WordBias: An Interactive Visual Tool for Discovering Intersectional Biases Encoded in Word Embeddings

Bhavya Ghai, Md Naimul Hoque, Klaus Mueller
Stony Brook University, USA

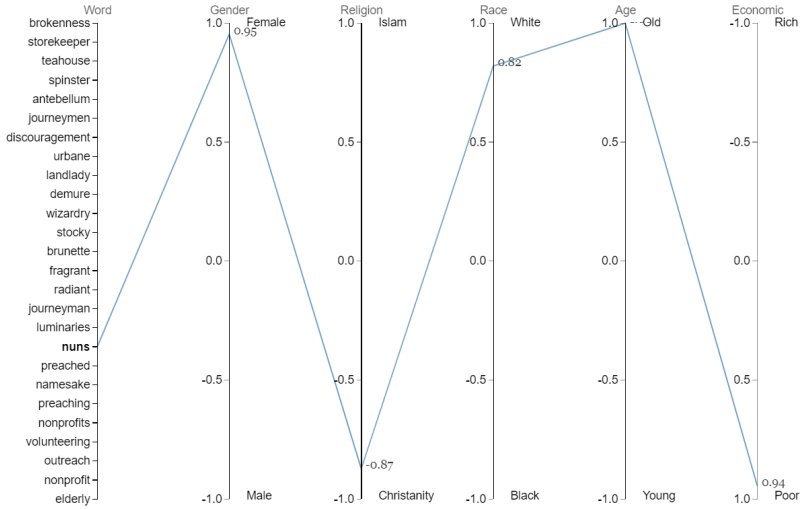
Abstract

Intersectional bias is a bias caused by an overlap of multiple social factors like gender, sexuality, race, disability, religion, etc. Word embedding models can be laden with biases against intersectional groups like African American females. We present WordBias, an interactive visual tool designed to explore biases against intersectional groups encoded in word embeddings. Identifying such biases will serve as the first step in deterring their spread and help develop counter strategies.

WordBias can be used as an *Auditing tool* by data scientists, *Educational tool* by students/non-experts and to expedite the bias discovery process by researchers.

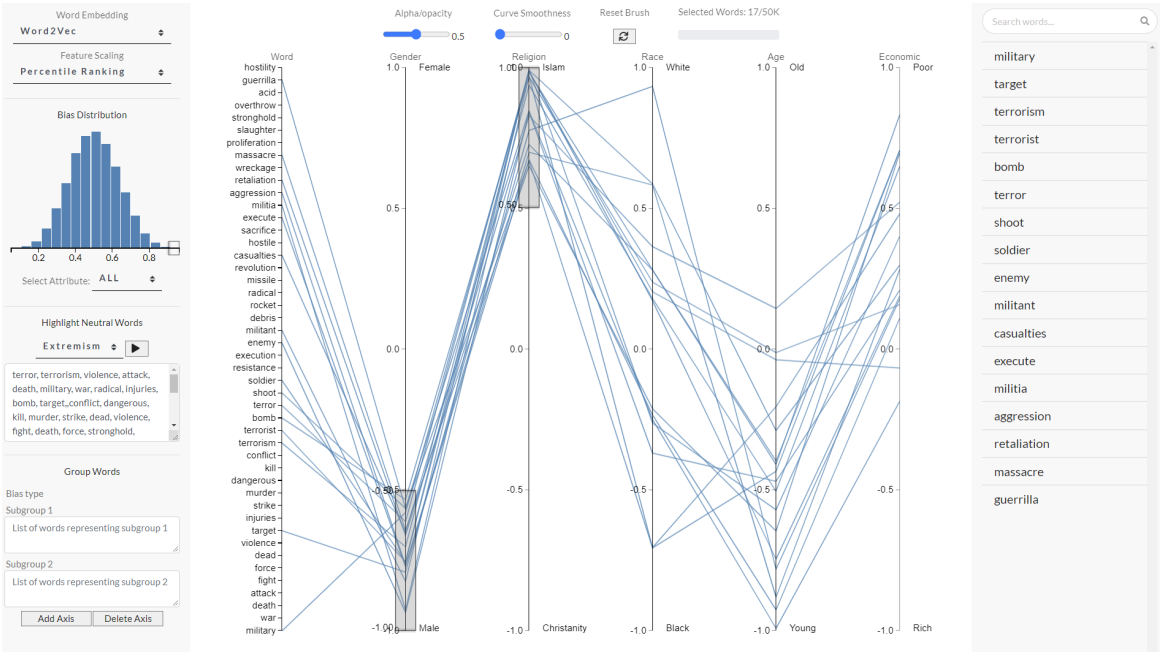
Design Choice

We used Parallel Coordinate plot to visualize associations (bias scores) of different words along different groups in the word embedding. Each polyline (blue) encodes a word, and each axis represents a sensitive attribute like gender, race, religion, etc.



On Hovering over the word 'nuns', its corresponding polyline gets highlighted. We can observe its association with 'Female', 'Christianity', 'White', 'Old' and 'Poor' subgroups.

Visual Interface



In the above figure, the user has brushed over 'Male' and 'Islam' subgroups. Words with strong association to both these subgroups are listed below the search box like military, terrorism, terrorist, soldier, aggression, retaliation, etc.

Results

Intersectional Groups	Associated Words
Poor - Young - Black	disaster, struggle, tackle, chaos, woes, hunger, uprising, desperation, insecurity, rampage, roadblocks
Rich - Old - White	formal, attractive, appealing, desirable, castle, desserts, seaside, golfing, cordial, bungalow, fanciful, warmly, salty
Black - Muslim - Male	gun, assassination, bullets, bribes, thugs, looted, dictators, electrocuted, cowards, agitating, storekeeper, looter, bleeping
Young - Christian - Male	career, dominant, brilliant, lone, terrific, heroes, superb, epic, monster, prowess, heavyweights, excelled, superstars

Stony Brook University