El rol de la IA en la sociedad

Sofía Trejo

El objetivo de esta sesión es hablar de los sesgos presentes IA y de los problemas sociales producto de los mismos.

Sesgos algorítmicos
algorithmic bias

Algoritmo
Es un juego de instrucciones que le dicen a una computadora como ejecutar determinada tarea. El algoritmos pueden ser comparados con una receta, ya que toman ciertos ingredientes y los transforma a través de pasos explicables en un resultado predecible. Usando la analogía de la receta podemos ver que los algoritmos pueden tener varios tipos de sesgos. Ya que la persona que cocina decide los ingredientes, y trabaja bajo su definición de cuál es un buen platillo. Sin embargo, otra persona puede no estar de acuerdo que el platillo es algo bueno para preparar, o que es asqueroso. En resumen, los algoritmos son opiniones transformadas en código.

El daño generado por algoritmos puede venir de distintas fuentes:

- El contexto social en el que se desarrollan los algoritmos : los algoritmos reflejan los valores y preferencias de su creador.
- Las restricciones técnicas: qué variables se consideran, datos que son usados para entrenar etc.
- La forma en que son usados en la práctica: un algoritmo puede ser usado para determinar qué tipo de ayuda social necesita una persona o para determinar los términos de su hipoteca.

Un serio problema es que los algoritmos y los datos usados para entrenarlos no son accesibles al público. Por lo tanto, es muy complicado saber si el algoritmo funciona correctamente o si está discriminando de manera sistemática a parte de la población. **Hay una falta gigantesca de auditoría y transparencia en el área de IA.**

Sesgos en los datos

Hay dos problemáticas principales referentes a los datos usados para entrenar IA.

Los datos no son representativos de la realidad

Uno de los grandes problemas con las tecnologías de datos es que se usan como proxies de las necesidades sociales. Cuando muchas personas en el mundo, sobre todo las de escasos recursos, no tienen acceso a dichas tecnologías. Esto significa que estas tecnologías tienen el potencial de exacerbar las desigualdades sociales ya existentes. Cuando se consideran bases de datos se debe considerar:

- ¿Qué personas son excluidas?
- ¿Qué lugares son visibles?
- ¿Qué pasa con las personas que no están el los datos?

Esto significa que para crear estos sistemas se deberían complementar los datos con estudios cualitativos rigurosos. Metodologías de ciencias sociales deben ser utilizadas para dar sentido a los datos y para entender el contexto de los mismos.

Genero y raza

Joy Buolamwini

Activista digital. Poet of Code.

Plática de TED sobre bias en algoritmos tiene más de un millón de visitas.

Fundadora de *Algorithmic Justice League* para luchar contra los sesgos en IA.

Gender shades

Elaborado en 2017, proyecto de Doctorado para MIT.

Difusión en 230 artículos en 37 países.

Data set: más de 1000 imágenes de políticos seleccionados de acuerdo a la representación de mujeres en el gobierno.

Rwanda 61.3% *
Bolivia 53.1 %
Cuba 48.9%
Islandia 47.9% *
Nicaragua 45.7%
Secia 43.6% *
Senegal 42.7 % *
México 42.6%
Finlandia 42.0% *
Sudafrica 41.5% *

Compañías:

- Microsoft
- Face ++ (Data set Chino)
- IBM

Puntaje general:

- Microsoft: 93.7%,
- Face ++ 90%
- IBM 87.9%

Diferencia de errores:

Hombres vs mujeres:

- Microsoft: 8.1%
- Face ++: 20.6%
- IBM : 14.7%

Raza:

- Microsoft: 12.2%
- Face ++: 11.8%
- IBM: 19.2%

Grupo con mayor bias:

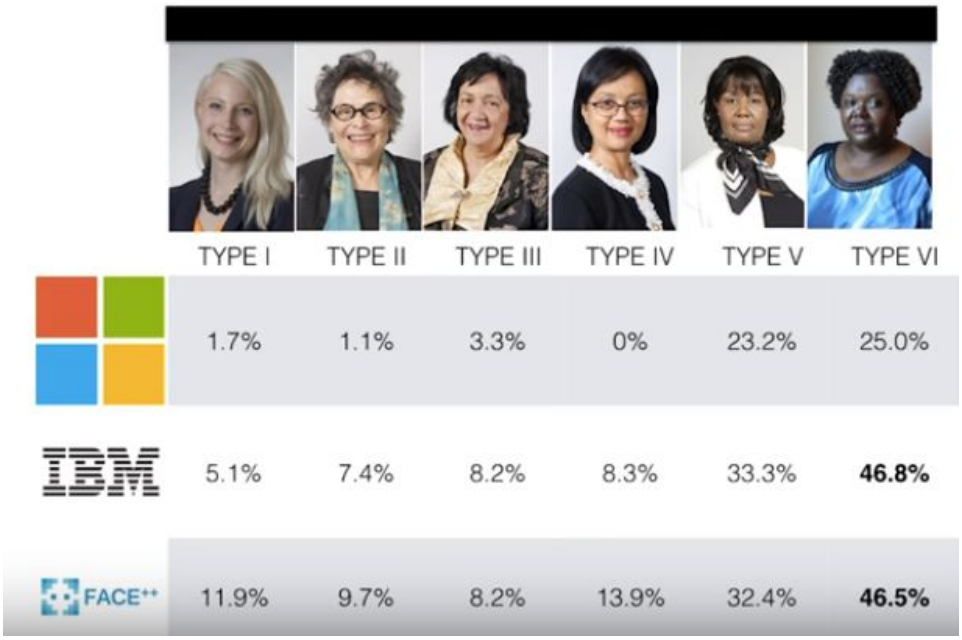
Mujeres negras

- 79.2%
- 66.5%
- 65.3%

Grupo con mejor puntaje:
Hombres blancos

- 100%
- 99.2%
- 99.7%

Muchos productos en línea fallan en 1 de 3 mujeres negras.
Mujeres de piel más oscura son clasificadas como hombres con una gran probabilidad.



Zonas urbanas privilegiadas

Streetbump

Boston tenía un problema de baches, arreglando aproximadamente 20,000 cada año. Para economizar recursos la ciudad hizo la app StreetBump, que registra la localización y la aceleración y ayuda a detectar baches de manera automática. El problema es que personas con bajos recursos no tienen el mismo acceso a tecnología, lo mismo que personas de la tercera edad donde el porcentaje de personas con smartphones es tan bajo como el 16%. Esto significa que los datos recolectados están

incompletos. Este problema se presenta, en general, con personas de bajos recursos.
Afortunadamente, los responsables del proyecto estaban conscientes de este problema y trabajan con académicos para compensar el problema de acceso igualitario a tecnología.

Huracán Sandy

El huracán Sandy generó más de 20 millones de tweets, entre el 27 de octubre y el 1 de noviembre. La mayor parte de dichos Tweets provenian de Manhattan, donde hay una gran concentración de smartphones y uso de Twitter. Pero esto da la impresión que Manhattan fue de los lugares más afectados por el huracán. Pocos mensajes se originaron en lugares que fueron severamente afectados como Breezy Point, Coney Island y Rockaway. Debido a falta de electricidad y falta de acceso a celulares. De hecho, había mucho ocurriendo fuera de las zonas urbanas privilegiadas que no aparece en los datos.

Sesgos en los modelos

Los atributos que se eligen tienen un impacto medible en la exactitud (accuracy) del sistema, lo que no es sencillo medir es la cantidad de sesgo generada por cada elección.

Puntajes de riesgo EUA

ProPublica

Es una organización sin fines de lucro (en NY) que tiene la finalidad de producir periodismo investigativo. En 2010 se convirtió en la primera publicación en línea en ganar un Pulitzer. Ha ganado más Pulitzers desde entonces, incluyendo uno en el 2019.

Mayo 2016 publicó el artículo: Machine Bias.

Puntajes de riesgo (risk scores): puntajes que intentan predecir qué tan probable es que una persona cometa nuevos crímenes.

Estos puntajes son usados para decidir qué criminales serán libe-

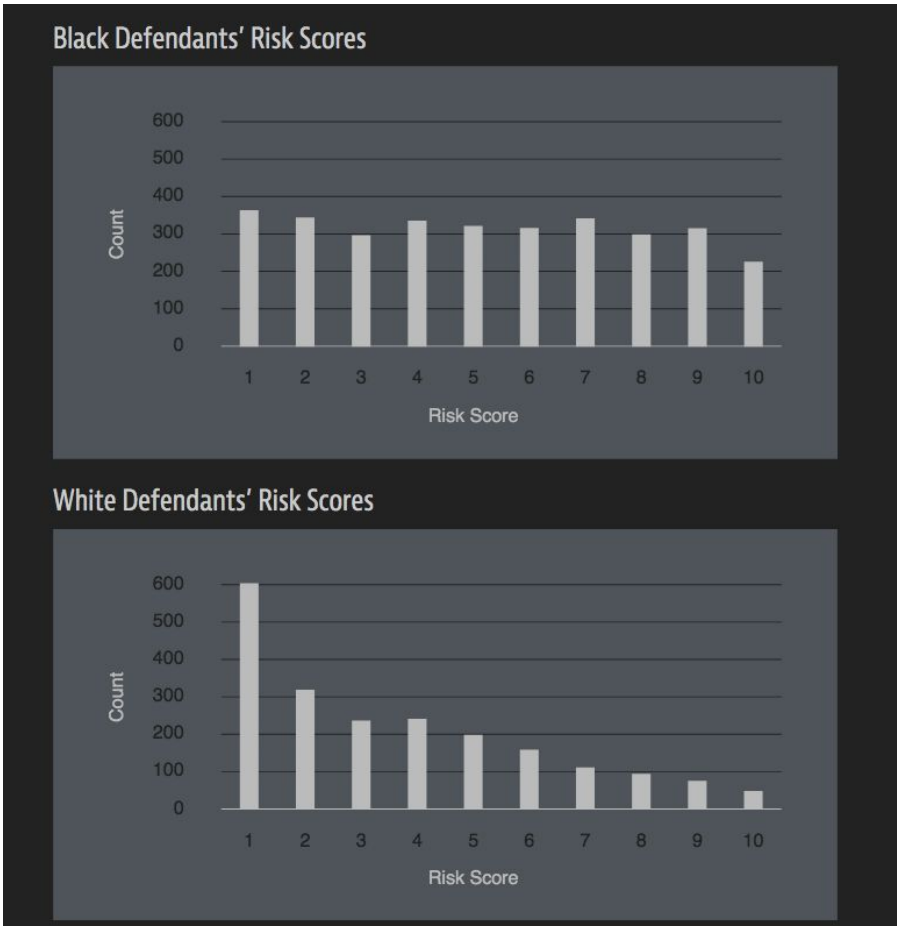
rados. Y para tomar decisiones a la largo de todo el proceso legal. Se supone que uno de sus mayores aplicaciones es para decidir el programa de rehabilitación más adecuado para la persona. No para dar sentencias de cárcel más largas.

En el 2016 ProPublica investigó el sistema COMPAS desarrollado por Northpointe que es usado en varios lugares en EUA. En particular en NY el sistema se comenzó a usar de manera piloto en el 2001 para personas en libertad condicional. Para el 2010 el sistema era implementado en el estado de manera oficial (excepto en NYC). Sin embargo, no publicaron ningún estudio estadístico comprensivo hasta el 2012.

El sistema tiene una exactitud de predicción de alrededor del 60%

ProPublica obtuvo los puntajes de más de 7,000 personas arrestadas en Florida entre 2013 y 2014 e investigaron cuantos fueron culpados por nuevos crímenes en un periodo de dos años, los mismos parámetros usados por los creadores del algoritmo.

- Al algoritmo es muy malo en predecir crímenes violentos: sólo el 20% de las personas quienes se esperaba que cometieron dichos crímenes lo hicieron.
- Sólo el 61% de las personas a quienes se les calificó como futuros criminales con alta probabilidad fueron arrestados dentro del periodo de 2 años.
- El sistema era propenso a etiquetar a personas de color como futuros criminales, con el doble de probabilidad que a personas blancas.
- Personas blancas eran incorrectamente etiquetadas como de bajo riesgo más que las personas de color.



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

El sistema funciona en base a 137 preguntas respondidas por el acusado o sacadas de su expediente. Las preguntas incluyen:

- ¿Alguno de tus padres ha sido enviado a prisión?
- ¿Cuántos de tus familiares/amigos consumen drogas de manera ilegal?
- Estas en acuerdo o en desacuerdo con: una persona con hambre tiene el derecho de robar.

De Acuerdo con el sistema es posible que un pederasta sea categorizado como de bajo riesgo porque tiene un trabajo fijo. Cuando una persona bebiendo en vía pública puede ser categorizada como de alto riesgo por vivir en la calle.

Follow up

Los creadores de COMPAS establecieron que crearon el sistema para predecir con la misma exactitud, alrededor del 60%, los puntajes de personas blancas y de color. Argumentando, que **si un test es correcto en la misma proporción para todos los grupos no puede estar sesgado.**

Medidas de justicia

COMPAS produce el puntaje de acuerdo a un cuestionario que explora el pasado criminal del acusado y sus actitudes ante el crimen. ¿Cómo es que esto produce sesgos?

Opciones al crear el sistema:

- Optimizar los positivos reales: identificar al mayor número de personas posibles quienes tienen un riesgo alto de cometer otro crimen. Un problema con este método es que tiene a incrementar los falsos positivos: personas injustamente clasificadas como futuros criminales.
- Minimizar los falsos positivos: en este caso el sistema crea más falsos negativos. I.e. personas con alto riesgo de cometer otro crimen que son tratadas como de bajo riesgo.

ProPublica comparó falsos positivos y falsos negativos entre distintas razas, y descubrió una preferencia hacia los blancos.

Northpointe comparó el PPV para distintas razas y las encontró similares. Esto ocurren en parte por que los scores de distintas razas son diferentes, entonces es matemáticamente probable que la proporción de positivos sea similar en todas las razas mientras que las los falsos negativos difieran.

Uno de los problemas

Política del departamento de policía de NY tiene la política de stop-and-frisk.

Entre enero 2004 y junio 2012 NYCP detuvo a 4.4 millones de personas. De estas detenciones:

- 88% no produjeron a ningún cargo.
- 83% fueron a personas de color o hispanos.

La pregunta hecha por académicos de Stanford, Cornell, Harvard, Google, entre otros era:

> *Ya que las personas de color son arrestadas con mayor frecuencia, ¿es posible crear una fórmula que predice igualitariamente todas las razas sin causar disparidad en quien sufre daño por predicciones incorrectas?*

Conclusión

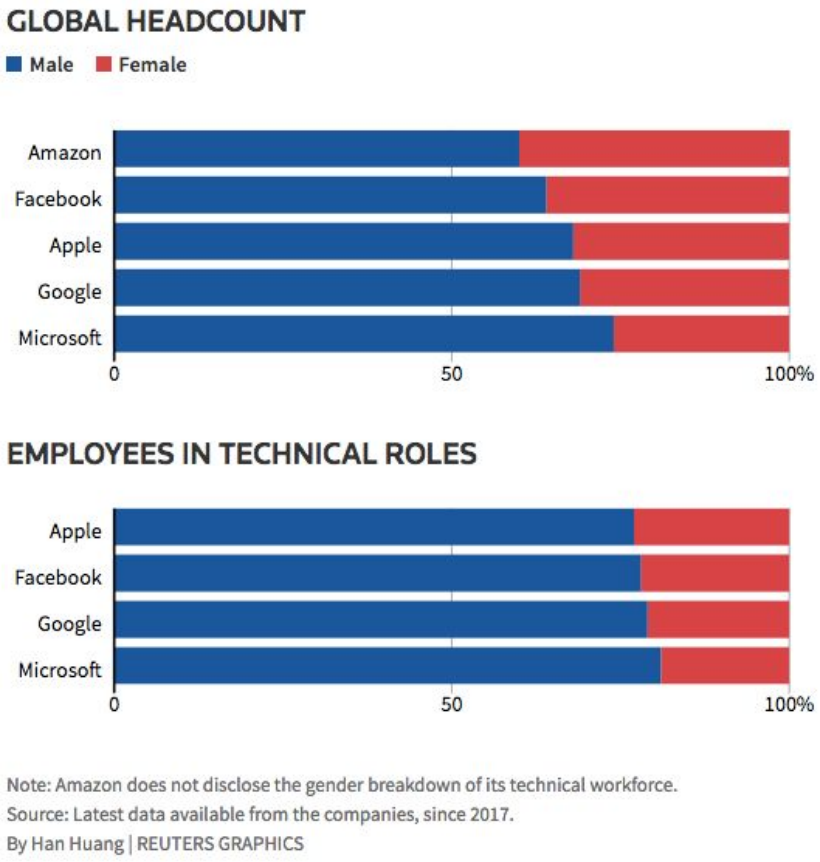
Es imposible crear un sistema que satisfaga las dos propiedades a la vez.

Los sistemas aprenden sesgos

Discriminación de género

En el 2014 Amazon comenzó a trabajar en la creación de un sistema para evaluar solicitudes de empleo. La idea era usar esta herramienta para evaluar a los candidatos y asignarles un ranking entre una y cinco estrellas.

En el 2015 la compañía se dió cuenta que el sistema no estaba evaluando a los candidatos de manera neutral respecto al sexo. En particular, las mujeres recibían bajas calificaciones para trabajos técnicos como Ingeniero de Software. El sistema había sido entrenado usando las aplicaciones a la compañía en los pasados 10 años. La mayoría de estas aplicaciones correspondían a hombres, un reflejo de la dominación masculina en el mundo del tech.



El sistema aprendió que los candidatos masculinos eran preferibles, discriminando aplicaciones donde aparecía la palabra "mujer" o "femenino" (equipo femenino de volleyball).

Amazon intentó arreglar el problema de Bias de género removiendo referencias como woman, woman's. El grupo creó 500 modelos y les enseñó a reconocer alrededor de 50,000 términos usados en aplicaciones anteriores. Cada sistema enfocado en distintas características laborales y localizaciones. Los algoritmos asignaron poco valor a habilidades que eran comunes en aplicaciones a IT, como la habilidad de escribir en varios lenguajes de programación. En lugar de ello, favorecian CV's usaban palabras como "ejecutar" y "capturar" que eran usadas generalmente por hombres.

Amazon terminó por desmantelar el equipo de investigación.

¿Qué tan serio es el problema?

Estudios realizados por compañías como CareerBuilder indican que alrededor del 55% de los departamentos de recursos humanos en EU tienen planeado usar IA en los próximos 5 años.

Problemas con modelos de encaje de palabras

Estos modelos convierten entradas de texto en vectores numéricos. En el proceso mapean palabras de manera semántica, lo que quiere decir que palabras similares son mapeadas unas cerca de las otras.

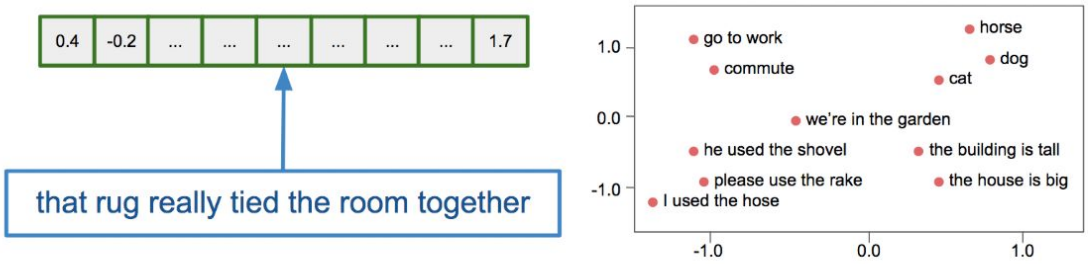


Figure 2: Text embeddings convert any text into a vector of numbers (left). Semantically similar pieces of text are mapped nearby each other in the embedding space (right).

Una vez que se ha entrenado el modelo, se puede medir la asociación entre palabras y frases. Se espera que estas asociaciones sean útiles para desarrollar PLN. Sin embargo, muchas de estas asociaciones pueden resultar nocivas. Por ejemplo, en *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* por T. Bolukbasi, K-Wi Chang, J. Zou, V. Saligrama, A. Kalai se descubrió que la relación entre "hombre" y "mujer" es similar a la relación entre "doctor" y "enfermera", o "tendero" (shopkeeper) y "ama de casa". Este popular modelo, llamado word2vec embedding, fue entrenado usando las Noticias de Google.

Formas de medición

Prueba WEAT: mide el grado en que un modelo asocia palabras con atributos.

Ejemplo:

Si se eligen como palabras: flores e insectos
Atributos: agradables (amor, paz)
Desagradables (feo, odio)

El puntaje refleja el grado en el cual las flores son asociadas con palabras agradables, relativo a los insectos.

Los puntajes están en un rango de [-2, 2]. Donde 2 es que las flores están mucho más asociadas a los atributos y -2 que los insectos lo están.

Las asociaciones son aprendidas de los datos usados para entrenar al modelo. Lo cual indica que los modelos están fortaleciendo sesgos humanos.

Targets (N)	Attributes (N)	GloVe*	word2vec	nnlm-en-dim50	nnlm-en-dim128	universal
Flowers vs Insects (25)	Pleasant vs Unpleasant (25)	**1.50***	**1.54***	**1.54***	**1.63***	**1.38***
Instruments vs Weapons (25)	Pleasant vs Unpleasant (25)	**1.53***	**1.63***	**1.66***	**1.55***	**1.44***
Eur-American vs Afr-American Names[6] (25)	Pleasant vs Unpleasant[6] (25)	**1.41***	**0.58***	**0.70***	0.04	0.36
Eur-American vs Afr-American Names[7] (18)	Pleasant vs Unpleasant[6] (25)	**1.50***	**1.24***	**1.04***	0.23	-0.37
Eur-American vs Afr-American Names[7] (18)	Pleasant vs Unpleasant[8] (8)	**1.28***	**0.72***	0.28	-0.09	0.72
Male vs Female names (8)	Career vs Family (8)	**1.81***	**1.89***	**1.45***	**1.70***	0.03
Math vs Arts (8)	Male vs Female (8)	1.06	0.97	**1.29***	1.07	0.59
Mental vs Physical Disease (6)	Temporary vs Permanent (7)	**1.38***	1.30	**1.35***	0.96	**1.60***
Science Arts (8)	Male vs Female (8)	**1.24***	**1.24***	**1.34***	1.19	0.24
Young vs Old Names (8)	Pleasant vs Unpleasant (8)	**1.21**	-0.08	0.75	-0.47	1.01

*Table 1: Word Embedding Association Test (WEAT) scores for different embedding models. Cell color indicates whether the direction of the measured bias is in line with (blue) or against (yellow) the common human biases recorded by the Implicit Association Tests. *Statistically significant ($p < 0.01$) using Caliskan et al. (2015) permutation test. Rows 3-5 are variations whose word lists come from [6], [7], and [8]. See Caliskan et al. for all word lists. * For GloVe, we follow Caliskan et al. and drop uncommon words from the word lists. All other analyses use the full word lists.*

Problemas con aplicaciones

Crear un chatbot. Usar un modelo para seleccionar una respuesta correcta (entre un grupo de respuestas).

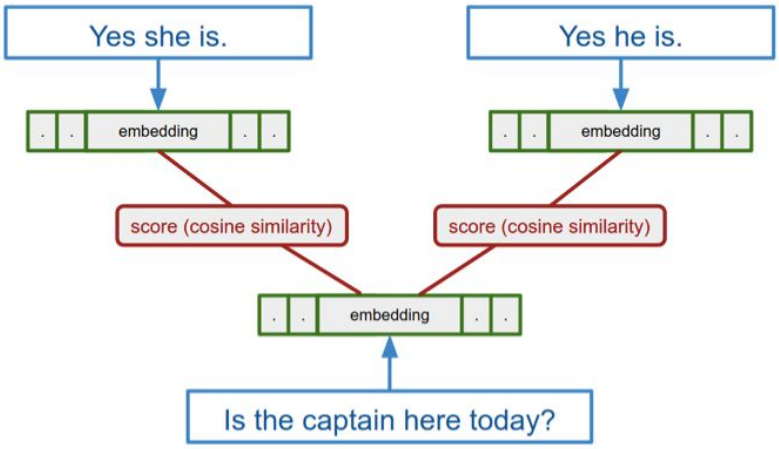
Se introduce la pregunta ---> se transforma en vector ---> se evalúa la distancia entre este vector y los vectores de las las respuestas usando la métrica en este espacio (cosine similarity).

Ejemplo de sesgo:

Pregunta: is the engineer available?
Respuestas: she is available
he is available

Sesgo: el sistema asume que los ingenieros son hombres.

Para medir el Sesgo:



	score
Yes she is	0.17
Yes he is	0.27
Bias score (difference)	-0.10

Para una ocupación, es sesgo es la suma de los sesgos para cada respuesta.

Highest female bias

occupation	bias	occupation	bias
maid	59.2	librarian	20.1
waitress	52.5	obstetrician	16.9
midwife	50.9	secretary	13.7
receptionist	50.2	socialite	12.1
nanny	47.7	therapist	10.2
nurse	45.4	manicurist	10.1
midwives	43.8	hairdresser	9.7
housekeeper	36.6	stylist	8.6
hostess	32	homemaker	6.9
gynecologist	31.6	planner	5.8

Highest male bias

occupation	bias	occupation	bias
undertaker	-73.4	captain	-53.4
janitor	-62.3	announcer	-51.1
referee	-60.7	architect	-50.7
plumber	-58	maestro	-50.6
actor	-56.9	drafter	-46.7
philosopher	-56.2	usher	-46.6
barber	-55.4	farmer	-45.4
umpire	-54.3	broadcaster	-45.2
president	-54	engineer	-45.1
coach	-53.8	magician	-44.8

Table 2: Occupations with the highest female-biased scores (left) and the highest male-biased scores (right).

Bibliografía:

- J. Buolamwini, *Gender Shades*, MIT media lab, https://www.media.mit.edu/projects/gender-shades/overview/.
- S. Danziger, J. Levav and L. Avnaim-Pesso, *Extraneous factors in judicial decisions*, National Academy of Sciences (2011).
- J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine bias*, ProPublica (Mayo 2016) https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
- J. Kleinberg, S. Mullainathan & M. Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, arxiv, (2016).
- J. Dastin, *Amazon scraps secret AI recruiting tool that showed bias against women,* Reuters, (octubre 2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-
- B Packer, Y Halpern, M Guajardo-Céspedes & M MitchelL, *Text Embedding Models Contain Bias. Here's Why That Matters*, Google Developers blog (2018), https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html
- T Bolukbasi, KW Chang, JY Zou, V Saligrama, AT Kalai, *Man is to computer programmer as woman is to homemaker? debiasing word embeddings*, Advances in neural information processing systems, 4349-4357 (2016).
- V. Sánchez Carmona, J. Mitchell, S. Riedel, *Behavior Analysis of NLI Models: Uncovering the Influence of Three Factors on Robustness*, NAACL (2018).
- A. Caliskan, J. Bryson, A. Narayananl, *Semantics derived automatically from language corpora contain human-like biases, Science* 356 : 183-186 (2017).