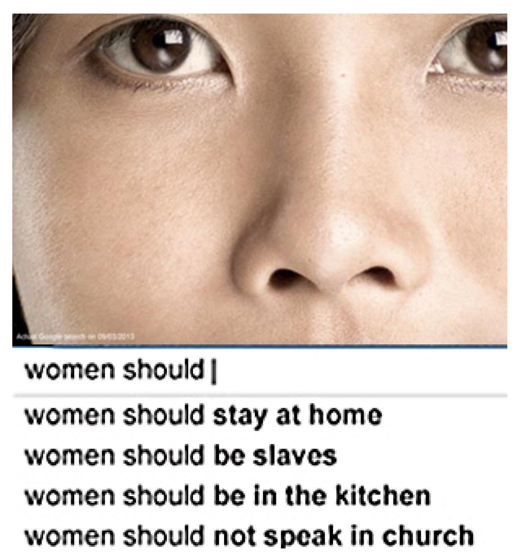
Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Wei Guo, Aylin Caliskan

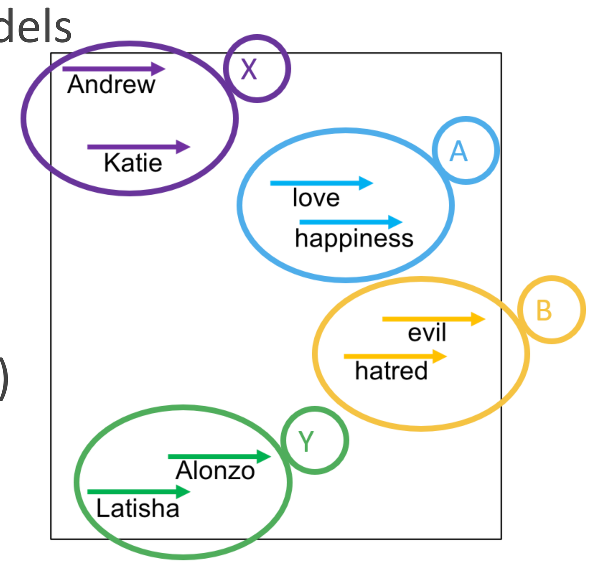
Motivation

- Bias in NLP **exacerbates bias**
- Cannot **automatically identify** bias
- **Incomprehensive measurement** of bias in contextualized word embeddings and neural language models
- Current work focus on a **single category or specific contexts**



Background

- Human-like biases embedded in word embeddings
- Social biases in SOTA neural language models
- Intersectional and emergent biases of the intersectional groups
- Emergent biases only associated with the intersectional group
- Word Embedding Association Test (WEAT) designed for static word embeddings



Two dimensional projection of the stimuli

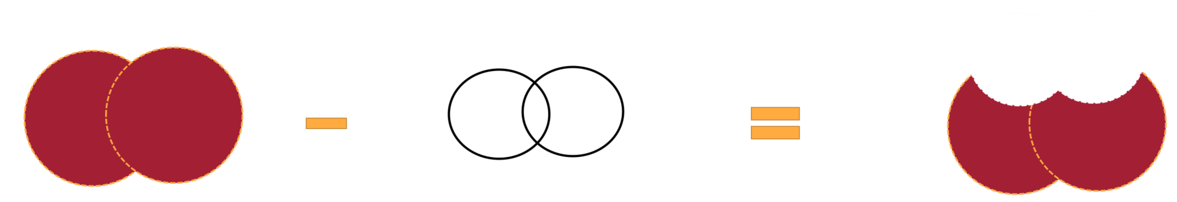
Approach

- **Intersectional Bias Detection** (IBD)
- Identify words associated with intersectional group members defined by two social categories

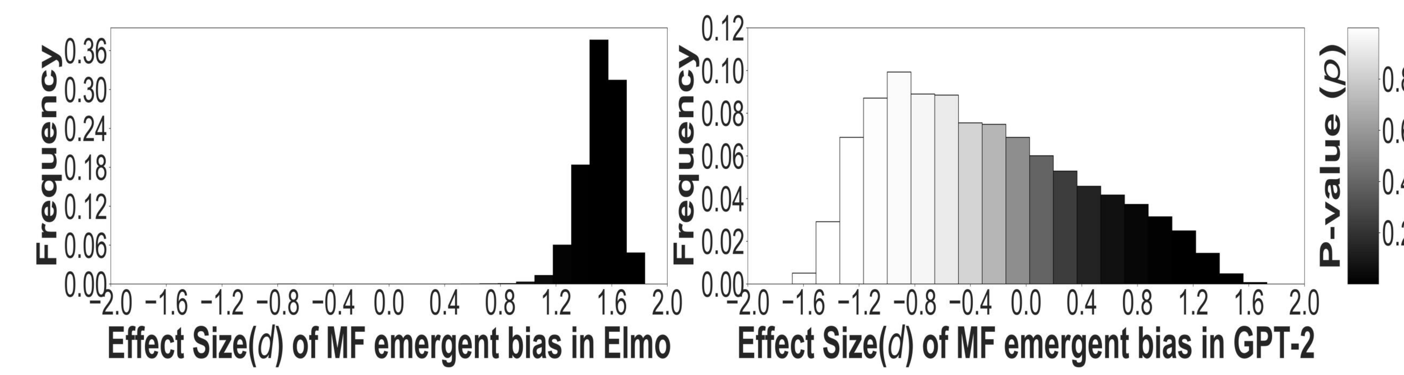
$$s(w,A,B)=\frac{\text{mean}_{a\in A}\cos(\vec{w},\vec{a})-\text{mean}_{b\in B}\cos(\vec{w},\vec{b})}{\text{std-dev}_{x\in A\cup B}\cos(\vec{w},\vec{x})}$$

The union set is the intersectional biases.

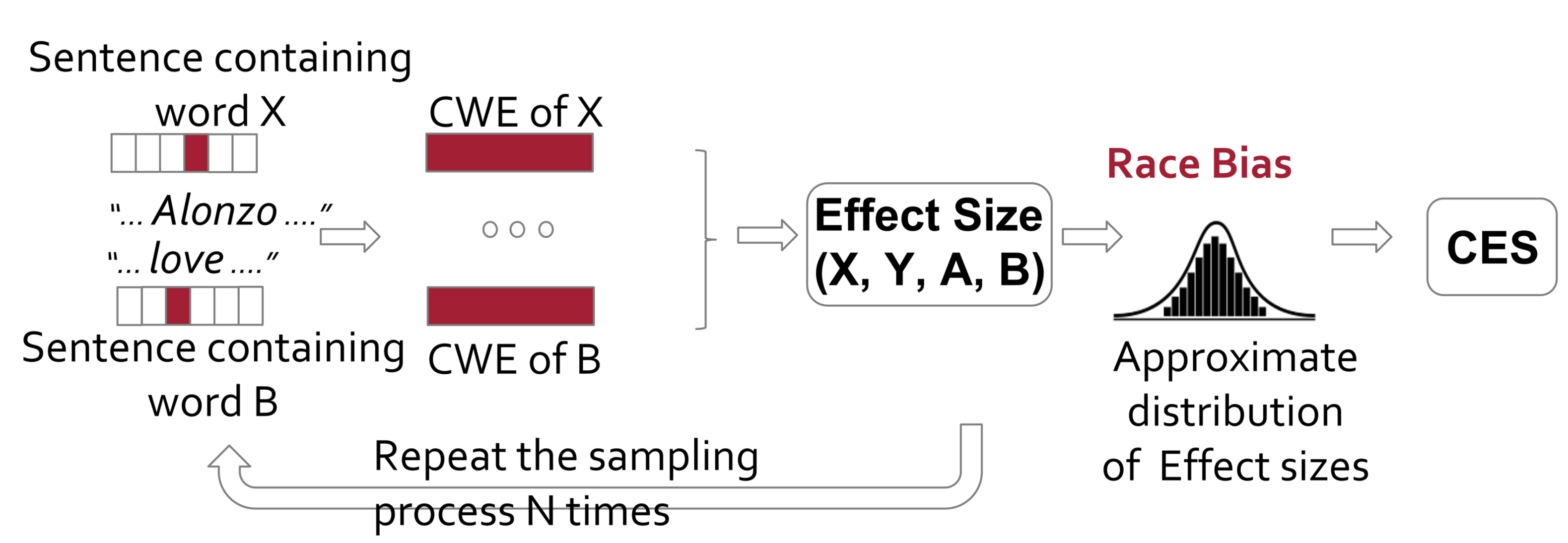
- **Emergent Intersectional Bias Detection** (EIBD)
- Identify words uniquely associated with intersectional group



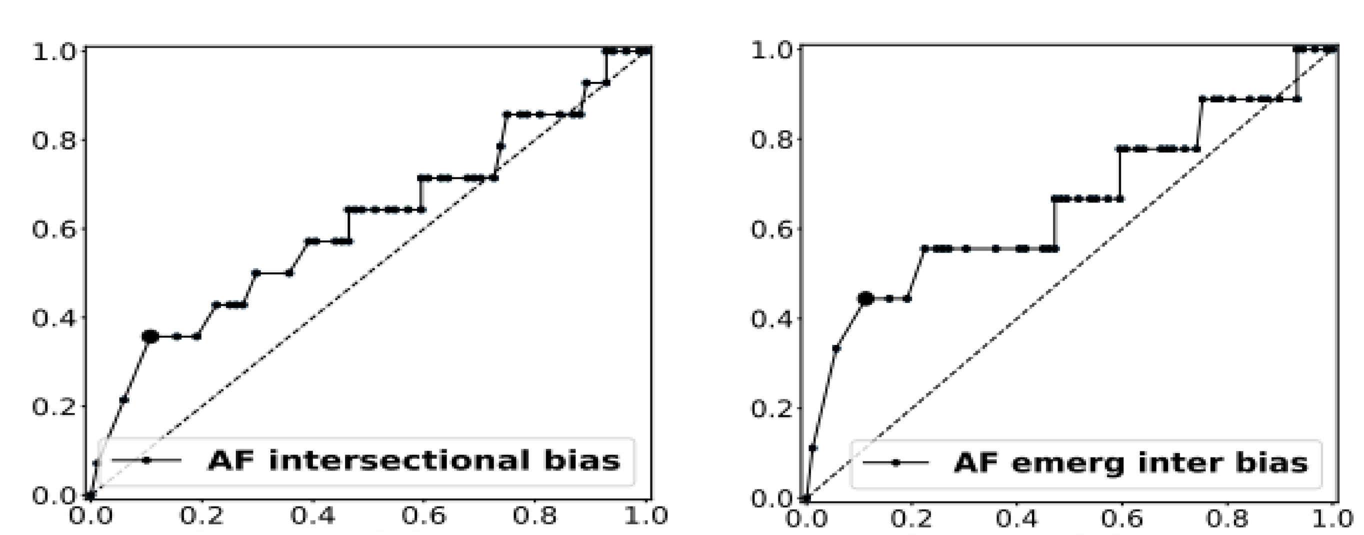
The magnitudes of social bias vary based on the level of contextualization in the neural language models.



Approach



- **Contextualized Embedding Association Test (CEAT)**
- Quantify social biases in contextualized embeddings
- **Random-effect model**
- Estimate the comprehensive summary statistics, combined effect size (CES) in CEAT



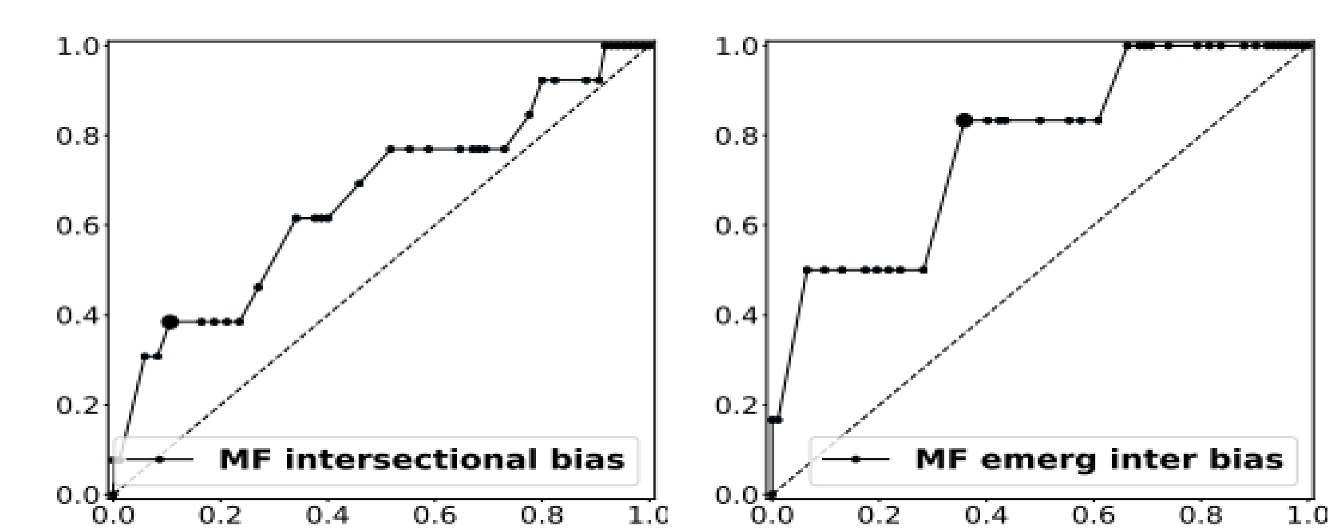
ROC Curve of IBD and EIBD for African American females

Results

- Intersectional biases have **high magnitude**
- ELMo is the **most biased**, followed by BERT, GPT, and GPT-2
- The overall magnitude of bias **negatively correlates** with the level of contextualization in the language model
- Accuracy of IBD: 81.6% and 82.7% (random correct rate: 14.3% and 13.3%)
- Accuracy of EIBD: 84.7% and 65.3% (random correct rate: 9.2% and 6.1%)

Bias Test		*d*
Flowers/Insects	Pleasant/Unpleasant	1.50
Instruments/Weapons	Pleasant/Unpleasant	1.53
European & African-American names	Pleasant/Unpleasant	1.41
Male/Female names	Career/Family	1.81
Math/Arts	Male/Female terms	1.06
Science/Arts	Male/Female terms	1.24
Mental/Physical disease	Temporary/Permanent	1.38
Young/Old people's names	Pleasant/Unpleasant	1.21
African females & European males	Intersectional attributes	1.64
African females & European males	Emergent attributes	1.69
Mexican females & European males	Intersectional attributes	1.71
Mexican females & European males	Emergent attributes	1.82

Intersectional biases
increased color density == increased bias magnitude



ROC Curve of IBD and EIBD for Mexican American females

Reference

- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
- Negin Ghavami and Letitia Anne Peplau. 2013. An intersectional analysis of gender and ethnicstereotypes: Testing three hypotheses. *Psychology of Women Quarterly* 37, 1 (2013), 113–127.

Github repository: https://github.com/weiguowilliam/CEAT