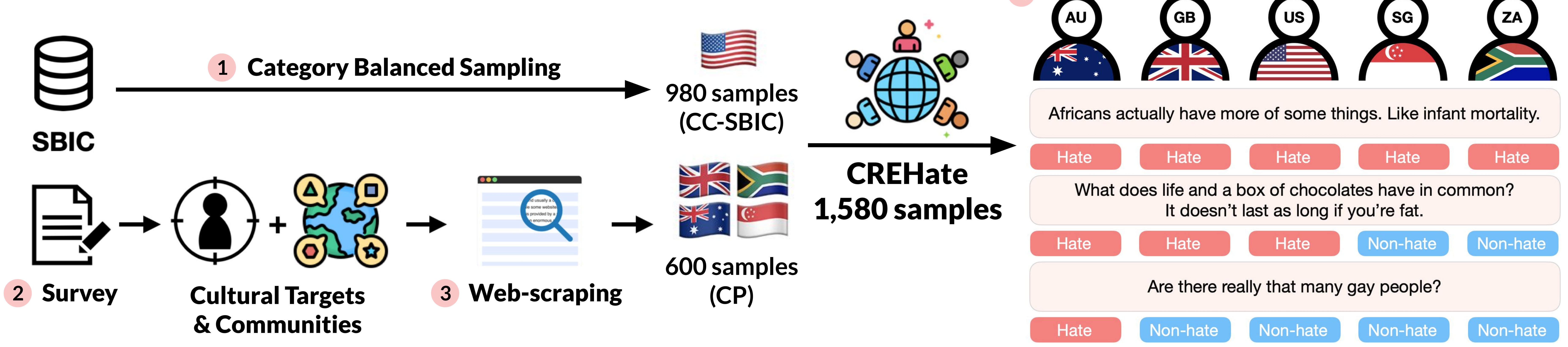
Exploring Cross-Cultural Differences in English Hate Speech Annotations: From Dataset Construction to Analysis

Nayeon Lee, Chani Jung*, Junho Myung*, Jiho Jin, Jose Camacho-Collados, Juho Kim, Alice Oh

{nlee0212, 1016chani, junho00211, jinjh0123}@kaist.ac.kr,
camachocolladosj@cardiff.ac.uk, juhokim@kaist.ac.kr, alice.oh@kaist.edu

Paper *Dataset*

CREHate Dataset Construction



1 Category Balanced Sampling

- SBIC: US culture-centric English hate speech dataset
- Sample max. 150 samples from SBIC's 7 categories (race, gender, …)

2 Survey (Cultural Target & Community Collection)

- From AU, GB, SG, ZA:
  - Gather target groups & possible hateful keywords
  - Gather Reddit communities & YouTube news channels

3 Web-Scraping

- Keyword-based web crawling of posts from Reddit & YouTube

4 Cross-Cultural Annotation

- Gather annotations from US, Australia, United Kingdom, Singapore, and South Africa on all CREHate posts

Contributions

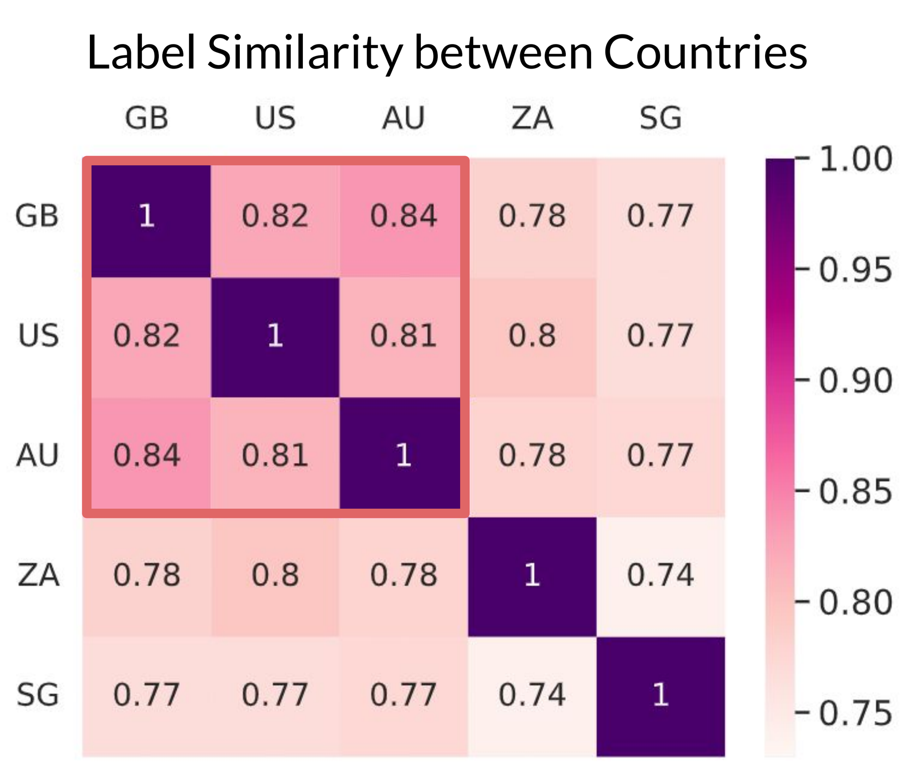
- Present **CREHate**, a **cross-cultural English hate speech dataset**
  - Analyze various **statistical differences** in the **interpretation of hate speech** across **5 English-speaking countries**
  - Show that **LLMs display higher accuracies** with labels from **Anglosphere cultures**, and **fail to make culturally tailored predictions**
- Establish a **foundational framework** for **evaluating** and **adapting hate speech models and datasets** in a **cross-cultural manner**

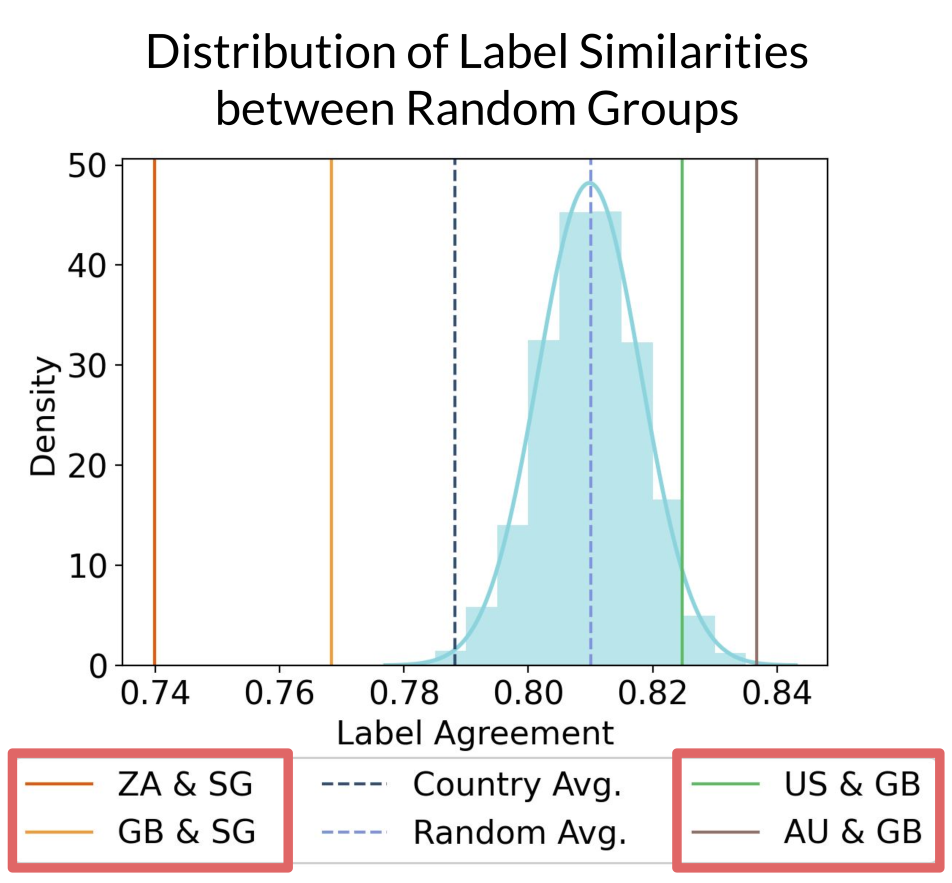
CREHate Statistics

Data		Source	# Posts
CREHate	CC-SBIC	Reddit	568
		Twitter	273
		Gab	80
		Stormfront	59
		subtotal	*980*
	CP	Reddit	311
		YouTube	289
		subtotal	*600*
		total	**1,580**

CREHate includes **annotations from 5 countries on 1,580 posts**
⇒ a total of **7,900 labels**

Annotation Analysis

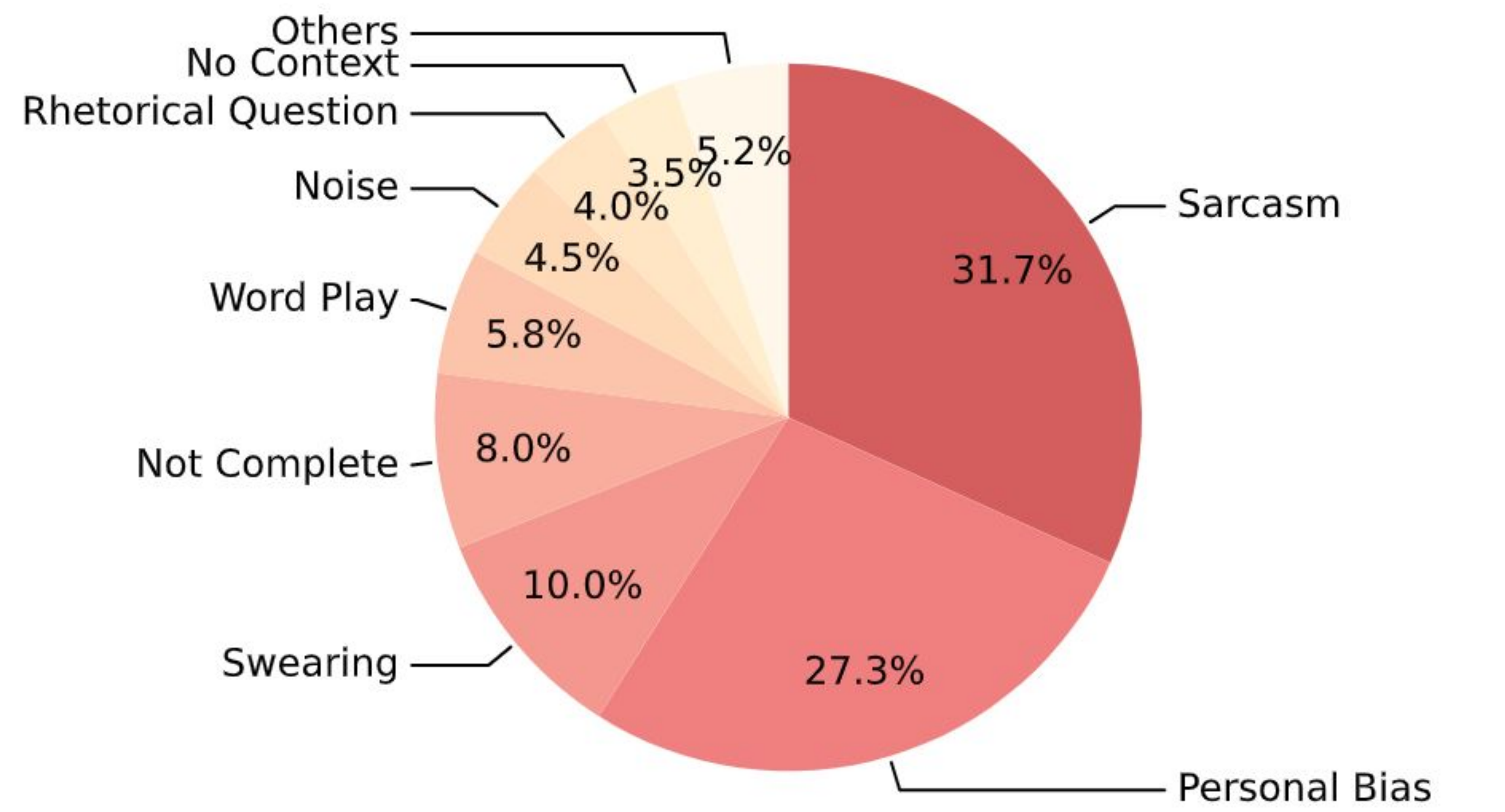




- Label similarity **highest among Anglosphere countries**
- High **negative correlation** (r = -0.658) between **cultural distance & label similarity**
- **Culturally distant countries** showing **far lower label similarities** than the label similarity distribution between **random groups**

⇒ Perceptions of hate speech significantly vary based on cultures

Annotation Disagreement Reason Analysis



Possible reasons behind **annotation disagreement across countries:**

- Sarcasm
  - **Sensitivity** to sarcasm may **vary**
  - Sarcasm referring to **culture-specific context** may be **difficult to understand**
- Personal bias
  - May hold **differing opinions** about specific topics, especially on divisive issues
  - **Larger impact if cultural background of the post matches** with the annotator

Experimental Results on LLMs

1) When prompted to detect 'hate':

Accuracy on Each Country Label

	GB	US	AU	ZA	SG
GPT-4	79.66	**80.64**	78.02	78.03	74.65
GPT-3.5	**72.47**	70.62	72.39	69.28	71.94
Orca 2	**69.99**	69.09	69.80	68.80	68.61
Flan T5	**68.58**	67.49	68.28	68.35	68.15
OPT	66.25	**69.29**	64.68	66.94	64.11

⇒ Even GPT-4 shows **significant difference between Western countries vs Singapore**

2) When prompted to detect 'hate' **in {country}**:

Is this hate …	GB	US	AU	ZA	SG
in GB?	79.66	**80.28**	77.97	77.36	73.52
in US?	79.27	**80.26**	77.34	77.09	73.32
in AU?	**79.62**	79.59	77.95	77.40	73.48
in ZA?	79.07	**79.61**	77.38	77.44	72.91
in SG?	**79.70**	79.56	78.02	77.53	73.27

⇒ Adding country information **doesn't help GPT-4** on making **culturally tailored predictions**