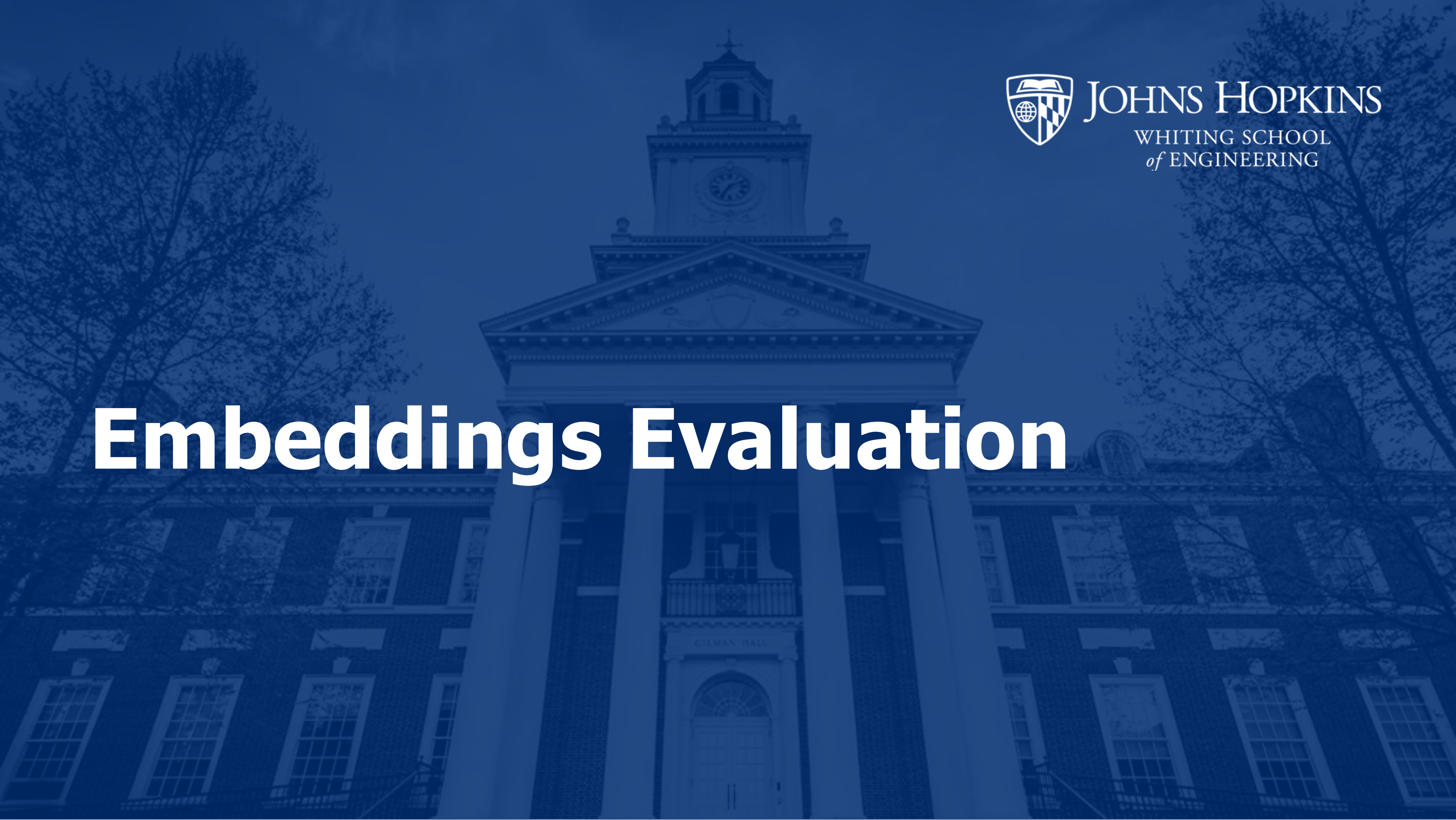
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Data Labeling

Announcements

- HW 1 due today
- HW 2 released today or tomorrow



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Embeddings Evaluation

Evaluation

- We’re using embeddings for analyzing data sets
- How do we know that the embeddings we trained are meaningful?
- How much do decisions like embedding model (word2vec-CBOW, word2vec-skipgram, fasttext), similarity metric, or seed words (man/woman) matter?

Evaluation: Intrinsic Metrics of Embedding Quality

- Test performance on similarity; correlation between an algorithm’s word similarity scores and word similarity ratings assigned by humans
  - WordSim-353 (Finkelstein et al., 2002): is ratings from 0 to 10 for 353 noun pairs; for example (plane, car) had an average score of 5.77.
  - SimLex-999 (Hill et al., 2015): more difficult dataset that quantifies similarity (cup, mug) rather than relatedness (cup, coffee), and including both concrete and abstract adjective, noun and verb pairs
  - TOEFL dataset (Landauer and Dumais, 1997): 80 questions, each consisting of a target word with 4 additional word choices; the task is to choose which is the correct synonym
- Data sets that incorporate context, such as sentence-level similarity (Huang et al., 2012; Pilehvar and Camacho-Collados, 2019)
- Analogy tasks (Turney and Littman, 2005)

Evaluation: Extrinsic Metrics of Embedding Quality

- Performance on downstream task when using embeddings in an NLP model
  - Useful for NLP models, less obviously indicative of analysis reliability

- Comparisons with external data
  - Occupation statistics from the census
  - Crowd-sourced annotations of stereotypes (note that we can crowd-source current stereotypes but it’s hard to crowd-source historical ones)

Evaluation: Capacity to capture social variables

- Do word embeddings reflect beliefs about people?
  - E.g. race and gender stereotypes
  - Dimension-level: how well do embeddings capture beliefs about gender relative to race?
  - Belief-level: how well do embeddings capture beliefs about potency (strength) of “children” vs “thugs”?

Methods

- Collect survey data from Amazon Mechanical Turk
  - Limiting assumption, how do we know if the survey data is reliable?

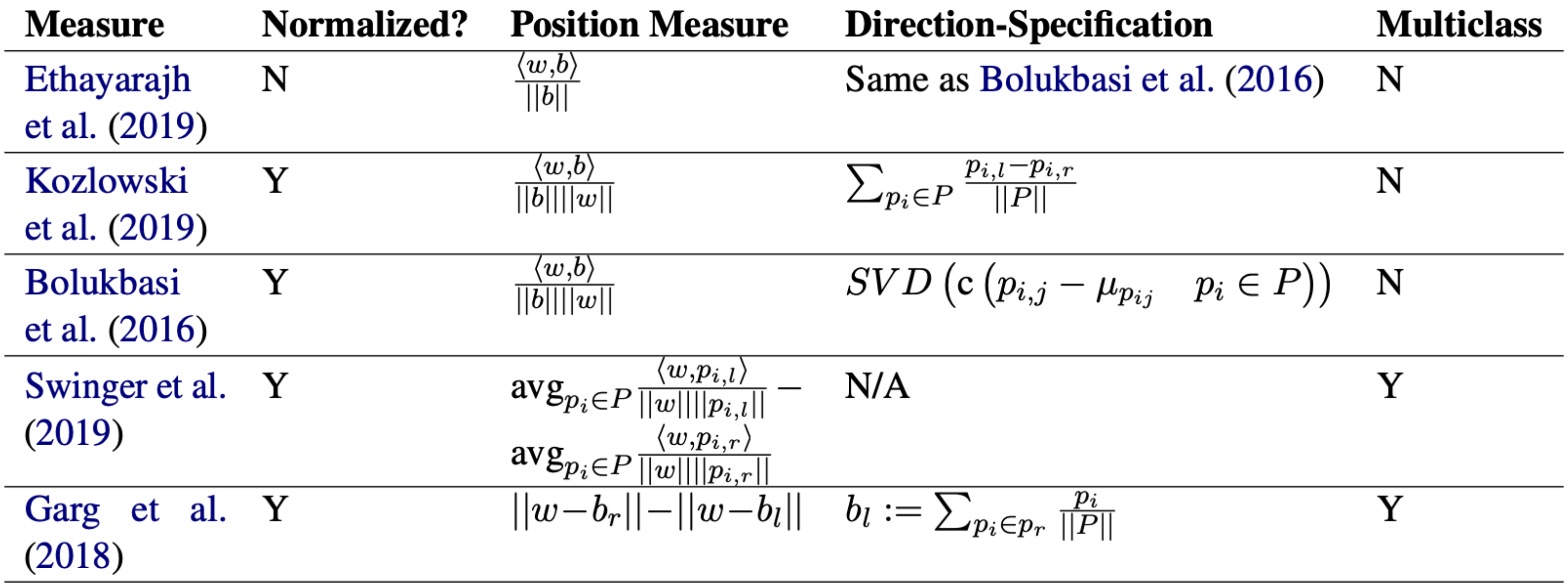
Joseph, Kenneth, and Jonathan Morgan. "When do Word Embeddings Accurately Reflect Surveys on our Beliefs About People?." ACL. 2020.

Evaluation: Specific Experimental Design Decisions

- Corpus/Embedding Selection
- Dimension Selection
  - Dimension-inducing word set
  - Methodology (average embeddings, PCA, etc)
- Word Position Measurement
  - E.g. projection, vector similarity metrics

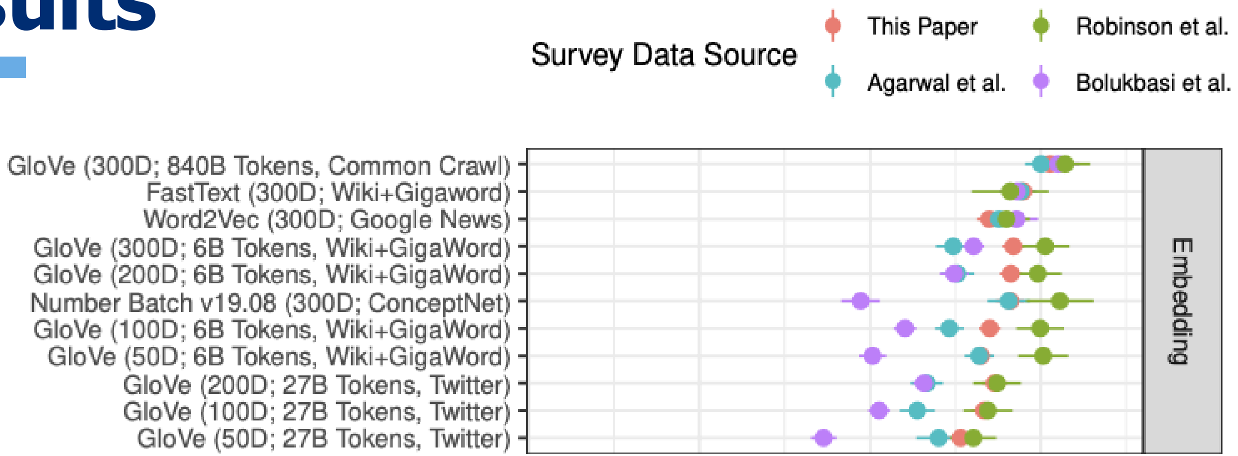
What approaches work best? How much do these choices matter?

Design Choices



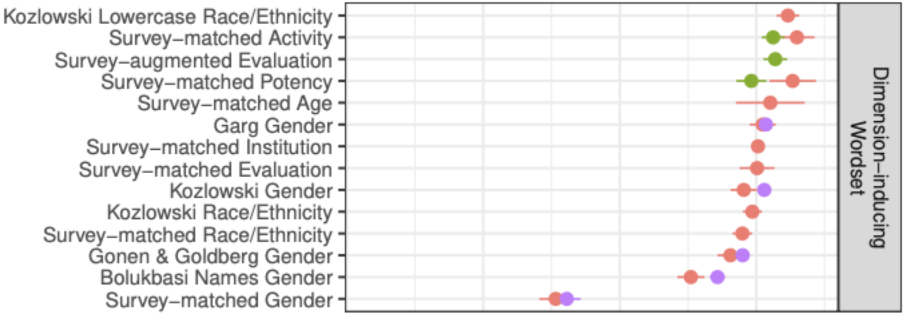
Measure	Normalized?	Position Measure	Direction-Specification	Multiclass
Ethayarajh et al. (2019)	N	$\frac{\langle w,b \rangle}{\lvert\lvert b \rvert\rvert}$	Same as Bolukbasi et al. (2016)	N
Kozlowski et al. (2019)	Y	$\frac{\langle w,b \rangle}{\lvert\lvert b \rvert\rvert \lvert\lvert w \rvert\rvert}$	$\sum_{p_i \in P} \frac{p_{i,l} - p_{i,r}}{\lvert\lvert P \rvert\rvert}$	N
Bolukbasi et al. (2016)	Y	$\frac{\langle w,b \rangle}{\lvert\lvert b \rvert\rvert \lvert\lvert w \rvert\rvert}$	$SVD\left(\mathrm{c}\left(p_{i,j} - \mu_{p_{ij}} \quad p_i \in P\right)\right)$	N
Swinger et al. (2019)	Y	$\mathrm{avg}_{p_i \in P} \frac{\langle w,p_{i,l} \rangle}{\lvert\lvert w \rvert\rvert \lvert\lvert p_{i,l} \rvert\rvert} -$ $\mathrm{avg}_{p_i \in P} \frac{\langle w,p_{i,r} \rangle}{\lvert\lvert w \rvert\rvert \lvert\lvert p_{i,r} \rvert\rvert}$	N/A	Y
Garg et al. (2018)	Y	$\lvert\lvert w - b_r \rvert\rvert - \lvert\lvert w - b_l \rvert\rvert$	$b_l := \sum_{p_i \in p_r} \frac{p_i}{\lvert\lvert P \rvert\rvert}$	Y

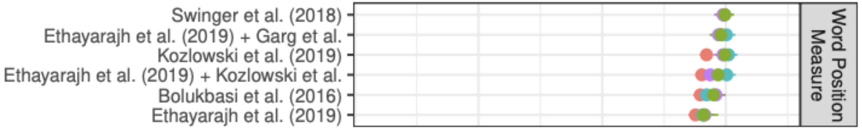
Results



- [Generally embedding results do correlate with survey results]
- Selection of embedding model can decrease correlation with survey results
- Less variation for 300D embeddings
- No embedding model is universally the best

Results



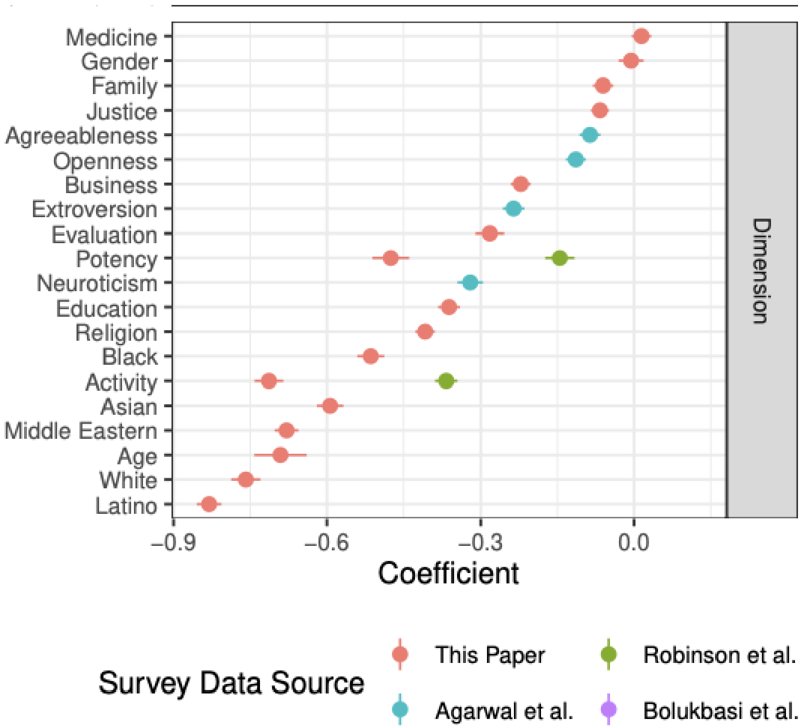


- Selection of dimension-inducing words doesn’t really matter (though you could make a particularly bad choice) [Note that other work has found more variance]
- Choice of position measure (e.g. similarity metric) has almost no effect

Antoniak, Maria, and David Mimno. "Bad seeds: Evaluating lexical methods for bias measurement." ACL 2021.

Results

- Correlations for some dimensions (e.g. gender) are much stronger than for others (e.g. race)!



Recap

- Example applications:
  - Measuring bias (gender bias / occupational stereotypes)
  - Measuring change in word meanings over time
  - Measuring stereotypes over time
- Embedding manipulation:
  - Cosine similarity, Euclidean distance
  - Gender subspace
  - Averaging keywords
- Evaluations:
  - Analogy tasks, similarity benchmarks, extrinsic metrics
  - Comparisons with hand-curated analyses or benchmarks
  - Comparisons with survey or crowd-sourced data

This class: Data annotating

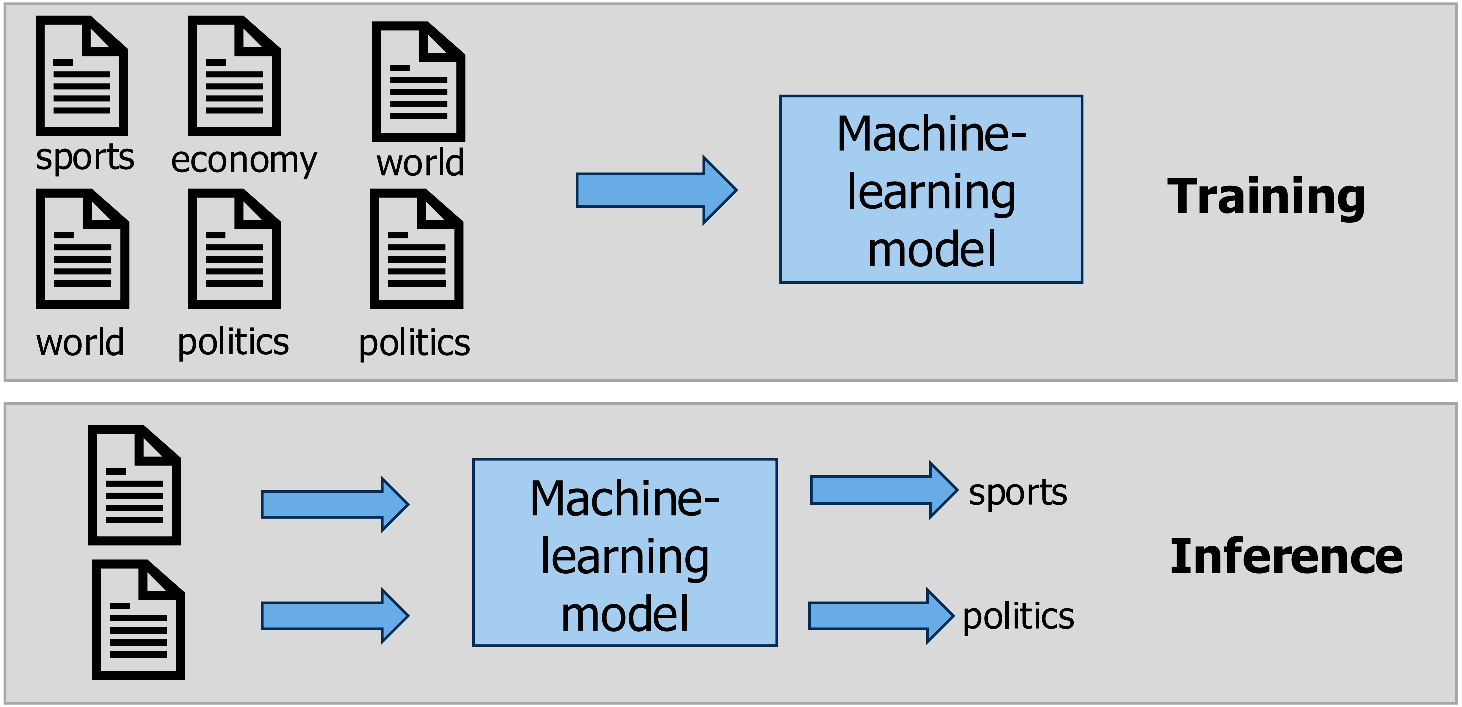
- Motivation
- Tips and tricks for components of annotation process
- Annotator agreement metrics

Background and Motivation

Methods of Data analysis

- We want to know if (and when and how) Republicans talk about taxes more than Democrats:
  1. We use word statistics to find if words like “taxes” and “spending” are more common in Republican speeches
  2. We can train a topic model, identify the tax-related topics and determine if that topic is more common in Republican vs. Democratic speech (or incorporate party affiliation as co-variate in STM)
  3. We could go through every speech by hand:
     - Label if each speech or sentence or word is related to taxes
     - Count if we labeled more Republican speech than Democratic speech
  4. We can automate #3 using machine learning

Supervised learning



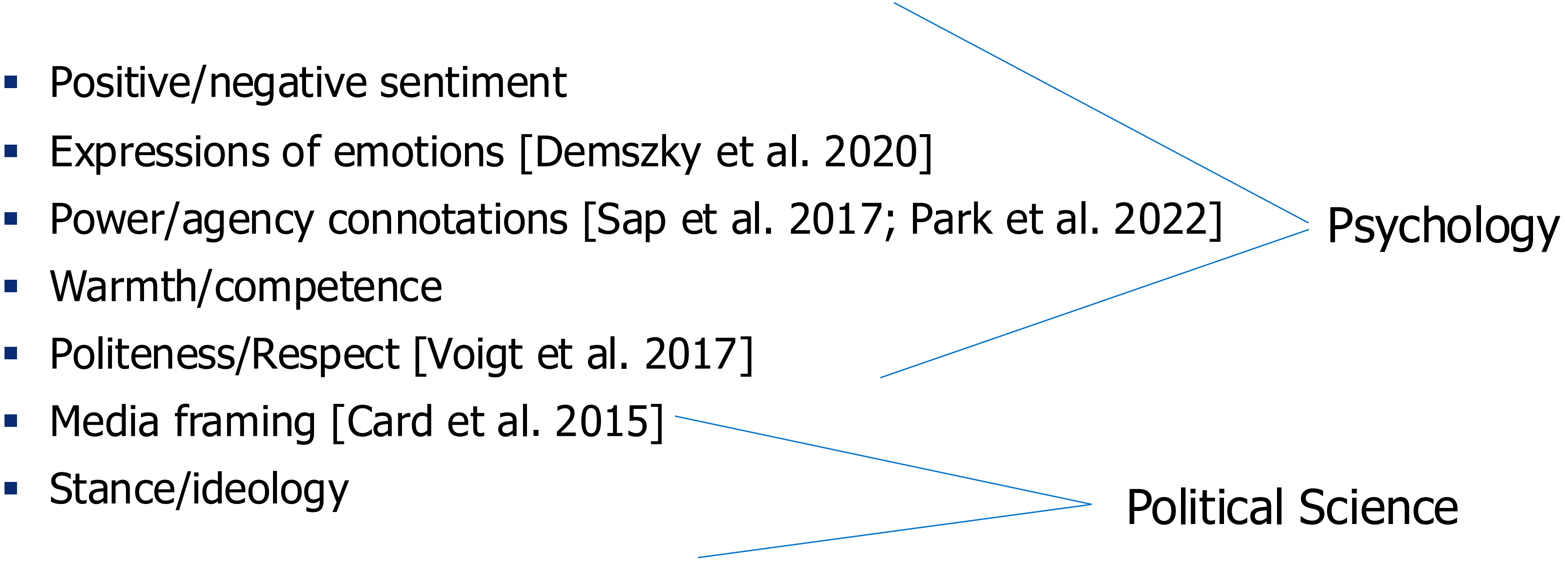
Why annotate data?

- Train machine learning models
  - [Allows us to analyze more data than we can annotated by hand]

- Evaluate machine learning models

- Direct analysis of annotations

Social-oriented data annotations tend to be particularly subjective



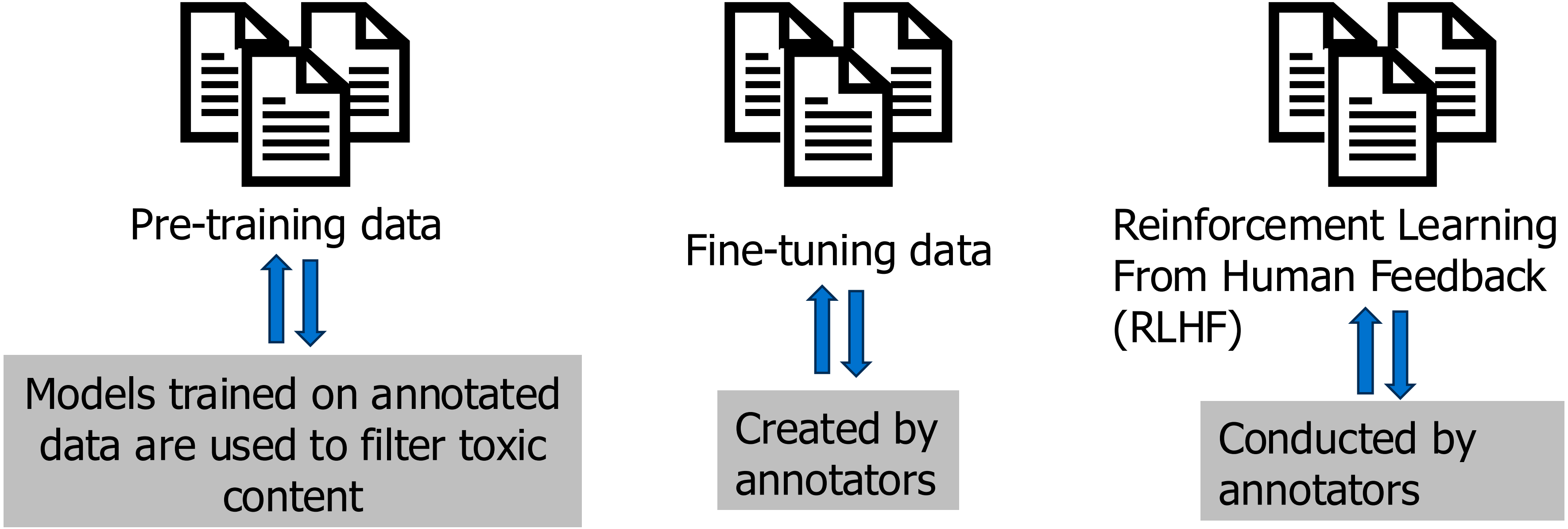
- Positive/negative sentiment
- Expressions of emotions [Demszky et al. 2020]
- Power/agency connotations [Sap et al. 2017; Park et al. 2022]
- Warmth/competence
- Politeness/Respect [Voigt et al. 2017]
- Media framing [Card et al. 2015]
- Stance/ideology

Psychology

Political Science

Can’t GPT-N code my data for me?

- Sometimes (more on this later), but how was GPT-N built?



Pre-training data

Models trained on annotated data are used to filter toxic content

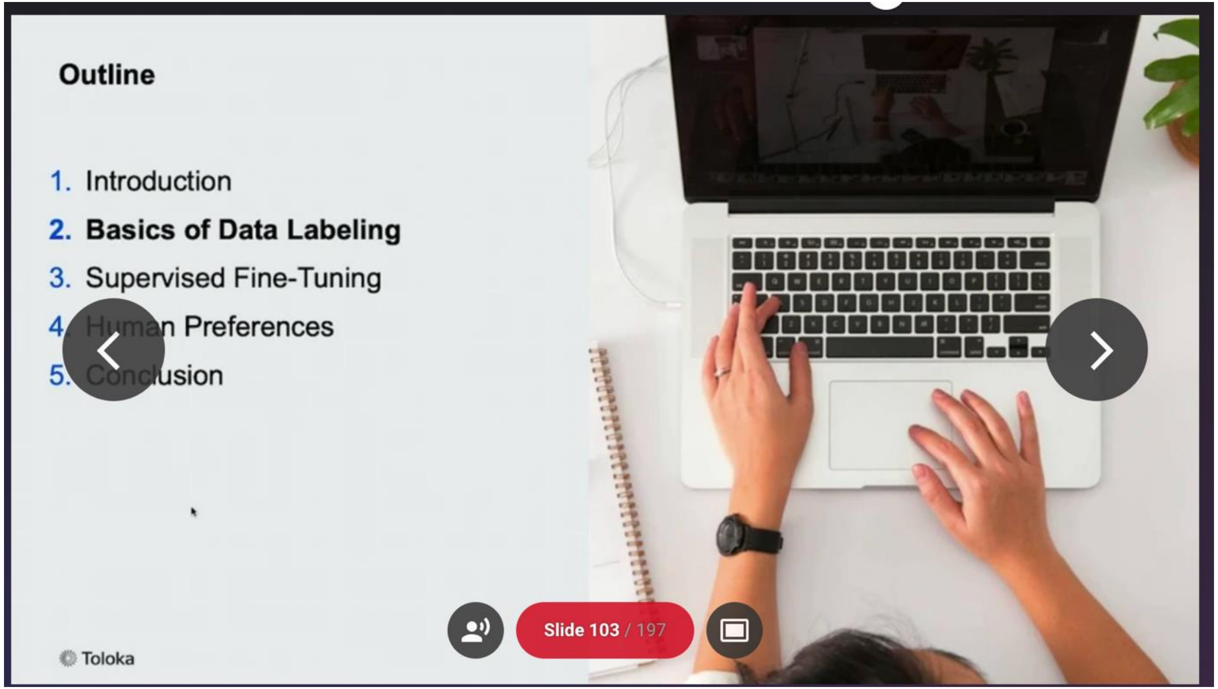
Fine-tuning data

Created by annotators

Reinforcement Learning From Human Feedback (RLHF)

Conducted by annotators

ICLR 2023 Tutorial on RLHF



Outline

1. Introduction
2. **Basics of Data Labeling**
3. Supervised Fine-Tuning
4. Human Preferences
5. Conclusion

Toloka

Half the tutorial was spent on data and annotating

Some Components of Data Annotation

Source data

Annotation scheme

Annotators

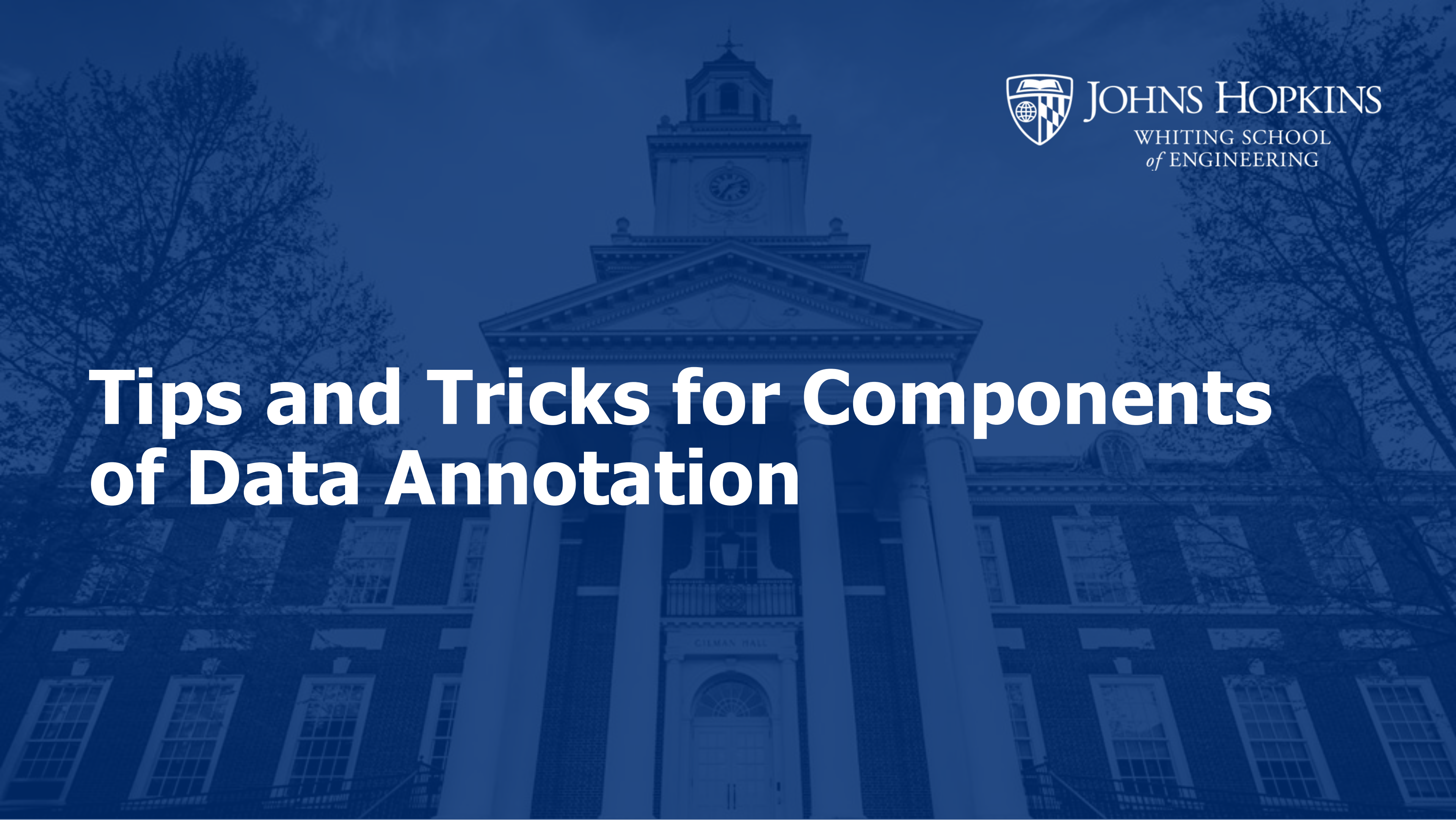
Budget

Annotation Interface

Quality Control

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Tips and Tricks for Components of Data Annotation



Running Example: Classifying hate speech or offensive language

- Goal:
  - Build a model to classify social media text as offensive or not offensive

- Use Cases:
  - Filter toxic data from model inputs
  - Filter toxic content from hosted feed
  - Social science goal: analyze what content people perceive as offensive

- Methods:
  - Collect annotated data to train and evaluate model

Choosing Data to Annotate

Source data

- Consider some questions:
  - Where will the model be used?
  - What data is representative of use cases?
  - Will models trained on Reddit data generalize to Twitter data?
  - Do we have access and appropriate permission for the ideal data?

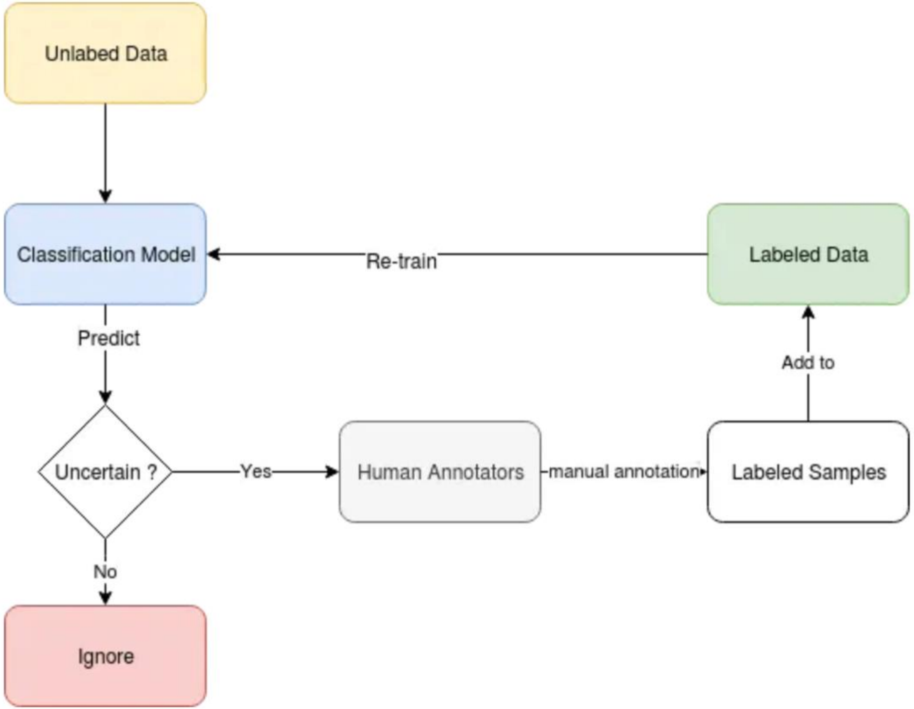
Choosing Data to Annotate

Source data

Budget

- Option 1: Randomly sample data
  - In the grand scheme of things, abusive tweets are quite rare (between 0.1% and 3%, depending on the label)” [Founta et al. 2018]
- Option 2: Pre-filtering
  - Keywords, rule-based or other “weak classifier”
  - “We choose tweets that, based on the sentiment analysis, show strong negative polarity (< −0.7) and contain at least one offensive word.” [Founta et al. 2018]
- Option 3: Active Learning

Active Learning



https://towardsdatascience.com/active-learning-for-an-efficient-data-annotation-strategy-4d007c5d7ed1 Image credit

Miller B, Linder F, Mebane WR. Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches. *Political Analysis*. 2020;28(4):532-551. doi:10.1017/pan.2020.4

Choosing Data to Annotate

Source data

Budget

- Option 1: Randomly sample data
  - In the grand scheme of things, abusive tweets are quite rare (between 0.1% and 3%, depending on the label)” [Founta et al. 2018]

Good enough in most cases

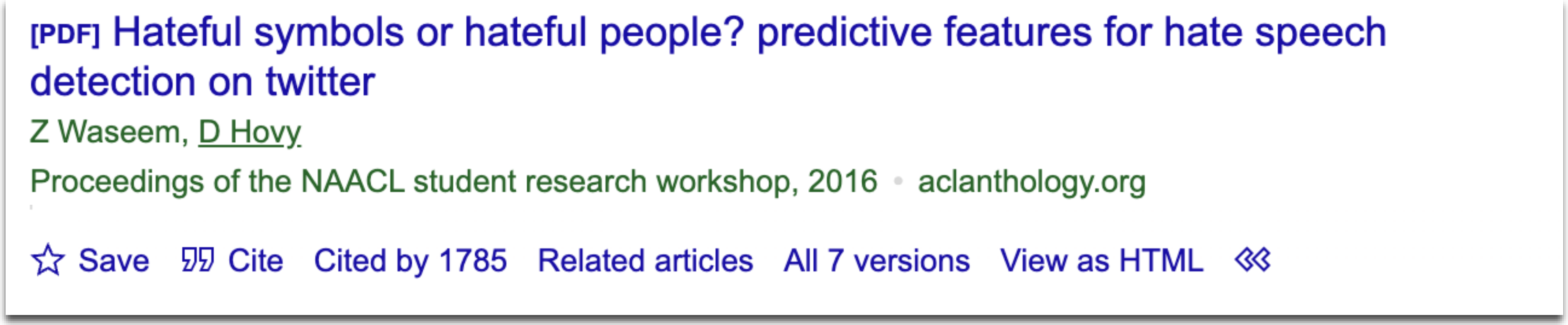
- Option 2: Pre-filtering
  - Keywords, rule-based or other “weak classifier”
  - “We choose tweets that, based on the sentiment analysis, show strong negative polarity (< −0.7) and contain at least one offensive word.” [Founta et al. 2018]

Probably most common for imbalanced data

- Option 3: Active Learning

Some research has shown promising results but isn’t that common in practice (probably performance improvements are often not worth the effort)

Example Pitfall [Perils of focused sampling]:



[PDF] Hateful symbols or hateful people? predictive features for hate speech detection on twitter

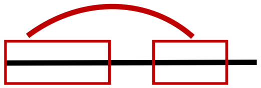
Z Waseem, D Hovy

Proceedings of the NAACL student research workshop, 2016 • aclanthology.org

- Detection of Abusive Language: the Problem of Biased Datasets (Wiegand et al., NAACL 2019)
  - 70% of the tweets annotated as sexist originate from the same two users
  - 99% of the tweets annotated as racist originate from a single user (i.e. Vile Islam).
- Can a model trained and evaluated on this data actually detect racism and sexism?
- Data can lead to wrong conclusions (e.g. that authorship information substantially improves model performance)

Takeaway: Look at your data!

Annotation Scheme: Design process

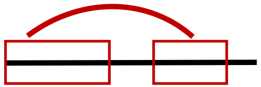


Annotation scheme

- Goal of task to be done
- Interface description
- Algorithm of required actions
- Examples of good and bad actions
- Algorithm and examples for rare cases →
- Reference materials

Toloka (ICML tutorial) suggest most failures occur here

Annotation Scheme: Design Process

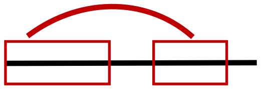


Annotation scheme

- Where do we find definitions of hate/offensive speech?
  - Where do we find categories like “racist”, “sexist”, “targeted/untargeted”?

- **Deductive coding** (top-down/prescriptive): use pre-defined scheme, for example, from existing social science literature!
  - Plutchik’s or Ekman’s emotion taxonomies
  - Affect Control Theory (Valence, Arousal, Dominance)
  - Stereotype Content Theory
- **Inductive Coding** (bottom-up/description): infer labels through multiple rounds of in-house annotations
  - E.g. Media Frames Corpus [Boystun 2014]
  - [Approach 1 can be starting point refined by approach 2]

Qualitative Data Analysis



Annotation scheme

- Inductive coding / thematic analysis

1. Prepare raw data files
2. Close reading of text: “until the evaluator is familiar with its content and gains an understanding of the themes and events covered in the text.”
3. Creation of categories: The evaluator identifies and defines categories or themes.
   - upper-level or more general categories are likely to be derived from the evaluation aims.
   - lower-level or specific categories will be derived from multiple readings of the raw data (in vivo coding), copying words or phrases from the data
4. Text may have multiple labels or no labels
5. Continuing revision and refinement of category system

Thomas, D. R. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, *27*(2), 237-246. https://doi.org/10.1177/1098214005283748

Qualitative Data Analysis

Annotation scheme

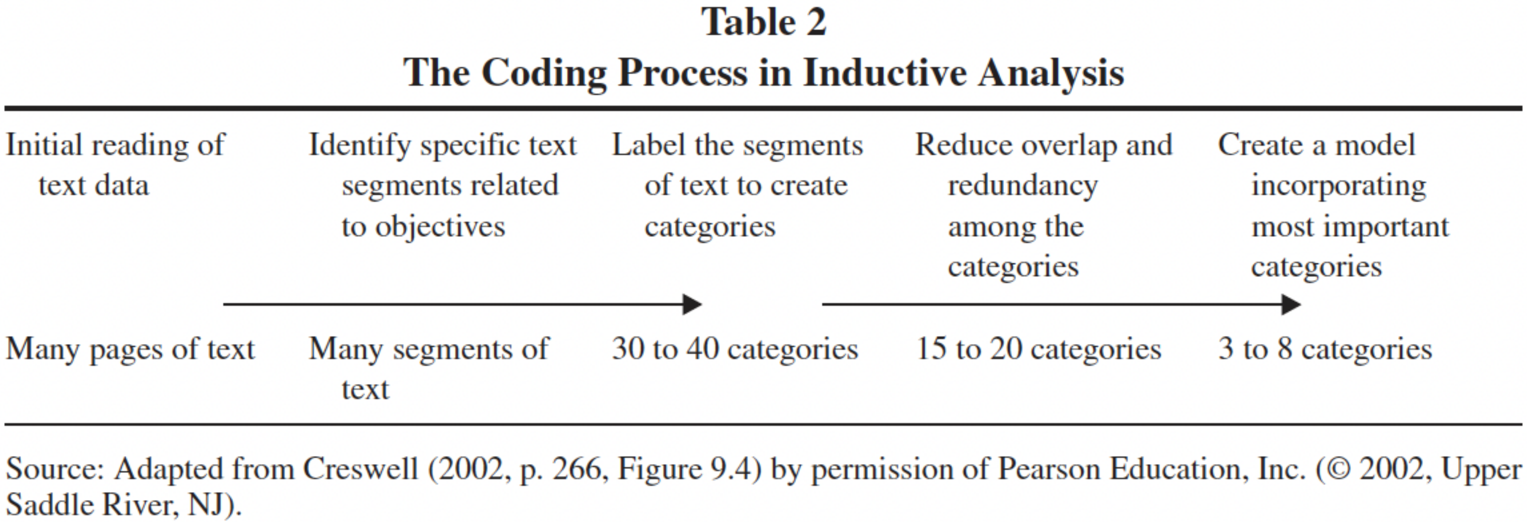


Table 2
The Coding Process in Inductive Analysis

Initial reading of text data	Identify specific text segments related to objectives	Label the segments of text to create categories	Reduce overlap and redundancy among the categories	Create a model incorporating most important categories
Many pages of text	Many segments of text	30 to 40 categories	15 to 20 categories	3 to 8 categories

Source: Adapted from Creswell (2002, p. 266, Figure 9.4) by permission of Pearson Education, Inc. (© 2002, Upper Saddle River, NJ).

Thomas, D. R. (2006). A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, *27*(2), 237-246. https://doi.org/10.1177/1098214005283748

Annotation Scheme: Design process

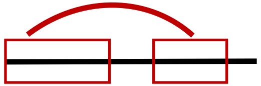
Annotation scheme

Instructions:
- Label each instance as to whether or not it contains hate speech.

It was just a joke! You're too sensitive.

- Does this instance contain hate speech?
  - ○ Yes
  - ○ No

Annotation Scheme: Design process



Instructions:
- Label each instance as to whether or not it contains hate speech.

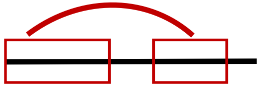
It was just a joke! You're too sensitive.

- Does this instance contain hate speech?
  - ○ Yes
  - ○ No

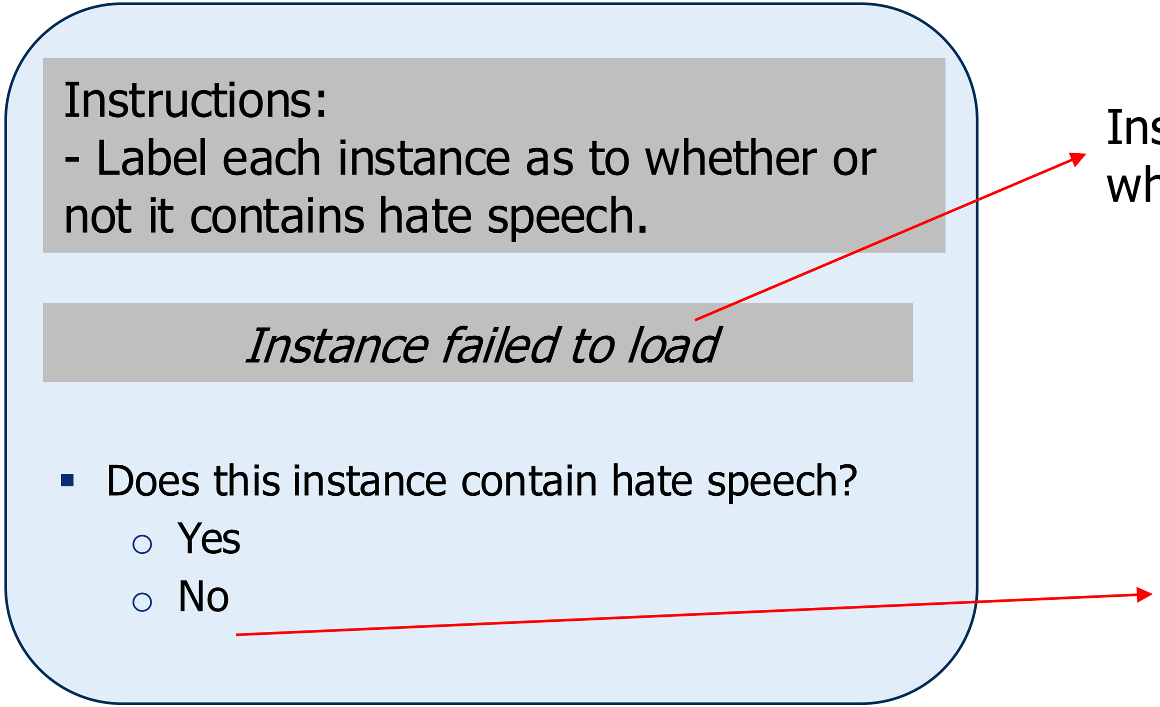
- We need to define hate speech:
- "language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group" [Davidson et al. 2017]

- Examples of what does and does not count

Annotation Scheme: Design process



Annotation scheme



Instructions:
- Label each instance as to whether or not it contains hate speech.

Instance failed to load

- Does this instance contain hate speech?
  - ○ Yes
  - ○ No

Instructions and/or examples of what to do in weird failures

Add “error” or “unable to determine” option

Decomposition

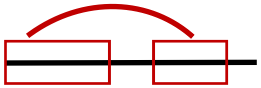
Instructions:
- Label each instance as to whether or not it contains hate speech.

It was just a joke! You're too sensitive.

- Does this instance contain hate speech?
- Does this instance contain sexism?
- Does this instance contain racism?
- Does this instance contain positive/negative/neutral sentiment?

- Best practice: Break complex questions into smaller simpler questions
- Run entirely separate annotation tasks for different dimensions

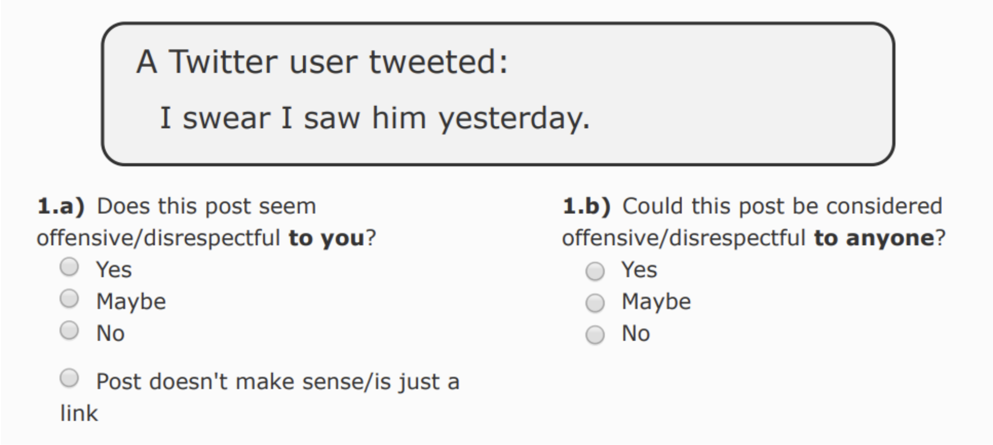
Context and Priming



Annotation scheme

- Contextual information, question ordering, question style can affect how annotators label data

- E.g., increasing evidence of *racial bias* in hate/offensive language detection
  - Models are more likely to label content as offensive if it contains African American English or identity terms [Davidson et al. 2019; Dixon et al. 2018]
  - Annotators are less likely to falsely flag content as offensive if they are told the dialect of the tweet or likely race/ethnicity of the user [Sap et al. 2019]



A Twitter user tweeted:

I swear I saw him yesterday.

1.a) Does this post seem offensive/disrespectful **to you**?

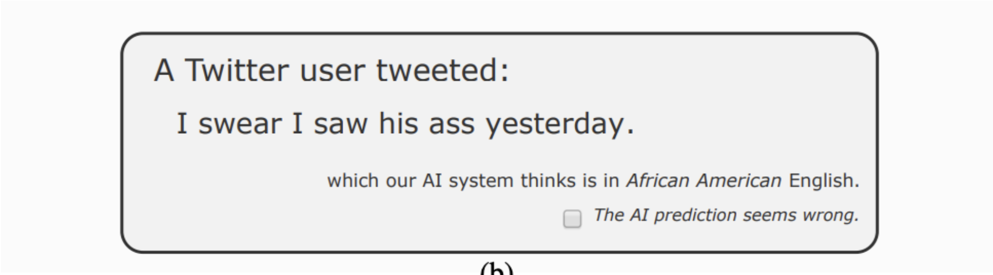
○ Yes
○ Maybe
○ No

○ Post doesn't make sense/is just a link

1.b) Could this post be considered offensive/disrespectful **to anyone**?

○ Yes
○ Maybe
○ No

(a)



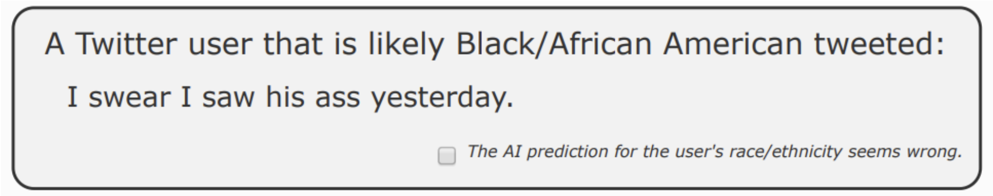
A Twitter user tweeted:

I swear I saw his ass yesterday.

which our AI system thinks is in *African American* English.

□ *The AI prediction seems wrong.*

(b)



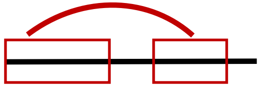
A Twitter user that is likely Black/African American tweeted:

I swear I saw his ass yesterday.

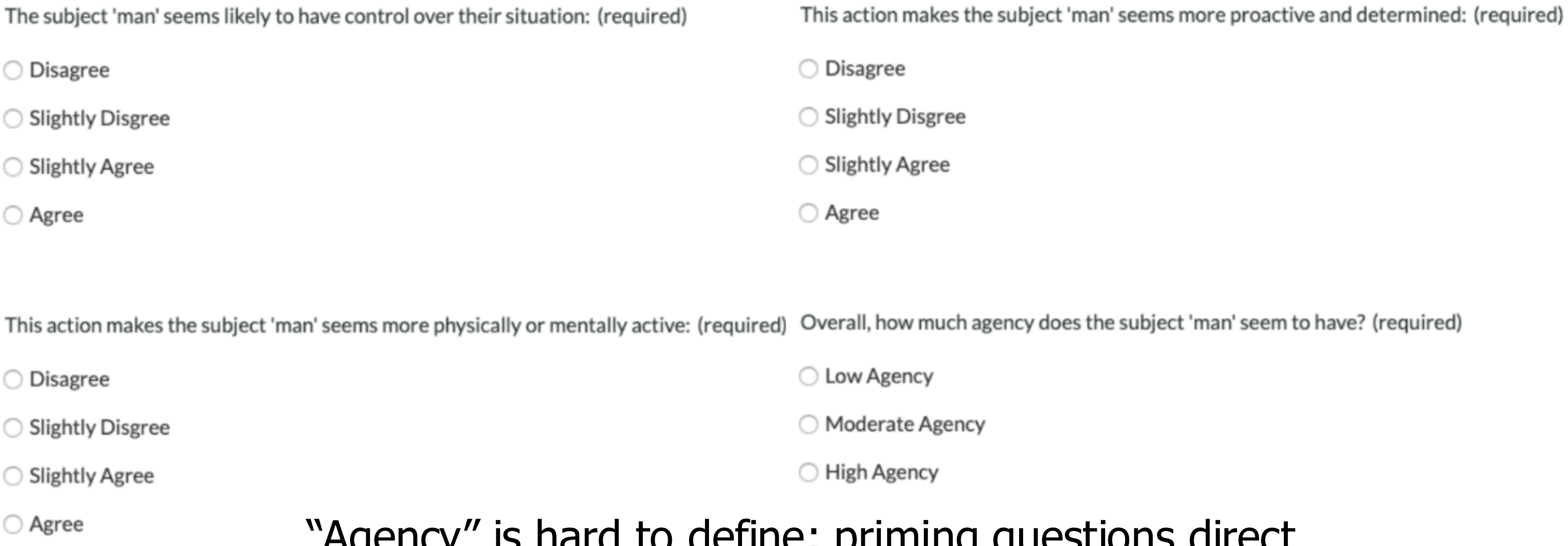
□ *The AI prediction for the user's race/ethnicity seems wrong.*

[Sap et al. 2019]

Context and Priming



Annotation scheme



The subject 'man' seems likely to have control over their situation: (required)

○ Disagree

○ Slightly Disgree

○ Slightly Agree

○ Agree

This action makes the subject 'man' seems more proactive and determined: (required)

○ Disagree

○ Slightly Disgree

○ Slightly Agree

○ Agree

This action makes the subject 'man' seems more physically or mentally active: (required)

○ Disagree

○ Slightly Disgree

○ Slightly Agree

○ Agree

Overall, how much agency does the subject 'man' seem to have? (required)

○ Low Agency

○ Moderate Agency

○ High Agency

"Agency" is hard to define: priming questions direct annotator's focus before actual annotation question

Park et al. 2022; Sap et al. 2017

Platforms

Annotation Interface

Hosted

- **Mechanical Turk**
- **Prolific**
- Toloka
- Surge
- Scale
- Sama
- …

On-Premise

- Label Studio
- CVAT
- Prodigy
- Excel & Co.
- WebAnno
- Jupyter Notebooks
- …

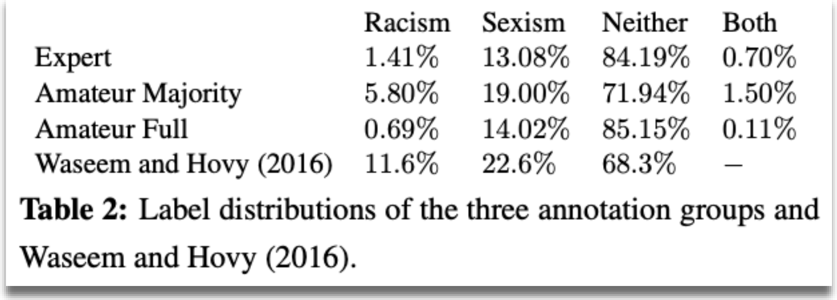
Some considerations:

1. Who the annotators are
2. Ease of designing task
3. Additional support (built-in metrics, quality control)
4. Whether or not you’ve used the platform before

Annotators of different backgrounds annotate differently

Annotators

- Ensuring annotators are qualified (e.g. fluent in the relevant language), understand the task, crowd-sourced vs. specific experts etc.



	Racism	Sexism	Neither	Both
Expert	1.41%	13.08%	84.19%	0.70%
Amateur Majority	5.80%	19.00%	71.94%	1.50%
Amateur Full	0.69%	14.02%	85.15%	0.11%
Waseem and Hovy (2016)	11.6%	22.6%	68.3%	–

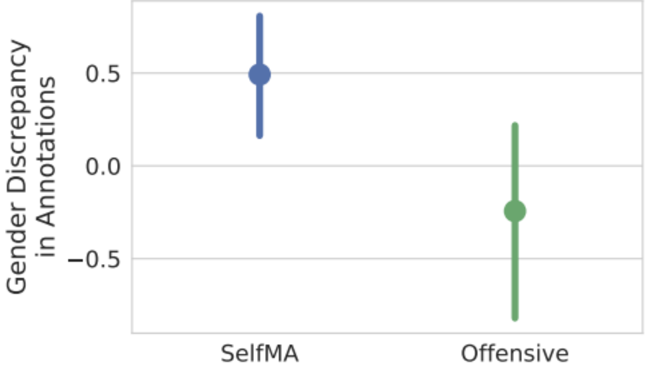
Table 2: Label distributions of the three annotation groups and Waseem and Hovy (2016).

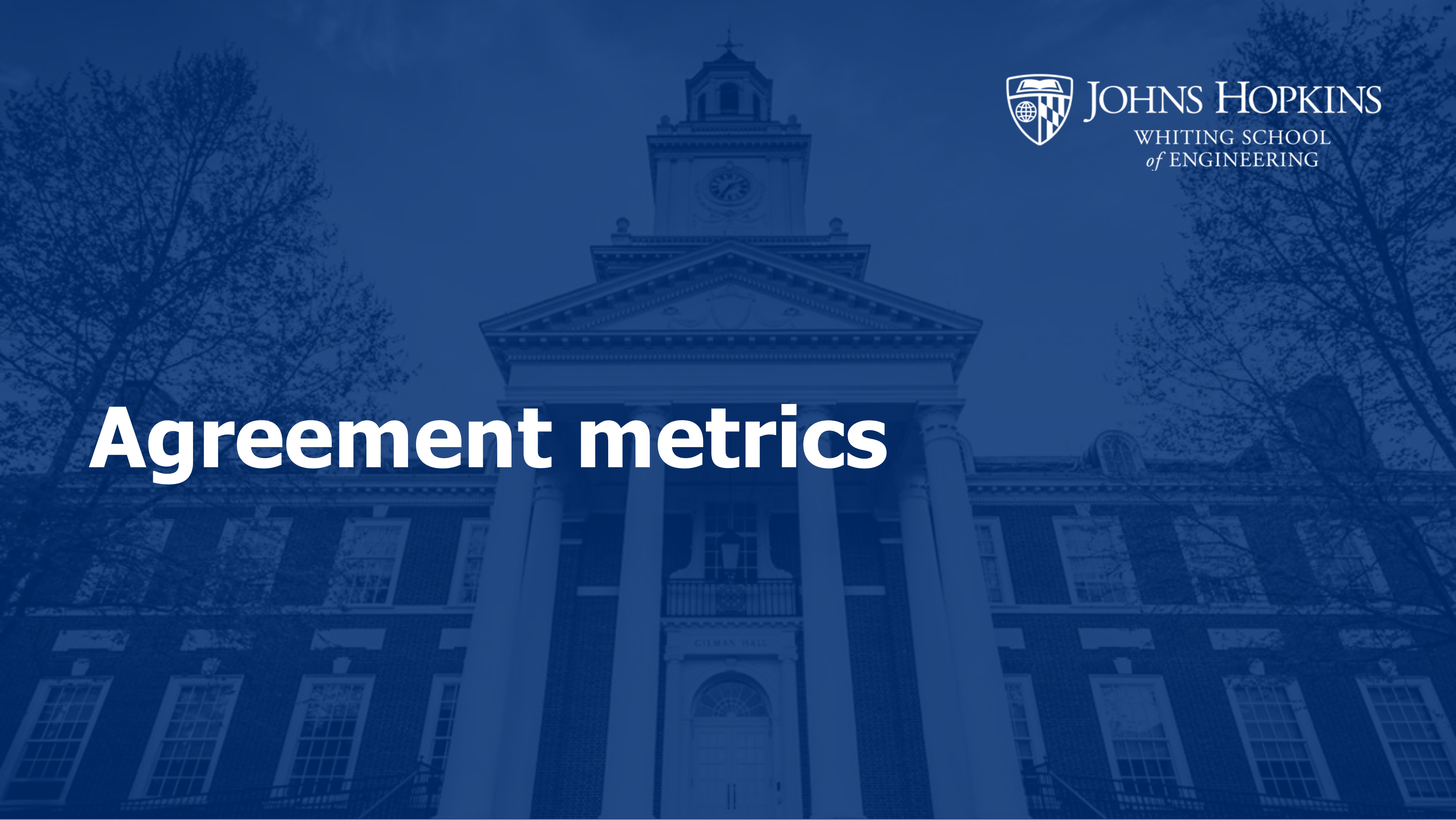
- Feminists and anti-racism activists label less content as racist/sexist than crowdworkers [Waseem 2016]

Annotators of different backgrounds annotate differently

Annotators

- Challenge: hate/offensive speech is already hard to define, how can we identify *microaggressions?*
  - ”Subtly or often unconsciously expresses a prejudiced attitude toward a member of a marginalized group such as a racial minority” [Merriam-Webster]
  - Example: “you’re too pretty to be a computer scientist!”
- Hypothesis: “there will be a discrepancy of perceived offensiveness between the dominant group and the marginalized groups for MAS [microagressions].” [Breitfeller 2019]





JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Agreement metrics

Inter-annotator Agreement

Quality Control

- How can we tell if annotations are reliable and high quality?
  - Standard metric: inter-annotate agreement
  - Each data point is annotated by multiple raters
  - If annotators didn’t agree on the label, maybe the instance was hard?
  - If annotators rarely agree on the label:
    - Task was hard or poorly defined
    - Annotators weren’t qualified (didn’t understand the task)

Inter-annotator Agreement

Quality Control

		Annotator 1		
		Not Offensive	**Offensive**	**Sum**
Annotator 2	**Not Offensive**	147	3	150
	Offensive	10	62	72
	Sum	157	65	222

Percent Agreement: $\frac{147+62}{222} = 0.94$

If each annotator selected randomly, they would have sometimes agreed by chance -- we need to correct for this

[McHugh 2012]

Cohen’s Kappa

Quality Control

		Annotator 1		
		Not Offensive	**Offensive**	**Sum**
Annotator 2	**Not Offensive**	147	3	150
	Offensive	10	62	72
	Sum	157	65	222

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

p_o = percent agreement
p_e = chance agreement

0 → agreement is random chance
- → agreement is worse than random

[McHugh 2012]

Cohen’s Kappa

Quality Control

		Annotator 1		
		Not Offensive	**Offensive**	**Sum**
Annotator 2	**Not Offensive**	147	3	150
	Offensive	10	62	72
	Sum	157	65	222

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$$

where n_{ki} = number of times annotator i picked category k

$$p_e = \left(\frac{157}{222}\right)\left(\frac{150}{222}\right) + \left(\frac{65}{222}\right)\left(\frac{72}{222}\right) = 0.573$$

Estimate of probability Annotator 2 selected “not offensive”

[McHugh 2012]

Cohen’s Kappa

Quality Control

		Annotator 1		
		Not Offensive	**Offensive**	**Sum**
Annotator 2	**Not Offensive**	147	3	150
	Offensive	10	62	72
	Sum	157	65	222

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

$$\kappa = \frac{0.94-0.573}{1-0.573} = 0.859$$

[McHugh 2012]

Agreement Metrics

Quality Control

- Percent Agreement
- Cohen’s Kappa
- Fleiss’ Kappa
  - Similar idea to Cohen’s Kappa but generalized to n annotators with different p_e formula
- Intraclass Correlation (ICC)
- Krippendorff’s Alpha

Krippendorff’s Alpha

$$\alpha = 1 - \frac{D_o}{D_e}$$

D_o = observed disagreement
D_e = disagreement attributable to chance

- Any number of annotators
- Any number of categories, scale values, or measures
- Any metric or level of measurement (nominal, ordinal, interval, ratio, and more)
- Incomplete or missing data
- Large and small sample sizes alike, not requiring a minimum

https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf

Other Tricks for Improving Quality

Quality Control

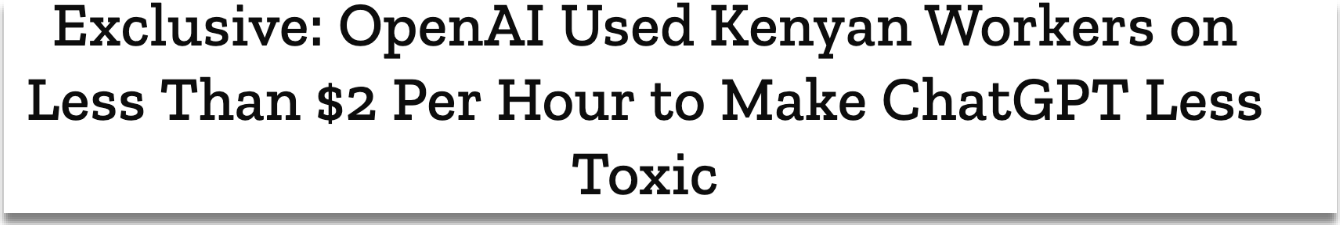
- Annotator qualifications
- Release data in small batches and continually refine annotation scheme and annotator pool
- Identify pool of annotators who are good at a task and ask them to keep doing it [depends on what you’re trying to capture!]
- “Gold tasks” / Quiz questions
- Lots of internal pilots

Ethics

- Is this data that we have permission to collect and annotate?
  - Social media users did not explicitly consent to this use of their data, even if it is within platform terms of service

- Asking annotators to repeatedly view toxic and offensive content can be mentally traumatic

- Annotator payment: local minimum wage? Impact on economy?



Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

https://time.com/6247678/openai-chatgpt-kenya-workers/

HW 2: Design an annotation scheme

- Group assignment
- Details
  - Annotate data under a scheme we give you
  - Revise and improve scheme
  - Re-annotate data
  - Conduct analysis of larger annotated data set

- No code submission: written report of your findings and revisions

Recap

- Why annotate data?
- Tips and tricks for components of annotation process
- Annotator agreement metrics
- Ethics of crowdsourcing

Next class:

- Lexicons (and examples of clever data annotating)

Acknowledgements and References

- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. ACL
- Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science* at ACL
- Davidson, Thomas, et al. "Automated hate speech detection and the problem of offensive language." *AAAI*
- Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. In EMNLP
- ICML Tutorial: https://slideslive.com/39004357/reinforcement-learning-from-human-feedback-a-tutorial-?ref=search-presentations-reinforcement+learning+from+human+feedback
- Amber E. Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues. APSA 2014 Annual Meeting Paper.
- Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. Quality and Quantity, 38(6):787–800

Break