You Are What You Annotate: Towards Better Models through Annotator Representations

Naihao Deng Siyang Liu Xinliang Frederick Zhang Winston Wu Lu Wang Rada Mihalcea
University of Michigan

(To appear at EMNLP 2023 Findings)

Inherent annotation disagreements

Friends QIA

Question: Did Rachel tell you we hired a male nanny?

Answer: I think that's great!

ANN ANSWER (1), NOT THE ANSWER (2), ANSWER SUBJECT TO SOME CONDITIONS (3), NEITHER (4), OTHER (5): 1, 1, 4

Pejorative

Text: @WORSTRAPLYRICS Everything Jay-Z writes is trash.

ANN PEJORATIVE (1) <–> NON-PEJORATIVE (0): 1, 0, 0

HS-Brexit

Text: RT <user>: Islam has no place in Europe #Brexit.

ANN NO HATE (1) <–> HATE (0): 1, 1, 1, 0, 0, 0

MultiDomain Agreement

Text: Please lost you yelling insanely at the sky on Nov 3 losers

ANN OFFENSIVE (1) <–> NOT OFFENSIVE (0): 1, 1, 1, 0, 0

Go Emotions

Text: This is how I feel when I use a crosswalk on a busy street

ANN POSITIVE (1), NEUTRAL (0), AMBIGUOUS (-1), NEGATIVE (-2): 1, 0

Humor

Text A: Being crushed by large objects can be very depressing.

Text B: As you make your bed, so you will sleep on it.

ANN WHICH IS FUNNIER, X MEANS A TIE: A, A, B, X, X

CommitmentBank

Premise: Meg realized she'd been a complete fool. She could have said it differently. If she'd said Carolyn had borrowed a book from Clare and wanted to return it they'd have given her the address.

Hypothesis: Carolyn had borrowed a book from Clare.

ANN ENTAIL (3) <–>CONTRADICT (-3): 3, 3, 3, 2, 0, -3, -3, -3

Sentiment Analysis

Text: Even hotel bar food is good in California...fresh avocados, old chicken, and reasonably recent greens. Mmmm. Really.

ANN POSITIVE (2) <–>NEGATIVE (-2) : 2, 2, 0, -1

Problematic to ignore such disagreement!

Hate speech Detection

aggregating labels → ignores the under-represented groups

Humor and Sentiment

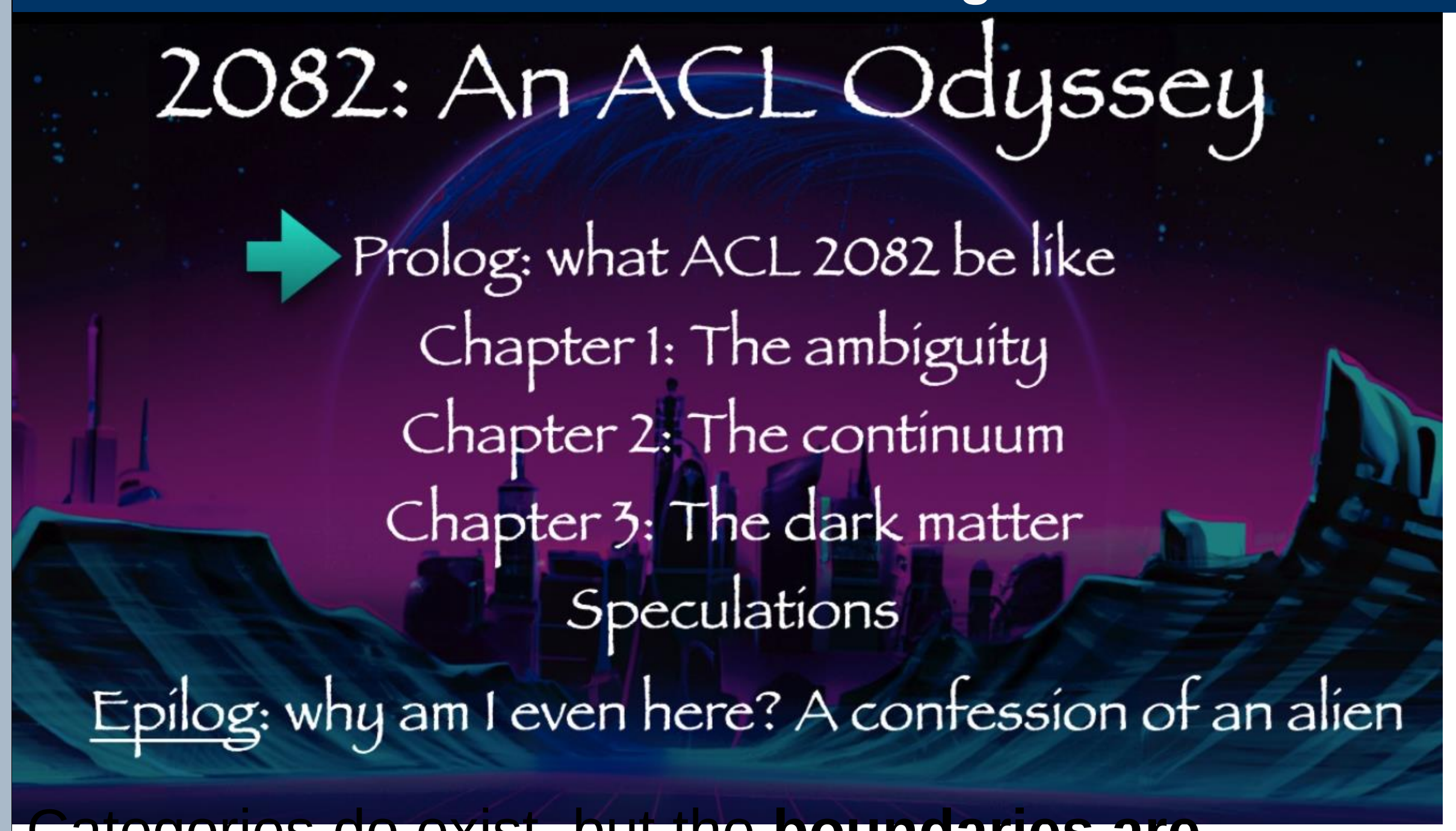
highly subjective

Natural Language Inference (NLI)

Previous studies showed the inherent annotation disagreements

Instead, Let models learn from data that has inherent disagreement!

Factors that cause annotation disagreements



2082: An ACL Odyssey

Prolog: what ACL 2082 be like

Chapter 1: The ambiguity

Chapter 2: The continuum

Chapter 3: The dark matter

Speculations

Epilog: why am I even here? A confession of an alien

Categories do exist, but the **boundaries are "squish".**

— Yejin Choi (University of Washington)

Babara Plank's survey:

- Differences in interpretation
- Certain preferences
- Difficult cases or multiple plausible answers

Is the dress white and gold or black and blue?

Qualia

Our Approach

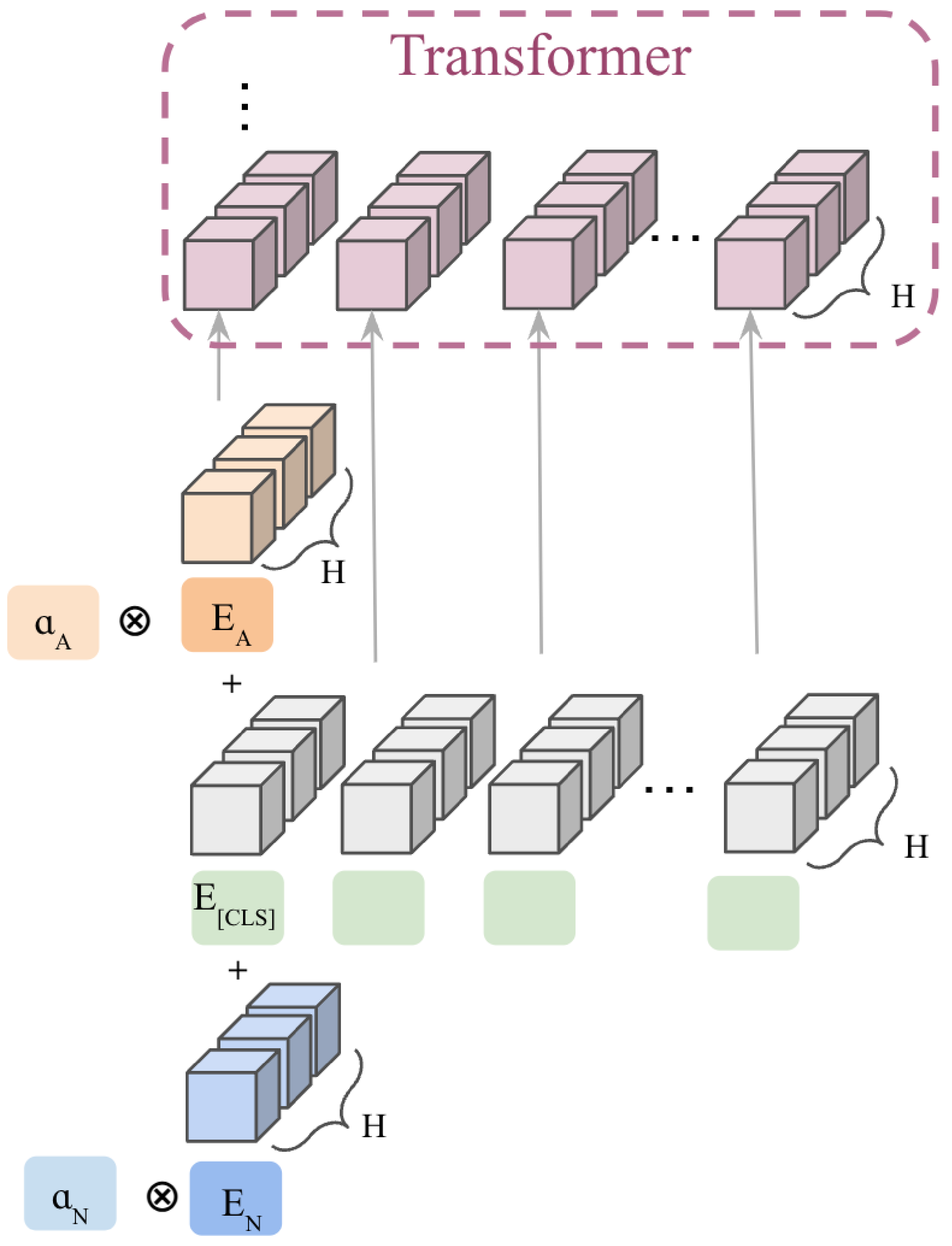
Two representations:

- **Annotator Embedding** (E_A): represent each annotator
- **Annotation Embeddings** (E_L): aggregate annotators' annotations on other examples

Two weights:

balance the effects of text and the embeddings

Our Approach (Continued)



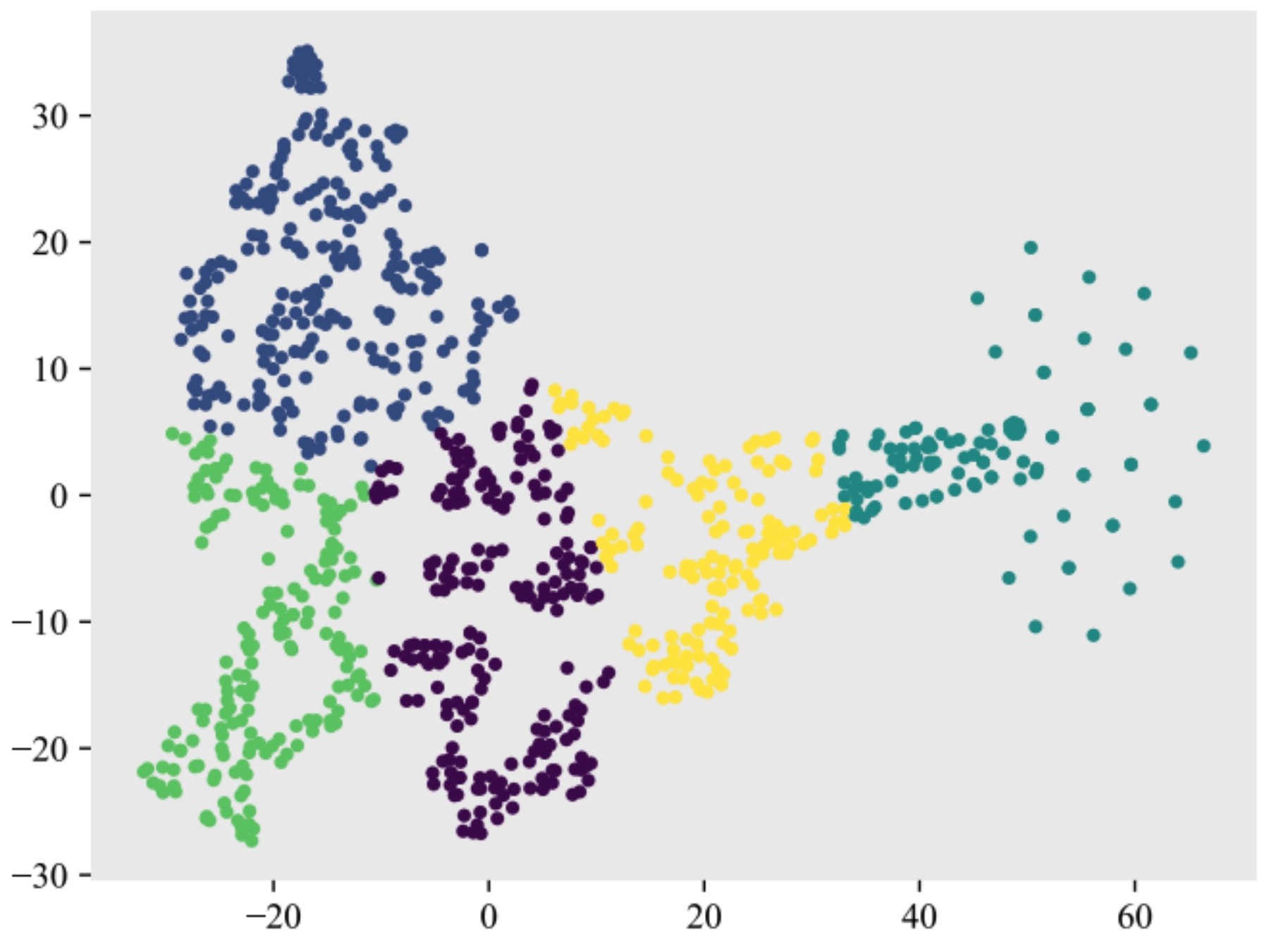
Annotator-based predictions

Text: We know it anecdotally from readers we've heard from who've been blatantly discriminated against because they're older.

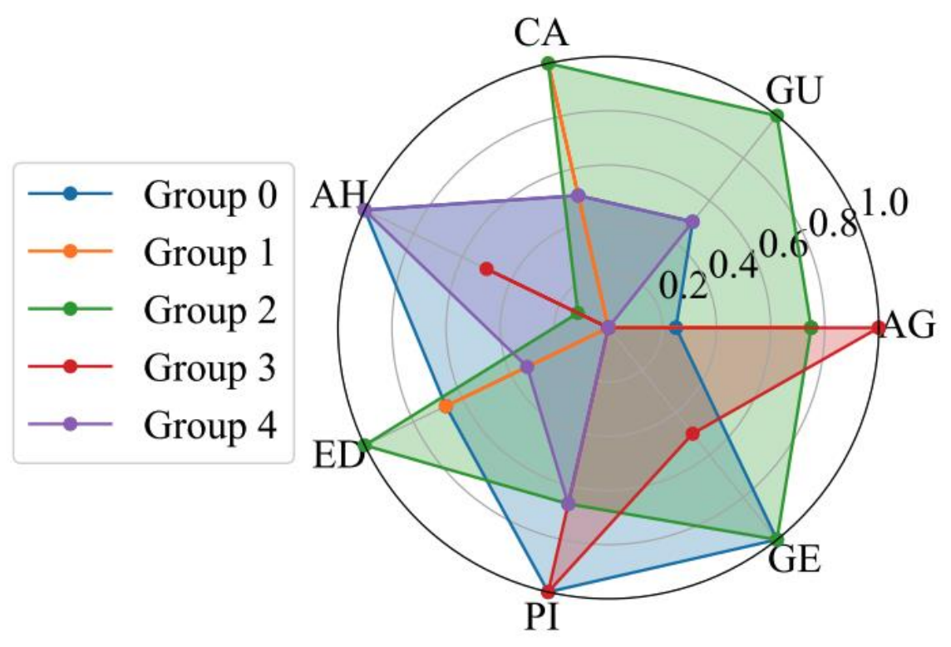
POSITIVE (2) <–> NEGATIVE (-2)

Annotator ID	1	2	3	4
Gold	-1	0	-2	-2
T	-1	-1	-1	-1
$\mathbf{E}_n + \mathbf{E}_a$	-1	0	-1	-2

Grounded to real-world demographic features



(a) Annotation Embedding



CA: current living area
GU: grew up area
AG: age
GE: gender
PI: political identification
ED: education
AH: annual household income