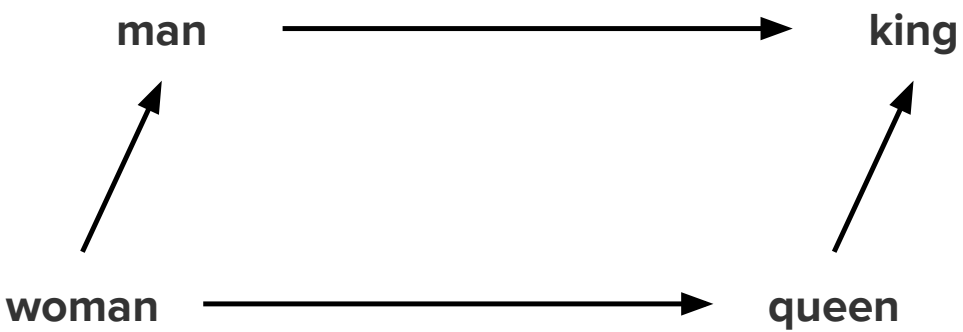
Agnostic debiasing of static embeddings: an approach to fairness in language models

Gianmarco Cafferata (UDESA), Mariano G. Beiró (UDESA-CONICET)

Universidad de San Andrés
CONICET
54 JAIIO
JORNADAS ARGENTINAS DE INFORMÁTICA

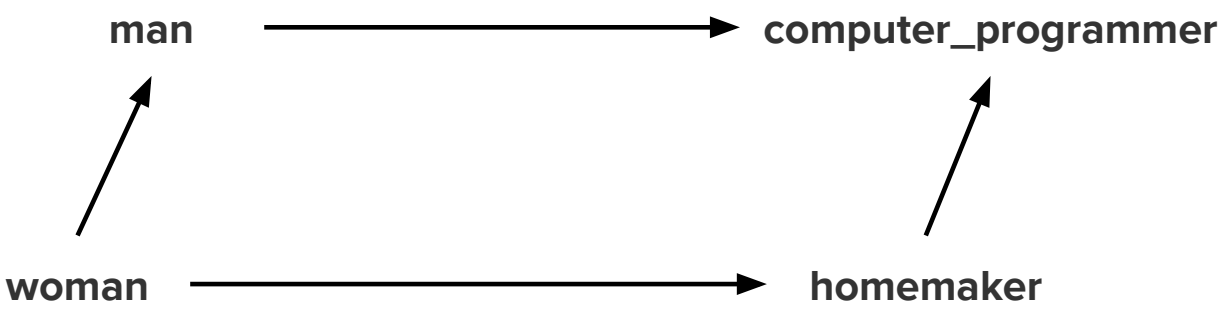
Introducción

Word embeddings



Trabajos previos

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi et al., NeurIPS 2016)



Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi et al., NeurIPS 2016)

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

“Hard debiasing”

$$\vec{w}_{debiased} = \vec{w} - \vec{w}_{bias}$$

“Hard debiasing”

- Requiere vocabulario específico para la dimensión a eliminar (por ejemplo un vocabulario para género, otro para religión, etc.)
- Elimina asociaciones que son útiles (ej: actress)
- Puede haber otras dimensiones que contengan información de género.

Semantics Derived Automatically from Language Corpora Contain Human-like Biases (Caliskan et al., Science 2017)

Crea un test basado en uno de psicologia: WEAT (Word Embedding Association Test). El test es un test de permutaciones entre 4 grupos.

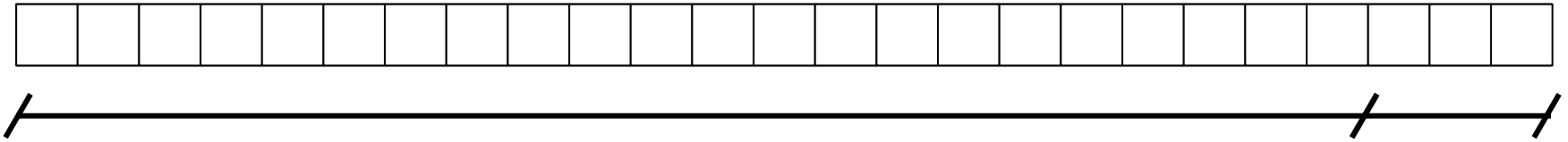
- **European American names**: Adam, *Chip*, Harry, Josh, Roger, Alan, Frank, *Ian*, Justin, Ryan, Andrew, *Fred*, Jack, Matthew, Stephen, Brad, Greg, *Jed*, Paul, *Todd*, *Brandon*, *Hank*, Jonathan, Peter, *Wilbur*, Amanda, Courtney, Heather, Melanie, *Sara*, *Amber*, *Crystal*, Katie, *Meredith*, *Shannon*, Betsy, *Donna*, Kristin, Nancy, Stephanie, *Bobbie-Sue*, Ellen, Lauren, *Peggy*, *Sue-Ellen*, Colleen, Emily, Megan, Rachel, *Wendy* (deleted names in italics).
- **African American names**: Alonzo, Jamel, *Lerone*, *Percell*, Theo, Alphonse, Jerome, Leroy, *Rasaan*, Torrance, Darnell, Lamar, Lionel, *Rashaun*, Tvree, Deion, Lamont, Malik, Terrence, Tyrone, *Everol*, Lavon, Marcellus, *Terryl*, Wardell, *Aiesha*, *Lashelle*, Nichelle, Shereen, *Temeka*, Ebony, Latisha, Shaniqua, *Tameisha*, *Teretha*, Jasmine, *Latonya*, *Shanise*, Tanisha, Tia, Lakisha, Latoya, *Sharise*, *Tashika*, Yolanda, *Lashandra*, Malika, *Shavonn*, *Tawanda*, Yvette (deleted names
- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison.

Semantics Derived Automatically from Language Corpora Contain Human-like Biases (Caliskan et al., Science 2017)

Hacen varios WEATs:

- Flowers vs Insects
- Musical instruments vs weapons
- European vs African names
- Male vs female names: Career or family
- Male vs female names: Math vs Arts

Learning Gender-Neutral Word Embeddings (Zhao et al., EMNLP 2018)



Embedding sin género

Embedding de género

$$J = J_G + \lambda_d J_D + \lambda_e J_E,$$

Learning Gender-Neutral Word Embeddings (Zhao et al., EMNLP 2018)

- Requiere vocabulario específico para cada demografía
- Requiere entrenar los embeddings desde el principio

Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them (Gonen & Goldberg, NAACL 2019)

Intenta probar que los métodos que eliminan la componente de género en realidad no lo hacen (GN-Glove y el de las proyecciones).

Toma la 2500 palabras más masculinas y femeninas y construye un clasificador antes y después del debiasing.

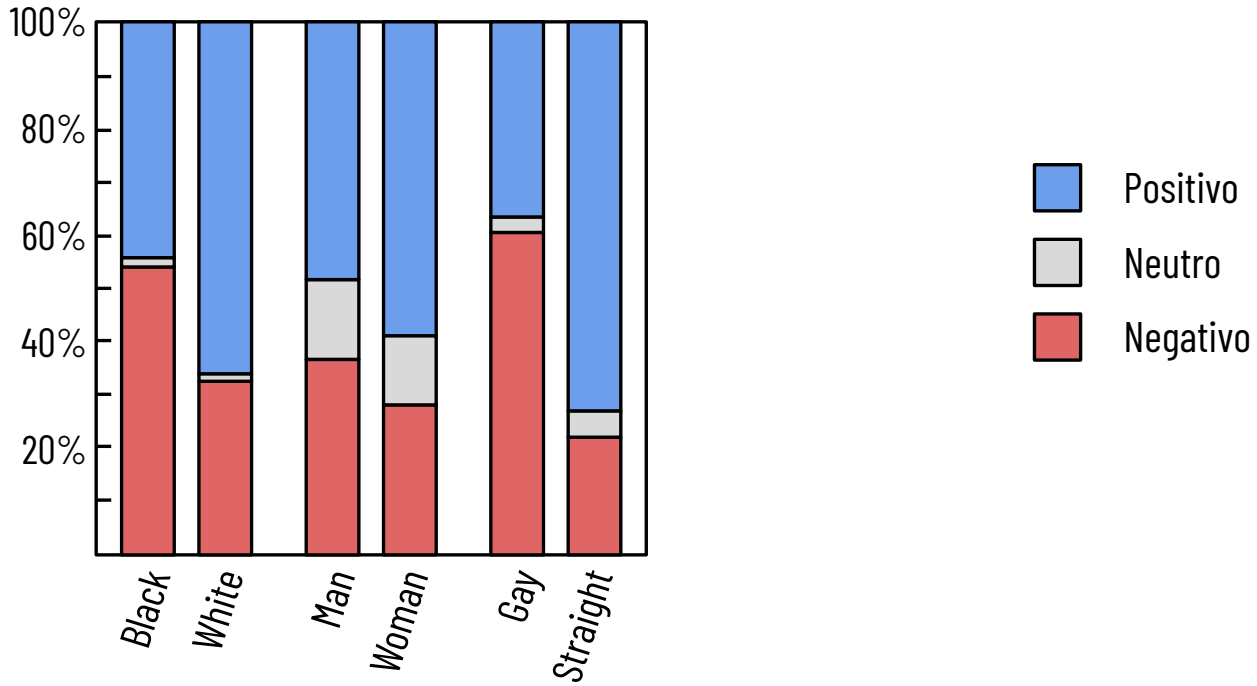
La accuracy por clustering después del debiasing es 0.99 para el método que elimina la proyección y 0.867 para GN-Glove. También se entrena una logistic regression.

El problema es que si bien sirve para probar que no se está aislando el género de las palabras, como métrica no sirve para nada porque entre esas 2500 palabras hay sesgos que no son nocivos (por ejemplo “bra” como femenino)

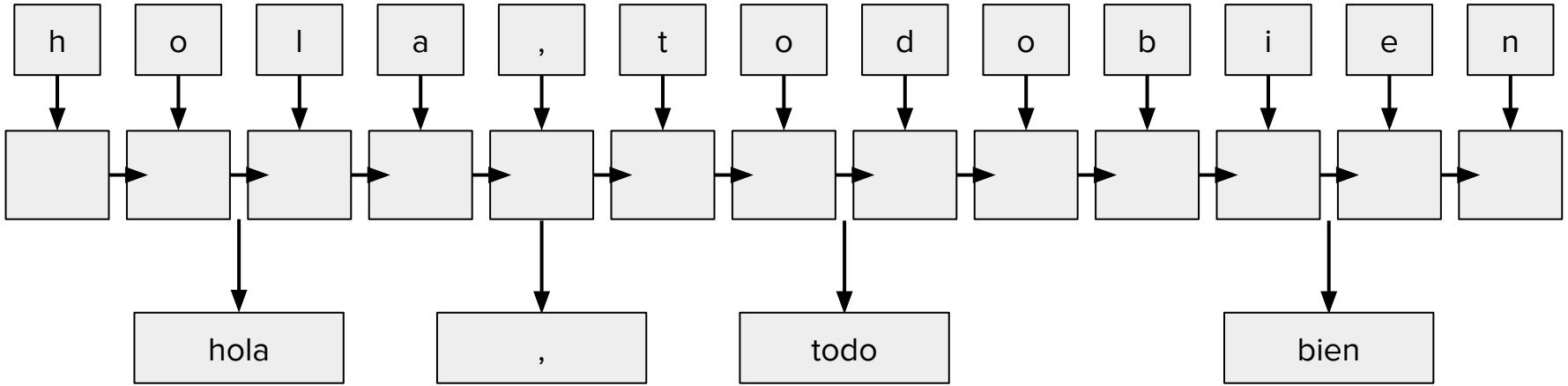
The Woman Worked as a Babysitter: On Biases in Language Generation (Sheng et al., EMNLP-IJCNLP 2019)

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

The Woman Worked as a Babysitter: On Biases in Language Generation (Sheng et al., EMNLP-IJCNLP 2019)



¿Qué tiene que ver una LLM con word embeddings?



Propuesta

Hipótesis

Si podemos extraer la dirección de género de pares de palabras (he-she, his-her, etc.) y además podemos medir los sesgos sobre otras demografías usando nombres, **¿por qué no podríamos obtener todas las dimensiones sensibles por medio de los nombres en general?**

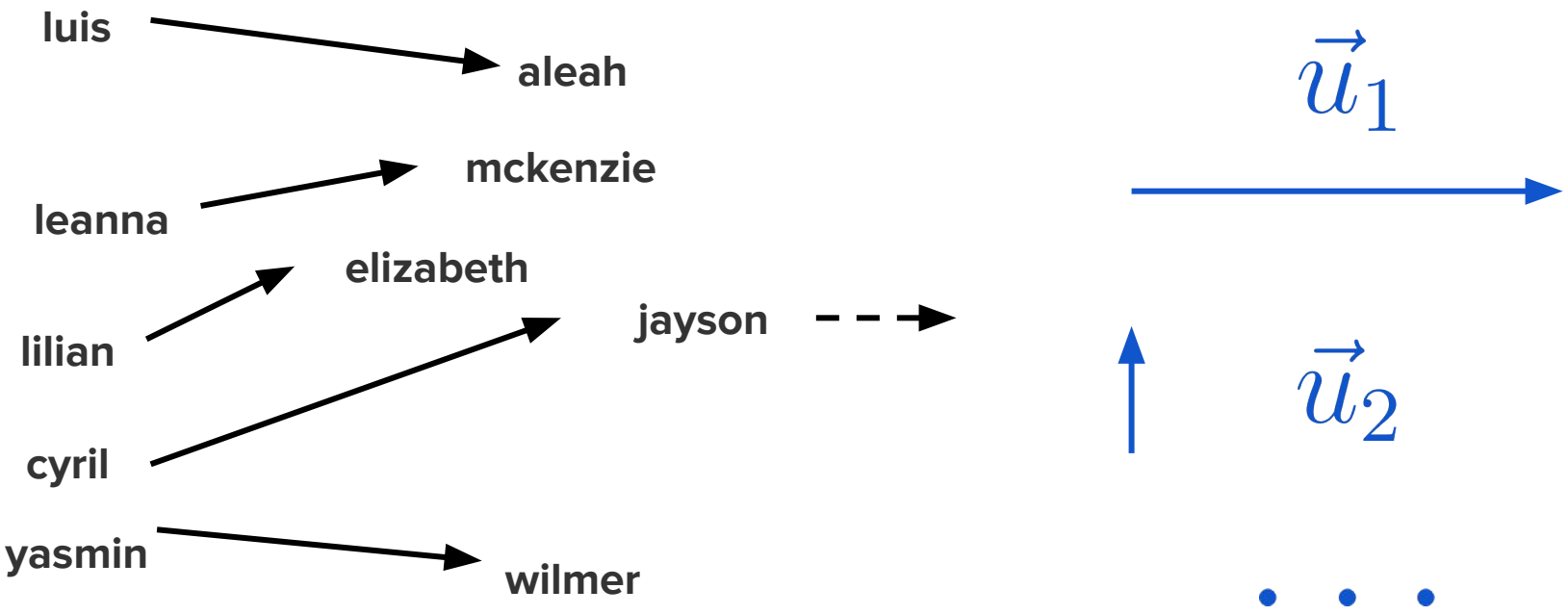
Metodología

Tomaremos miles de nombres frecuentes y sus restas entre sí.

De allí extraemos las componentes principales con PCA y eliminamos la proyección de todos los embeddings sobre las componentes que acumulan 35% de varianza explicada de una forma similar al “hard debiasing”.

Este método no requiere reentrenar los embeddings y no necesita vocabularios particulares para cada grupo a insesgar.

Metodología



Nuevos WEATs

Necesitamos medir los resultados en tests que no dependan de nombres, vamos a definir nuevos tests:

- Western vs. asian associations
- Latin American vs. Anglo-American Cultural Terms
- Heteronormative vs. Queer Associations
- Young vs Old Associations
- Christian vs. Muslim
- Caucasian vs. Black

Nuevos WEATs: Ejemplo

Group 1: caucasian, opera, french, german, italian, swiss, christmas, rock, latin, Michelangelo, Alps, Madonna, zulu, celtic, bard, slavic, folk, Kennedy, Siberia, prairie, druid

Group 2: black, gospel, creole, haitian, jamaican, Gullah, kwanzaa, hip-hop, ebonics, Basquiat, Appalachians, Beyoncé, viking, igbo, maasai, Mandinka, blues, Obama, Sahara, savanna, shaman

Resultados

Resultados

	Wordsim	Simlex	Rarewords	Card660	SimVerb
Glove (vanilla)	0.80	0.44	0.45	0.53	0.29
Gender Hard-Debias	0.80	0.44	0.45	0.53	0.29
GN-GloVe	0.72	0.38	0.39	0.44	0.22
Name-based SVD debiasing (ours)	**0.81**	**0.49**	**0.51**	**0.61**	**0.35**

Table 1: **Embedding quality**. Results of the word similarity benchmarks for our model, as compared previous ones in the literature. The values represent Pearson correlations between the human similarity scores and the cosine similarity of the word embeddings.

Resultados

	GloVe (vanilla)	Gender Hard-Debias	GN-GloVe	Name-based SVD Debiasing (ours)
Gender: Math vs. Arts	0.20	**0.02**	0.16	0.03
Gender: Science vs. Arts	0.35	-0.01	0.27	**0.00**
Western vs. Asian Associations	0.29	0.31	**0.13**	0.27
Latin American vs. Anglo-American Cultural Terms	0.38	0.41	0.53	**0.03**
Heteronormative vs. Queer Associations	0.33	0.34	**0.11**	0.26
Young vs. Old Associations	0.26	0.24	0.23	**-0.08**
Christian vs. Muslim	1.21	1.26	0.90	**0.43**
Caucasian vs. Black	0.80	0.81	**0.22**	0.46

Table 5: **Name-independent WEATs**. Statistics for the name-independent WEAT tests. Values close to zero ensure that the biases have been removed. We highlight in bold the best method for each test.

Resultados

	Gender clustering accuracy	Gender classifier accuracy	Gender-Professions neighbor correlation
Glove	0.989	0.999	0.800
Gender Hard-Debias	0.914	0.975	0.713
GN-GloVe	**0.866**	0.998	0.786
Name-based SVD Debiasing	0.945	**0.973**	**0.701**

Table 6: **Preservation of gender information**. In the classification experiments (first two column) a higher accuracy indicates larger prevalence of gender information in the embeddings after debiasing. In the gender-professions neighbor correlation experiment (last column) higher values indicate larger harmful gender bias over profession embeddings. We highlight in bold the best method for each test.

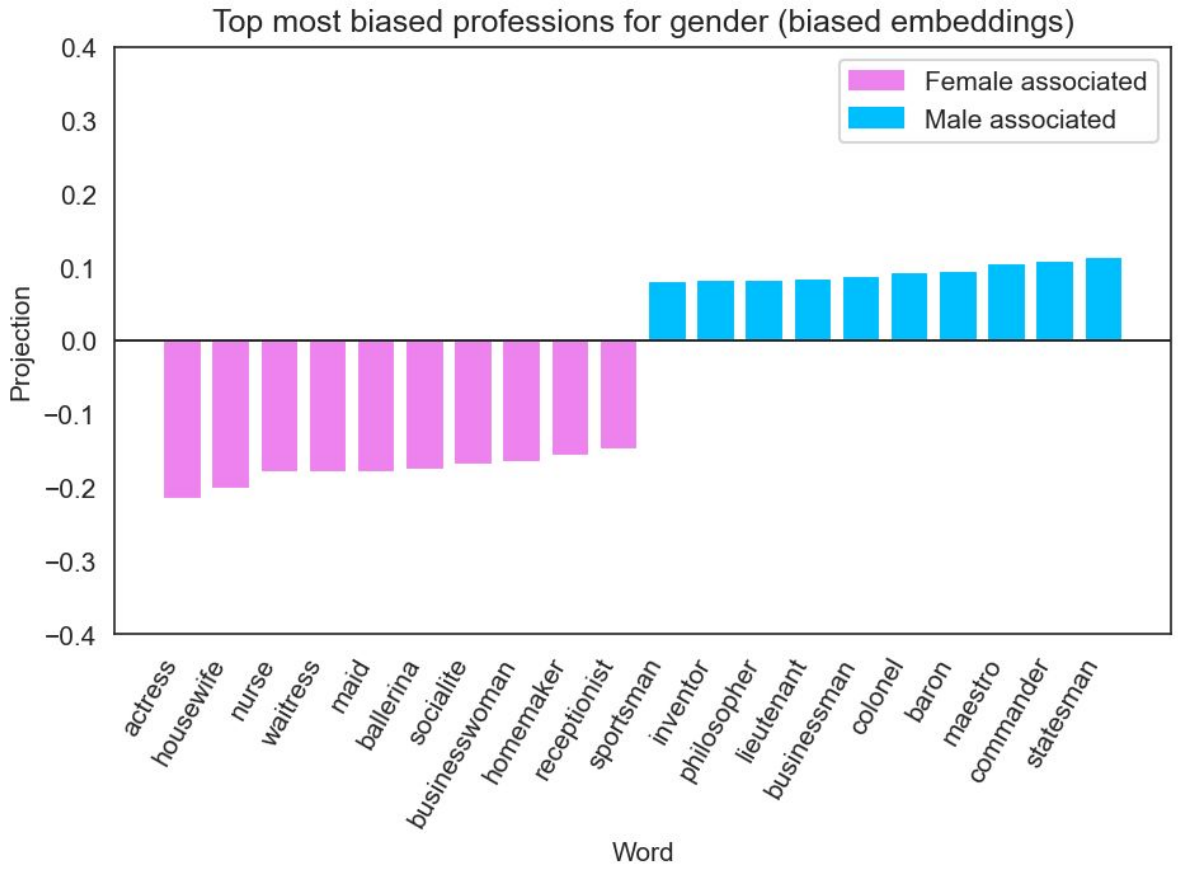
Cambios para distintas demografías

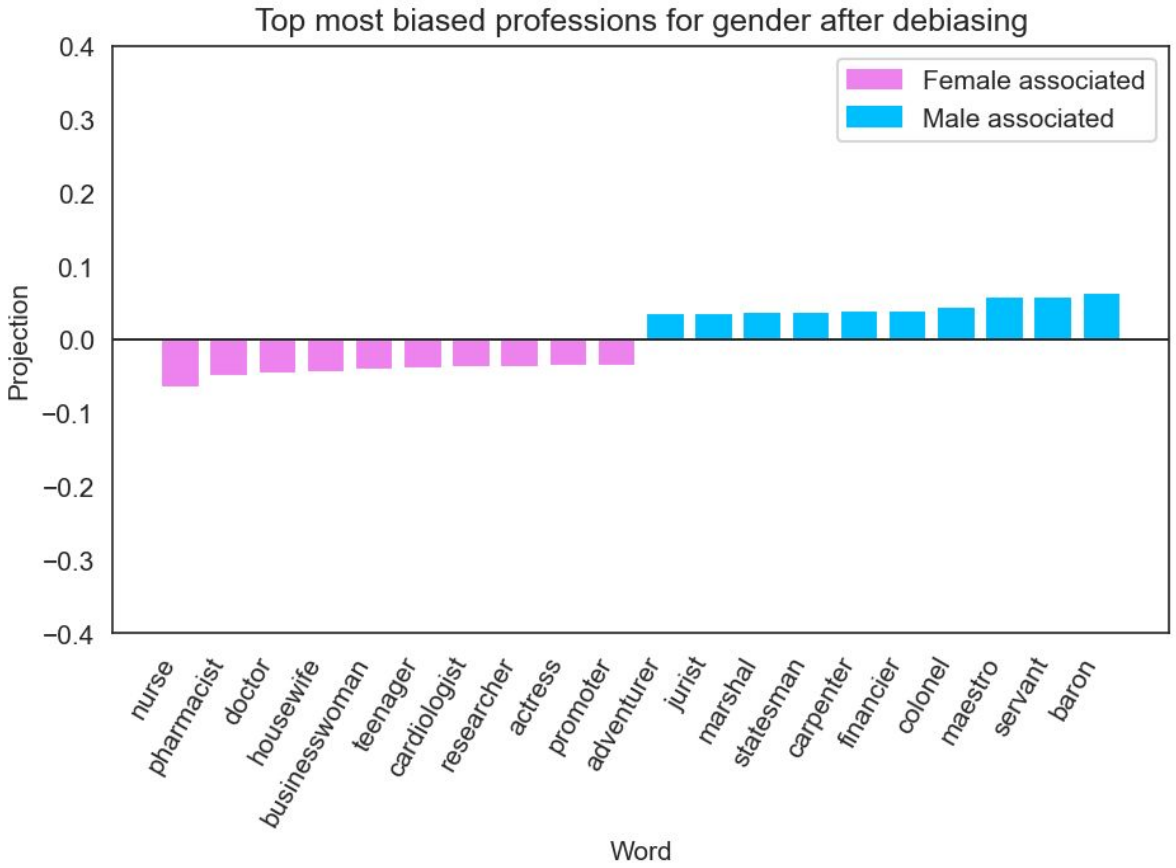
Podemos usar las palabras de los WEATs creados con el producto cartesiano de las restas para generar una componente para cada grupo.

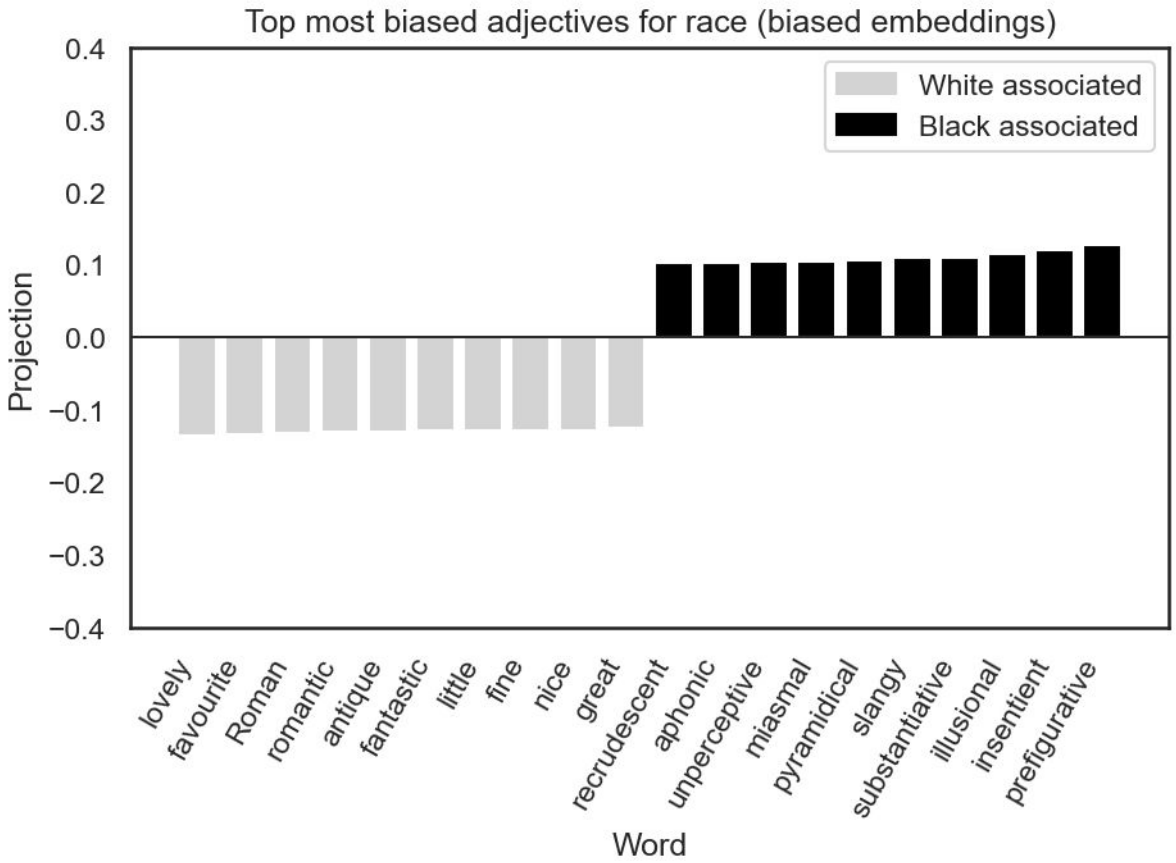
Con esta componente podemos estudiar los efectos de nuestro método sobre cada población tomando un set de profesiones y de adjetivos.

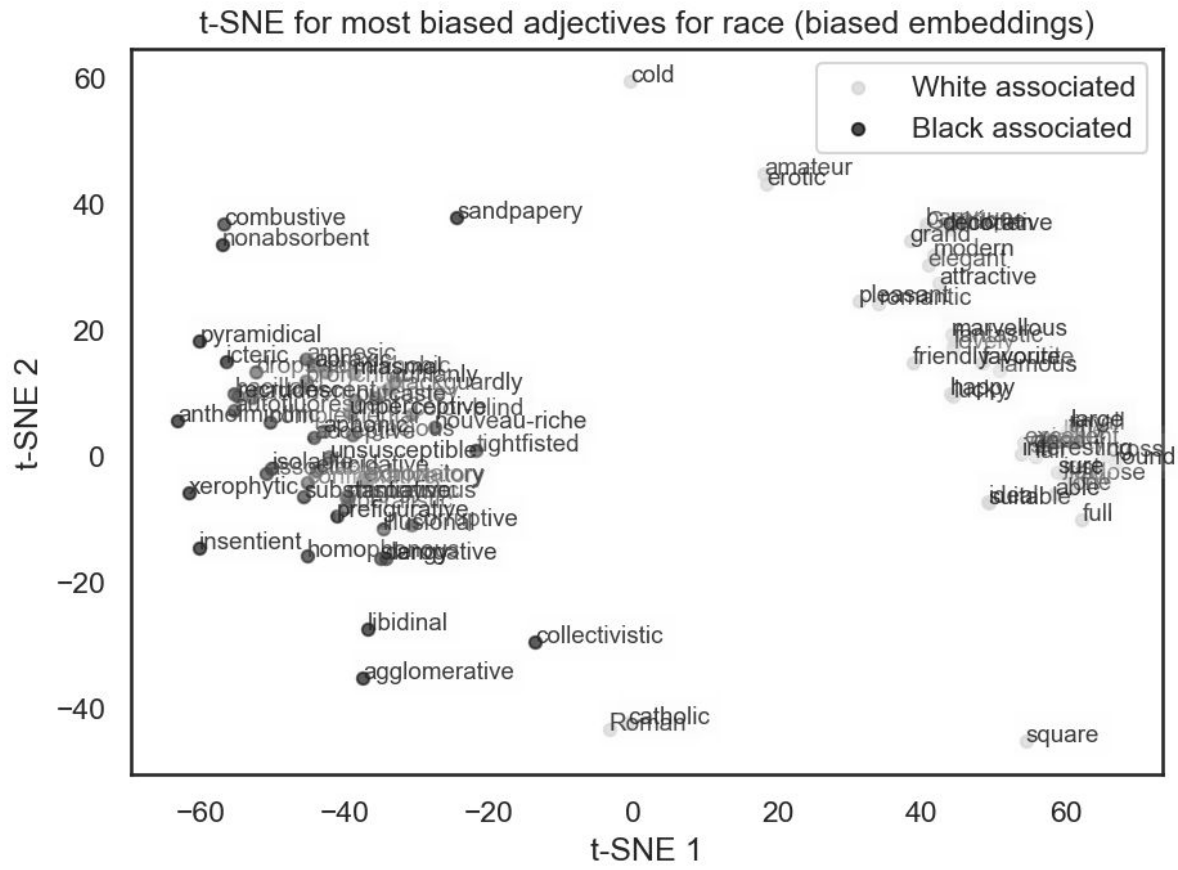
Vamos a tomar **3 grupos**:

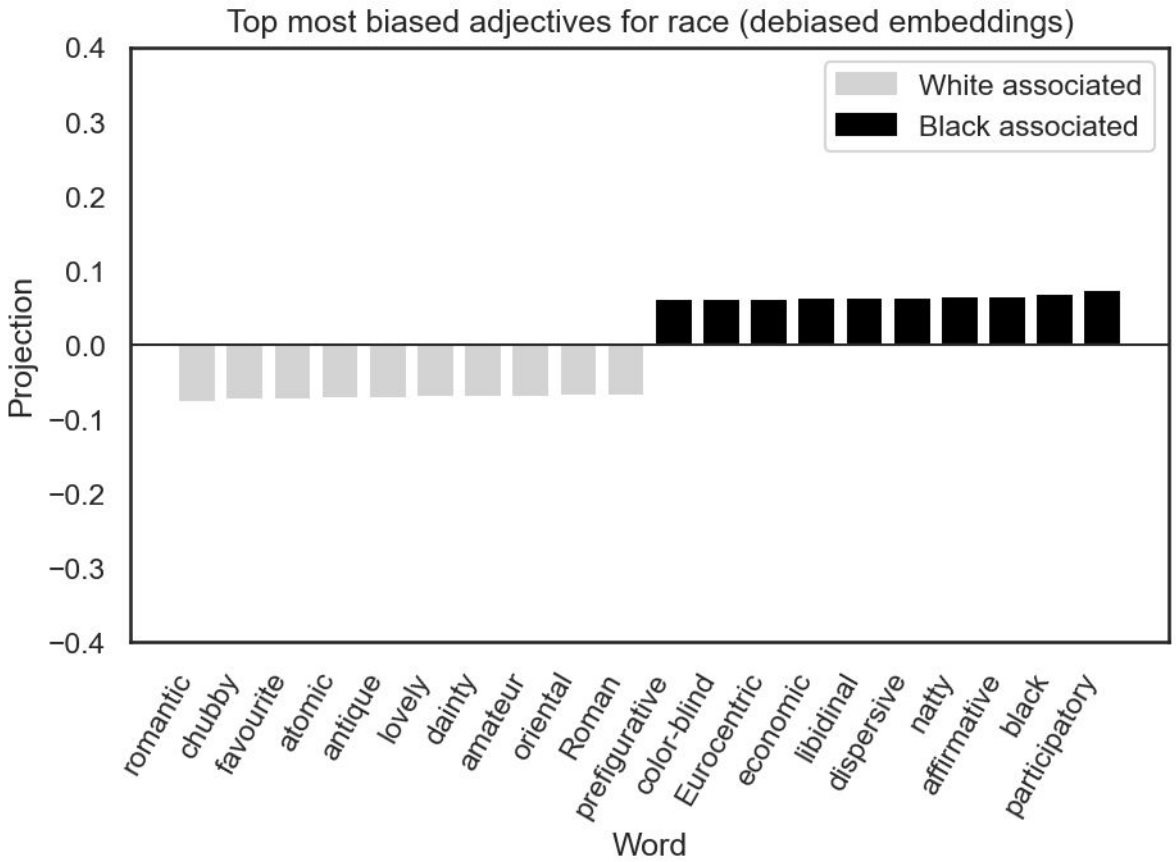
- Hombre y mujer
- Blanco y negro
- Musulmanes y cristianos

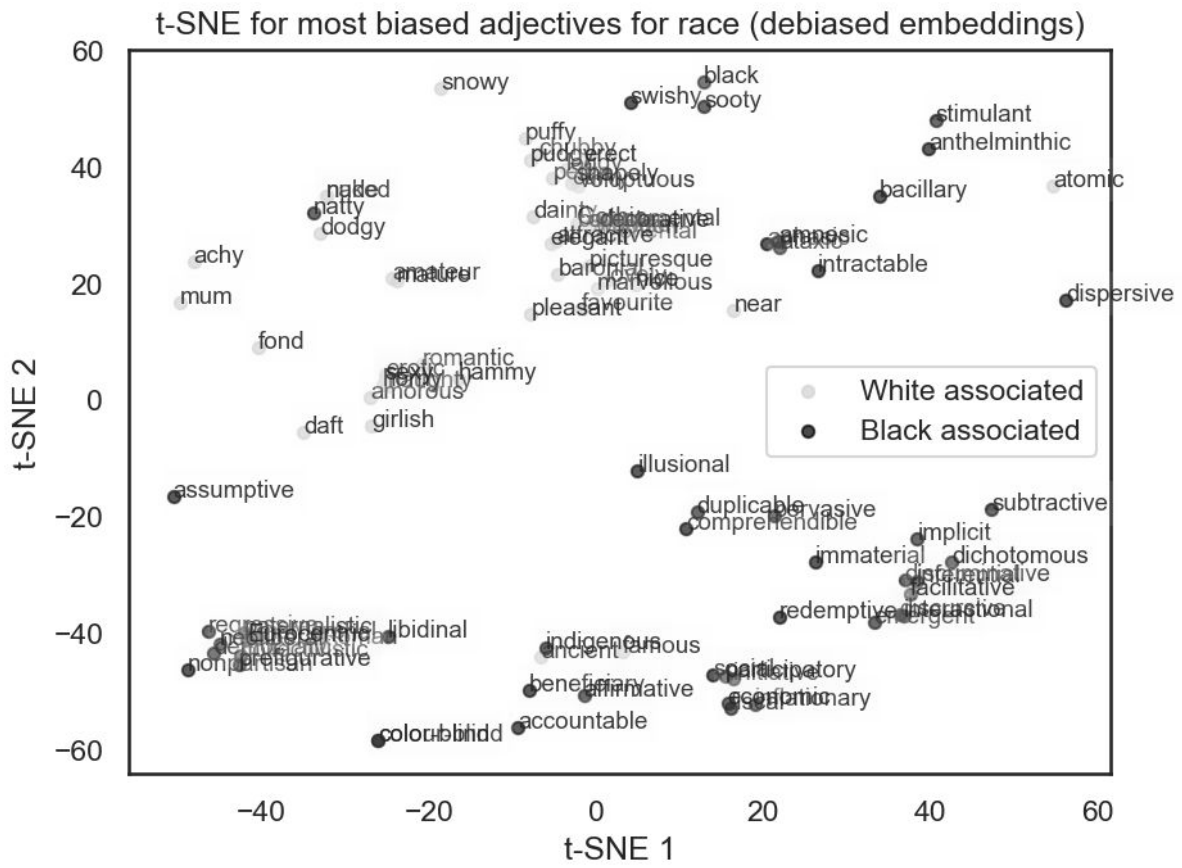

Top most biased professions for gender (biased embeddings)
Female associated
Male associated
Projection
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
actress
housewife
nurse
waitress
maid
ballerina
socialite
businesswoman
homemaker
receptionist
sportsman
inventor
philosopher
lieutenant
businessman
colonel
baron
maestro
commander
statesman
Word

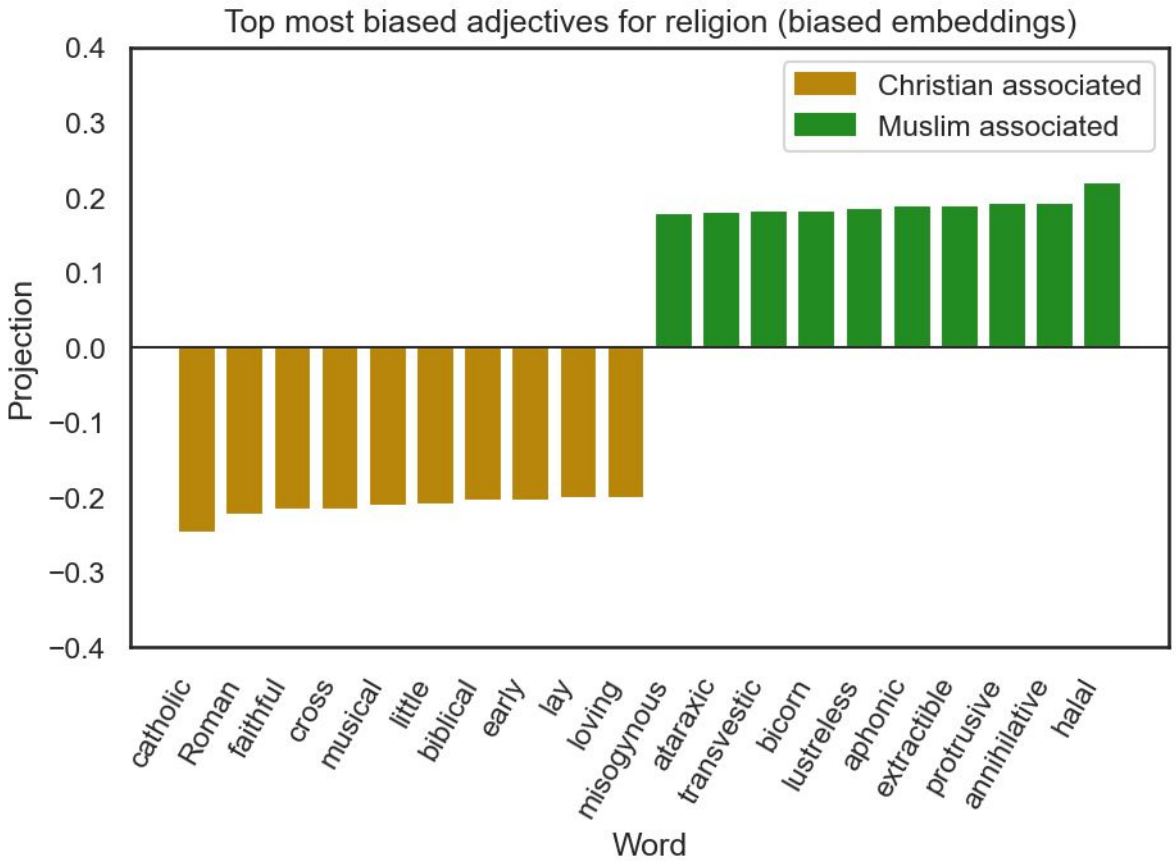

Top most biased professions for gender after debiasing
Female associated
Male associated
Projection
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
nurse
pharmacist
doctor
housewife
businesswoman
teenager
cardiologist
researcher
actress
promoter
adventurer
jurist
marshal
statesman
carpenter
financier
colonel
maestro
servant
baron
Word

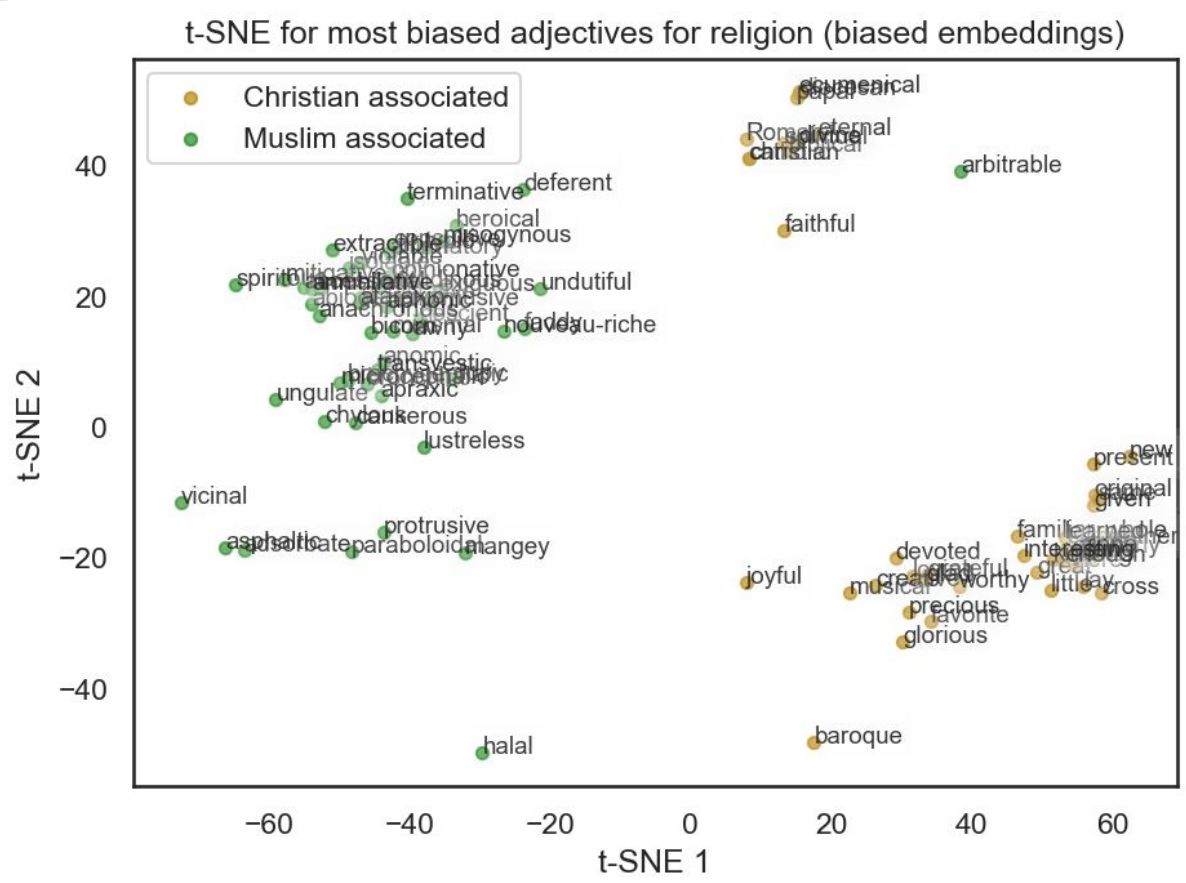

Top most biased adjectives for race (biased embeddings)
White associated
Black associated
Projection
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
lovely
favourite
Roman
romantic
antique
fantastic
little
fine
nice
great
recrudescent
aphonic
unperceptive
miasmal
pyramidical
slangy
substantiative
illusional
insentient
prefigurative
Word

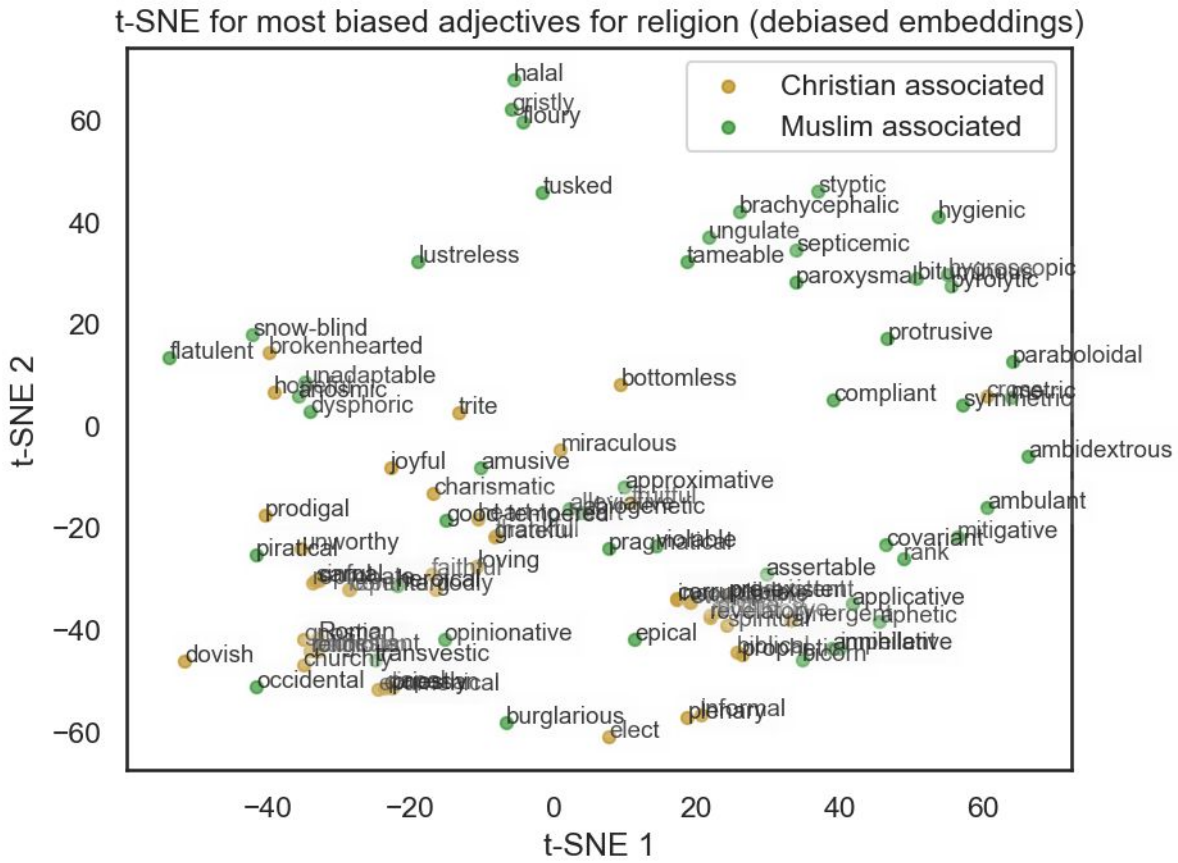

t-SNE for most biased adjectives for race (biased embeddings)
White associated
Black associated
cold
amateur
erotic
sandpapery
combustive
nonabsorbent
grand
modern
elegant
attractive
pleasant
pyramidical
marvellous
friendly
favorite
famous
happy
lucky
unperceptive
nouveau-riche
tightfisted
unsusceptible
xerophytic
suitable
full
insentient
libidinal
collectivistic
agglomerative
Roman
catholic
square
t-SNE 1
t-SNE 2
60
40
20
0
−20
−40
−60
−40
−20
0
20
40
60


Top most biased adjectives for race (debiased embeddings)
White associated
Black associated
Projection
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
romantic
chubby
favourite
atomic
antique
lovely
dainty
amateur
oriental
Roman
prefigurative
color-blind
Eurocentric
economic
libidinal
dispersive
natty
affirmative
black
participatory
Word


t-SNE for most biased adjectives for race (debiased embeddings)
t-SNE 1
t-SNE 2
White associated
Black associated
snowy
black
swishy
sooty
stimulant
anthelminthic
puffy
atomic
bacillary
natty
dodgy
achy
picturesque
intractable
mum
dispersive
pleasant
near
fond
romantic
hammy
amorous
daft
girlish
illusional
assumptive
subtractive
implicit
immaterial
dichotomous
facilitative
redemptive
libidinal
indigenous
ancient
beneficiary
accountable
color-blind


Top most biased adjectives for religion (biased embeddings)
Christian associated
Muslim associated
Projection
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
catholic
Roman
faithful
cross
musical
little
biblical
early
lay
loving
misogynous
ataraxic
transvestic
bicorn
lustreless
aphonic
extractible
protrusive
annihilative
halal
Word


t-SNE for most biased adjectives for religion (biased embeddings)
Christian associated
Muslim associated
t-SNE 1
t-SNE 2
40
20
0
−20
−40
−60
−40
−20
0
20
40
60
ecumenical
papacy
eternal
divine
christian
arbitrable
faithful
deferent
terminative
heroical
misogynous
undutiful
bicorn
nouveau-riche
anomic
transvestic
apraxic
ungulate
lustreless
vicinal
protrusive
asphaltic
paraboloidal
mangey
new
present
original
given
devoted
joyful
musical
precious
favorite
glorious
worthy
cross
halal
baroque


t-SNE for most biased adjectives for religion (debiased embeddings)
Christian associated
Muslim associated
t-SNE 1
t-SNE 2
halal
gristly
joury
tusked
styptic
brachycephalic
hygienic
ungulate
septicemic
lustreless
tameable
paroxysmal
pyrolytic
snow-blind
protrusive
flatulent
brokenhearted
paraboloidal
unadaptable
bottomless
compliant
symmetric
anosmic
dysphoric
trite
miraculous
ambidextrous
joyful
amusive
charismatic
approximative
prodigal
ambulant
piratical
unworthy
mitigative
covariant
rank
loving
assertable
applicative
opinionative
epical
dovish
transvestic
churchly
occidental
burglarious
elect
plenary
informal

Conclusiones

Conclusiones

Nuestro método:

- Funciona para múltiples grupos al mismo tiempo
- No requiere un vocabulario específico para cada par de grupos
- No requiere saber cuales son los grupos a tratar
- Se puede hacer después del entrenamiento
- Los embeddings no pierden calidad
- Tiene mejor resultado que los métodos previos

Preguntas?