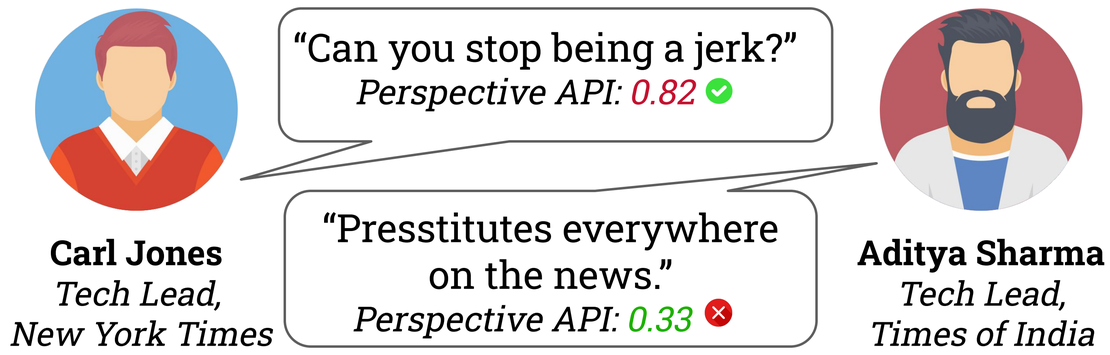
# NLPositionality: Characterizing Design Biases of Datasets and Models

Sebastin Santy[1], Jenny T. Liang[2], Ronan Le Bras[3], Katharina Reinecke[1], Maarten Sap[2,3]
*University of Washington[1], Carnegie Mellon University[2], Allen Institute for AI[3]*

## Background

**Design biases** are when datasets and models exhibit *disparities in performance or representativeness* for different populations. **Positionality** is the perspectives people hold due to their demographics, identity, and life experiences. Positionality of NLP researchers can influence the design decisions researchers make and introduce design biases in language technologies.




**Design bias example:** For identifying toxic content, Perspective API works better for Carl Jones from the U.S. than Aditya Sharma from India, as it it does not understand offensive terms used in Indian contexts.

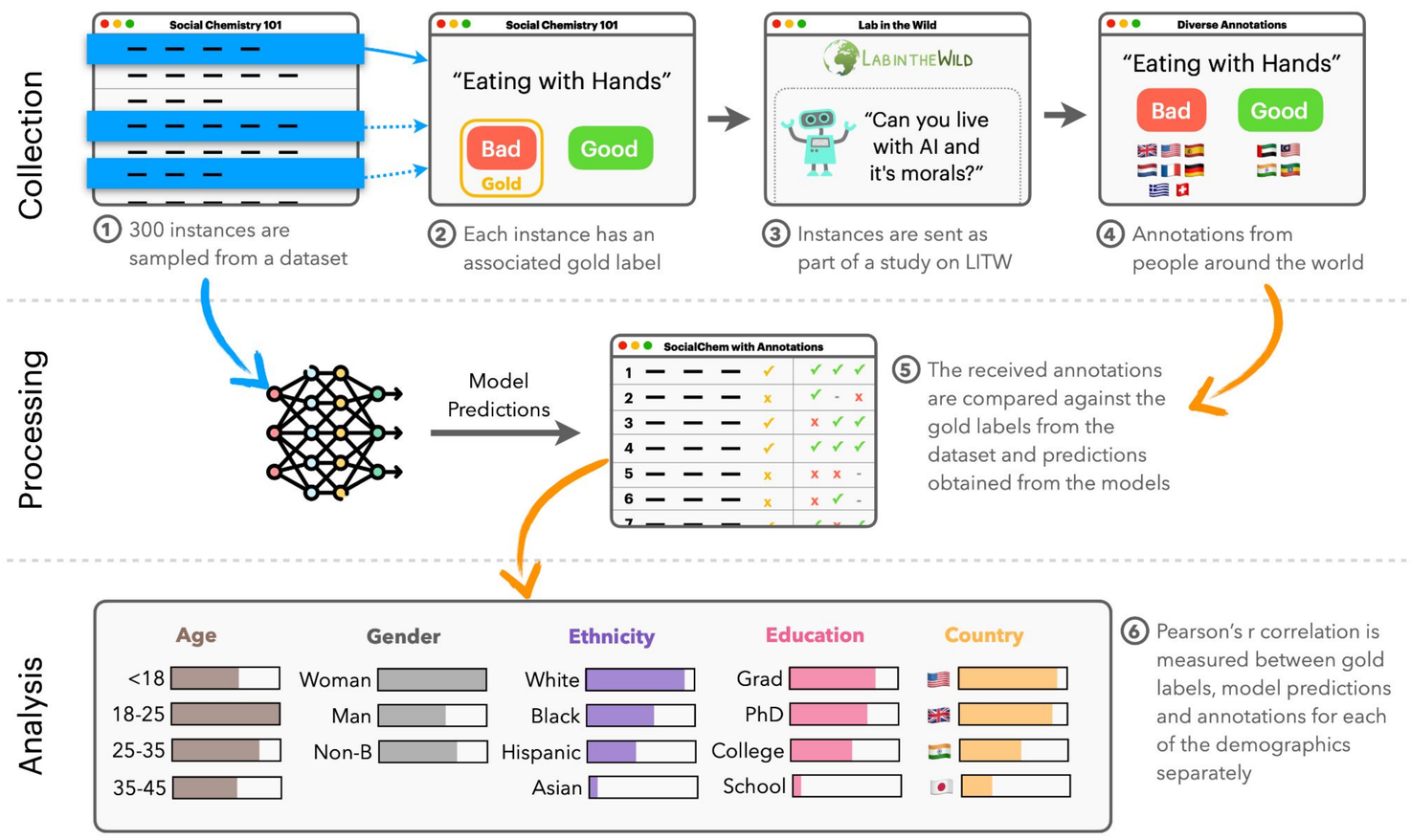
**Positionality example:** The U.S.-based developers who made Perspective API used toxicity datasets that primarily was based on American English.

## NLPositionality

**NLPositionality** is a framework for characterizing design biases and positionality of NLP datasets and models. We have collected **16,299 annotations** from **1,096 annotators** from **87 countries**.

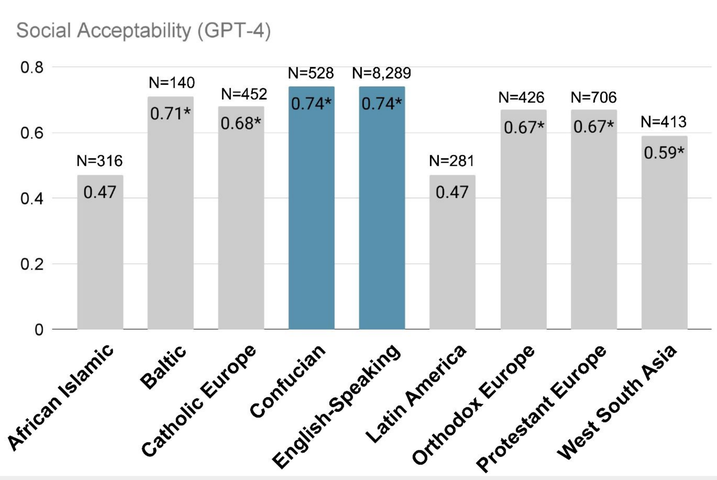
**Collection (steps 1-4):** A subset of datasets' instances are re-annotated on a platform called LabintheWild, which has more diverse annotators.

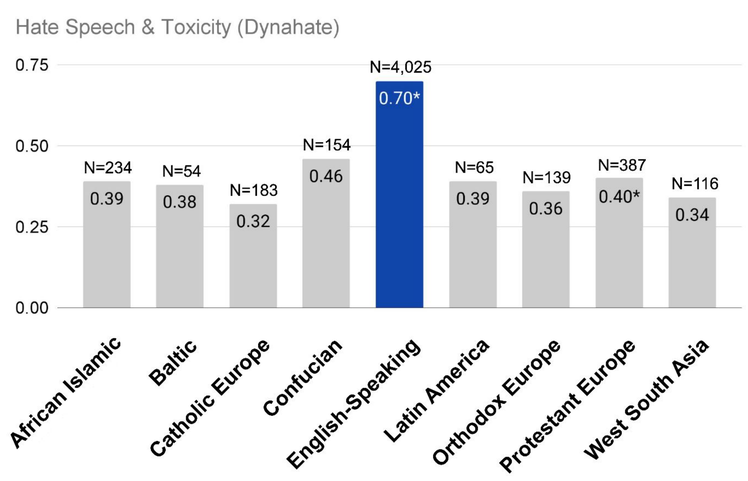
**Processing & Analysis (steps 5-6):** We compute the Pearson's r correlation between the LabintheWild annotations by demographic for the dataset's original labels and the models' predictions.



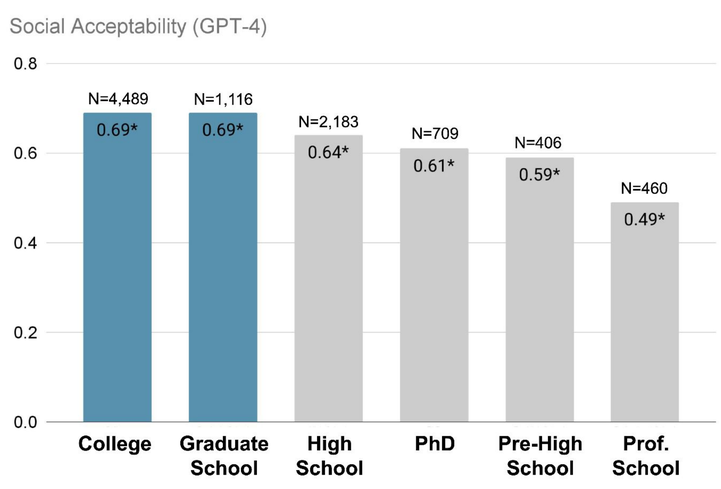

## Analysis & Results

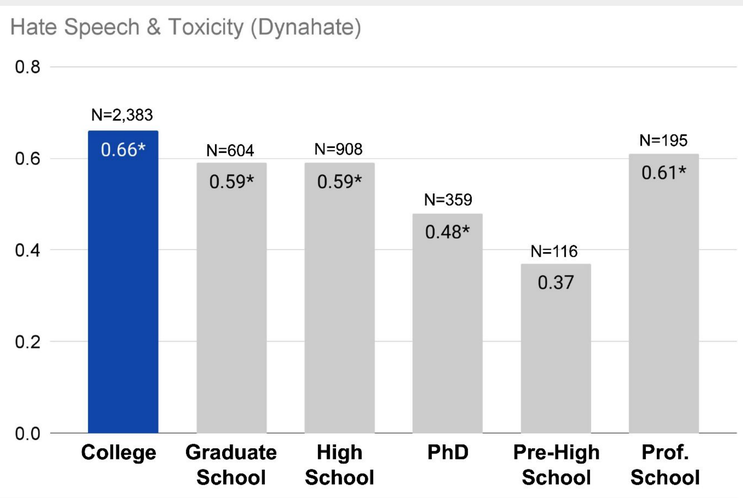
We apply NLPositionality to two tasks: **social acceptability** (Social Chemistry, Delphi, GPT-4) and **hate speech detection** (Dynahate, Perspective API, Rewire API, Hate RoBERTa, GPT-4).




**Finding 1:** Datasets and models align the most with **people from English-Speaking countries**.




**Finding 2:** Datasets and models align the most with **people with college education**.

## Discussion

**Takeaway:** There is positionality in NLP, and it tends to be Western-Centric. But,.some populations are left behind.

**Recommendation 1:** Record all relevant design choices made while building datasets or models.

**Recommendation 2:** Do NLP research through the lens of perspectivism.

**Recommendation 3:** Building specialized datasets and models for specific communities is valuable for inclusive NLP (e.g., Masakhane initiative).

## Learn more

**Paper** bit.ly/NLPositionality-Paper

**Project website** nlpositionality.cs.washington.edu/

61 ACL 2023