doi:10.5477/cis/reis.172.41

Sesgos de género ocultos en los macrodatos y revelados mediante redes neurales: ¿hombre es a mujer como trabajo es a madre?

Hidden Gender Bias in Big Data as Revealed Through Neural Networks: Man is to Woman as Work is to Mother?

Capitolina Díaz Martínez, Pablo Díaz García y Pablo Navarro Sustaeta

Palabras clave

Encaje de palabras
- Macrodatos
- Red neural
- Sesgos de género
- Wikipedia

Resumen

Los actos sociales se convierten en *big data*. El análisis de *big data* se convierte en conocimiento sobre la sociedad. Si los *big data* están sesgados, el sesgo se transmite al análisis y a nuestro conocimiento. Proponemos una herramienta para descubrir los sesgos de género y, potencialmente, eliminarlos de los *big data* antes del análisis. Utilizamos la técnica de análisis neural mediante el procedimiento de encaje de palabras. Es la primera vez que esta técnica se prueba con un cuerpo de datos en español. Como prueba de concepto, la red neural analiza la mitad de la Wikipedia en español. Más de 28 millones de palabras.

Se describen las técnicas y los conocimientos especializados necesarios para poder discernir los sesgos de género y se evalúa si es posible dividir el trabajo de análisis en microtareas externalizables.

Key words

Words Embedding
- Big Data
- Neural Network
- Gender Bias
- Wikipedia

Abstract

Social events become big data. The big data analysis becomes knowledge about society. If big data is biased, the bias is transmitted to the analysis and to our knowledge. We propose here a tool to discover gender biases and, potentially, eliminate them from big data before analysis. We use the neural network analysis and the words embedding. This is the first time that this technique is tested on a body of data in Spanish. As proof of concept, the neural network was fed with half of Wikipedia in Spanish. More than 28 million words.

We describe the techniques and specialized knowledge necessary to discern gender and it is evaluated whether it is possible to divide the analysis work into externalizable microtasks.

Cómo citar

Díaz Martínez, Capitolina; Díaz García, Pablo y Navarro Sustaeta, Pablo (2020). «Sesgos de género ocultos en los macrodatos y revelados mediante redes neurales: ¿hombre es a mujer como trabajo es a madre?». *Revista Española de Investigaciones Sociológicas,* 172: 41-60. (http://dx.doi.org/10.5477/cis/reis.172.41)

La versión en inglés de este artículo puede consultarse en http://reis.cis.es

Capitolina Díaz Martínez: Universidad de Valencia | capitolina.dm@gmail.com

Pablo Díaz García: Telice Comet | pablo.diaz.13@gmail.com

Pablo Navarro Sustaeta: Universidad de Valencia | Pablo.Navarro@uv.es

Introducción

Los macrodatos (*big data*) tienen el potencial de cambiar y mejorar nuestro mundo; sin embargo, frente a ellos nos encontramos, al menos, con dos tipos de problemas inmediatos. El primero es que disponer de estos conjuntos masivos de datos complejos, y manipularlos de manera epistémicamente productiva, no es simple (Navarro y Ariño, 2015) —y menos aún para las personas no especializadas—. Descubrir sus sesgos o taras es más difícil todavía. Esta dificultad suele provocar la desafección de las personas ajenas al campo de la informática. De ahí que los avances en este dominio hayan exigido la aparición de un nuevo campo del saber: Visualización de Información (*Information Visualization*). El objeto de esta subdisciplina es que la visualización de los *big data* los haga más compresibles y, consecuentemente, permita tomar decisiones mejor informadas sobre la base de tales datos. Esta mejor información redundaría en una reducción de la brecha social entre quienes pueden y no pueden interpretarlos. En este nuevo campo de la visualización de la información destaca, por ejemplo, Sheelagh Carpendale (Lam *et al.*, 2011), científica informática pionera en este ámbito y que, entre otras actividades, ha colaborado con la Organización Mundial de la Salud para hacer más asequible a la ciudadanía la presentación «visualizada» (en mejor castellano, «visibilizada») de su 11.ª Clasificación Internacional de Enfermedades (ICD11), de 2018. Relacionado con el problema de los sesgos en la visualización, está el caso de los sesgos en los *big data* de rostros humanos. En este ámbito destaca el trabajo de Joy Buolamwini y Timnit Gebru (2018) que han realizado una investigación con macrobases de datos de imágenes faciales estadounidenses. En esa investigación prueban cómo estas macrobases son menos sensibles a la hora de reconocer rostros de piel oscura y rostros de mujer.

El segundo gran problema, que por su trascendencia debiera ser el primero, es que la producción de los *big data* puede traducirse en una visión distorsionada y, con frecuencia, interesada del fenómeno que representan. Esta distorsión sería el reflejo de una sociedad dividida por múltiples fracturas y desigualdades (económicas, de género, educativas, étnicas, de salud, interseccionales, etc.), y traduciría los propios intereses de quienes elaboran los datos en cuestión. Campos de conocimiento como la agnotología (Proctor y Shiebinger, 2008) y la epistemología de la ignorancia (Tuana y Sullivan, 2006) han mostrado los numerosos sesgos que taran los datos convencionales. Cuando los datos son complejos y masivos, como es el caso de los *big data*, el problema de los sesgos es más profundo: esos sesgos no aparecen meramente en la superficie de los datos, sino en la estructura profunda de los mismos, es decir, en las relaciones implícitas que mantienen. Esas relaciones solo son accesibles estadísticamente, en la medida en que solo se revelan a través de correlaciones complejas a través de grandes conjuntos de datos, y permanecen invisibilizadas —como la redondez de la Tierra— para el observador micro.

El problema de los sesgos en los datos ha sido trabajado desde hace tiempo, tanto en la inteligencia artificial (IA) como en las ciencias sociales y en particular en los STS (Estudios de Ciencia, Tecnología y Sociedad). Desde la IA, por ejemplo, se ha tratado de descubrir primero, y evitar después, los sesgos en los conceptos relativos a seres humanos y que pueden resultar moral, social y políticamente dañinos: Swinger *et al.* (2018); Caliskan *et al.* (2017), entre otros.

En psicología y, por mencionar solo un ejemplo, Grenwald, McGhee y Schwarz (1998) desarrollaron el Test de Asociaciones Implícitas (IAT) con el que constataron sesgos sexistas y racistas cuando pidieron

a los sujetos de estudio que asociaran entre sí nombres propios, imágenes o adjetivos de mujeres y hombres, o bien de personas de diferente aspecto racial.

Por su parte, la teoría feminista ha puesto de manifiesto, en abundantes investigaciones, la *generización* de la ciencia. Así, por ejemplo, Sandra Harding (1996: 36) señala que el simbolismo de género, la estructura generizada de la ciencia y las identidades, y conductas masculinas de los científicos individuales, han dejado su huella en los problemas, los conceptos, las teorías, los métodos, las interpretaciones, la ética, los significados y los objetivos de la ciencia. Un ejemplo de esta generización de la ciencia es el que nos ofrece Diana Maffía (2001), quien revisando a Linneo nos indica que en el mismo volumen en el que introdujo el término *mammalia*, también introdujo *homo sapiens.* De manera tal que la pareja humana, según Lineo, queda compuesta por una mamífera y un *homo sapiens*. Las implicaciones y consecuencias de tamaña distinción creemos que no se le escapan a nadie.

Para entender la generización epistemológica de la ciencia nos conviene prestar atención a los numerosos y fundamentados estudios de Ciencia, Tecnología y Género: (Keller, 1991; Harding, 1991, 1996; Healy, 1991; Haraway, 1995; Longino, 2002; García y Romero, 2018; Shiebinger, 2004; Sousa, 2010; Nikhil *et al.*, 2018). Además, es obligado considerar lo que Robert Proctor (1995, 2008) ha llamado *agnotología* y lo que Nancy Tuana llama *epistemologías de la ignorancia*. Proctor (1995: 8) nos dice que debemos

> […] estudiar la construcción social de la ignorancia. La persistencia de la controversia no es a menudo consecuencia de un conocimiento imperfecto, sino una consecuencia política de conflictos de intereses y apatías estructurales. La controversia puede ser diseñada: la ignorancia y la incertidumbre pueden ser fabricadas, sostenidas y diseminadas.

En la misma línea, Tuana y Sullivan (2006) dicen que «las prácticas de ignorancia están, a menudo, entremezcladas con sistemas de opresión y exclusión». En un artículo individual, en el mencionado volumen de 2006, Nancy Tuana señala hasta cinco tipos de ignorancia epistemológica, desde «saber que no se sabe, sin que importe» hasta la ignorancia voluntaria, «el no querer saber». Naturalmente, esos tipos de ignorancia derivan, en su mayoría, de relaciones de poder y exclusión.

Con mucha frecuencia, los sesgos de género y raza van muy próximos y operan con dispositivos epistémicos similares, como se pone de manifiesto en los trece capítulos del volumen *Race and Epistemologies of Ignorance* editado por Shannon Sullivan y Nancy Tuana en 2007.

Los sesgos de género en el lenguaje también han sido estudiados desde hace tiempo, si bien desde un punto de vista mayormente gramatical e intuitivo. Por «intuitivo» se entiende que quien investiga se convierte en intérprete natural y omnisciente del lenguaje, utilizando simplemente su comprensión propia y espontánea de él. Esta aproximación intuitiva al hecho del lenguaje resulta objetable en la medida en que carece de un respaldo independiente que apoye sus conclusiones. Efectivamente, conocemos las distintas formas morfosintácticas de la representación del género a través del lenguaje; los procesos cognitivos afectados por el lenguaje sexista y las imágenes mentales estereotipadas o prejuiciosas que este genera; el análisis concreto del lenguaje de los textos escolares, radiofónicos, etc.; y, para no hacer demasiado larga esta lista, las distintas formas de uso cotidiano del lenguaje propias de mujeres y hombres (Bergvall *et al.*, 1996; Goddard y Patterson, 2005; Lakoff, 2004; Bengoechea, 2000, entre muchas otras obras).

Asimismo, en el mundo del análisis de los macrodatos de texto crecientemente disponibles, también son numerosas las publicaciones referidas a todo tipo de relaciones sociolingüísticas. Sin embargo, hasta hace no mucho no se habían revelado los potenciales sesgos sexistas implícitos en las enormes bases de textos susceptibles de ser analizadas por medio de las potentes técnicas de tratamiento de los *big data* que están apareciendo en gran número y que tienen una influencia cada vez mayor.

Lo cierto es que, en el análisis de bases de macrodatos de texto, como demuestran matemáticamente Bolukbasi *et al.*, en su inspirador artículo «Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings», la geometría de los pares de palabras o de los «encajes[1] de palabras» (*word embeddings*) de dichas bases de datos incluye sesgos de género. Estos sesgos, aunque se limitan a reflejar el persistente sexismo de nuestras sociedades, no por eso dejan de ser particularmente perniciosos en el día a día de nuestras sociedades. El sesgo de género detectable en estos macrodatos revela la condición estructuralmente sexista de cada lenguaje natural —una condición que traduce y, a la vez, reproduce el sexismo ambiente de la vida social—. En este artículo vamos a presentar un modesto estudio de caso, concebido en el nivel de «prueba de concepto»[2], en el que trataremos de desvelar la semántica implícita en una muestra razonablemente representativa de la lengua española, la Wikipedia en español. Esta muestra constituirá una base empírica traducible en términos de *big data* a la que «haremos hablar» para que nos muestre, gracias a una técnica ya probada de inteligencia artificial —el uso de redes neurales— qué tipos de errores y sesgos se pueden detectar (inducidos socialmente y, al tiempo, generadores de efectos sociales) en esa macrobase de datos.

La aproximación metodológica que hemos elegido en este artículo para sondear ese fenómeno de la distorsión semántica de género descansa en el uso de la técnica del encaje de palabras. Es esta una poderosa técnica de aprendizaje automático mediante la cual se representa cada palabra de un idioma como un vector. La relación geométrica entre estos vectores captura relaciones semánticas significativas entre las palabras correspondientes. Dichas relaciones emergen a través de una evidencia básica: en multitud de pares de palabras (palabras expresivas, se supone, de conceptos) aparecen unas analogías estereotipadas, y moral, social y políticamente dañinas, como las que encontró Tolga Bolukbasi en pares tales como: *A es a B, como X es a Z.*

Estas analogías se concretan en pares tales como:

Hombre —(es a)— mujer = programador informático —(es a)— ama de casa

Hombre —(es a)— mujer = médico —(es a)— enfermera

Realmente, resulta muy llamativo que un algoritmo tan perceptivo e «inteligente» como para ser capaz de captar relaciones del tipo *si París es a Francia, Tokio es a X*, y con «conocimiento» suficiente como para encontrar que *X* es *Japón,* cuando se le pregunta *si Hombre es a Mujer, Programador informático es a X,* arroje para *X* el resultado *Ama de Casa*.

[1] Traducimos *embedding* como «encaje» porque este es el término usual en matemáticas, aunque otras posibles traducciones podrían ser «empotre», «encastre» o «incrustación».

[2] Según la Wikipedia: «Una prueba de concepto o PoC (por sus siglas en inglés) es una implementación, a menudo resumida o incompleta, de un método o de una idea, realizada con el propósito de verificar que el concepto o teoría en cuestión es susceptible de ser explotado de una manera útil». https://es.wikipedia.org/wiki/Brecha_de_g%C3%A9nero_en_Wikipedia

Sobre la base de esas evidencias, en el citado artículo, Tolga Bolukbasi nos alerta del «flagrante sexismo de los encajes de palabras y, consecuentemente, del riesgo de introducir sesgos de varios tipos en sistemas del mundo-real» (*op. cit.*: 1). No se nos oculta que, si en la base de tres millardos de datos de Google News utilizada por Bolukbasi, este y su equipo encontraron asociaciones de palabras como la que da título a su artículo, casi cualquier uso de esa base de datos no solo reproducirá, sino que incrementará los sesgos de género de la misma. Tanto más cuanto ese uso va a estar cada vez más mediado por técnicas o algoritmos de inteligencia artificial. Imaginemos, por ejemplo, que una empresa necesita contratar personas expertas en programación y, con este fin, le pide a Google unas decenas de nombres. Este motor de búsqueda obtendrá tales nombres a partir de la macrobase de datos de Google News. Ahora bien, en esa enorme base de datos, no aparece «Mujer» como par compatible con «Programador informático», al menos en términos de probabilidad estadística. Por consiguiente, ninguna mujer figurará en la lista que Google venda a la empresa cliente, y ninguna mujer llegará a ser contratada. Se ampliaría así, a través de ese voluminoso rodeo de los macrodatos, el círculo vicioso de la discriminación de género.

La extraordinaria potencia de la red neural word2vec con la que Bolukbasi ha analizado los 3 millardos de datos de Google News es lo que nos ha llevado a pensar en la pertinencia de una investigación que realice un análisis similar sobre alguna base de macrodatos en español.

La investigación que da base a este artículo se limita a bosquejar una descripción de cómo hemos adaptado el word2vec a una base de macrodatos en español y a dar cuenta de las pruebas realizadas, con esa aplicación y sobre esa base empírica, para comprobar su funcionamiento y adecuación para el propósito indicado: analizar los sesgos semánticos encontrados en esa macrobase de datos. Como se detallará más adelante, hemos utilizado, como base de macrodatos, una parte (la mitad) de la versión en español de la Wikipedia de 2006.

Antes de seguir adelante, hemos de reconocer que nos acercamos a la Wikipedia sabiendo que en ella íbamos a encontrar sesgos de género. Esperábamos, sin embargo, que con el uso de la tecnología de redes neurales seríamos capaces de descubrir sesgos de género más profundos que los obvios (entre estos se contarían, digamos, la mayor presencia de personajes masculinos frente a la minorización de los femeninos). Lo que hemos encontrado son sesgos de género incrustados (*embedded*) en la propia y compleja estructura semántica de todo el espacio lingüístico cubierto por, y materializado en, la Wikipedia en español.

Varias publicaciones nos alertaron de los sesgos de género presentes en la Wikipedia. Entre estas merecen ser citados los propios estudios de esta, como el WikiProyecto *Countering Systemic Bias*, que dedica informes (2008 y 2009) al análisis de los sesgos de género de esta enciclopedia *online*, así como la propia entrada de la enciclopedia sobre sesgos de género[3]. En esos informes los sesgos de género se explican, sobre todo, por factores como la peculiar demografía de los editores, que con sus aportaciones construyen globalmente la Wikipedia. Esta empresa colectiva global, de hecho, está desarrollando su propia Meta-Wikipedia. Un ejemplo de este espacio autorreflexivo lo proporciona la página https://meta.wikimedia.org/wiki/Gender_gap que está abierto a la inclusión de todos los estu-

[3] https://es.wikipedia.org/wiki/Brecha_de_g%C3%A9nero_en_Wikipedia

dios que traten los sesgos de género en la Wikipedia[4]. Fuera del ámbito estricto de la Wikipedia, Josep Reagle y Lauren Rhue (2011), en su comparación entre la Wikipedia en inglés y la versión en línea de la *Enciclopedia Británica*, encontraron que la Wikipedia, aun con menos artículos sobre mujeres que sobre hombres, incluye más artículos sobre mujeres, en términos absolutos, que la *Enciclopedia Británica*. Los estudios de Glott *et al.* (2010) revelaron que las mujeres constituyen alrededor del 13% de los wikipedistas. Según Benjamin Mako Hill y Aaron Shaw (2013) esta cifra podría ser algo más alta: al revisar las encuestas analizadas por Ghost y sus colegas, los anteriores autores hallaron que la proporción de mujeres editoras estadounidenses subía a un 22,7% y la proporción total de mujeres editoras alcanzaba el 16,1%. El estudio más completo realizado hasta el momento es el de Wagner *et al.* (2015), que analiza sesgos en la cobertura (proporción de artículos sobre mujeres y hombres notables), sesgos estructurales (probabilidad de que un artículo sobre una persona de un sexo esté enlazado con otro de una persona del otro sexo), sesgos léxicos y sesgos de visibilidad (proporción de mujeres y hombres en las páginas iniciales de la Wikipedia en inglés).

Pero lo que no sabíamos aún, y hemos tratado de comprobar con la aplicación de la red neural word2vec, es si la Wikipedia mostraba sesgos de género semántico-estructurales, similares a los encontrados en Google News por Bolukbasi. En efecto, hemos corroborado, aunque sea en el formato mínimo exigible en una prueba de concepto, que al menos en la Wikipedia en español del año 2006, esos sesgos de género se encuentran presentes. No podemos decir que el hecho constituyera para nosotros una sorpresa, pero es muy llamativo hallar, en nuestra considerable muestra de más de 28 millones de palabras, que el algoritmo de la red neural empleada, asociaba pares como: *Hombre es a Experto como Mujer es a Sabelotodo*. Este fue el primer par analógico que nos escandalizó. Al revelar el análisis de la semántica implícita en la Wikipedia en casos como este, esta macroenciclopedia *online* no hace más que seguir la tradición de otras enciclopedias convencionales. Así, por ejemplo, y como señala la historiadora Gillian Thomas en su estudio sobre la 11.ª edición de la *Enciclopedia Británica*, las mujeres que aparecen en ella suelen ser «percibidas como siervas pedantes ante el amplio alcance de la inteligencia masculina» (Thomas, 1992: 18-26).

La red neural word2vec

Word2vec es una red neural, desarrollada originalmente por Google, que procesa texto. Se puede aplicar a lenguaje natural escrito, pero también a genes, códigos, gustos, listas de música y en general a cualquier serie verbal o simbólica de la que se puedan extraer patrones.

[4] En el espacio de la Wikimedia, https://meta.wikimedia.org/wiki/Gender_gap, en una muestra clara de su preocupación por los sesgos de género, se citan los siguientes artículos seleccionados: «Unlocking the Clubhouse: Five Ways to Encourage Women to Edit Wikipedia», *Sue Gardner's Blog*, 14 de noviembre de 2010; Noam Cohen, «Gender Gap? Look Up Wikipedia's Contributor List», *New York Times*, 30 de enero de 2011; «Where Are the Women in Wikipedia? Debate with a Number of Debaters», *New York Times*, 2 de febrero de 2011; «Wikipedia(EN) Signpost Issue on Gender Gap», 7 de febrero de 2011; «Top 10 Reasons to Encourage more Women Participation in Wikipedia here on Meta», 8 de febrero de 2011; «Nine Reasons Women Don't Edit Wikipedia (in their own words)», *Sue Gardner's Blog*, 19 de febrero de 2011; Adrianne Wadewitz, «Wikipedia's Gender Gap and the Complicated Reality of Systemic Gender Bias», *HASTAC*, 26 de julio de 2013; *Gender Gap Manifesto*, creado en marzo de 2011 por siete editores/as de Wikimedia; *Charting Diversity: Working Together Towards Diversity in Wikipedia,* Wikimedia Alemania en colaboración con la Universidad Beuth de Ciencia Aplicada.

Una red neural es un modelo informático automatizado de aprendizaje inspirado en los sistemas nerviosos biológicos. Llamamos aprendizaje a la transformación en el comportamiento de las neuronas como consecuencia del procesamiento de los estímulos que inciden en la red. Por ejemplo, en una red neural entrenada para reconocer fotos de gatos, la entrada solo reconocería ciertos ángulos y colores; sin embargo, cuando los patrones se repiten, va reconociendo poco a poco pelo, ojos, orejas, cola, etc. En el estadio final, cuando todas las neuronas correspondientes a la estructura de un gato se encuentran activadas, la neurona de salida (en nuestra hipotética red, solo habría una neurona correspondiente a un gato completo) se encendería haciéndonos saber que un gato ha sido reconocido.

Word2vec analiza y convierte en números (un vector) las palabras que se le introducen a través de conjuntos de frases. Además, busca la probabilidad de que los estados discretos[5] que se dan en la red co-ocurran. Consigue esto repartiendo los vectores que crea —que son llamados encajes neurales de palabra (*neural word embeddings,* en inglés—) en un espacio vectorial de *n* dimensiones.

En nuestro caso la dimensión de los vectores es de 500, y cada palabra se define en su interacción en esas 500 dimensiones (es decir, no una a una). Todas las dimensiones definen todas las palabras, ya que cada palabra está situada en el espacio 500-dimensional y el vector no es más que sus coordenadas.

Para entender qué es el vector de una palabra hemos de pensar que cada elemento en el vector está asociado a una palabra del vocabulario del cuerpo de datos que vamos a analizar. En nuestro caso, como hemos dicho, la dimensión de cada vector es 500. En este vector, cada palabra se representa por una distribución de pesos a través de esas dimensiones. Así que, en lugar de una asignación uno a uno entre un elemento en el vector y una palabra, la representación de una palabra se extiende a través de todos los elementos en el vector, y cada elemento en el vector contribuye a la definición de muchas palabras.

Los vectores de palabras ayudan a los ordenadores a «aprender» del texto. De este modo, el programa se convierte en un *intérprete semántico artificial*, que de alguna manera «emula» a los *intérpretes semánticos naturales*, que somos los seres humanos en nuestro uso espontáneo de los lenguajes naturales que conocemos y en los que estamos socializados.

Los encajes: una definición matemática

Un encaje es una instancia de una estructura matemática contenida dentro de otra. Para que se dé un encaje, nuestra estructura *X*, que queremos encajar en *Y*, tiene que ser insertada de manera inyectiva con una función tal que $f: X \rightarrow Y$.

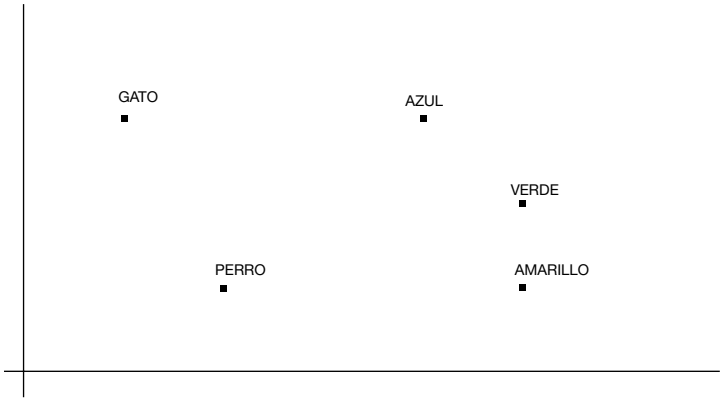
Cada una de las palabras que analiza la red neural no se analiza por separado, sino en grupos definidos; en el caso del lenguaje natural, esos grupos son frases. Si, por ejemplo, se entrena una red de estas características con las siguientes frases:

— Mi perro es azul.
— Mi gato es azul.
— Mi perro es verde.
— Mi perro es amarillo.

En el espacio resultante, las palabras «azul», «verde» y «amarillo» se encontrarán agrupadas entre ellas, estando todas a la misma distancia de «perro»; pero solo «azul» estará cerca de «gato», a su vez agrupada con «perro».

[5] Un sistema discreto es aquel con un número contable de estados, en nuestro caso cada una de las palabras del diccionario.

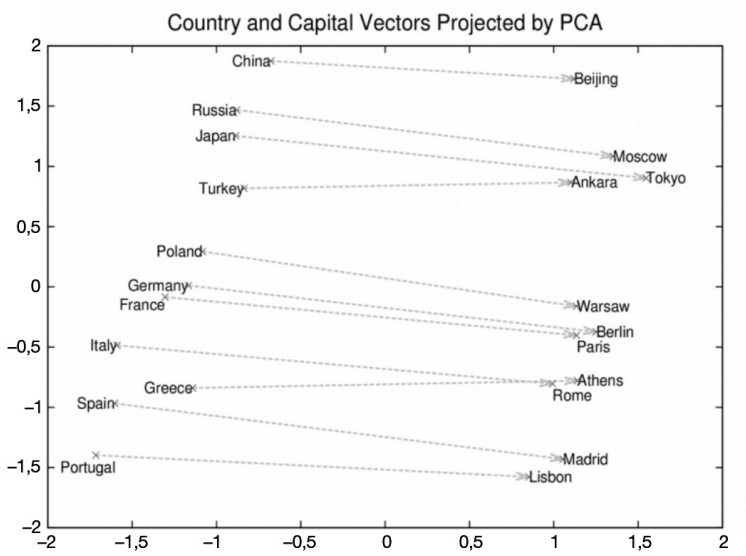
GRÁFICO 1. *Representación bidimensional de un modelo simple de red neural*



Fuente: Gráfico propio.

A lo largo de muchas iteraciones con mucha cantidad de frases, las palabras similares van agrupándose y distanciándose de manera congruente gracias a la magia de las matemáticas, como se puede ver en el siguiente conjunto real de países frente a capitales. Obsérvese que la red neural «ha aprendido» no solo cuáles son las capitales de cada país, sino que las ha posicionado en un orden aproximadamente geográfico (de este a oeste y de norte a sur).

GRÁFICO 2. *Vectores de país y capital proyectados por distancias*



Fuente: https://deeplearning4j.org/word2vec[6]

[6] El análisis de componentes principales (PCA) es un procedimiento estadístico que utiliza una transformación ortogonal para convertir un conjunto de observaciones de las variables posiblemente correlacionadas en un conjunto de valores de variables no correlacionadas linealmente llamadas componentes principales.

Conviene recordar que estas representaciones son proyecciones en dos dimensiones y no concuerdan con el modelo real subyacente, que en este caso es 500-dimensional. Para traducir los resultados de nuevo al terreno humano, como representación de la distancia entre vectores se usa el coseno (cos (90) es 0, cos (0) es 1).

En nuestro modo de empleo, las palabras a analizar provienen de textos en lenguaje natural, como una enciclopedia, los artículos de Google News o *La Biblia*, siempre que tales textos estén digitalizados. Tras el proceso de aprendizaje, la red neural ha transformado dicho texto a números, con los cuales el ordenador puede realizar operaciones matemáticas a gran escala; ya sean sumas, restas o, lo más interesante, sistemas de ecuaciones, que permiten obtener relaciones de palabras tan sugerentes como las que siguen:

— Geopolítica: *Iraq – (menos) Violencia = Jordania*
— Distinciones: *Humano – Animal = Ética; Presidente – Poder = Primer Ministro; Biblioteca – Libros = Sala*
— Analogías: *Bolsa de Valores ≈ Termómetro*

Chris Moody, en su didáctico artículo «Una palabra vale mil vectores»[7], parte de otro ejemplo del lenguaje natural en inglés

King – man + (más) women = queen (rey – hombre + mujer = reina)

Este autor nos presenta el siguiente ejemplo: un ser humano hizo una pregunta a un ordenador: ¿*qué es rey – hombre + mujer*? (se trata de la ecuación matemática: «rey menos hombre más mujer es igual a X»). Y el ordenador resolvió esta ecuación respondiendo: *reina*. La máquina «comprende» que la mayor diferencia entre las palabras hombre y mujer es el género. Si se añade la diferencia de género a «rey», se obtiene «reina».

Este resultado es llamativo, porque ¡nunca se había enseñado explícitamente a la máquina nada sobre el género! De hecho, nunca se había dado al ordenador nada como un diccionario, un tesauro o una red de relaciones entre palabras. Simplemente se había puesto una montaña de texto en word2vec y se esperó a que la máquina, a través del correspondiente algoritmo, «aprendiera» del contexto de cada palabra. Lo que hace el algoritmo, para cada palabra, es tratar de predecir las palabras que acompañan a otras en una oración. O, mejor dicho, al representar internamente las palabras como vectores, y dado un vector de palabra, trata de predecir los otros vectores de palabras en el texto cercano.

Los algoritmos de word2vec, finalmente, escudriñan tantos ejemplos que pueden inferir el género de una sola palabra, igual que pueden inferir que *The Times* y *The Sun* son periódicos, que *Matrix* es una película de ciencia ficción, y que el estilo de una prenda de vestir podría ser *hippy* o formal. Que esos vectores de palabra representen una gran parte de la información disponible en una definición de diccionario es un efecto secundario —conveniente y casi milagroso— de esa tarea algorítmica que consiste en intentar predecir el contexto de una palabra[8].

Sesgo de género en word2vec

Al analizar los textos de manera ciega y sin ningún apriorismo, word2vec ofrece la *ventaja* de ser en cierto modo «objetivo», al tiempo que tiene la desafortu-

[7] http://multithreaded.stitchfix.com/blog/2015/03/11/word-is-worth-a-thousand-vectors/ (descargado 20 octubre, 2016). La descripción del ejemplo es una traducción adaptada y simplificada de la explicación de Chris Moody.

[8] Traducción libre y simplificada de la explicación de Chris Moody (*op. cit.*).

nada desventaja de arrastrar y posiblemente multiplicar cualquier sesgo, sexista o de otro tipo, presente en los textos originales con los que el programa es alimentado y entrenado. Bolukbasi y sus colegas, analizando la información contenida en Google News, identificaron los sesgos de género adjudicados a las palabras «él» y «ella», como puede verse en el gráfico siguiente.

GRÁFICO 3. *Proyección bidimensional de distancias semánticas entre palabras*

Fuente: Bolukbasi (*op. cit.*: 11).

El gráfico 3 corresponde a una serie de palabras seleccionadas y proyectadas sobre dos ejes (*X* e *Y*). *X* es una proyección de la diferencia entre los encajes de las palabras «ÉL» (*he*) y «ELLA» (*she*). Por su parte, *Y* es una dirección aprendida en el encaje que captura la neutralidad de género, situando las palabras neutras en relación al género por encima de *Y*, y las no neutras por debajo.

Nuestro proceso de análisis con word2vec consistirá en utilizar la red neural implementada para identificar, diagnosticar y denunciar posibles sesgos presentes en el corpus de base (la parte de la Wikipedia en español que será procesada).

Construcción de nuestra red neural word2vec

La versión de word2vec que utilizamos en esta prueba de concepto es la disponible en el proyecto de software libre deeplearning4j, mantenido por la corporación Skymind[9].

Antes de proceder a la construcción de la red neural, y ya que queremos analizar textos en castellano, debemos adaptar el código de word2vec, originalmente ideado para procesar textos en inglés. Para ello es necesario saber un poco más sobre cómo funciona esta aplicación y sobre el procedimiento que emplea para analizar el texto que se le introduce.

Cuando analiza texto, word2vec distingue entre dos unidades básicas: palabras sueltas o *tokens,* y agrupaciones de palabras en frases (*sentences*). El programa dispone de unos clasificadores básicos, que le permiten identificar cada grupo de caracteres entre dos espacios como un *token* y

[9] Disponible en: https://deeplearning4j.org/word2vec

cada grupo de *tokens* entre dos pasos de carro como una frase. Esta clasificación es importante, ya que el aprendizaje está basado en cómo se localizan las palabras dentro de una frase.

En nuestro caso, hemos utilizado párrafos enteros como frases, y palabras sueltas (unidades entre dos espacios) como *tokens*. Para minimizar duplicidades e inconsistencias hemos filtrado y descartado cualquier carácter que no fuera una letra o un guion (-), con el objeto de evitar, por ejemplo, que «punto» y «punto.» resultaran dos entradas de vocabulario distintas.

Esta adecuación del corpus a estudiar depende en gran medida de la sintáctica específica que tenga dicho corpus; si conocemos con gran precisión el contenido del corpus, se puede afinar mucho este procesamiento previo, con el objetivo de eliminar ambigüedades como plurales, nombres compuestos y nombres propios que queramos destacar o ignorar.

Una vez construidos estos clasificadores (*tokens* y *sentences*), solo falta entrenar la red neural. Para ello se deben definir ciertos atributos de los cuales los más interesantes en este estudio serían:

Frecuencia mínima de palabras: indica cuántas veces debe aparecer un *token* en el corpus para ser reconocido. Ajustar este aspecto ayuda a filtrar palabras superfluas.

Tamaño de ventana: indica el número de pasos hacia delante y hacia atrás que se consideran dentro de la frase a la hora de comparar y analizar cada *token*. En principio, cuanto más grande sea la ventana, mejor; pero si nuestras frases son pequeñas o disponemos de poco poder computacional es mejor reducir este tamaño de ventana. En nuestro caso, y después de varias pruebas, ese tamaño fue establecido en ocho.

Iteraciones: indican el número de veces que el algoritmo recorre y analiza secuencialmente el corpus. Aunque técnicamente se distinguen diferentes maneras de calificar estos barridos, en nuestro caso podemos generalizarlas al número de veces que se hace una lectura completa del texto.

Como en el caso anterior, cuanto más altos sean los valores asignados a las iteraciones, más precisa será la red. Conviene advertir, sin embargo, que la asignación de valores altos a las iteraciones multiplica exponencialmente el tiempo de entrenamiento de la red.

Análisis de un resultado del empleo de word2vec

En primera instancia se intentó analizar un corpus compuesto por 15 páginas de noticias de la FECyT (Federación Española de Ciencia y Tecnología). Sin embargo, pese a contener 28.089 palabras, la red obtenida era de muy baja calidad. La razón de este hecho hay que achacarla a la falta de alcance del vocabulario del corpus y a la limitación sintáctica de las indicadas noticias, escritas todas, seguramente, por la misma persona y bajo las mismas pautas[10]. Este primer resultado nos llevó a buscar un corpus más grande, concretamente uno extraído de la Wikipedia en español de 2006, cortesía de la Universidad Politécnica de Catalunya[11]. La ventaja de usar este corpus específico, aparte de la gran cantidad de información que contiene, estriba en que la UPC la ofrece ya reducida a texto plano, lo cual nos ahorraba un esfuerzo considerable. Por otra parte, y como se ha indicado arriba, la Fundación Wikimedia ha reconocido la existencia de sesgos de género y de

[10] Efectivamente, Chris Moody (*op. cit.*) advierte que la vectorización de palabras requiere una gran cantidad de ellas y que se pueden descargar palabras de cualquier texto, pero si este tiene un vocabulario muy especializado, se necesita una gran cantidad de texto para entrenar a los vectores. Normalmente, esto significa cientos de millones de palabras.

[11] Disponible en: http://www.cs.upc.edu/~nlp/wikicorpus/

falta de diversidad en su enciclopedia. Este hecho incrementaba de entrada las posibilidades de éxito de nuestro escrutinio.

Debido a limitaciones de tiempo y poder computacional, se entrenó la red neural usando solo la mitad del corpus indicado, 28.291.729 palabras. Se practicaron 10 iteraciones. El tiempo de entrenamiento, con estos parámetros, ascendió a 38 horas. El peso total de la base de datos tras el entrenamiento fue de 1,16 GB, un tamaño no despreciable. Una vez entrenada la red neural, se procedió al análisis de los resultados arrojados por esta.

Técnicas de análisis

Como hemos visto anteriormente, la red neural genera una colección de encajes que representan cada una de las palabras del vocabulario del corpus, ordenadas entre sí de tal manera que las distancias entre ellas indican la proximidad de su uso en el lenguaje con el que se entrenó dicha red.

Con estos encajes realizamos una *concatenación de operaciones* como la que sigue:

palabra2 + palabra3 – palabra1 = ?

Esta operación, traducida a lenguaje natural, es una comparación del tipo «uno es a dos como tres es a...». El resultado de la operación presenta en pantalla las diez palabras más cercanas a la ecuación. Por ejemplo:

Hombre es a actor como mujer es a:
[actor, actriz, mujer, nominada, isbert, trudie, oscar, hodiak, haymes, karina]

Aunque hemos usado un corpus relativamente grande (más de 28 millones de palabras), las carencias en iteraciones suficientes al entrenar la red saltan a la vista en los resultados generados automáticamente por el algoritmo. Por ello es necesario que personas (entrenadas exprofeso para ello) interpreten, en calidad de expertas, cada salida. Los criterios de interpretación que hemos establecido son los siguientes:

«Bueno», cuando el resultado incluye mayoría de palabras lógica y semánticamente correctas y que no muestran sesgos de género evidentes. El ejemplo anterior sería «bueno» porque «actriz» aparece en segundo lugar (aunque no en primero) y, en general, la lista de conceptos asociados no contradice la consistencia semántica propia del español. Obsérvese que si «actor» aparece antes que «actriz» es, probablemente, por el carácter mayoritariamente inclusivo (del femenino) que damos al masculino.

«Sesgado», cuando el resultado es lógica y semánticamente correcto, pero evidencia un sesgo de género claro, ya sea directo o de ámbito[12].

«No válido», cuando el resultado es absurdo.

Una vez definidos estos criterios, se procedió al análisis del modelo usando la estructura *«Hombre es a palabra como mujer es a…»*, y viceversa, *«mujer es a palabra como hombre es a…»* . El conjunto de palabras usadas fueron 110, distribuidas en los siguientes dominios: profesiones, actitudes, objetos y trayectoria de vida.

Análisis de resultados generales

Como ya se indicó anteriormente, es importante destacar que los resultados aquí obtenidos son solo ilustrativos de una prueba de concepto acerca de las posibilidades de uso de una red neural como

[12] Por «ámbito» aquí se entiende el conjunto semántico creado por las diez palabras que aparecen como resultado de la ecuación. A menudo nos hemos encontrado con que buena parte de las palabras asociadas a mujer son de ámbito familiar, lo cual es un indudable sesgo de género porque tal asociación no se da en el caso de la palabra «hombre». A este sesgo, como más adelante veremos, le hemos llamado «familización».

herramienta de análisis de las diferencias sociosemánticas de género (Díaz, 1996, 2000)[13]. Unos resultados más refinados podrían haberse obtenido si hubiéramos empleado la técnica de los microtrabajos utilizada por el equipo de Bolukbasi en su obra ya citada. Esta prueba de concepto tiene limitaciones vinculadas al hecho de que los resultados obtenidos fueron analizados solo por dos personas, a lo largo de dos sesiones de tres horas. Bolukbasi, por el contrario, a través de la técnica de los microtrabajos[14], pudo contar con cientos de colaboradores que aportaron una validez adicional a sus resultados.

La primera ronda de preguntas «*Hombre es a X como mujer es a..*» produjo un resultado con un 71% de coherencia (esto es, resultados buenos y sesgados a la vez), de los cuales, el 38,2% eran sesgados. La segunda ronda, «*Mujer es a X como hombre es a...*» produjo un 63,4% de coherencia, pero solo un 19,6% de sesgo.

La falta de precisión en las preguntas femeninas (*Mujer es a X*…), viene dada por la escasa presencia en el corpus de los calificativos femeninos y por la poca densidad de estos respecto a los masculinos. Vale la pena señalar que a este sesgo lo llamamos *sesgo de omisión*, como se expondrá en lo que sigue.

Análisis de sesgos

El número y la variedad de sesgos detectados en nuestro análisis otorga a los resultados un nivel de complejidad que aconseja intentar una clasificación de tales sesgos. Podemos distribuirlos en tres categorías: directos, semánticos y de omisión.

Sesgos directos

Llamamos sesgos directos a aquellos que se reflejan en resultados que claramente indican una fuerte «desigualación semántica de género»[15], al arrojar términos sorprendentes e incluso ofensivos. Algunos ejemplos obtenidos son:

- *Hombre es a experto como mujer es a sabelotodo*
- *Hombre es a fidelidad como mujer es a obediencia*
- *Hombre es a trabajo como mujer es a madre*
- *Hombre es a muebles como mujer es a calzado, textiles*
- *Hombre es a inteligencia como mujer es a lucirse*
- *Mujer es a abogada como hombre es a estilista*

Este tipo de sesgo es fácilmente identificable y su interpretación no requiere conocimiento previo alguno del corpus, por lo cual se podría recurrir a una red de microtrabajos para posterior valoración y clasificación en «bueno», «sesgado» o «no válido».

El análisis sociológico de cada uno de estos encajes puede resultar muy enriquecedor, y ofrece pistas bastante contundentes sobre los estereotipos de género que infiltran no solo nuestro lenguaje sino nuestra visión y prácticas sociales. En esta investigación, sin embargo, no nos detendremos a realizar dicho análisis propiamente

[13] Estas dos publicaciones desarrollan una aproximación a las diferencias sociosemánticas similar a la que aquí se presenta, pero que utiliza métodos y técnicas bien distintas (entre otras cosas, porque entonces las redes neurales no se habían inventado todavía).

[14] Carecemos de espacio para explicar esta técnica de externalización del trabajo, por lo que referimos al mencionado artículo de este autor.

[15] Proponemos utilizar los términos «igualación/desigualación semántica de género», en lugar de los más obvios «igualdad/desigualdad semántica de género», para indicar el carácter no estático, sino dinámico, acumulativo e intencional de los procesos de diferenciación por género que operan en el nivel semántico de nuestros lenguajes naturales.

sociológico. El reciente descubrimiento por nuestra parte de esta metodología no nos ha permitido avanzar más que en la puesta en funcionamiento de la red neural, y en la realización de la prueba de concepto que aquí estamos presentando. Por ello, haremos solo unas breves anotaciones en relación con los ejemplos seleccionados, deteniéndonos en su interpretación semántica, cuando esta interpretación no sea obvia. Así, de los seis ejemplos anteriores, puede que el cuarto y el sexto necesiten alguna aclaración. En efecto, en el cuarto ejemplo: *«Hombre es a muebles como mujer es a calzado, textiles»*, entendemos que se trasluce, como mínimo, la tradicional división sexual del trabajo que está en la base de la masculinización o feminización de ciertas profesiones. Los hombres aparecen asociados a la producción de muebles, a la carpintería, mientras las mujeres mantienen una relación privilegiada con los textiles y su producción, confección, etc. Obsérvese pues cómo, incrustada en el lenguaje, aparece una diferencia ocupacional obvia para cualquier persona que conozca la distribución de hombres y mujeres en el mundo laboral. Resulta pues que la red neural atesora, inopinadamente, un «conocimiento» que, en el mundo de la «inteligencia natural» (de las personas de carne y hueso) solo atribuiríamos a profesionales de la sociología del trabajo o la sociología del género.

Nos quedaría pendiente un análisis detallado del calzado el cual, al no ser, a primera vista, tan transparente, requeriría que hiciéramos otras preguntas a la red neural para buscar nuevas concatenaciones, proximidades, etc., que dieran sentido a tal asociación.

El sexto ejemplo, *«Mujer es a abogada como hombre es a estilista»*, es particularmente interesante y, sin duda, exige que dirijamos a la red más preguntas clarificadoras. En la fase actual de desarrollo del análisis, nos atrevemos a decir que esta asociación indica un *sesgo de género inverso*: mientras que generalmente son los hombres quienes aparecen asociados a profesiones de mayor prestigio que las asociadas a mujeres, en este caso, son los hombres quienes aparecen asociados a una profesión (estilista) aparentemente menos prestigiosa que la que se vincula a las mujeres (abogada). Probablemente esta anomalía que llamaremos «inversión de género» esté ocasionada por el hecho de que en nuestro corpus aparecen muy pocas mujeres abogadas, y estas suelen estar asociadas a palabras también con pocas ocurrencias, como la de estilista. En cualquier caso, apenas hemos encontrado ejemplos de ese tipo, lo que en sí mismo sería una prueba del carácter anómalo y muy minoritario de esta llamada inversión de género. En todo caso, se trata de un fenómeno que convendría explorar y acotar.

La investigación que proponemos abordar en el futuro se basa en una elaboración más amplia y profunda en la línea recién sugerida. Esta línea deberá avanzar a través de la formulación de nuevas preguntas y/o de otras formulaciones matemáticas más complejas, que no hemos podido llevar a cabo en este primer esbozo de investigación. Nos referimos al estudio de las distancias absolutas entre palabras, o al listado de palabras más cercanas, técnicas que nos permitirían dar una razón más cabal de los sesgos de género presentes en la Wikipedia en español de 2006 o en otros corpus análogos.

Sesgos semánticos

Estos tipos de sesgos son más sutiles y difíciles de observar. Sin embargo, es aquí donde la red neural brilla en especial. El sesgo más general y extendido que hemos encontrado en nuestro estudio es el de la *familización de las mujeres*: el término «mujer» casi siempre aparece rodeado de términos que tienen que ver con la familia, mien-

tras que el término «hombre» aparece como una entidad independiente. Por ejemplo:

Hombre es a amor como mujer es a:
[madre, hija, pareja, hijos, esposa, hermana]

Mujer es a amor como hombre es a:
[espíritu, dios, mundo, deseo]

Hombre es a casa como mujer es a:
[madre, familia, hija, esposa, hermana]

Mujer es a casa como hombre es a:
[pueblo, tiempo, vida]

Estos ejemplos evidencian, de manera ciertamente contundente, eso que hemos llamado «sesgo semántico», un sesgo que distorsiona fuertemente nuestro lenguaje (en cierto modo como la masa distorsiona el espacio-tiempo einsteniano). Aunque estos cuatro reveladores resultados merecerían una glosa mucho más pormenorizada, apuntemos simplemente unas pocas observaciones que deben entenderse como pautas para ese examen que se apunta como más reposado y sociológicamente fundamentado.

Es fácil apreciar que los términos y conceptos asociados a «hombre» tienen un carácter marcadamente general y abstracto frente a la naturaleza concreta y centrada en el ámbito familiar de los conceptos y términos vinculados a «mujer». Esa mayor abstracción de los conceptos masculinos va emparejada, en buena parte de los casos (espíritu, dios, deseo, etc.), con la condición inmaterial de estos. Los conceptos vinculados a las mujeres son, en contraste, mucho más tangibles, y en este sentido, materiales. Es llamativo que los conceptos masculinos no incluyan personas concretas, sino entidades impersonales, mientras que los femeninos se refieren siempre a personas concretas, físicas y conocidas. El contraste sociosemántico de género a todas luces más revelador y que conjuga y sintetiza los anteriormente señalados es el par «hombre socio-centrado/ mujer familia-centrada» (Díaz, *op. cit.*). Por socio-centrado entendemos «centrado» en el mundo más allá de los límites familiares (pueblo, mundo), además de entidades abstractas. La condición familia-centrada es tan clara que no requiere mayor comentario. En consecuencia, con lo anterior, la dimensionalidad de términos o conceptos de «hombre» aparece como superior a la de «mujer», encerrada en una cierta unidensionalidad familista.

Otra modalidad de sesgo semántico que hemos podido observar es la *sexualización de los ámbitos femeninos* en comparación con los masculinos.

Mujer es a lesbiana como hombre es a:
[fetiche, fiery, musings, spotless, libertine]

Esta sexualización de ámbitos femeninos es, sin duda inducida por los hombres, y parece estar relacionada con la difusión de la industria pornográfica.

Un sesgo semántico todavía más sutil se puede encontrar en el siguiente par:

Mujer es a embajadora como hombre es a:
[unicef, acnur, naciones(unidas), microcrédito]

Hombre es a embajador como mujer es a:
[tymoshenko, vizcondesa]

El par anterior nos indica que cuando le preguntamos a la red neural por una mujer embajadora la única asociación con hombre que encuentra es cuando estos trabajan en organizaciones de buena voluntad o parecidas, como Naciones Unidas u organizaciones benéficas, no comparable al «embajador» masculino, representante de potencias estatales. La red neural solo cita a una persona en un papel análogo, «tymoshenko» (suponemos que será Yulia Tymoshenko), o un (inidentificado) título nobiliario. En este par, pues, y sin más información, se dan sesgos de género semánticos pero también sesgos por omi-

sión, porque no se encuentra un encaje lo bastante saturado como para resultar representativo (la red no encuentra un número de relaciones suficiente).

Sesgos por omisión

Como ya se anticipó, estos sesgos se dan cuando el peso de una de las partes en el corpus es tan bajo que el hecho de operar con el encaje correspondiente apenas resulta significativo por carencia de casos (es lo que llamamos falta de saturación). Esta carencia se puede detectar en relación con conceptos como «física», «matemática», «química»… en los cuales los resultados siempre se refieren al significado de «ciencia física», «ciencia matemática», «ciencia química», etc., pero no a la acepción «mujer (profesional de la) física, matemática, química, etc.».

Este sesgo por omisión se puede observar también, por ejemplo, en la siguiente operación:

Mujer es a reina como hombre es a:
[rey, amidala, príncipe, naboo]

Hombre es a rey como mujer es a:
[Hija, mujer, esposa]

A primera vista puede parecer que el primer componente del par arroja un resultado adecuado (rey es la primera respuesta y príncipe la tercera), sin embargo, dos de los cuatro conceptos que nos devuelve la red tienen que ver con la saga de películas *Star Wars*, lo cual sugiere que la mayor parte de las instancias de «reina» aparecen en artículos sobre uno de los personajes de dicha saga, la reina Amidala. Esta extraña interferencia de una heroína cinematográfica, con un concepto mucho más amplio como el de reina, puede deberse al perfil del colectivo de autores de la Wikipedia (con toda probabilidad mayoritariamente socializados en la mencionada saga).

En el segundo par, cuando preguntamos a la inversa («hombre es a rey como mujer es a…»), el algoritmo ni siquiera devuelve «reina», si no que directamente habla de «hija», «mujer» y «esposa». Este resultado, aparte de ejemplificar de nuevo la familización detectada en el punto anterior, indica que, en los artículos de la Wikipedia, cuando los hombres son reyes, las mujeres de su entorno aparecen solo como familiares suyos. Encontramos la misma relación de omisión al preguntar por «profesora», obteniendo solo personajes de la saga de libros de Harry Potter.

Conclusiones

De los resultados obtenidos es fácil inferir que existe un *sesgo de género global* en la Wikipedia en español de 2006 y, en especial, hay una gran *omisión y familización de las mujeres* en los artículos publicados. Por contra, los hombres suelen aparecen como entidades sustantivas en su individualidad.

De forma tentativa, cabe señalar que la presencia de conceptos relacionados con el mundo de la fantasía juvenil y de adulto joven nos permitiría arriesgar una estimación demográfica sobre las personas que editaron esa versión de la Wikipedia: gran mayoría de varones adultos de entre 20 y 30 años de edad. Estas conclusiones no se alejan mucho de la realidad de la Wikipedia actual, harto más la del 2006[16]. Entre las conclusiones extraídas hay algunas que coinciden con las que la propia Wikipedia ha aportado sobre sí misma y que trata de solucionar, según hemos señalado en la introducción a este artículo.

Creemos que, disponiendo de poder computacional suficiente, el análisis de

[16] https://es.wikipedia.org/wiki/Sesgo_de_g%C3%A9nero_en_Wikipedia

una red neural puede facilitar mucho el diagnóstico de problemas de sesgos, y no solo de género, dentro de un corpus suficientemente amplio y variado. En la era digital en que nos encontramos, las posibilidades en este sentido son innumerables: publicaciones de un grupo editorial concreto, bibliografía de un individuo o un grupo, personas partidarias de un partido político en redes sociales, etc. Podríamos incluso analizar todas las publicaciones científicas de un mismo ámbito (o, ya puestos, de todos los ámbitos).

Finalmente, y resumiendo lo dicho hasta ahora, esta investigación muestra la posibilidad cierta de utilizar la red neural word2vec u otras herramientas similares para el análisis de macrodatos textuales en español. Sobre ellos se pueden realizar las operaciones más simples y que requieren menor potencia computacional, como la concatenación de operaciones que hemos presentado sumariamente, o bien operaciones más gravosas en poder de cómputo que actúen a partir de las distancias absolutas entre palabras, listados de palabras más cercanas, etc.

Las técnicas de análisis concretos de tales operaciones pueden variar según el objetivo del proyecto que nos planteemos o las características de los datos que nos interesen. Esas técnicas pueden generar desde las representaciones gráficas a través de distancias usando la similitud coseno entre palabras (del tipo del gráfico 2) al uso de distinciones (restas) y analogías (sumas). Debido a la naturaleza de la red neural, la interpretación de los resultados arrojados por estas últimas modalidades de análisis es más difícil de trasladar al lenguaje natural, y requeriría de una metodología y planificación especiales que incluyan aportes de la *visualización de la información*. En resumen, el abanico de opciones de investigación usando word2vec sobre un gran cuerpo de datos es notablemente amplio.

El campo de la aplicación de instrumentos de inteligencia artificial al análisis sociosemántico podría convertirse en una subdisciplina de indudable interés sociológico. Esa subdisciplina rendiría resultados objetivamente inatacables acerca de la subjetividad social. Estos resultados, en efecto, iluminarían no solo la estructura de los lenguajes naturales y prevalentes en una determinada sociedad o dominio social, sino que también nos brindarían información relevante de la composición y estructura social misma de los sujetos que utilizan tales lenguajes.

Bibliografía

Bengoechea, Mercedes (2000). «Historia (española) de una sugerencia para evitar el androcentrismo lingüístico». *Revista Iberoamericana de Discurso y Sociedad*, 2(3): 33-58.

Bergvall, Victoria; Bing, Janet M. y Freed, Alice F. (eds.) (1996). *Rethinking Language and Gender Research: Theory and Practice*. London: Addison Wesley Longman.

Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James; Saligrama, Venkatesh y Kalai, Adam (2016). «Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings». Disponible en: https://arxiv.org/abs/1607.06520-https://arxiv.org/abs/1607.06520m, acceso el 5 agosto de 2016.

Buolamwini, Joy y Gebru, Timnit (2018). «Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification». *PMLR*, 81: 77-91. Disponible en: https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms/transcript, acceso el 20 de enero de 2020.

Caliskan, Aylin; Bryson, Joanna J. y Narayanan, Arvind (2017). «Semantics Derived Automatically from Language Corpora Contain Human-like Biases». *Science,* 356(6334): 183-186.

Díaz Martínez, Capitolina (1996). *El presente de su futuro. Modelos de autopercepción y de vida entre los adolescentes españoles.* Madrid: Siglo XXI.

Díaz Martínez, Capitolina (2000). «El análisis sociosemántico en la psicología social: una propuesta teórica y una técnica para su aplicación». *Psicothema,* 12(3): 451-457.

Dunlop, Claire A. (2013). «Epistemic Communities». En: Howlett, M.; Fritzen, S.; Xun, W. y Araral, E. (eds.). *Routledge Handbook of Public Policy*. London: Routledge.

García Dauder, S. y Romero Bachiller, Carmen (2018). «De epistemologías de la ignorancia a epistemologías de la resistencia: correctores epistémicos desde el conocimiento activista». En: Cordero, M.ª T. (comp.). *Discusiones sobre investigación y epistemología de género en la ciencia y la tecnología*. San José: Universidad de Costa Rica, pp. 145-164.

Garg, Nikhil; Schiebinger, Londa; Jurafsky, Dan y Zou, James (2018). «Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes». *Proceedings of the National Academy of Sciences*, 115(16).

Glott, Ruedige; Ghosh, Rishab y Schmidt, Philipp (2010). «Wikipedia Survey. Technical Report, UNU-MERIT». Disponible en: http://wikipediasurvey.org/, acceso el 4 de abril de 2019.

Goddard, Angela y Patterson, Lindsey M. (2005). *Lenguaje y Género*. Cuenca: Ediciones de la Universidad de Castilla La Mancha.

Greenwald, Anthony G.; McGhee, Debbie E. y Schwartz, Jordan L. K. (1998). «Measuring Individual Differences in Implicit Cognition: the Implicit Association Test». *Journal of Personality and Social Psychology* 74(6): 1464-1480.

Haraway, Donna J. (1995). *Ciencia, cyborgs y mujeres. La reinvención de la naturaleza*. Madrid: Cátedra.

Harding, Sandra (1991). *Whose Science? Whose Knowledge? Thinking from Women's Lives.* Milton Keynes: Open University Press.

Harding, Sandra (1996). *Ciencia y feminismo*. Madrid: Morata.

Hass, Peter (2016). *Epistemic Communities, Constructivism, and International Environmental Politics*. London: Routledge.

Healy, Bernadine (1991). «The Yentl Syndrome». *New England Journal of Medicine*, 325(4): 274-276.

Lakoff, Robin T. (2004). *Language and Woman's Place.* Oxford: Oxford Universty Press.

Keller, Evelyn F. (1991). *Reflexiones sobre género y ciencia*. Valencia: Edicions Alfons el Magnànim.

Lam, Heidi; Bertini, Enrico; Isenberg, Petra; Plaisant, Catherine y Carpendale, Sheelagh (2011). «Empirical Studies in Information Visualization: Seven Scenarios». *IEEE Transactions on Visualization and Computer Graphics*, 18(9): 1520-1553.

Longino, Hellen E. (2002). *The Fate of Knowledge.* Princeton: Princeton University Press.

Maffía, Diana H. (2001). «El sexo oculto de la ciencia. Historia de la ciencia y política sexual». En: Pérez-Sedeño, E. y Cortijo, P. (coords.). *Ciencia y Género*. Madrid: UCM, pp. 407-416.

Mako Hill, Benjamin y Shaw, Aaron (2013). «The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation». Disponible en: http://www.oalib.com/paper/ 3023720#. WA1GXeCLQ2w, acceso el 6 de julio de 2018.

Navarro, Pablo y Ariño, Antonio (2015). «La investigación social ante su segunda revolución digital». En: García Ferrando, M.; Alvira, F. R.; Alonso, L. E. y Escobar, M. (coords.). *El análisis de la realidad social. Métodos y técnicas de investigación*. Madrid: Alianza Editorial, pp. 110-141.

Proctor, Robert N. (1995). *Cancer Wars: How Politics Shapes What We Know and Don't Know About Cancer*. New York: Basic Books.

Proctor, Robert N. y Schiebinger, Londa (2008). *Agnotology: the Making and Unmaking of Ignorance*. Stanford, California: Stanford University Press.

Reagle, Joseph y Rhue, Lauren (2011). «Gender Bias in Wikipedia and Brittanica». *International Journal of Communication,* 5: 1138-1158. Disponible en: http://ijoc.org., acceso el 7 de agosto de 2016.

Shiebinger, Londa (2004). *¿Tienes sexo la mente?* Madrid: Cátedra.

Sousa Santos, Boaventura (2010). *Descolonizar el saber, reinventar el poder*. Montevideo: Trilce.

Sullivan, Shannon y Tuana, Nancy (2007). *Race and Epistemologies of Ignorance.* New York: State University of New York Press.

Swinger, Nathaniel; Arteaga, Maria de; Heffernan, Neil Thomas IV; Leiserson, Mark D. M. y Tauman, Kalai Adam (2018). «What are the biases in my word embedding? ». *Proc. of the AAAI/ ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*. Disponible en: https://doi.org/10.1145/3306618.3314270, acceso el 7 de marzo de 2019.

Thomas, Gillian (1992). *A position to command respect: Women and the Eleventh Britannica.* Metuchen. New Jersey: The Scarecrow Press.

Tuana, Nancy y Sullivan, Shannon (2006). «Introduction: Feminist Epistemologies of Ignorance». *Hypatia*, 21(3): 1-19.

Wagner, Claudia; García, David; Jadidi, Mohsen y Strohmauier, Markus (2015). «It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia». *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media*.

Wikipedia (2008). *Wikipedia: WikiProject countering systemic gender bias*. Disponible en: http://en.wikipedia.org/?oldid=183541656 Wikipedia, acceso el 11 de enero.

Wikipedia (2009). *Wikipedia: WikiProject gender studies/countering systemic gender bias.* Disponible en: http://en.wikipedia.org/?oldid=2746106583, acceso el 11 de marzo.

RECEPCIÓN: 21/09/2019
REVISIÓN: 16/12/2019
APROBACIÓN: 25/02/2020

doi:10.5477/cis/reis.172.41

Hidden Gender Bias in Big Data as Revealed by Neural Networks: Man is to Woman as Work is to Mother?

Sesgos de género ocultos en los macrodatos y revelados mediante redes neurales: ¿hombre es a mujer como trabajo es a madre?

Capitolina Díaz Martínez, Pablo Díaz García and Pablo Navarro Sustaeta

Key words

Word Embedding
- Big Data
- Neural Network
- Gender Bias
- Wikipedia

Abstract

Social events become big data. The analysis of big data becomes knowledge about society. If the big data is biased, this bias is transmitted to the analysis and to our knowledge. We propose a tool to discover gender biases and, potentially eliminate them from big data prior to analysis. We use neural network analysis via word embedding. This is the first time that this technique has been tested on a Spanish data body. As proof of concept, the neural network was fed with half of the Wikipedia in Spanish. More than 28 million words.
We describe the techniques and specialized knowledge necessary to discern gender bias and examined whether it is possible to divide the analysis work into externalizable microtasks.

Palabras clave

Encaje de palabras
- Macrodatos
- Red neural
- Sesgos de género
- Wikipedia

Resumen

Los actos sociales se convierten en *big data*. El análisis de *big data* se convierte en conocimiento sobre la sociedad. Si los *big data* están sesgados, el sesgo se transmite al análisis y a nuestro conocimiento. Proponemos una herramienta para descubrir los sesgos de género y, potencialmente, eliminarlos de los *big data* antes del análisis. Utilizamos la técnica de análisis neural mediante el procedimiento de encaje de palabras.
Es la primera vez que esta técnica se prueba con un cuerpo de datos en español. Como prueba de concepto, la red neural analiza la mitad de la Wikipedia en español. Más de 28 millones de palabras.
Se describen las técnicas y los conocimientos especializados necesarios para poder discernir los sesgos de género y se evalúa si es posible dividir el trabajo de análisis en microtareas externalizables.

Citation

Díaz Martínez, Capitolina; Díaz García, Pablo and Navarro Sustaeta, Pablo (2020). "Hidden Gender Bias in Big Data as Revealed by Neural Networks: Man is to Woman as Work is to Mother?". *Revista Española de Investigaciones Sociológicas,* 172: 41-60. (http://dx.doi.org/10.5477/cis/reis.172.41)

Capitolina Díaz Martínez: Universidad de Valencia | capitolina.dm@gmail.com
Pablo Díaz García: Telice Comet | pablo.diaz.13@gmail.com
Pablo Navarro Sustaeta: Universidad de Valencia | Pablo.Navarro@uv.es

Introduction

Big data has the potential to change and improve our world; however, it poses at least two potential and immediate problems. The first is that, the availability of these massive sets of complex data and handling them in an epistemically productive manner is not easy (Navarro and Ariño, 2015) —especially for those who are not specialists in this sort of work. Discovering their biases or flaws is even more challenging. This difficulty often leads to disaffection by those who are not experts in computers. Thus, advances in this area have forced the creation of a new field of knowledge, the so-called *Information Visualization.* The aim of this sub-discipline is to make the visualization of big data be more compressible and, therefore, permit better informed decision making based on this data. This improved information results in a decrease in the social gap existing between those who can and cannot interpret it. In this new field of information visualization, the work of Sheelagh Carpendale (Lam *et al.*, 2011) is especially noteworthy. This pioneering computer scientist, among other things, collaborated with the World Health Organization to make the 2018 "visualized" presentation of its 11th International Classification of Diseases (ICD-11) more accessible to the public. Closely related to the problem of bias in visualization is the issue of bias in big data of human faces. Here, the works of Joy Buolamwini and Timnit Gebru (2018) are of great interest. They have conducted research on US data facial image macrobases. This research reveals that these macrobases are less sensitive in terms of recognizing dark skinned or female faces.

The second major problem, which, given its significance, should be considered with priority, is that the production of big data may result in a distorted and often biased view of the represented phenomenon. This distortion is the reflection of a society that is plagued with numerous divides and inequalities (economic, gender-based, educational, ethnic, health-based, intersectional, etc.), and it may translate into the very interests of those producing the data at hand. Fields of knowledge such as agnotology (Proctor and Shiebinger, 2008) and epistemology of ignorance (Tuana and Sullivan, 2006) have been found to have numerous biases that plague conventional data. When the data are complex and massive, as is the case with big data, the bias issue is even greater: these biases are not found only in the surface of the data, but rather, they extend to its deep structure, that is, the implicit relationships. These relationships are only accessible statistically, since they are only revealed through complex correlations via large sets of data, and they remain invisible to the micro observer, like the roundness of the earth.

The bias problem in data has been considered for some time now, both in artificial intelligence (AI) and the social sciences (especially in the Studies of Science, Technology and Society (STS)). In AI, for example, an attempt has been made to first discover and then avoid bias in concepts related to humans and those that may be morally, socially and politically damaging: Swinger *et al.* (2019); Caliskan *et al.* (2017), etc.

In the field of psychology, for example (Grenwald, McGhee and Schwarz, 1998), the Implicit Associations Test (IAT) was created to determine sexist and racist biases revealed when asking study subjects to associate proper names, images or adjectives to women and men, or to individuals of different races.

Based on numerous studies, feminist theory has stated a *gender-based bias* of science. For example, Sandra Harding (1996: 36) argues that the symbolism of gender, the gender-based structure of science and of the identities and masculine identities and behaviors of scientists, has

left its mark on the issues, concepts, theories, methods, interpretations, ethic, meanings and objectives of science. One example of this gender-based bias of science has been provided by Diana Mafía (2001), who reviewed Linneo, indicating that in the same volume in which the term *mammalia* was introduced, the term *homo sapiens* was also introduced. So, the human couple, according to Linneo, consists of a female mammal and a male *homo sapiens*. The implications and consequences of this distinction are quite apparent.

In order to understand the epistemological gender-based bias of the sciences, it is useful to pay attention to the numerous and well-founded studies of Science, Technology and Gender (Keller, 1991; Harding, 1991, 1996; Healy, 1991; Haraway, 1995; Longino, 2002; García and Romero, 2018; Shiebinger, 2004; Sousa, 2010; Nikhil *et al.,* 2018). Furthermore, it is necessary to consider what Robert Proctor (1995 and 2008) called *agnotology* and what is referred by Nancy Tuana as *epistemologies of ignorance*. Proctor (1995:8) said that we should

> […] study the social construct of ignorance. The persistence of the controversy is often not a result of imperfect knowledge, but rather, a political consequence of conflicts of interest and structural apathies. The controversy may be designed: ignorance and uncertainties may be manufactured, sustained and disseminated.

Along these lines, Nancy Tuana and Sannon Sullivan (2006: i) stated that "the practices of ignorance are often intertwined with practices of oppression and exclusion". In an individual article in the mentioned volume from 2006, Nancy Tuana suggested up to five types of epistemological ignorance. From "knowing that one does not know, without caring" to voluntary ignorance, "not wanting to know". Naturally, these types of ignorance tend to be derived from relationships of power and exclusion.

Quite often, gender and race-based biases go hand in hand and operate with similar epistemological devices, as revealed in the thirteen chapters of the volume *Race and Epistemologies of Ignorance* edited by Shannon Sullivan and Nancy Tuana in 2007.

Gender bias in language has also been studied for some time now, although mainly from a grammatical and intuitive perspective. By intuitive, we mean that the researcher becomes a natural and omniscient interpreter of the language, simply using his/her own understanding and spontaneous response. This intuitive approach to the language is objectionable since it lacks the independent support to back up the researcher's conclusions. In fact, we know the distinct morphosyntactic forms of the representation of gender through language; the cognitive processes affected by sexist language and the stereotyped or prejudicial mental images resulting from the same; the specific analysis of language of school texts, radio programs, etc.; and, to limit this list, the distinct forms of everyday use of the languages of men and women (Bergvall, 1996; Goddard and Patterson, 2005; Lakoff, 2004; Bengoechea, 2000, among many others).

Similarly, in a world of increasingly available macrodata text analysis, many publications refer to all types of sociolinguistic relationships. However, until quite recently, the potential sexist bias implicit in the huge text bases under analysis via big data treatment techniques had not been revealed. These bases are appearing in large volumes and have an ever increasing influence.

In fact, in the analysis of macrodata text bases, as Tolga Bolukbasi *et al.* demonstrated mathematically in his inspiring article "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings", the geometry of word pairs

or word embedding in these databases includes gender-based biases. These biases, although limited to reflecting the persistent sexism of our societies, continue to be quite detrimental in our everyday societies. The gender bias that is detectable in these macrodata reveals the structurally sexist condition of each natural language —a condition that translates, and at the same time, reproduces the environmental sexism of social life—. This article presents a modest case study, conceived at a "proof of concept" level[1], in which we will attempt to unveil the implicit semantics from a reasonably representative sample of the Spanish language, the Spanish Wikipedia. This sample consists of an empirical base that is translatable in terms of big data with which we "will converse". So, using an already proven artificial intelligence technique —the use of neural networks— it reveals which types of errors and biases can be detected (socially induced and at the same time, generators of social effects) in these data macrobases.

The methodological approach that we have selected for this article to examine the phenomenon of semantic distortion of gender relies on the use of the word embedding technique. This is a powerful automatic learning technique in which each word of a language is represented as a vector. The geometric relationship between these vectors captures significant semantic relationships between the corresponding words. These relationships emerge from basic evidence: in many word pairs (supposed expressive words, of concepts) some stereotypical and morally, socially and politically harmful analogies appear, such as those found by Tolga Bolukbasi, in the following pairs:

A is to B, as X is to Z. These analogies appear in pairs, such as:

Man —(is to)— woman = computer programmer —(is to)— homemaker

Man —(is to)— woman = physician —(is to)— nurse

In fact, it is quite noteworthy that such a perceptive and "intelligent" algorithm which may be capable of capturing the following type of relationships: *if Paris is to France, Tokyo is to X*, and with sufficient "knowledge" to find that *X is Japan,* will respond, when asked *if man is to Woman, Computer programmer is to X,* suggesting that the result for *X* is *Homemaker*.

Based on this type of evidence, in the cited article, Tolga Bolukbasi warns us of "flagrant sexism of the word embedding and, therefore, of the risk of introducing biases of several types into real world systems" (*op. cit.*, 1). We do not concern ourselves with whether or not, in the three-million-word database of Google News, used by Bolukbasi, he and his team found word associations such as that used in the title of the article, since almost any use of this database will not only reproduce, but increase the gender bias of the same. Even more so when this is increasingly mediated by artificial intelligence techniques/algorithms. We can imagine, for example, that a company needs to hire experts in programming, and for this, Google is requested to provide a few dozen names. This search engine obtained such names from the macrobase of data from Google News. However, in this huge database, the word "Woman" does not appear as being compatible with "Computer programmer", at least in terms of statistical probability. Therefore, no woman appears in the Google list that is sold to a client company and no woman will be hired. In this way, the huge detour of microdata manages to extend the

[1] According to the Wikipedia: "A proof of concept or PoC is a realization of a method or idea in order to demonstrate its feasibility, or a demonstration in principle with the aim of verifying that some concept or theory has practical potential". https://en.wikipedia.org/wiki/Proof_of_concept

vicious cycle of gender-based discrimination.

The extraordinary potential of the word2vec with which Bolukbasi analyzed the 3 billion words of data from Google News has led us to believe that a similar study of a Spanish macrodata base would be relevant.

The study conducted for this article is limited to providing a description of how we have adapted word2vec to a macrodata base in Spanish and considers the tests performed, with this application and on this empirical base, to verify its functioning and appropriateness for the indicated purpose: to analyze semantic bias found in the data macrobase. As detailed below, we have used, as a macrodata base, part (half) of the Spanish language version of the Wikipedia from 2006.

Before continuing, we must recognize that we consider the Wikipedia with the knowledge that it includes gender bias. We expect, however, that with the use of the neural networking technology, we will be able to discover gender biases that are more profound than the obvious ones (including the counting of a larger presence of masculine personalities as compared to the minoritization of the feminine ones). We have found embedded gender biases in the complex semantic structure of all of the linguistic space covered by and appearing in the Spanish language Wikipedia.

Various publications warn us of the gender biases found in the Wikipedia. These include studies conducted by the very Wikipedia organization, such as the WikiProject "Countering Systemic Bias", which includes reports (2008 and 2009) of the analysis of gender bias of this on-line encyclopedia, as well as the very encyclopedia entry on gender bias[2]. In these reports the gender bias is explained, above all, by factors such as the unique demographic of the editors who create the Wikipedia through their contributions. In fact, this global collective company is creating its own Meta-Wikipedia. An example of this space for self-reflection is provided by the page at https://meta.wikimedia.org/wiki/Gender_gap which is open to the inclusion of all studies examining gender bias in the Wikipedia[3]. Apart from the strict Wikipedia environment, Josep Reagle and Lauren Rhue (2011), in their comparison between the Wikipedia in English and the on-line version of the Encyclopaedia Britannica, found that the Wikipedia, although having fewer articles on women than men, includes more articles on women in absolute terms, as compared to the Encyclopaedia Britannica. Studies by Glott *et al.* (2010) reveal that women make up approximately 13 % of all Wikipedia users. According to Benjamin Mako Hill and Aaron Shaw (2013), this figure could be even higher: upon reviewing the surveys analyzed by Ghost and colleagues, the previous authors found that the percentage of female US editors reached 22.7 % and the total percentage of female editors was 16.1 %. Currently, the most thorough study to have been conducted is that of Claudia

[2] https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia

[3] In the Wikimedia space https://meta.wikimedia.org/wiki/Gender_gap, in a clear demonstration of its concern over gender bias, the following articles are cited: *Unlocking the Clubhouse: Five ways to encourage women to edit Wikipedia*, Sue Gardner's Blog, November 14, 2010; *Gender Gap? Look Up Wikipedia's Contributor List*, Noam Cohen in the New York Times, January 30, 2011; *Where Are the Women in Wikipedia? Debate with a number of debaters*, New York Times, February 2, 2011; *Wikipedia(EN) Signpost Issue on Gender Gap*, February 7, 2011; *Top 10 Reasons to Encourage more Women Participation in Wikipedia here on Meta*, February 8, 2011; *Nine Reasons Women Don't Edit Wikipedia (in their own words)*, Sue Gardner's Blog, February 19, 2011; *Wikipedia's gender gap and the complicated reality of systemic gender bias*, Adrianne Wadewitz in HASTAC, July 26, 2013; *Gender Gap Manifesto*, created March 2011 by seven Wikimedia editors; *Charting Diversity: Working together towards diversity in Wikipedia,* Wikimedia Deutschland cooperated with the Beuth University of Applied Science.

Wagner *et al.* (2014), which analyzes biases in coverage (proportion of articles on noteworthy women and men), structural biases (the probability that an article about an individual of one sex will be linked to another about an individual of another sex), lexicon biases and visibility biases (proportion of women and men in the initial pages of the Wikipedia in English).

But what we do not yet know and are attempting to determine through the application of the word2vec neural network, is whether or not the Wikipedia has gender, semantic-structural biases, similar to those found by Bolukbasi in Google News. In fact, we have corroborated, although in the minimal format required for a proof of concept, that, at least in the Spanish language Wikipedia from 2006, these gender biases are found. This comes as no great surprise; but it is interesting to note that in our considerable sample of over 28 million words, the neural network algorithm used associated pairs such as: Man is to Expert as Woman is to Know-to-all. This was the first shocking analogue pair. Upon revealing the implicit semantic analysis of Wikipedia cases such as this one, this online macro encyclopedia continues with the tradition of the other conventional encyclopedias. So, for example, and as detailed by historian Gillian Thomas in her study on the 11th edition of the Encyclopaedia Britannica, women appearing in it tend to be "perceived as pedantic servants given the broad scope of the masculine intelligence" (Thomas, 1992: 18- 26).

The word2vec neural network

Word2vec is a neural network, originally created by Google, which processes text. It may be applied to the natural written language, or to genes, codes, likings, music lists, and generally speaking, to any verbal or symbolic series from which patterns may be extracted.

A neural network is an automated computer model of learning inspired by the biological nervous systems. We call learning the transformation of the behavior of the neurons as a result of the processing of stimuli acting on the network. For example, in a neural network trained to recognize photos of cats, the input will only recognize certain angles and colors; however, when the patterns repeat, little by little it will recognize fur, eyes, ears, tail, etc. In the final state, when all of the neurons corresponding to the cat structure are activated, the output neuron (in our network hypothesis, there will only be one neuron corresponding to a complete cat) will turn on, informing us that a cat has been recognized.

Word2vec analyzes and converts the words introduced through sets of sentences into numbers (a vector). In addition, it seeks the probability that the discreet states[4] found in the network coexist. It manages this by dividing the vectors that it creates —the so-called neural word embeddings— into a vectorial space of *n* dimensions.

In our case, the dimension of the vectors is 500, and each word is defined in its interaction in these 500 dimensions (that is, not one by one). All of the dimensions define all of the words, since each word is situated in the 500-dimension space and the vector is no more than their coordinates.

To understand just what a vector of a word is, we must think that each element in the vector is associated with a vocabulary word from the corpus of data that we are going to analyze. In our case, as we have stated, the dimension of each vector is 500. In this vector, each word is represented by a distribution of weights via these dimensions. So, instead of a one-by-one assignment between one element in the vector and a word, the representation of a word

[4] A discreet system is one with a countable number of states; in our case, each of the words in the dictionary.

extends through all of the elements in the vector, and each element in the vector contributes to the definition of many words.

The word vectors help the computers to "learn" the text. In this way, the program becomes an *artificial semantic interpreter*, which in some manner "emulates" the *natural semantic interpreters*, which are the humans, in our spontaneous use of the natural languages that we know and in which we are socialized.

The embeddings: a mathematical definition

An embedding is an instance of a mathematical structure contained within another.

In order for an embedding to arise, our structure X, which we wish to embed in Y, must be inserted invectively, with a function such as $f: X \rightarrow Y$.

Each of the words analyzed by the neural network cannot be analyzed separately, but rather, in defined groups; in the case of the natural language, these groups are sentences. If, for example, there is a network of these characteristics with the following sentences:

— My dog is blue.

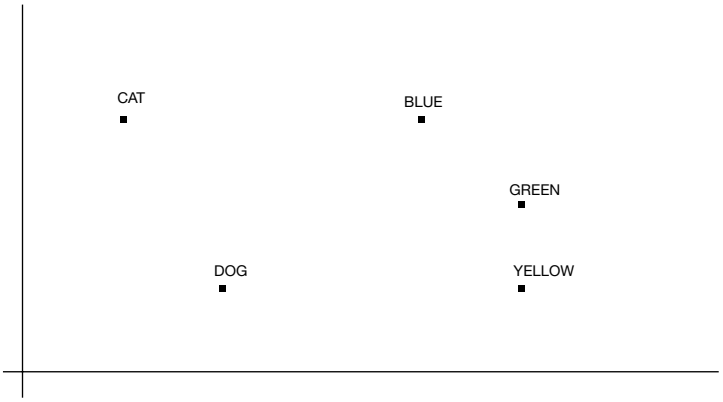
— My cat is blue.

— My dog is green.

— My dog is yellow.

In the resulting space, the words "blue", "green" and "yellow" are grouped between them, all at the same distance from "dog"; but only "blue" will be close to "cat", which is grouped with "dog".

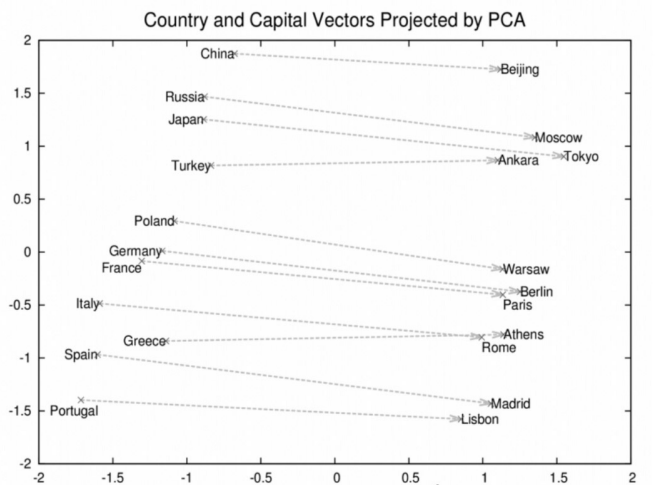
GRAPH 1. *Bi-dimensional representation of a simple neural network model*



Source: Own graphic.

Throughout the many iterations with numerous sentences, similar words are grouped together and distanced from one another in a congruent manner, thanks to the magic of mathematics. This is evident in the following real set of countries in front of the capitals. Observe how the neural network "has learned" not only which are the capitals of each country, but has also positioned them in approximately geographical order (from east to west and north to south).

GRAPH 2. *Country and capital vectors projected by distances*



Source: https://deeplearning4j.org/word2vec[5]

It may be useful to recall that these representations are two-dimensional projections and do not agree with the real underlying model, which, in this case, is 500-dimensional. To translate the results once again to the human terrain, as representation of the distance between vectors, the following cosine is used: (cos (90) is 0, cos (0) is 1).

In our instructions for use, the words to analyze come from texts in the natural language, such as an encyclopedia, Google News articles or the Bible, assuming that these texts are digitalized. After the learning process, the neural network has transformed the text into numbers, with which the computer may carry out large scale mathematical operations; be it sums, subtractions or, more interestingly, systems of equations that permit relationships between such suggestive words as:

— Geopolitics: *Iraq – (minus) Violence = Jordan*
— Distinctions: *Human – Animal = Ethics; President – Power = Prime Minister; Library – Books = Room*
— Analogies: *Stock Market ≈ Thermometer*

Chris Moody, in his didactic article "A word is worth a thousand vectors"[6], provides another example of the natural language in English

king–man+(plus) women = queen (king-man + woman = queen)

This author presents us with the following example: a human asked a computer a question: What is king – *man* + *woman*? (this is a mathematical equation: king minus man plus woman is equal to X). And the computer solved this equation by responding: *queen*. The machine "understands" that the biggest difference between the words man and woman is the gender. If the gender difference is added to "king", "queen" is obtained.

[5] Principal Component Analysis (PCA) is a statistical procedure in which orthogonal transformation is used to convert a set of observations on potentially correlated variables into a set of linearly uncorrelated variables, the so-called "principal components".

[6] http://multithreaded.stitchfix.com/blog/2015/03/11/word-is-worth-a-thousand-vectors/ (downloaded 20 October, 2016). The description of the example is an adapted and simplified translation of the explanation by Chris Moody.

GRAPH 3. *Bi-dimensional projection of semantic distance between words*

Source: Bolukbasi (*op. cit.*: 11).

The results are noteworthy, since the machine had *never* been explicitly taught anything about gender. In fact, the computer had never been provided with anything such as a dictionary, thesaurus or network of relationships between words. A lot of text had simply been entered into word2vec, waiting to see what the machine "would learn" from the context of each word, through the corresponding algorithm. For each word, the algorithm attempts to predict the words that accompany the others in a sentence. Or, better said, it internally represents the words as vectors, and given a word vector, attempts to predict the other word vectors in the nearby text.

Ultimately, the word2vec algorithms scan examples that may interfere with the gender of a single word, just as they can infer that *The Times* and *The Sun* are newspapers, that *Matrix* is a science fiction movie and that the style of a piece of clothing may be considered *hippy* or *formal*. The fact that these vectors represent a large part of the available information from a dictionary definition is a secondary effect —convenient and almost miraculous— of this algorithmic task which consists of attempting to predict the context of a word"[7].

Gender bias in word2vec

By blindly analyzing the texts without an *a priori* perspective, word2vec has the *advantage* of being somewhat "objective", although at the same, it has the disadvantage of dragging and possibly multiplying any bias (sexist or other) that may be present in the original texts with which the program is fed and trained. Bolukbasi *et al.,* analyzing the information from Google News, identified the gender bias attributed to the words "he" and "she", as revealed in the following graph.

Graph 3 corresponds to a series of words selected and projected over two axes (X and Y). X is a projection of the difference between the embedding of the

[7] Free and simplified translation of the explanation by Chris Moody *(op. cit.)*.

words “HE” and “SHE”. On the other hand, Y is a learned direction in the embedding that captures the gender neutrality, situating the neutral words in relation to the gender above the Y axis, and the non-neutral ones below it.

Our analysis process with word2vec consists of using the neural network implemented to identify, diagnose and denounce possible bias present in the corpus of the base (the part of the Wikipedia in Spanish that will be processed).

Creation of our Word2vec Neural Network

The version of word2vec that we use in this proof of concept is available in the *deeplearning4j* free software project maintained by the Skymind corporation[8].

Before proceeding with the creation of a neural network, and since we wish to analyze texts in Spanish, we should adapt the word2vec code, originally designed to process texts in English. For this, it is necessary to know a bit more about how this application functions and about the procedure that is used to analyze the text entered.

When analyzing text, word2vec distinguishes between two basic units: single words or *tokens,* and words grouped in sentences. The program has some basic classifiers that permit the identification of each group of characters between two spaces, such as a *token* and each group of tokens between two spaces as a sentence. This classification is important since the learning is based on how the words are positioned in a sentence.

In our case, we have used entire paragraphs as sentence and single words (units between two spaces) as *tokens*. To minimize duplicates and inconsistencies, we have filtered and removed any character that was not a letter or a dash (-), in order to prevent, for example, “dot” and “grammar dot” from being considered two distinct vocabulary entries.

The adaptation of the corpus to be studied depends in large part on the specific syntax of the corpus; if we know the content of the corpus with great precision, many of these previous processes can be refined, in order to eliminate ambiguities such as plurals, compound names and first names that we wish to highlight or ignore.

After building these classifiers (*tokens* and *sentences*), it is only necessary to train the neural network. To do so, certain attributes should be defined, including the following which are of special interest in this study:

Minimal frequency of words: Indicates how many times a token should appear in the corpus in order to be recognized. Adjusting this aspect helps filter superfluous words.

Window size: Indicates the number of steps forward and backwards that are considered within a sentence when comparing and analyzing each *token*. Initially, the larger the window, the better; but if sentences are small or we have limited computational power, it may be better to reduce this window size. In our case, and after various tests, this size was set to eight.

Iterations: They indicate the number of times that the algorithm runs and sequentially analyzes the corpus. Although technically speaking, different ways of qualifying these sweeps are determined, in our case we can generalize them to the number of times that a full reading of the text is made.

As with the previous case, the higher the values assigned to the iterations, the more precise the network. However, it should be noted that the assignment of high values to the iterations exponentially increases the time of network training.

[8] Available at https://deeplearning4j.org/word2vec

Analysis of a result of the use of word2vec

At first, an attempt was made to analyse a corpus consisting of 15 pages of news from the FECyT (Spanish Foundation of Science and Technology). However, despite containing 28,089 words, the network obtained was of very poor quality. This was due to the limited scope of the vocabulary of the corpus and the limited syntax of the indicated news, all surely written by the same individual and based on the same guidelines[9]. This initial result led us to seek out a larger corpus, specifically, one that was extracted from the Spanish language Wikipedia from 2006, courtesy of the *Universidad Politécnica de Catalunya*[10]. The advantage of using this specific corpus, aside from the large quantity of information that it contains, lies in the fact that the UPC offers it already reduced in plain text, saving us considerable work. On the other hand, as indicated above, the Wikimedia Foundation has recognized the existence of gender bias and a lack of diversity in its encyclopedia. This, from the onset, increases the possibilities of success for our scrutiny.

Given the limitations of time and computational power, the neural network was trained using only half of the indicated corpus, 28,291,729 words. Ten iterations were performed. The time of training, using these parameters, totaled 38 hours. The total weight of the database after the training was 1.16 GB, a significant size. Once the neural network was trained, an analysis of the results was performed.

[9] Indeed, Chris Moody (op. cit) warned that the vectorization of words requires a large quantity of the same and that words can be downloaded from any text; but if it has a very specialized vocabulary, a large quantity of text will be required in order to train the vectors. Normally, this means billions of words.

[10] Available at http://www.cs.upc.edu/~nlp/wikicorpus/

Analysis techniques

As previously shown, the neural network generates a collection of embeddings that represent each of the vocabulary words from the corpus, ordered in such a way that the distance between them indicates the proximity of their use in the language with which the network is trained.

With these embeddings, we perform a *chain of operations* like the following:

$word_2 + word_3 - word_1 = ?$

This operation, translated into natural language, is a comparison of the following type: “one is to two as three is to...”. The result of the operation presents the 10 closest words to the equation. For example:

Man is to actor as woman is to:

(actor, actress, woman, nominee, isbert, trudie, oscar, hodiak, haymes, karina)

Although we have used a relatively large corpus (more than 28 million words), deficiencies clearly exist in terms of having sufficient iterations when training the network, as revealed by the results that are automatically generated by the algorithm. So, it is necessary for people (with special training) to interpret, as experts, each output. We established the following interpretation criteria:

“*Good*”*,* when the result includes a majority of logical and semantically correct words and does not have any evident gender bias. The previous example would be “good” because “actress” appears in second place (although not in the first) and, in general, the list of associated concepts does not contradict the semantic consistency of the Spanish language. It is seen that if “actor” appears before “actress” this is most likely due to the predominantly inclusive nature (of the feminine) given to the masculine form.

"Biased", when the result is logical and semantically correct, but there is evidence of a clear gender bias, be it direct or specific to the area[11].

"Not valid", when the result is absurd.

Having defined these criteria, an analysis is carried out on the model, using the following structure: *"Man is to word as woman is to…*", and vice versa, *"woman is to word as man is to…"* The set of words used totaled 110, distributed amongst the following domains: professions, attitudes, objects and life trajectory.

Analysis of general results

As previously mentioned, it is important to note that the results obtained here are only illustrative of a proof of concept regarding the possibilities of use of a neural network as a tool for the analysis of socio-semantic gender differences (Díaz, 1996, 2000)[12]. More refined results may have been obtained if we had used the microtask technique used by Bolukbasi's team in the previously mentioned work. This proof of concept has its limitations, since the obtained results were analyzed by only two individuals, throughout two three-hour sessions. Bolukbasi, on the other hand, used the microtasks technique[13] and relied on hundreds of collaborators who offered additional validity to his team's results.

The first round of questions: *"Man is to X as woman is to..."* produced a result having 71 % coherence (that is, good results which, at the same time, were biased), of which, 38.2 % were biased. The second round, *"Woman is to X as man is to..."* produced 63.4 % coherence, but only 19.6 % bias.

The lack of precision in the feminine questions (*Woman is to X*…), results from the scarce presence of feminine qualifiers in the corpus and the low density of these as compared to the masculine ones. It is important to note that this bias, will be called *bias of omission*, as detailed below.

Analysis of biases

The number and variety of biases detected in our analysis leads to a level of complexity of the results which should be classified based on these biases. We can sort them in three categories: direct, semantic and omission-based.

Direct bias

Direct bias are those that are reflected in the results, which clearly indicate a strong "semantic gender inequality"[14], since they include surprising and even offensive terms. Some of the examples obtained are:

— *Man is to expert as woman is to know-to-all*

— *Man is to fidelity as woman is to obedience*

— *Man is to work as woman is to mother*

[11] By "area" here we refer to the semantic set created by the ten words that appear as a result of the equation. We have often found that many of the words associated with *woman* are of the family area, a clear gender bias since this association is not found in the case of the word *man*. This bias, as we will see later, has been called "*familization*".

[12] These two publications offer an approximation of the socio-semantic differences similar to those presented here, but that use quite different methods and techniques (among other things, because the neural networks had not yet been invented).

[13] We lack the sufficient space to explain this outsource pay-per-task technique, so we refer to the mentioned article by this author.

[14] We propose the use of the terms "igualación/desigualación semántica de género" (*gender semantics equalization/inequalization*) instead of the more obvious "igualdad/desigualdad semántica de género" (gender semantics equality/inequality), to indicate the non-static and dynamic, cumulative and intentional nature of the gender differentiation processes operating at a semantic level in our natural languages.

— *Man is to furniture as woman is to shoes, textiles*
— *Man is to intelligence as woman is to showing off*
— *Woman is to attorney as man is to stylist*

This type of bias is easily identifiable and its interpretation does not require any previous knowledge of the corpus, so, a network of microtasks may be used for its subsequent assessment and classification as "good", "biased" or "not valid".

The sociological analysis of each of these embeddings may be quite rewarding, and offers very forceful clues as to the gender stereotypes that infiltrate not only our language, but also our social view and practices. In this study, however, we do not perform this sociological analysis. Our recent discovery of this methodology does not allow us to advance further than the operationalization of the neural network, and the implementation of the proof of concept that is presented here. For this, we offer only some brief notations related to the selected examples, considering their semantic interpretation when this interpretation is not obvious. So, of the six previous examples, the fourth and sixth may require some clarification. In fact, in the fourth example: *"Man is to furniture as woman is to shoes, textiles"*, we understand that this reveals, at least, the traditional sexual division of work that is based on the masculinization or feminization of certain professions. Men appear to be associated with the production of furniture and carpentry, while women have a privileged relationship with clothing, their production, creation, etc. It may be observed that a clear occupational distinction is embedded in the language, evident to anyone who knows the distribution of men and women in the labor market. So, the neural network unexpectedly hoards a "knowledge" which, in the "natural intelligence" world (of real life individuals), we only tend to attribute to labor or gender sociologists.

A detailed analysis of *shoes* is necessary, which, initially appearing quite transparent, requires that further questions be asked to the neural network in order to determine new linkages, relationships, etc. that may offer added meaning to this association.

The sixth example, *"Woman is to attorney as man is to stylist"*, is of special interest, and without a doubt, requires further clarification from the network. In the current phase of analysis, we do not attempt to state that this association indicates a *reverse gender bias*: while typically men are associated with the more prestigious professions as compared to women, in this case, men appear associated with a profession (stylist) which is apparently less prestigious than that linked to women (attorney). This anomaly is probably what we would call a "gender reversal" caused by the fact that in our corpus very few female attorneys appear, and these tend to be associated with words that appear infrequently such as stylist. In any case, we have barely found examples of this type, which in itself would be a proof of anomaly and very much the minority of this so-called gender reversal. In any case, it is a phenomenon that should be explored and outlined.

The study that we propose considers that in the future, a larger and more in-depth examination will be carried out in the same line. This progression should rely on the creation of new questions and/or other more complex mathematical formulations that this initial work could not carry out. We refer to the study of absolute distances between words or a listing of closer words, techniques that allow us to offer a more thorough reasoning to the gender bias present in the Spanish language Wikipedia from 2006 or analogous works.

Semantic bias

This type of bias is quite subtle and difficult to observe. However, it is here where the neural network is especially noteworthy. The most generalized and extensive bias that we have found in our study is that of *familization of women*: the term "woman" almost always appears surrounded by terms that relate to the family, while the term "man" appears as an independent entity. For example:

Man is to love as woman is to:
(mother, daughter, partner, children, wife, sister)

Woman is to love as man is to:
(spirit, god, world, desire)

Man is to house as woman is to:
(mother, family, daughter, wife, sister)

Woman is to house as man is to:
(town, time, life)

These examples overwhelmingly reveal the so-called "semantic bias", a bias that strongly distorts our language (somewhat like mass distorts Einstein's space-time). Although these revealing four results deserve a much more detailed examination, we offer a few observations that should be considered guidelines for a more in-depth and sociologically based examination.

It is quite evident that the terms/concepts associated with "man" are quite general and abstract as compared to the specific and family-centered ones that tend to be linked to "woman". This greater abstraction of the masculine concepts is often paired with others of the immaterial nature of the same (spirit, god, desire, etc.). Concepts related to women, on the other hand, tend to be much more tangible and, in this sense, material. It is of interest to note that the masculine concepts do not include specific people, but rather, impersonal entities, whereas the feminine ones always refer to specific, physical and well-known individuals. The socio-semantic gender contrast that is clearly more revealing and which brings together and synthesizes the previously mentioned ones, is the "man socio-centered/ female family-centered" pair (Díaz, *op. cit.*). By socio-centered we mean "focused" on the world beyond the family (town, world), in addition to abstract entities. The family-centered condition is sufficiently clear so as to warrant no further comments. Therefore, with the previous, the dimensionality of terms/concepts of the "man" appears to be superior to that of the "woman", who is enclosed in a family–based single dimension.

Another modality of semantic bias that has been observed is the *sexualization of the feminine environments* as compared to the masculine ones.

Woman is to lesbian as man is to:

(fetish, fiery, musings, spotless, libertine)

This sexualization of the feminine environments is, without a doubt, induced by men, and appears to be related to the dissemination of the pornography industry.

An even more subtle semantic bias may be seen in the following pair:

Woman is to ambassador as man is to:
(Unicef, acnur, (United) Nations, microcredit)

Man is to ambassador as woman is to:
(tymoshenko, viscountess)

The previous pair indicates that when we ask neural network about a female ambassador, the only association with men that it finds is when these men work in good will organizations or similar, such as the United Nations or non-profit organizations, not comparable to the masculine "ambassador" who represents State powers. The sole neural network is made to a persona of an analogue role, "tymoshenko" (supposedly Yulia Tymoshenko), or an (unidentifiable) title of nobility. In this pair, and without further information, semantic gender biases arise as well

as biases by omission, since not enough saturated embedding is found so as to be representative (the network does not find a sufficient number of relationships).

Bias by omission

As previously mentioned, this type of bias arises when the weight of one of the parts of the corpus is so limited that the fact that it operates with the corresponding embedding results barely significant due to a lack of cases (the so-called *lack of saturation*). This deficiency may be detected with regard to concepts such as "physician", "mathematician", "chemist", etc. in which the results always refer to the meaning of "physical science", "mathematical science", "chemical science", etc., but not to the meaning of "woman (professional of) physics, mathematics, chemistry, etc.".

This bias by omission may also[15] be observed, for example, in the following operation:

Woman is to queen as man is to:
(king, amidala, prince, naboo)

Man is to king as woman is to:
(daughter, woman, wife)

On initial glance, it may appear that the first component of the pair offers an appropriate result (king is the first response and prince is the third), however, two of the four concepts that are returned by the network refer to the *Star Wars* films, suggesting that the majority of the instances of "queen" appearing in articles refer to the queen of this movie series, Queen Amidala. This strange interference by a film heroine in a much broader concept such as that of queen, may be due to the collective profile of the Wikipedia authors, all of whom are most likely socialized in the mentioned films.

In the second pair, when asking the reverse (man is to king as woman is to …), the algorithm fails to return the term "queen", but rather, directly speaks of "daughter", "woman" and "wife". This result, apart from once again exemplifying the *familization* that was detected in the previous point, also indicates that in the Wikipedia articles, when men are kings, the women in their environment appear only as their family members. We find this same relationship of omission when asking for "profesora" (female professor), obtaining only personalities from the Harry Potter book saga.

Conclusions

From our results, it is easy to infer that there is a global gender bias in the Spanish language Wikipedia from 2006 and, especially, a large omission and *familization* of women in the published articles. On the other hand, men appear as substantive entities in their individuality.

It should be tentatively stated that the presence of concepts related to the world of youthful and young adult fantasy allows us to risk making a demographic estimate of the individuals who edited this Wikipedia version: largely males aged between 20 and 30. These conclusions are not far from the reality of the current Wikipedia, which is much more complex than that of 2006[16]. These conclusions are quite in line with those made by the very Wikipedia, which attempts to resolve these issues, as mentioned in the introduction to this article.

With sufficient computational power, we believe that the analysis of a neural network can greatly facilitate the diagnosis of bias problems (not only gender bias) within a large and varied corpus. In the digital age in which we live, there are numerous possibilities in this sense: publications by a specific

[15] In Spanish the name of the professions has grammatical gender. Here we refer to "she-physician", "she-mathematician", etc.

[16] https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia

editorial group, bibliography of an author or group of authors, followers of a political party in the social networks, etc. We may even analyze all of the scientific publications of a same area (or, of all areas).

Finally, and in summary, this study demonstrates the clear possibility of using the word2vec neural network, as well as other similar tools, to analyze the textual macrotexts in the Spanish language. With these, it may be possible to carry out the simplest operations requiring less computational power, such as the interlinking of operations presented in this abbreviated manner, or more onerous operations (in terms of computational power), acting based on the absolute distances between words, lists of closest words, etc.

The specific analysis techniques used in these operations may vary depending on the project's proposed objective or the characteristics of the data of interest. These techniques may be generated from graphic representations via distances using the cosine similarity between words (as in graph 2), or through the use of distinctions (subtractions) and analogies (sums). Given the nature of the neural network, the interpretation of the results produced by these latter analysis modalities is more difficult to translate to the natural language, and will require special methodology and planning, including contributions from *Visualization of Information*. In summary, the range of research options using word2vec on a large body of data is quite extensive.

The field of application of artificial intelligence instruments to the socio-semantic analysis may become a sub-discipline of great sociological interest. This sub-discipline will objectively render indisputable results on social subjectivity. And these results may illuminate, not only the structure of the natural and prevalent languages in a specific society or social domain, but they may also provide relevant information on the social composition and structure of the individuals using these languages.

Bibliography

Bengoechea, Mercedes (2000). "Historia (española) de una sugerencia para evitar el androcentrismo lingüístico". *Revista Iberoamericana de Discurso y Sociedad*, 2(3): 33-58.

Bergvall, Victoria; Bing, Janet M. and Freed, Alice F. (eds.) (1996). *Rethinking Language and Gender Research: Theory and Practice*. London: Addison Wesley Longman.

Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James; Saligrama, Venkatesh and Kalai, Adam (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". Available at: https://arxiv.org/abs/1607.06520https://arxiv.org/abs/1607.06520m, access August 5, 2016.

Buolamwini, Joy and Gebru, Timnit (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". *PMLR*, 81: 77-91. Available at: https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms/transcript, access January 20, 2020.

Caliskan, Aylin; Bryson, Joanna J. and Narayanan, Arvind (2017). "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases". *Science,* 356(6334): 183-186.

Díaz Martínez, Capitolina (1996). *El presente de su futuro. Modelos de autopercepción y de vida entre los adolescentes españoles.* Madrid: Siglo XXI.

Díaz Martínez, Capitolina (2000). "El análisis sociosemántico en la psicología social: una propuesta teórica y una técnica para su aplicación". *Psicothema,* 12(3): 451-457.

Dunlop, Claire A. (2013). "Epistemic Communities". In: Howlett, M.; Fritzen, S.; Xun, W. and Araral, E. (eds.). *Routledge Handbook of Public Policy*. London: Routledge.

García Dauder, S. and Romero Bachiller, Carmen (2018). "De epistemologías de la ignorancia a epistemologías de la resistencia: correctores epistémicos desde el conocimiento activista". In: Cordero, M.ª T. (comp.). *Discusiones sobre investigación y epistemología de género en la ciencia y la tecnología*. San José: Universidad de Costa Rica, pp. 145-164.

Garg, Nikhil; Schiebinger, Londa; Jurafsky, Dan and Zou, James (2018). "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes". *Proceedings of the National Academy of Sciences*, 115(16).

Glott, Ruedige; Ghosh, Rishab and Schmidt, Philipp (2010). "Wikipedia survey. Technical report, UNU-MERIT". Available at: http://wikipediasurvey.org/, access April 4, 2019.

Goddard, Angela and Patterson, Lindsey Miard (2005). *Lenguaje y Género*. Cuenca: Ediciones de la Universidad de Castilla La Mancha.

Greenwald, Anthony G.; McGhee, Debbie E. and Schwartz, Jordan L. K. (1998). "Measuring Individual Differences in Implicit Cognition: the Implicit Association Test". *Journal of Personality and Social Psychology,* 74(6): 1464-1480.

Haraway, Donna J. (1995). *Ciencia, cyborgs y mujeres. La reinvención de la naturaleza.* Madrid: Cátedra.

Harding, Sandra (1991). *Whose Science? Whose Knowledge? Thinking from Women's Lives.* Milton Keynes: Open University Press.

Harding, Sandra (1996). *Ciencia y Feminismo*. Madrid: Morata.

Hass, Peter (2016). *Epistemic Communities, Constructivism, and International Environmental Politics*. London: Routledge.

Healy, Bernadine (1991). "The Yentl Syndrome". *New England Journal of Medicine*, 325(4): 274-276.

Keller, Evelyn Fox (1991). *Reflexiones sobre género y ciencia*. Valencia: Edicions Alfons el Magnànim.

Lakoff, Robin T. (2004). *Language and Woman's Place.* Oxford: Oxford Universty Press.

Lam, Heidi; Bertini, Enrico; Isenberg, Petra; Plaisant, Catherine and Carpendale, Sheelagh (2011). "Empirical Studies in Information Visualization: Seven Scenarios". *IEEE Transactions on Visualization and Computer Graphics*, 18(9): 1520-1553.

Longino, Hellen E. (2002). *The Fate of Knowledge.* Princeton: Princeton University Press.

Mako Hill, Benjamin and Shaw, Aaron (2013). "The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation". Available at: http://www.oalib.com/paper/3023720#. WA1GXeCLQ2w, access July 6, 2018.

Maffía, Diana H. (2001). "El sexo oculto de la ciencia. Historia de la ciencia y política sexual". In: Pérez-Sedeño, E. and Cortijo, P. (coords.). *Ciencia y Género*. Madrid: UCM, pp. 407-416.

Navarro, Pablo and Ariño, Antonio (2015). "La investigación social ante su segunda revolución digital". In: García Ferrando, M.; Alvira, F. R.; Alonso, L. E. and Escobar, M. (coords.). *El análisis de la realidad social. Métodos y técnicas de investigación*. Madrid: Alianza Editorial, pp. 110-141.

Proctor, Robert N. (1995). *Cancer Wars: How Politics Shapes what We Know and Don't Know About Cancer*. New York: Basic Books.

Proctor, Robert N. and Schiebinger, Londa (2008). *Agnotology: the Making and Unmaking of Ignorance*. Stanford, California: Stanford University Press.

Reagle, Joseph and Rhue, Lauren (2011). "Gender Bias in Wikipedia and Brittanica". *International Journal of Communication,* 5: 1138-1158. Available at: http://ijoc.org, access August 7, 2016.

Shiebinger, Londa (2006). *¿Tiene sexo la mente?* Madrid: Cátedra.

Sousa Santos, Boaventura (2010). *Descolonizar el saber, reinventar el poder*. Montevideo: Trilce.

Sullivan, Shannon and Tuana, Nancy (2007). R*ace and Epistemologies of Ignorance*. New York: State University of New York Press.

Swinger, Nathaniel; Arteaga, Maria de; Heffernan, Neil Thomas IV; Leiserson, Mark D. M. and Tauman, Kalai Adam (2018). "What are the biases in my word embedding?". *Proc. of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*. Available at: https://doi.org/10.1145/3306618.3314270, access March 7, 2019.

Thomas, Gillian (1992). *A Position to Command Respect: Women and the Eleventh Britannica.* Metuchen, New Jersey: The Scarecrow Press.

Tuana, Nancy and Sullivan, Shannon (2006). "Introduction: Feminist Epistemologies of Ignorance". *Hypatia*, 21(3): 1-19.

Wagner, Claudia; García, David; Jadidi, Mohsen and Strohmauier, Markus (2015). "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia". *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media*.

Wikipedia (2008). *Wikipedia: WikiProject countering systemic gender bias*. Available at: http://en.wikipedia.org/?oldid=183541656 Wikipedia, access January 11, 2008.

Wikipedia (2009). *Wikipedia: WikiProject gender studies/countering systemic gender bias.* Available at: http://en.wikipedia.org/?oldid=2746106583, access March 11, 2009.

RECEPTION: September 21, 2019
REVIEW: December 16, 2019
ACCEPTANCE: February 25, 2020