click the category that best applies: - Hate speech / Offensive language / Neither	***Condition 3:*** **2 items, HS first** Does this tweet contain hate speech? - yes/no **If no:** Does this tweet contain offensive language? - yes/no	***Condition 5:*** **2 items, OL first** Does this tweet contain offensive language? - yes/no **If yes:** Does this tweet contain hate speech? - yes/no
Condition 2: Same, with DK response option	***Condition 4:*** Same, with DK response option	***Condition 6:*** Same, with DK response option

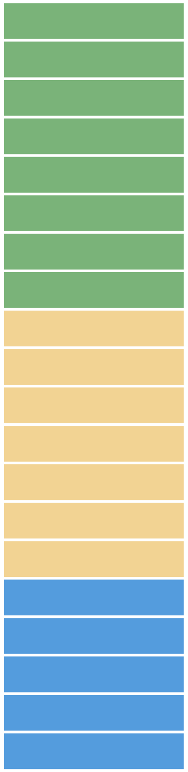
Results: Condition 1



Results: Conditions Differ



Some Evidence for Order Effects



Less hateful

Middle tweets

More hateful

Order	Tweet	% labelled hate speech
1	Less hateful	
2	Middle tweets	51%

Order	Tweet	% labelled hate speech
1	More hateful	
2	Middle tweets	33%

Annotator Effects

Annotators explain 3% of variability in labels

- Models learn annotators' quirks
- More annotators labelling fewer tweets preferred

Implications & Next Steps

- Task Structure matters
  - Transparency in label collection
- Order matters
  - Purposeful ordering may backfire
- Annotators matter
  - Carefully select annotators & collect annotator characteristics
  - Watch out for predatory inclusion

Next steps:
- **More experiments**
- **Impact on models**

“Everyone want
the model
da

Sambasivan et al, 20 10.1145/3411764.3445518

We want to do
the Data Work

Stephanie Eckman
Fellow, RTI International

@stephnie
stepheckman.com

Extended abstract:
https://osf.io/hqj43/

Condition 1

Click the category that best applies

?

At this rate, I'd cheer for the awful New York Yankees over the St. Louis Cardinals.

hate speech	offensive language	neither
○	○	○

Condition 3

Does this tweet contain offensive language?

At this rate, I'd cheer for the awful New York Yankees over the St. Louis Cardinals.

?

Yes No

Does this tweet contain hate speech?

At this rate, I'd cheer for the awful New York Yankees over the St. Louis Cardinals.

?

Yes No

Condition 5

Does this tweet contain hate speech? ?

At this rate, I'd cheer for the awful New York Yankees over the St. Louis Cardinals.

Yes No

Does this tweet contain offensive language? ?

At this rate, I'd cheer for the awful New York Yankees over the St. Louis Cardinals.

Yes No