



**UNIVERSIDAD DE BUENOS AIRES**

FACULTAD DE CIENCIAS EXACTAS Y NATURALES

# **Estimación de sesgos en textos con Pointwise Mutual Information (PMI)**

Tesis Presentada para Optar al Título de Magíster en Explotación de  
Datos y Descubrimiento del Conocimiento

**Tesista:** Lic. Francisco Tomás Valentini

**Director:** Dr. Edgar Altszyler

**Co-Director:** Dr. Germán Rosati

**Lugar de trabajo:** Laboratorio de Inteligencia Artificial Aplicada (LIAA) en el Instituto de Ciencias de la Computación (ICC)

Buenos Aires, 16 de enero de 2024

# Prefacio

Esta tesis de maestría está basada en los hallazgos presentados en la siguiente publicación, la cual es el resultado de las investigaciones realizadas durante el desarrollo de la tesis:

Valentini, F., Rosati, G., Blasi, D., Fernandez Slezak, D., y Altszyler, E. (2023). On the Interpretability and Significance of Bias Metrics in Texts: a PMI-based Approach. En *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, Canada. Association for Computational Linguistics

# Resumen

En los últimos años se ha extendido el uso de los *word embeddings* para medir sesgos y estereotipos sociales en textos. Las métricas basadas en *word embeddings* han demostrado su eficacia en la detección de una amplia variedad de sesgos pero carecen de transparencia e interpretabilidad. En esta tesis introducimos y analizamos una métrica alternativa basada en *Pointwise Mutual Information* (PMI) para medir sesgos en textos. Mostramos que esta métrica, a diferencia de las métricas basadas en *word embeddings*: (1) puede expresarse como una función de probabilidades condicionales, lo que proporciona una interpretación sencilla en términos de coocurrencias de palabras, y (2) permite estimar intervalos de confianza y la significación estadística de los resultados paramétricamente. Realizamos un conjunto de experimentos para comparar la métrica basada en PMI con las métricas basadas en *word embeddings* en tres dimensiones: estimación de la variabilidad, correlación con el juicio humano e interpretabilidad. Los resultados sirven para ilustrar las ventajas del método basado en PMI, así como también la diferencia fundamental en el tipo de asociaciones semánticas que capturan. El código usado para realizar esta tesis está disponible en <https://github.com/ftvalentini/tesis-SesgoPMI>.

# Abstract

In recent years, the use of *word embeddings* to measure social biases and stereotypes in texts has become widespread. Metrics based on *word embeddings* have been shown to be effective in detecting a wide variety of biases but lack transparency and interpretability. In this thesis we introduce and analyse an alternative metric based on *Pointwise Mutual Information* (PMI) to measure bias in texts. We show that this metric, unlike metrics based on *word embeddings*: (1) can be expressed as a conditional probability function, which provides a simple interpretation in terms of word co-occurrences, and (2) allows estimating confidence intervals and the statistical significance of the results parametrically. We conducted a set of experiments to compare PMI-based metrics with metrics based on *word embeddings* along three dimensions: variability

estimation, correlation with human judgement, and interpretability. The results serve to illustrate the advantages of the PMI-based method, as well as the fundamental difference in the type of semantic associations they capture. The code used for this thesis is available at <https://github.com/ftvalentini/tesis-SesgoPMI>.

# Índice general

|  |           |
|--|-----------|
| Índice de figuras  | v         |
| Índice de tablas   | vi        |
| <b>1. Introducción</b>   | <b>1</b>  |
| <b>2. Medición de sesgos textuales</b>   | <b>4</b>  |
| 2.1. <i>Word embeddings</i> estáticos . . . . .  | 4         |
| 2.1.1. <i>Skip-gram with negative sampling</i> (SGNS) . . . . .                              | 5         |
| 2.1.2. FastText . . . . .  | 6         |
| 2.1.3. GloVe . . . . .   | 7         |
| 2.2. Medición de sesgos con <i>word embeddings</i> . . . . .                                 | 8         |
| 2.2.1. Estimación de la variabilidad de métricas basadas en <i>word embeddings</i> . . . . . | 11        |
| <b>3. Medición de sesgos con PMI</b>   | <b>13</b> |
| 3.1. Antecedentes . . . . .  | 13        |
| 3.2. Métrica de sesgo basada en PMI . . . . .  | 15        |
| 3.3. Estimación de la variabilidad del sesgo basado en PMI . . . . .                         | 17        |
| <b>4. Experimentos</b>   | <b>20</b> |
| 4.1. Aspectos metodológicos . . . . .  | 20        |
| 4.1.1. Corpus . . . . .  | 20        |
| 4.1.2. Medición de sesgos . . . . .  | 21        |
| 4.1.3. Detalles de implementación . . . . .  | 22        |
| 4.2. Estimación de la variabilidad . . . . .   | 23        |
| 4.3. Correlación con el juicio humano . . . . .  | 26        |
| 4.4. Interpretación de las estimaciones . . . . .  | 32        |
| <b>5. Conclusiones</b>   | <b>37</b> |
| <b>Bibliografía</b>  | <b>39</b> |

# Índice de figuras

|  |    |
|--|----|
| 4.1. p-valores en función del valor del sesgo para cada tipo de sesgo y cada método de medición . . . . .  | 24 |
| 4.2. Amplitud de los intervalos de confianza al 95 % de BiasWE en función del valor del sesgo, para cada tipo de sesgo y cada método de entrenamiento de <i>embeddings</i> . . . . . | 25 |
| 4.3. Relación entre el sesgo de género textual y de acuerdo al juicio humano en las palabras de Lewis y Lupyan [2020] .  | 28 |
| 4.4. Relación entre el sesgo de sentimiento textual y de acuerdo al juicio humano en las palabras de Toney y Caliskan [2021] . . . . .   | 29 |
| 4.5. Relación entre el sesgo étnico textual y de acuerdo al juicio humano en las palabras de Kozlowski <i>et al.</i> [2019] . .  | 30 |

# Índice de tablas

|      |  |    |
|------|--|----|
| 3.1. | Conteos de co-ocurrencias de los contextos $A$ y $B$ con la palabra $x$ y con el resto del vocabulario $\bar{x}$ . . . . .             | 16 |
| 4.1. | Porcentaje de p-valores menores a 0,10 para cada métrica de sesgo en cada experimento . . . . .  | 23 |
| 4.2. | Coeficientes de correlación de Pearson entre el sesgo de acuerdo al juicio humano y el sesgo textual medido con cada métrica . . . . . | 27 |
| 4.3. | Sesgo de género de las 20 <i>stopwords</i> más frecuentes de las palabras del experimento de Glasgow [Scott <i>et al.</i> , 2019]      | 36 |

# Capítulo 1

## Introducción

El campo de investigación de la equidad en la Inteligencia Artificial (*fairness in AI*) ha recibido una gran atención en los últimos años, impulsado por la necesidad de que los algoritmos basados en aprendizaje automático tomen decisiones imparciales y no discriminatorias. Los sesgos en los sistemas de Inteligencia Artificial emergen por la **replicación y amplificación de los sesgos presentes en los datos de entrenamiento**. Por consiguiente, es fundamental estudiar exhaustivamente estos sesgos y desarrollar métodos computacionales dentro del dominio del Procesamiento del Lenguaje Natural (NLP) que puedan medir eficaz y fiablemente los sesgos de los textos usados para entrenar estos algoritmos.

Además, la medición de sesgos mediante técnicas computacionales es de gran importancia en las **ciencias sociales computacionales** porque permite analizar cómo se representan los distintos grupos sociales en productos culturales como libros, películas, revistas, diarios y redes sociales. Un estudio riguroso de los métodos de cuantificación de sesgos en textos es fundamental para entender la reproducción de estereotipos relacionados con el género, la nacionalidad y otras características.

La medición de los sesgos en los textos, entendida como **una tarea diferente a la medición de los sesgos de los modelos de aprendizaje automático**, tiene su propia importancia. Mientras que una parte importante de las investigaciones anteriores se han centrado predominantemente en medir los sesgos de los modelos (Blodgett *et al.*, 2020, Bolukbasi *et al.*, 2016, Bordia y Bowman, 2019, Gonen y Goldberg, 2019, Kiritchenko y Mohammad, 2018, Lu *et al.*, 2020, Zhao *et al.*, 2018, por mencionar algunos ejemplos ampliamente citados), es esencial comprender y cuantificar los sesgos inherentes a los propios textos. Esta tesis pretende proporcionar una herramienta analítica fiable para abordar esta tarea, la cual es especialmente valiosa para los estudios de ciencias sociales computacionales.

En este campo, se ha vuelto extendido el uso de *word embeddings* para medir sesgos y estereotipos sociales en *corpora*. Los

*embeddings*, representaciones de palabras como vectores densos, capturan asociaciones semánticas entre palabras basadas en sus patrones de coocurrencia en el *corpus*. Aunque han demostrado ser útiles para detectar una amplia variedad de sesgos, las métricas basadas en *embeddings* carecen de transparencia e interpretabilidad. Cuando se usan *embeddings* para medir sesgos, es difícil determinar si los resultados se deben a asociaciones de primer orden generalizadas o si se derivan de asociaciones de orden superior poco claras. Esta falta de interpretabilidad dificulta la comprensión de los aspectos específicos del *corpus* que contribuyen a las mediciones de sesgo. Además, las métricas existentes basadas en *embeddings* no proporcionan un medio eficiente y confiable de estimar los intervalos de confianza o la significación estadística.

Para abordar estas limitaciones, **presentamos una métrica alternativa basada en *Pointwise Mutual Information* (PMI) para medir sesgos en textos**. PMI es una medida de asociación de primer orden entre dos palabras que compara sus probabilidades de ocurrencia individuales con su probabilidad de coocurrencia. La métrica basada en PMI tiene una interpretación transparente y ofrece la posibilidad de estimar intervalos de confianza y evaluar la significación estadística de manera confiable. Esto permite a los investigadores comprender mejor las relaciones semánticas específicas que dan lugar a las estimaciones de sesgo, así como también obtener conclusiones más sólidas acerca de la presencia y la magnitud de los sesgos en los textos.

Los **objetivos de esta tesis** son dos. En primer lugar, buscamos analizar la métrica basada en PMI para medir sesgos en textos, estudiando sus propiedades estadísticas y sus ventajas de interpretabilidad, que hasta ahora se habían pasado por alto. En segundo lugar, nos proponemos evaluar las diferencias, ventajas y desventajas del método basado en PMI frente a las técnicas existentes basadas en *embeddings*. Para ello, realizamos un conjunto de experimentos en la Wikipedia en inglés que apuntan a comparar ambas métricas en tres dimensiones: estimación de la variabilidad, correlación con el juicio humano e interpretabilidad.

La organización de esta tesis es la siguiente. El **capítulo 2** hace una revisión de la literatura sobre la medición de sesgos en textos con herramientas de NLP, enfocándose predominantemente en las métricas basadas en *word embeddings*. Describimos los métodos más difundidos para entrenar *embeddings* y las métricas típicamente usadas para medir sesgos con éstos. El **capítulo 3** presenta los fundamentos teóricos de la métrica de medición de sesgos basada en PMI, explicando su interpretación en términos de coocurrencias. También presenta la estimación de

los intervalos de confianza y la significación estadística dentro de este marco. En el **capítulo 4** realizamos experimentos orientados a ilustrar las propiedades de la métrica basada en PMI en términos de estimación de la variabilidad, correlación con el juicio humano e interpretabilidad. Describimos la configuración experimental, los conjuntos de datos usados, y las metodologías de evaluación empleadas, y luego presentamos los resultados. Finalmente, el **capítulo 5** concluye la tesis, resumiendo las contribuciones y discutiendo las implicaciones de nuestros hallazgos.

El código utilizado en esta tesis está disponible en <https://github.com/ftvalentini/tesis-SesgoPMI> para su uso y exploración.

# Capítulo 2

## Medición de sesgos textuales

A los fines de este trabajo, definimos el **sesgo textual** como el grado en que el lenguaje usado para describir grupos o cosas es diferente [Hoyle *et al.*, 2019]. Algunos de estos sesgos pueden ser moralmente neutrales, como por ejemplo, el sesgo según el cual los insectos son relativamente desagradables, mientras que las flores son agradables. En cambio, los sesgos derivados de aspectos de la cultura humana que pueden conducir a comportamientos dañinos, como los sesgos de género o de nacionalidad, pueden ser problemáticos y llevan el nombre de sesgos estereotipados o directamente **estereotipos** [Caliskan *et al.*, 2017].

Típicamente, la literatura sobre sesgos en el NLP se centra en el estudio de estereotipos porque su reproducción es potencialmente perjudicial para la sociedad. La metodología más usada para medir sesgos textuales son las métricas basadas en *word embeddings* estáticos, los cuales describimos a continuación.

### 2.1 *Word embeddings* estáticos

Los *embeddings* son **representaciones vectoriales densas** de las palabras de un corpus que tienen dimensión relativamente baja (usualmente entre 50 y 1000 dimensiones). Los *embeddings* que usamos en este trabajo y que se usan típicamente para medir sesgos textuales son estáticos. Esto significa que cada palabra del vocabulario se representa con un **único vector fijo**, diferenciándose de las representaciones contextualizadas que se desarrollaron posteriormente, donde las palabras tienen *embeddings* distintos según el contexto en el que se encuentran (por ejemplo, los *embeddings* BERT de Devlin *et al.*, 2019).

Los espacios vectoriales de los *embeddings* estáticos se generan a partir de la distribución de las palabras en el *corpus*, bajo la hipótesis de que las palabras que aparecen en contextos similares suelen tener un contenido semántico similar. Con esta metodología, el significado de las palabras intenta aprenderse a partir de sus **coocurrencias**, es decir, de

la frecuencia de palabras que aparecen en su contexto cercano. Durante el proceso de aprendizaje de los vectores, se busca que los vectores de palabras semánticamente asociadas estén relativamente cerca en el espacio vectorial, y los de palabras no asociadas, relativamente lejos.

A continuación describimos los tres métodos de generación de *word embeddings* que usamos en este trabajo.

### 2.1.1. *Skip-gram with negative sampling* (SGNS)

La metodología de *Skip-gram with negative sampling* (SGNS) [Mikolov *et al.*, 2013] representa a cada palabra del vocabulario con dos vectores de igual dimensión: un **vector objetivo**  $v_w$  y un **vector de contexto**  $v_c$ . Los parámetros que se buscan aprender son entonces dos matrices,  $W$  y  $C$ , cada una de las cuales contiene en cada fila el *embedding* o vector de cada una de las palabras del vocabulario.

Estos parámetros se ajustan para optimizar una función de pérdida donde los pares de palabras que efectivamente coocurren en el corpus se toman como ejemplos positivos, mientras que como ejemplos negativos se toman muestras aleatorias del vocabulario, llamadas **muestras negativas**.

Más específicamente, los ejemplos positivos son los pares de palabras  $(w, c)$  que surgen de considerar una ventana de tamaño  $2T$  alrededor de cada palabra  $w$ , donde  $w$  es la palabra central,  $T$  es el tamaño de la ventana, y las palabras en la ventana son las palabras de contexto  $c$ . Por cada ejemplo positivo  $(w, c)$  se muestrean  $k$  palabras usando la distribución de probabilidad de ocurrencia de cada palabra estimada con

$$P_\alpha(i) = \frac{f(i)^\alpha}{\sum_{j \in V} f(j)^\alpha} \quad (2.1)$$

donde  $f(i)$  es la cantidad de veces que una palabra  $i$  coocurre con cualquier otra palabra en el corpus,  $V$  es el vocabulario y  $\alpha$  es un parámetro de suavizado. Estos  $k$  ejemplos son los ejemplos negativos o *palabras de ruido*. El parámetro  $0 < \alpha < 1$  suaviza la distribución, incrementando la probabilidad de muestrear palabras con baja frecuencia porque  $P_\alpha(i) > P(i)$  para palabras relativamente poco frecuentes.

Para entrenar los parámetros  $W$  y  $C$ , para cada ejemplo positivo  $(w, c)$  junto con sus  $k$  ejemplos negativos se minimiza una **función de pérdida** definida como:

$$L(w, c) = -\log \sigma(v_c^T v_w) - \sum_{i=1}^k \log \sigma(-v_{n_i}^T v_w) \quad (2.2)$$

$\sigma$  es la función sigmoidea, de modo que  $\sigma(v_c^T v_w)$  modela la probabilidad de que la palabra  $c$  sea una palabra de contexto real para la palabra objetivo  $w$ , mientras que la probabilidad de que una palabra  $c$  no sea una palabra de contexto real para  $w$  se modela como  $\sigma(-v_c^T v_w)$ . Entonces, la función de pérdida  $L$  arroja valores bajos cuando la probabilidad asignada a que  $c$  sea una palabra de contexto real para  $w$  es alta, y cuando las probabilidades asignadas a que los ejemplos negativos  $n_i$  sean palabras de contexto reales para  $w$  son bajas.

Esta función de pérdida se minimiza con **descenso por el gradiente** tomando como parámetros ajustables los vectores  $v_w$  y  $v_c$ , almacenados en las matrices  $W$  y  $C$ . Más específicamente, se usan los ejemplos positivos y negativos como datos de entrenamiento, se inicializan aleatoriamente  $W$  y  $C$ , y se recorren los datos de entrenamiento aplicando descenso por el gradiente, ajustando los pesos de  $W$  y  $C$  de manera tal que la similitud  $\sigma(v_c^T v_w)$  de los pares  $(w, c)$  que efectivamente ocurren en el corpus tienda a maximizarse, a la vez que se minimice la similitud de los pares que no ocurren en el corpus (ejemplos negativos).

Como resultado, una vez alcanzado un número máximo de pasadas por el *corpus*, las palabras que tienden a coocurrir en las ventanas de coocurrencia tienden a tener representaciones cercanas en el espacio vectorial, mientras que las palabras que aparecen en contextos diferentes tienden a representarse con vectores que están relativamente lejos.

En las librerías más populares que implementan SGNS (como `gensim` de Řehůřek y Sojka, 2010, que usamos en este trabajo), se tiende a usar únicamente el vector  $v_w$  como representación vectorial final de cada palabra, mientras que el vector  $v_c$  se descarta.

### 2.1.2. FastText

Los *embeddings* generados con SGNS sólo están definidos para palabras que son parte de  $V$ , el vocabulario de entrenamiento. Es decir, no es posible obtener *embeddings* para palabras que no estén en el *corpus* de entrenamiento (**palabras OOV**, *out-of-vocabulary*) pero que puedan ser de interés, como las distintas formas o inflexiones asociadas a verbos y sustantivos.

Para resolver esto, la metodología FastText [Bojanowski *et al.*, 2017] representa a cada palabra como sí misma más el conjunto de n-gramas o **subpalabras** que la constituyen. Para ello, se añaden símbolos especiales < y > al comienzo y final de la palabra antes de hacer la partición en subpalabras. Luego se aprenden vectores para cada subpalabra con la misma metodología de entrenamiento que SGNS. En general se consideran como subpalabras todos los n-gramas de entre 3 y 6 caracteres que componen a la palabra.

Finalmente, para obtener el vector final de palabras OOV se calcula el promedio de los vectores de las subpalabras que la constituyen que existen en el vocabulario, mientras que para palabras que sí están en  $V$  se incluye también en el promedio el vector de la palabra completa. Por ejemplo, si la palabra *gato* está en el vocabulario, su vector final será el promedio de los vectores de los 3-gramas <*ga*, *gat*, *ato*, *to*>, los 4-gramas <*gat*, *ato*, *to*>, y los 5-gramas <*gato*, *ato*>, y el vector de la palabra completa <*gato*>. En el presente trabajo consideramos únicamente palabras que están en  $V$ .

Al igual que en SGNS, se suelen considerar únicamente los vectores  $v_w$ , mientras que los vectores  $v_c$  se descartan.

### 2.1.3. GloVe

En lugar de recorrer todos los pares de coocurrencias a la hora de aprender los *embeddings*, la metodología GloVe (*Global Vectors for Word Representation*, Pennington *et al.*, 2014) busca aprovechar los conteos globales almacenados en la **matriz de coocurrencias**  $M$ , la cual se precalcula antes de iniciar el entrenamiento. La misma almacena en cada celda  $M_{ij}$  la cantidad de veces que la palabra  $i$  aparece en el contexto de la palabra  $j$  en el *corpus* de entrenamiento al considerar una ventana de tamaño  $2T$  alrededor de cada palabra.

La función de pérdida GloVe para el *corpus* en su conjunto viene dada por

$$L = \sum_{i,j=1}^V f(M_{ij})(v_{w_i}^T v_{c_j} + v_{b_w} + v_{b_c} - \log M_{ij})^2 \quad (2.3)$$

donde  $M_{ij}$  indica el número de veces que la palabra  $j$  aparece en el contexto de la palabra  $i$ ,  $v_{w_i}$  es el vector objetivo de la palabra  $i$ ,  $v_{c_j}$  es el vector de contexto de la palabra  $j$ , y  $v_{b_w}$  y  $v_{b_c}$  son escalares específicos para cada palabra que funcionan como interceptos.  $f(M_{ij})$  es

una función de ponderación de las coocurrencias que se define como

$$f(x) = \begin{cases} \left(\frac{x}{x_{\max}}\right)^{\alpha} & \text{si } x < x_{\max} \\ 1 & \text{si } x \geq x_{\max} \end{cases} \quad (2.4)$$

Con esta función se reduce el peso de las coocurrencias poco frecuentes (porque  $f(x) < 1$  para  $x < x_{\max}$ ). Pennington *et al.* [2014] usan  $\alpha = 0.75$ , lo cual incrementa ligeramente el peso de las coocurrencias más pequeñas (porque  $f(x)_{\alpha=0.75} > f(x)_{\alpha=1}$  para  $x < x_{\max}$ ), de manera similar a como lo hace el parámetro de *smoothing*  $\alpha$  de SGNS. En este trabajo mantenemos este valor.

La pérdida de la ecuación 2.3 se minimiza cuando  $v_{w_i}^T v_{c_j} + v_{b_w} + v_{b_c} = \log M_{ij}$ . Para optimizar la función se muestran aleatoria e iterativamente tandas de elementos no nulos de la matriz  $M$  para calcular los gradientes de los parámetros  $v_{w_i}$ ,  $v_{c_j}$ ,  $v_{b_w}$  y  $v_{b_c}$ , y luego actualizarlos con el algoritmo AdaGrad [Duchi *et al.*, 2011]. El entrenamiento finaliza cuando se alcanza un número máximo de pasadas completas por la matriz de coocurrencias.

Al igual que SGNS y FastText, el modelo genera dos conjuntos de vectores de palabras,  $W$  y  $C$ . Dado que  $M$  es simétrica,  $W$  y  $C$  son conceptualmente equivalentes y sólo difieren por tener inicializaciones aleatorias distintas. Por lo tanto, por defecto, los *embeddings* GloVe de una palabra  $i$  se obtienen sumando los vectores objetivo y de contexto de la misma ( $v_{w_i} + v_{c_i}$ ).

## 2.2 Medición de sesgos con *word embeddings*

Dado que miden la similitud semántica entre las palabras de un corpus, los *word embeddings* son ampliamente utilizados para **cuantificar sesgos textuales de corpora específicos**. La metodología consiste en entrenar *embeddings* sobre el *corpus* que se desea estudiar, y luego computar una **métrica de medición de sesgo textual**.

En la versión más general, se conforman dos conjuntos de **palabras de contexto  $A$  y  $B$** , y un conjunto de **palabras objetivo  $X$** . El sesgo textual de las palabras  $X$  en relación a los atributos  $A$  y  $B$  en un *corpus* dado se mide calculando la diferencia de similitudes entre  $X$  con respecto a  $A$  y  $B$ :

$$\text{Bias}(X, A, B) = \text{sim}(X, A) - \text{sim}(X, B) \quad (2.5)$$

Las medidas de sesgo textual cuantifican, entonces, cuánto más se asocian las palabras de  $X$  con las de  $A$  que con las de  $B$ . La similitud semántica se puede cuantificar con la **similitud coseno** entre *word embeddings* (WE), de manera que el sesgo para una palabra objetivo individual  $x$  queda definida como:

$$\text{BiasWE}(x, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(v_x, v_a) - \frac{1}{|B|} \sum_{b \in B} \cos(v_x, v_b) \quad (2.6)$$

donde  $v_i$  es el vector de la palabra  $i$  y  $\cos(v_i, v_j)$  es la similitud coseno entre vectores [Lewis y Lopyan, 2020].

Por ejemplo, siguiendo a Lewis y Lopyan [2020], para medir el BiasWE de género binario (femenino vs masculino) en un *corpus* en inglés, los conjuntos  $A$  y  $B$  se pueden conformar con palabras que representan a los géneros femenino y masculino, respectivamente:  $A = \{\text{female}, \text{woman}, \text{she}, \dots\}$  y  $B = \{\text{male}, \text{man}, \text{he}, \dots\}$ . Las palabras  $x$ , por su parte, son aquellas donde es de interés medir estereotipos, por ejemplo, ocupaciones como *nurse*, *doctor*, *engineer*, etc. En el caso en que se desea medir el sesgo conjuntamente para un conjunto de palabras  $X$ , se toma el promedio de BiasWE a lo largo de las  $x$  que conforman el conjunto.

En la literatura que estudia sesgos lingüísticos, múltiples estudios han usado variantes de la ecuación 2.6 para **analizar corpus específicos**. Por ejemplo, Garg *et al.* [2018] entrenaron *embeddings* GloVe en el *New York Times Annotated Corpus* para cuantificar los cambios en los estereotipos hacia las mujeres y las minorías étnicas en los Estados Unidos a lo largo del siglo XX. Para este análisis, usaron métricas basadas en la distancia euclíadiana y la similitud coseno muy similares a la ecuación 2.6, y explicaron que ambas arrojaban resultados similares.

Kozlowski *et al.* [2019], por otro lado, entrenaron *embeddings* SGNS en libros digitalizados disponibles en Google Ngrams para examinar la evolución de los sesgos de etnia, género y clase a lo largo del tiempo. La métrica que utilizaron primero calcula la diferencia promedio entre los vectores de  $N$  palabras de contexto apareadas (es decir,  $d = \frac{1}{N} \sum_{i=1}^N v_{a_i} - v_{b_i}$ ), y luego toma la similitud coseno de esta dirección con el vector de la palabra objetivo,  $v_x$  i.e.  $\cos(d, v_x)$ .

Lewis y Lopyan [2020] midieron sesgos de género con la ecuación 2.6 usando *embeddings* FastText entrenados en Wikipedias y subtítulos de 25 idiomas. Descubrieron que los sesgos de género medidos en pruebas de asociaciones psicológicas implícitas están estrechamente relacionados con los sesgos de género textuales de la lengua que hablan los participantes de las pruebas.

Otro estudio, el de Charlesworth *et al.* [2021], midió los estereotipos de género relacionados con las ocupaciones y los rasgos de personalidad con *embeddings* FastText en *corpora* de diversos dominios (por ejemplo, conversaciones de niños y adultos, libros, películas, televisión). Encontraron que los sesgos eran estables a pesar de las diferencias en los *corpora*.

Otro tipo de estudios han usado ***embeddings* pre-entrenados sobre grandes volúmenes de texto, en lugar de entrenar desde cero *embeddings* sobre *corpora* de interés**. Esto les ha permitido estudiar los sesgos que podrían existir potencialmente en el *corpus* de entrenamiento.

Uno de los trabajos más destacados de este tipo es el de Caliskan *et al.* [2017], quienes midieron los sesgos de género en vectores GloVe pre-entrenados en el *corpus* Common Crawl, obtenido de un barrido de la web a gran escala [Pennington *et al.*, 2014]. Encontraron una correlación entre los sesgos de género de los *embeddings* y la distribución del género en los nombres de personas y en las ocupaciones en Estados Unidos. Para realizar este análisis, utilizaron la métrica SC-WEAT (*Single-Category Word Embedding Association Test*, Toney y Caliskan, 2021), que agrega el desvío estándar de las similitudes de  $v_x$  con respecto al conjunto de vectores de  $A \cup B$  como denominador de la ecuación 2.6. Esta misma medida también se usó en el estudio de Charlesworth *et al.* [2021].

En otro estudio relevante, Garg *et al.* [2018] analizaron las tendencias de los estereotipos a lo largo de la historia utilizando los *embeddings* HistWords pre-entrenados con *Google Books* y el *Corpus of Historical American English* [Hamilton *et al.*, 2016]. De manera similar, Jones *et al.* [2020] utilizaron los HistWords para analizar la trayectoria en el tiempo de las asociaciones estereotipadas de género en la lengua inglesa escrita desde el 1800 hasta el 2000. Para medir los sesgos, usaron una métrica parecida a la ecuación 2.6, pero adaptada a múltiples palabras objetivo: la similitud entre  $X$  y las palabras de contexto, por ejemplo  $\text{sim}(X, A)$ , se computa tomando el promedio de las similitudes de todas las combinaciones posibles de pares  $(x_i, a_i)$ , donde  $x_i \in X$  y  $a_i \in A$ .

Por otro lado, DeFranza *et al.* [2020] usaron *embeddings* FastText pre-entrenados en las Wikipedias y Common Crawls de 45 idiomas distintos para analizar los pensamientos voluntarios de las personas expresados en los textos de cada idioma. La métrica que usaron para medir sesgos textuales es similar a la de la ecuación 2.6, pero adaptada al caso en que se cuenta con dos grupos de palabras objetivo. Esto es útil en los idiomas en los que las palabras tienen género gramatical. Sus resultados mostraron que los sesgos textuales de género son más frecuentes en las

lenguas con género grammatical que en las que no lo tienen.

En el presente trabajo usamos la especificación de la ecuación 2.6 por su flexibilidad: no requiere palabras de contexto apareadas, admite grupos de contexto de distinta longitud y puede calcularse si los grupos de contexto están conformados por una sola palabra cada uno (esto último no es posible en una métrica como el SC-WEAT). Además, no precisamos extenderla para el caso de palabras con género grammatical porque trabajamos con el idioma inglés, en el que las palabras no tienen género grammatical.

### 2.2.1. Estimación de la variabilidad de métricas basadas en *word embeddings*

Las métricas basadas en *embeddings* admiten el uso de técnicas de remuestreo para estimar la variabilidad del estadístico específico que se esté calculando. Esto permite evaluar la robustez de los resultados obtenidos.

Por una parte, la literatura ha usado **tests de permutaciones para calcular la significancia estadística** de BiasWE o métricas similares [Caliskan *et al.*, 2017, Charlesworth *et al.*, 2021]. Los tests de permutaciones consisten en asignar aleatoriamente las palabras de contexto entre entre los grupos *A* y *B* en múltiples iteraciones y calcular la métrica de sesgo en cada iteración. Con los valores del sesgo de cada iteración se construye la distribución nula del sesgo, y se calcula el p-valor a dos colas como la fracción de veces que el valor absoluto del sesgo de la distribución nula es igual o mayor que el observado [North *et al.*, 2002].

Por otro lado, se ha usado **bootstrap para obtener intervalos de confianza** con un enfoque de remuestreo parecido a los tests de permutaciones [Garg *et al.*, 2018]. En este caso, en cada iteración de bootstrap se muestran las palabras de contexto *A* y *B* por separado con reemplazo y se calcula la métrica de sesgo. Con los valores del sesgo de cada iteración se construye la distribución de bootstrap del sesgo. El error estándar del sesgo se estima luego como la desviación estándar de la distribución de bootstrap, y los cuantiles de la distribución se utilizan para obtener intervalos de confianza [Davison y Hinkley, 1997].

Destacamos que, en rigor, los intervalos de confianza computados por Garg *et al.* [2018] hacen bootstrap sobre las palabras objetivo y estiman entonces la variabilidad del sesgo cuando se computa para muchas

palabras objetivo en simultáneo. Sin embargo, este enfoque se puede extender a las palabras de contexto cuando se computa el sesgo de una sola palabra objetivo, como en nuestro caso.

Otro enfoque, propuesto por Kozlowski *et al.* [2019], consiste en estimar la variabilidad del sesgo que se origina en la variabilidad propia de los *embeddings*. El método consiste en entrenar múltiples conjuntos de *embeddings* sobre subconjuntos del *corpus* (e.g. 20) y estimar el desvío estándar e intervalo de confianza en base a estas realizaciones. Si bien consideramos que este enfoque es válido, no lo implementamos por ser computacionalmente costoso para *corpora* relativamente grandes, como el que usamos en esta tesis.

En resumen, la estimación de la variabilidad de las métricas sesgo textual es fundamental para evaluar la robustez de los resultados obtenidos. Los métodos de remuestreo, como los tests de permutaciones y el bootstrap, son las herramientas típicamente usadas para lograr este objetivo en el caso de métricas basadas en *word embeddings*.

# Capítulo 3

## Medición de sesgos con PMI

Hemos mostrado que los enfoques basados en *embeddings* estáticos se han usado ampliamente para detectar y cuantificar sesgos en *corpora*. Estos vectores pueden usarse para medir la similitud semántica entre palabras y, por lo tanto, detectar patrones de sesgo en el uso de ciertas palabras. Si bien los *embeddings* son una herramienta poderosa para medir similitud, y por lo tanto, sesgos, no son la única. El **Pointwise Mutual Information (PMI)** es una medida que también puede usarse para medir la similitud semántica entre palabras.

A continuación, describiremos la medida de PMI y presentaremos una métrica para cuantificar sesgos textuales basado en esta medida. Como veremos en los capítulos subsiguientes, la introducción de esta métrica se justifica por su mayor transparencia e interpretabilidad.

### 3.1 Antecedentes

El PMI entre dos palabras  $x$  y  $y$  se define como

$$\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \quad (3.1)$$

donde  $p(x, y)$  es la probabilidad de coocurrencia de las palabras  $x$  y  $y$  en una ventana de  $2T$  palabras en un *corpus*, y  $p(x)$  y  $p(y)$  representan las probabilidades de ocurrencia individuales de las palabras  $x$  e  $y$  en cualquier contexto en el mismo *corpus* [Church y Hanks, 1990].

El PMI compara la probabilidad de ocurrencia conjunta de dos palabras con su probabilidad de ocurrencia independiente. Específicamente, el PMI nos indica, en una escala logarítmica, cuántas veces más probable es que dos palabras aparezcan juntas en un *corpus* (numerador de la ecuación 3.1) en comparación con lo que se esperaría por azar (denominador).

Por este motivo, el PMI es una medida de **asociación de primer orden** entre dos palabras: dos palabras tienen una asociación de primer orden

(también llamada asociación sintagmática) si están típicamente cerca una de la otra [Jurafsky y Martin, 2009]. Cuanto más alto es el PMI, más probable es que dos palabras coocurran en un *corpus* en relación a lo que se esperaría si fueran independientes.

Las probabilidades de la ecuación 3.1 pueden estimarse por máxima verosimilitud usando los conteos de una **matriz de coocurrencias** simétrica  $M$  donde cada entrada indica el número de veces que una palabra aparece en el contexto de otra (en ventanas de  $2T$  palabras). En particular:

$$\hat{p}(x, y) = \frac{M_{xy}}{N} \quad \hat{p}(x) = \frac{M_{x\cdot}}{N} \quad \hat{p}(y) = \frac{M_{\cdot y}}{N} \quad (3.2)$$

donde  $N = \sum_x \sum_y M_{xy}$  es el número total de coocurrencias en el *corpus*,  $M_{xy}$  es el número de veces que la palabra  $x$  aparece en el contexto de la palabra  $y$ ,  $M_{x\cdot} = \sum_y M_{xy}$  es el número de veces que la palabra  $x$  aparece en cualquier contexto, y  $M_{\cdot y} = \sum_x M_{xy}$  es el número de veces que la palabra  $y$  aparece en cualquier contexto.

Estimamos el PMI entonces con

$$\text{PMI}(x, y) = \log \frac{\hat{p}(x, y)}{\hat{p}(x)\hat{p}(y)} = \log \frac{M_{xy} \cdot N}{M_{x\cdot} \cdot M_{\cdot y}} \quad (3.3)$$

El PMI también puede usarse para calcular asociaciones entre listas de palabras  $X$  y  $Y$ . En este caso,  $p(X, Y)$  es la probabilidad de coocurrencia entre cualquier palabra de  $X$  con cualquier otra de  $Y$ . Del mismo modo,  $p(X)$  y  $p(Y)$  son la probabilidad de aparición de cualquier palabra de  $X$  y cualquier palabra de  $Y$ , respectivamente. Para estimar las probabilidades, debemos sumar las coocurrencias de las palabras individuales i.e.  $M_{XY} = \sum_{x \in X} \sum_{y \in Y} M_{xy}$ ,  $M_{X\cdot} = \sum_{x \in X} M_{x\cdot}$  y  $M_{\cdot Y} = \sum_{y \in Y} M_{\cdot y}$ .

Una manera útil de reexpresar el PMI es:

$$\text{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x|y)p(y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \quad (3.4)$$

siguiendo la definición de probabilidad conjunta  $p(x, y) = p(x|y)p(y) = p(y|x)p(x)$ .

En la ecuación 3.4,  $p(x|y)$  es la probabilidad de que la palabra  $x$  aparezca en el contexto de la palabra  $y$ . Podemos interpretar el PMI, entonces, como la relación entre la probabilidad de que  $x$  aparezca en el contexto de  $y$  y la probabilidad de que  $x$  aparezca en cualquier contexto, o bien, entre

la probabilidad de que  $y$  aparezca en el contexto de  $x$  y la probabilidad de que  $y$  aparezca en cualquier contexto.

Por poner un ejemplo, si definimos  $X = \{\text{hoja}\}$  y  $Y = \{\text{árbol}\}$  y obtenemos un valor de PMI de 0.2, esto significa que la probabilidad de que la palabra *hoja* aparezca en el contexto de la palabra *árbol* es  $\exp(0.2) \approx 1.22$  veces la probabilidad de que aparezca en cualquier contexto; es decir, es un 22 % más probable. En cambio, si el PMI fuera 0, sería igual de probable que *hoja* aparezca en el contexto de *árbol* que en cualquier contexto.

En el ámbito del estudio de sesgos textuales, se ha usado la medida de PMI, aunque con menor popularidad que los *word embeddings*. Un ejemplo de su aplicación es el estudio llevado a cabo por Gálvez *et al.* [2019], quienes analizaron la presencia de estereotipos de género en los subtítulos de películas en inglés. En este trabajo se usó la medida de PMI para medir la asociación la asociación de palabras asociadas a los géneros femenino ( $A$ ) y masculino ( $B$ ) con palabras asociadas a la inteligencia ( $X$ ). En resumidas cuentas, descubrieron que  $\text{PMI}(A, X)$  tenía que ser menor que  $\text{PMI}(B, X)$ , lo cual indicaría que el estereotipo que asocia la inteligencia con la masculinidad está presente en las películas occidentales.

Insipirada en este enfoque, la métrica que presentamos a continuación se construye a partir de dos medidas de PMI, y tiene ventajas que desarrollaremos en las secciones subsiguientes.

## 3.2 Métrica de sesgo basada en PMI

La métrica que proponemos surge de tomar la expresión genérica de sesgo de la ecuación 2.5 y usar el PMI como medida de similitud entre palabras, lo que da lugar a la expresión:

$$\text{Bias}_{\text{PMI}}(X, A, B) = \text{PMI}(X, A) - \text{PMI}(X, B) \quad (3.5)$$

Usando la reexpresión de la ecuación 3.4, y considerando el caso de una sola palabra objetivo  $x$ , podemos reescribir la ecuación 3.5 como:

$$\text{Bias}_{\text{PMI}}(x, A, B) = \log \frac{p(x|A)}{p(x)} - \log \frac{p(x|B)}{p(x)} = \log \frac{p(x|A)}{p(x|B)} \quad (3.6)$$

Es decir, BiasPMI indica, en escala logarítmica, **cuánto más probable es encontrar la palabra  $x$  en el contexto de las palabras  $A$  que en el contexto de las palabras  $B$** . Al igual que el PMI, este cociente de probabilidades condicionales puede estimarse por máxima verosimilitud con los conteos de coocurrencias del *corpus* almacenados en la matriz  $M$ ,

$$\text{BiasPMI}(x, A, B) = \log \frac{\hat{p}(x|A)}{\hat{p}(x|B)} = \log \frac{\frac{M_{x,A}}{M_{\cdot A}}}{\frac{M_{x,B}}{M_{\cdot B}}} = \log \frac{\frac{M_{x,A}}{M_{x,A} + M_{\bar{x},A}}}{\frac{M_{x,B}}{M_{x,B} + M_{\bar{x},B}}} \quad (3.7)$$

recordando que  $M_{\cdot A} = \sum_{a \in A} M_{\cdot a}$  y sabiendo que  $\hat{p}(x|A) = \frac{\hat{p}(x,A)}{\hat{p}(A)} = \frac{M_{x,A}}{M_{\cdot A}}$ , y análogamente para  $\hat{p}(x|B)$ .

En la ecuación 3.7  $M_{x,A}$  y  $M_{x,B}$  representan el número de veces que la palabra  $x$  aparece en el contexto de las palabras en  $A$  y  $B$ , respectivamente, y  $M_{\bar{x},A}$  y  $M_{\bar{x},B}$  representan la cantidad de veces que todas las palabras menos  $x$  aparecen en el contexto de las palabras en  $A$  y  $B$ , respectivamente. La tabla de contingencia 3.1 representa estos conteos.

|     | $x$       | $\bar{x}$       | <b>Total</b>                            |
|-----|-----------|-----------------|---|
| $A$ | $M_{x,A}$ | $M_{\bar{x},A}$ | $M_{\cdot A} = M_{x,A} + M_{\bar{x},A}$ |
| $B$ | $M_{x,B}$ | $M_{\bar{x},B}$ | $M_{\cdot B} = M_{x,B} + M_{\bar{x},B}$ |

Tabla 3.1: Conteo de co-ocurrencias de los contextos  $A$  y  $B$  con la palabra  $x$  y con el resto del vocabulario  $\bar{x}$ .

Cuando no hay coocurrencias entre la palabra objetivo  $x$  y cualquiera de las palabras de contexto ( $M_{x,A} = 0$  o  $M_{x,B} = 0$ ), la métrica BiasPMI no está definida. En estos casos, se puede usar la versión suavizada de la métrica, que consiste en sumar previamente un pequeño valor  $\epsilon$  a todas las coocurrencias [Jurafsky y Martin, 2009].

La expresión de las ecuaciones 3.5 y 3.6 tiene **antecedentes en la literatura**. En primer lugar, Turney [2002] propuso una medida de *Semantic Orientation* (SO) para bigramas, que es equivalente a la métrica BiasPMI presentada de la ecuación 3.5. La SO de un bigrama  $x$  se define como la diferencia entre el PMI de  $x$  con la palabra *excellent* y el PMI de  $x$  con la palabra *poor*. En este caso, los PMI se computan a partir de la cantidad de resultados que devuelve un motor de búsqueda al buscar el bigrama  $x$  y las palabras *excellent* o *poor*. El autor propone usar el SO para clasificar reseñas de productos como positivas o negativas.

Por otro lado, Bordia y Bowman [2019] utilizaron una expresión matemática para cuantificar sesgos de género que es equivalente a BiasPMI. Este puntaje de sesgo se calcula para cada palabra de un *corpus*. En el estudio, los autores toman el promedio de los puntajes a lo largo de las palabras de un *corpus* de entrenamiento de modelos de lenguaje y de *corpora* generados por estos modelos, con el objetivo de evaluar la eficacia de distintas metodologías para reducir el sesgo de género en modelos de lenguaje.

Por último, BiasPMI también es equivalente al PMI<sub>gap</sub> propuesto por Aka *et al.* [2021]. En este estudio se usó esta métrica en un contexto más general que el NLP, específicamente, para medir los sesgos que un modelo de aprendizaje automático puede haber aprendido en relación con diferentes etiquetas en un problema de clasificación supervisada.

Si bien el PMI se ha usado para estudiar sesgos y patrones de orientación semántica de las palabras, acá proponemos usar la diferencia de PMIs como métrica para medir sesgos específicamente en el contexto de las ciencias sociales computacionales. Además, estudiamos por primera vez las propiedades estadísticas de esta métrica, lo cual presentamos en la siguiente sección.

### 3.3 Estimación de la variabilidad del sesgo basado en PMI

En las aplicaciones que son de interés en este trabajo, los grupos de palabras de contexto  $A$  y  $B$  están conformados típicamente por palabras que aluden a grupos sociales, como  $\{he, man, she, woman, \dots\}$  en el caso del género, mientras que  $X$  refiere a palabras específicas donde interesa medir un sesgo (en general trabajaremos con  $|X| = 1$ , y entonces  $X = x$ ).

Considerando esto,  $M_{\bar{x},C}$  i.e. las coocurrencias entre palabras que no están en un grupo  $C$  (la mayor parte del vocabulario) y una palabra específica  $x$  son considerablemente mayores que  $M_{x,C}$  i.e. las coocurrencias entre  $x$  y las palabras de  $C$ . Más precisamente:

$$M_{\bar{x},A} \gg M_{x,A} \quad \text{y} \quad M_{\bar{x},B} \gg M_{x,B} \tag{3.8}$$

Por ejemplo, si  $A = \{ping\}$  y  $X = \{pong\}$ ,  $\bar{X}$  representa el resto de palabras del vocabulario que no son *pong*. La aproximación 3.8 dice que

*ping* coocurre considerablemente menos con *pong* que con el resto de palabras del vocabulario. Aunque es probable que *ping* y *pong* tengan muchas coocurrencias, *ping* coocurrirá más con el resto de palabras del vocabulario dentro de una ventana móvil de palabras.

Cuando se cumple la condición de la ecuación 3.8, la ecuación 3.7 puede aproximarse mediante

$$\text{BiasPMI} = \log \frac{\frac{M_{x,A}}{M_{x,A} + M_{\bar{x},A}}}{\frac{M_{x,B}}{M_{x,B} + M_{\bar{x},B}}} \approx \log \frac{\frac{M_{x,A}}{M_{\bar{x},A}}}{\frac{M_{x,B}}{M_{\bar{x},B}}} \approx \log \text{OR} \quad (3.9)$$

Es decir, BiasPMI **se puede aproximar como un log odds ratio (OR)**.

La distribución del log odds ratio converge a la normalidad [Agresti, 2003]. Por ende es sencillo evaluar la hipótesis nula de que el log odds ratio es 0 (ausencia de sesgo) mediante una **prueba paramétrica**. En particular, obtenemos el p-valor a dos colas con  $2P(Z < -|\text{BiasPMI}|/SE)$ , donde  $Z$  es una variable aleatoria normal estándar, y el desvío estándar  $SE$  se estima mediante

$$SE = \sqrt{\frac{1}{M_{x,A}} + \frac{1}{M_{x,B}} + \frac{1}{M_{\bar{x},A}} + \frac{1}{M_{\bar{x},B}}} \quad (3.10)$$

A su vez, el intervalo de confianza del 95 % viene dado por

$$CI_{95\%}(\text{BiasPMI}) = \text{BiasPMI} \pm 1.96 SE \quad (3.11)$$

Los p-valores e intervalos de confianza de BiasPMI se basan en estimar una variabilidad que es fundamentalmente distinta a la que se estima con las permutaciones o bootstrap de BiasWE presentados en la sección 2.2.1.

En particular, la incertidumbre asociada a BiasPMI medida por medio del test de log odds ratio captura la **variabilidad del proceso generador de datos subyacente**, es decir, la variabilidad debida a que los conteos de coocurrencias son variables aleatorias. En cambio, los p-valores de permutaciones e intervalos de bootstrap de BiasWE sólo consideran la **variabilidad de los grupos de palabras de contexto**. Esto significa que deben elegirse varias palabras de contexto para poder realizar la inferencia.

En el límite, si  $A$  y  $B$  fueran listas de una sola palabra, no hay forma de estimar la incertidumbre para BiasWE con estos métodos, mientras que

es perfectamente factible para BiasPMI. Si en alguna aplicación queremos medir sesgos con listas de una sola palabra, los procedimientos de remuestreo empleados con BiasWE son inútiles, mientras que la prueba de log odds ratio funciona perfectamente bien para BiasPMI.

Otra ventaja del test parámetrico de log odds ratio para BiasPMI es que es computacionalmente barato en comparación con los procedimientos no parámetricos de bootstrap y permutaciones requeridos para BiasWE. Éstos pueden ser muy lentos si queremos hacer inferencia sobre muchas palabras objetivo.

Para ilustrar la diferencia entre los dos métodos, en la sección 4.2 compararemos la variabilidad estimada para BiasWE y BiasPMI en un experimento con datos reales.

# Capítulo 4

## Experimentos

En este capítulo hacemos una comparación empírica entre Bias<sub>PMI</sub> y Bias<sub>WE</sub> a fin de ilustrar las principales diferencias entre las métricas. En particular, nos proponemos comparar los métodos en las siguientes tres dimensiones:

- **Variabilidad:** ¿Cómo difieren los métodos de medición de la variabilidad de Bias<sub>PMI</sub> y Bias<sub>WE</sub>? (sección 4.2)
- **Correlación con el juicio humano:** ¿En qué medida las estimaciones de Bias<sub>PMI</sub> y Bias<sub>WE</sub> correlacionan con el juicio humano de los sesgos? (sección 4.3)
- **Interpretabilidad:** ¿Qué tipos de asociaciones semánticas capturan Bias<sub>PMI</sub> y Bias<sub>WE</sub>? (sección 4.4)

### 4.1 Aspectos metodológicos

#### 4.1.1. Corpus

Para los experimentos usamos un *corpus* en inglés construido a partir de *English Wikipedia* de agosto de 2014 (<https://archive.org/download/enwiki-20141208>) al que llamamos, de aquí en más, Wikipedia.

En el **preprocesamiento** se eliminan los artículos con menos de 50 tokens, se convierten los tokens a minúsculas, se eliminan los símbolos no alfanuméricos y se aplica una partición en oraciones (*sentence splitting*), de modo que una oración equivalga a un documento. Tras aplicar estos pasos, el corpus de Wikipedia consta de 1.200 millones de tokens y 53,9 millones de documentos.

#### 4.1.2. Medición de sesgos

Para cuantificar los **sesgos textuales**, computamos BiasWE usando SGNS, FastText y GloVe según la ecuación 2.6, mientras que usamos la ecuación 3.7 para BiasPMI.

Nos enfocamos en medir sesgos que ya han sido estudiados por la literatura existente, la cual usamos como referencia para definir las listas de palabras de contexto  $A$  y  $B$ . Asimismo, medimos cada uno de los sesgos en palabras objetivo cuyos **sesgos de acuerdo al juicio humano** han sido medidos por estudios previos, en experimentos independientes de la Wikipedia. De esta manera nos aseguramos de estar midiendo los sesgos en palabras donde es razonable estudiarlos en el contexto de las ciencias sociales computacionales.

En particular, medimos los siguientes sesgos:

**Sesgo de género.** Aquí  $A=\{\text{female}, \text{woman}, \text{girl}, \text{sister}, \text{she}, \text{her}, \text{hers}, \text{daughter}\}$  y  $B=\{\text{male}, \text{man}, \text{boy}, \text{brother}, \text{he}, \text{him}, \text{his}, \text{son}\}$  [Caliskan *et al.*, 2017]. Valores positivos (negativos) indican que la palabra objetivo se asocia relativamente más con los términos femeninos (masculinos).

Medimos el sesgo de género en las palabras de las *Glasgow Norms*, un conjunto de 5.500 palabras en inglés con puntajes de género que resumen las respuestas de los participantes a los que se pidió que valoraran la asociación de género de cada palabra [Scott *et al.*, 2019]. Los participantes midieron el grado en que cada palabra se asocia a un comportamiento masculino o femenino en una escala de 1 (muy femenino) a 7 (muy masculino). Siguiendo a Lewis y Lupyán [2020], promediamos las normas de los homónimos y calculamos  $8 - \text{puntaje}$  para invertir la escala de los puntajes de modo que representen la feminidad según el juicio humano. 4.661 palabras de la lista original coinciden con el vocabulario de Wikipedia.

**Sesgo étnico.** Usamos  $A=\{\text{black}, \text{blacks}, \text{african}, \text{afro}\}$  y  $B=\{\text{white}, \text{whites}, \text{european}, \text{anglo}\}$  [Kozlowski *et al.*, 2019]. Valores positivos (negativos) indican que la palabra objetivo se asocia relativamente más con los términos de etnia negra (blanca).

Kozlowski *et al.* [2019] seleccionaron 60 palabras de siete ámbitos temáticos (ocupaciones, alimentos, ropa, vehículos, géneros musicales, deportes y nombres de pila) y pidieron a 398 encuestados de Amazon Mechanical Turk que valoren cómo calificarían cada palabra en una escala de 0 (*very African American*) a 100 (*very white*). La valoración promedio de estas

respuestas representa la asociación relativa *black-white* de cada palabra según el juicio humano. Medimos el sesgo étnico en las 59 palabras que coinciden con el vocabulario de nuestro *corpus*.

**Sesgo de sentimiento.** Aquí  $A=\{caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation\}$  y  $B=\{abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison\}$  [Caliskan *et al.*, 2017]. Valores positivos (negativos) indican que una palabra objetivo se asocia relativamente más con los términos agradables (desagradables). En la bibliografía en inglés este sesgo se suele denominar *valence bias* [Toney y Caliskan, 2021].

En este caso consideramos las palabras del estudio de Warriner *et al.* [2013], en el que 1.827 trabajadores de Amazon Mechanical Turk calificaron 13.915 palabras de tópicos diversos en una escala de 1 (*unhappy*) a 9 (*happy*). El sentimiento de cada palabra de acuerdo al juicio humano se computa como el promedio de estas calificaciones. En este trabajo medimos el sesgo de sentimiento en las 13.565 palabras que están en el vocabulario de Wikipedia.

#### 4.1.3. Detalles de implementación

En todas las metodologías utilizadas (SGNS, FastText, GloVe y PMI), contamos las coocurrencias a partir de una ventana de tamaño  $\pm 10$  dentro de cada oración ( $T = 10$ ). Consideramos a los tokens con menos de 100 apariciones como OOV (fuera del vocabulario) y los eliminamos antes de computar la matriz de coocurrencias.

Para entrenar SGNS y FastText, usamos la implementación de Gensim [Řehůřek y Sojka, 2010], mientras que para GloVe usamos la implementación de Pennington *et al.* [2014]. En los tres casos los vectores tienen 300 dimensiones y usamos los hiperparámetros por defecto. En el caso de PMI, contamos las coocurrencias con el módulo de GloVe [Pennington *et al.*, 2014], y establecemos el parámetro de suavizado  $\epsilon$  en 0,01.

Todos los experimentos se realizaron en una máquina de escritorio con un procesador Intel Core i5-4460 de 4 núcleos a 3,20 GHz y 32 GB de RAM.

## 4.2 Estimación de la variabilidad

Para cada uno de los sesgos medidos en sus respectivas palabras objetivo, calculamos el **p-valor de permutaciones** de BiasWE (10.000 permutaciones) y el **p-valor del test de log odds ratio** de BiasPMI, considerando la hipótesis nula de ausencia de sesgo. Aplicamos la corrección de Benjamini-Hochberg a los p-valores de cada método de medición de sesgo por separado para ajustar por las comparaciones múltiples [Benjamini y Hochberg, 1995]. También computamos los **intervalos de confianza** de 95 % para cada estimación, usando bootstrap en el caso de BiasWE (2.000 permutaciones) y el método parámetrico en el caso de BiasPMI.

|                      | PMI     | GloVe   | SGNS    | FastText |
|----------------------|---------|---------|---------|----------|
| Sesgo de género      | 82.51 % | 0.30 %  | 0.00 %  | 0.32 %   |
| Sesgo de sentimiento | 66.44 % | 29.72 % | 33.85 % | 21.33 %  |
| Sesgo étnico         | 77.97 % | 0.00 %  | 0.00 %  | 0.00 %   |

Tabla 4.1: Porcentaje de p-valores menores a 0,10 para cada métrica de sesgo en cada experimento.

La **cantidad de palabras que se identifican con un sesgo significativamente distinto de 0 es muy distinta entre BiasPMI y BiasWE** (Tabla 4.1). Por ejemplo, en el caso del sesgo de género, sólo 14 palabras de entre 4661 aparecen con BiasWE (SGNS) significativamente diferente de cero a un nivel de significatividad de 0,10, mientras que alrededor de 82 % de las palabras tienen un BiasPMI significativamente distinto de cero. Aquellas que aparecen con BiasPMI no significativo son las que tienden a tener valores de sesgo cercanos a cero (ver Figura 4.1).

Esto se debe a que **los procedimientos de cálculo de los p-valores para cada tipo de métrica capturan esencialmente distintos tipos de variabilidad**, como explicamos en la sección 3.3. La incertidumbre cuantificada para BiasPMI por medio del test de log odds ratio captura la variabilidad del proceso generador de datos subyacente, es decir, la variabilidad debida al hecho que los conteos de coocurrencias son variables aleatorias. En cambio, los p-valores de permutaciones de BiasWE sólo consideran la variabilidad de los conjuntos de palabras de contexto.

Esta diferencia fundamental se ve reflejada en que a medida que se usan menos palabras en los grupos de contexto, la proporción de palabras que

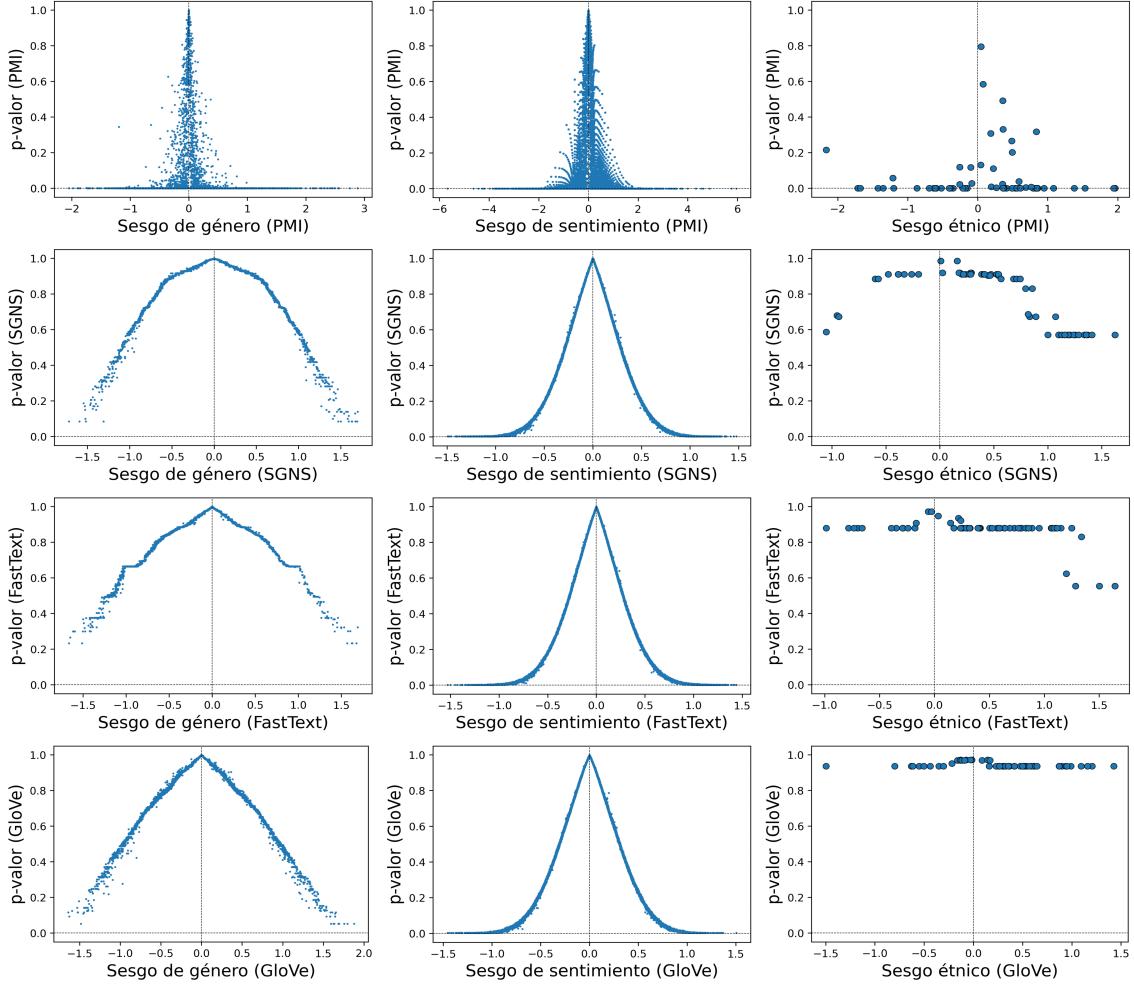


Figura 4.1: p-valores (eje vertical) en función del valor del sesgo (eje horizontal) para cada tipo de sesgo y cada método de medición.

aparecen con sesgo significativamente distinto de cero es menor: alrededor de 30 % de las palabras tienen BiasWE de sentimiento significativo y ninguna palabra tiene BiasWE étnico significativo. De hecho, los p-valores de BiasWE étnico tienden a ser particularmente altos (ver Figura 4.1).

Esta diferencia tan grande se debe a que el sesgo de sentimiento se computa con conjuntos de contexto de 25 palabras cada uno, mientras que el de etnia usa solo 4 palabras en cada grupo. En el límite, si  $A$  y  $B$  fueran listas de una sola palabra, no hay forma de estimar la incertidumbre para BiasWE con estos métodos, mientras que es perfectamente factible para BiasPMI. Este ejemplo ilustra que estos dos métodos de inferencia miden tipos de variabilidad completamente distintos.

Por el mismo motivo, la **amplitud de los intervalos** de BiasWE es relativamente alta en el sesgo étnico, intermedia en el sesgo de género

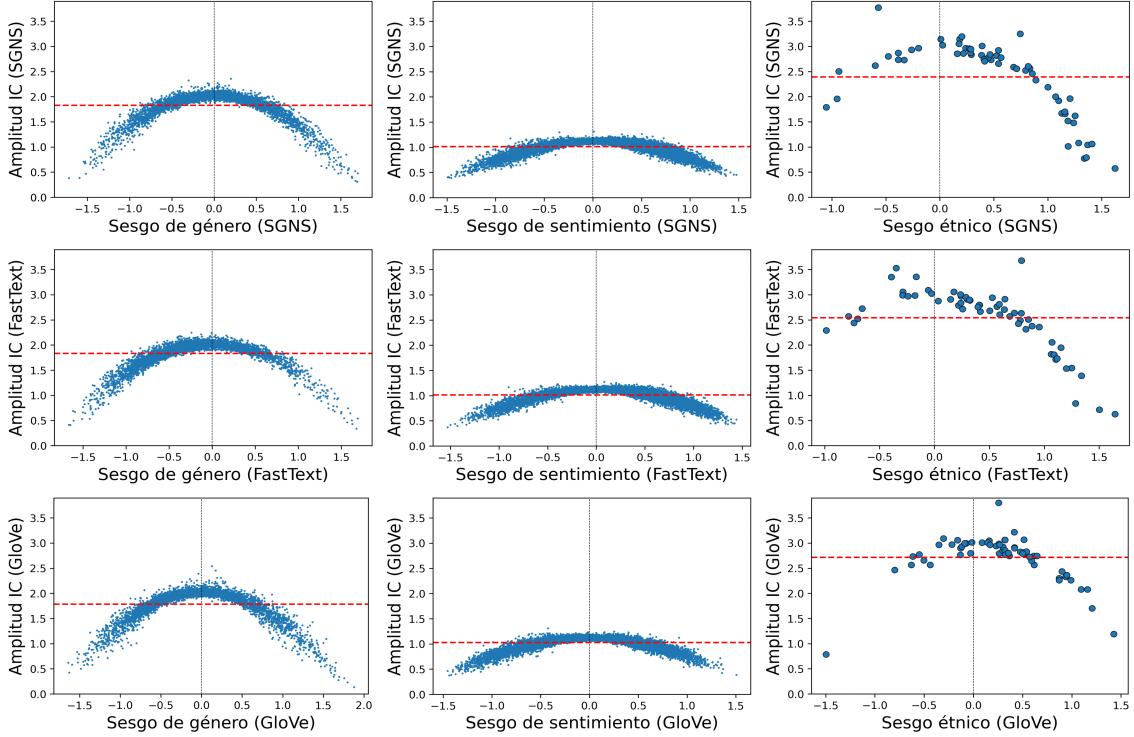


Figura 4.2: Amplitud de los intervalos de confianza al 95 % de BiasWE (eje vertical) en función del valor del sesgo (eje horizontal), para cada tipo de sesgo y cada método de entrenamiento de *embeddings*. En rojo se indica el promedio.

y baja en el sesgo de sentimiento (ver Figura 4.2). Como hemos visto, esto no indica, por ejemplo, que haya relativamente poca evidencia a favor de que haya sesgo étnico en las palabras objetivo elegidas, sino sencillamente que estamos empleando pocas palabras en los grupos de contexto para estimar el sesgo étnico. En cambio, las diferencias en la amplitud de los intervalos de BiasPMI sí puede atribuirse a diferencias en la evidencia disponible en el *corpus* que estamos estudiando.

Una de las desventajas de los métodos de estimación de la variabilidad de BiasWE es, entonces, que para detectar diferencias sistemáticas entre los grupos de contexto para una palabra objetivo determinada, requerimos listas de palabras relativamente grandes en cada grupo de contexto; pero utilizar listas más grandes, con palabras semánticamente asociadas a las palabras de contexto de interés, podría ir en detrimento de la interpretabilidad del sesgo que estamos midiendo.

Otra ventaja del test de log odds ratios para BiasPMI es que es **computacionalmente barato** en comparación con los procedimientos de bootstrap y permutaciones requeridos para BiasWE. Estos pueden ser muy lentos cuando se hace inferencia sobre muchas palabras objetivo, como

es el caso de las estimaciones de sesgo de género y sentimiento en este trabajo.

Por ejemplo, para obtener los p-valores e intervalos de confianza para el conjunto de alrededor de 18,000 palabras objetivo de los tres experimentos, el tiempo de cómputo es despreciable para BiasPMI, mientras que los tests de permutaciones y los intervalos de bootstrap demoran en el orden de 5 horas con el hardware que empleamos.

Ilustramos la ventaja de BiasPMI con algunos casos individuales de medición de sesgo étnico. El BiasPMI étnico del nombre propio *shanice* arroja un valor de -2,17 lo cual indica que es una palabra relativamente más asociada con *white* que con *black* en el corpus. Sin embargo, obtenemos un intervalo de confianza 95 % de (-5,40; 1,07) y el p-valor para la hipótesis nula de ausencia de sesgo es de 0,20. Esto nos dice que la incertidumbre de la estimación puntual es alta y las diferencias entre *A* y *B* no son sistemáticas; la baja cantidad de coocurrencias en el corpus no nos permite concluir que existe un sesgo. En cambio, para la palabra *wine* obtenemos BiasPMI = -1,71 con IC (-1,80; -1,62) y p-valor < 10<sup>-10</sup>. En este caso podemos estar más seguros de que las coocurrencias del corpus indican que existe un sesgo étnico para *wine*, que tiende a estar más asociado en primer orden con *white* que con *black*.

Si usamos, en cambio, por ejemplo, BiasWE con SGNS, obtenemos BiasWE(*shanice*) = 1,07 con IC = (-0,19; 1,81) y p-valor = 0,24; y BiasWE(*wine*) = -0,57 con IC = (-1,95; 1,82) y p-valor = 0,44. Los p-valores altos e intervalos de confianza amplios que incluyen al 0 no necesariamente quieren decir que no existen sesgos sistemáticos para estas palabras en el *corpus*. Por el contrario, esto ocurre porque estamos computando el sesgo con listas de palabras pequeñas que no permiten sacar conclusiones fiables acerca de la variabilidad del sesgo, como delineamos más arriba.

### 4.3 Correlación con el juicio humano

Para los tres sesgos que analizamos, evaluamos la **correlación entre las métricas de sesgo textual y el sesgo de acuerdo al juicio humano**. Medimos la correlación con el coeficiente  $r$  de Pearson. También calculamos un  $r$  de Pearson ponderado, que tiene en cuenta el error estándar de cada estimación de sesgo textual y reduce la influencia de las estimaciones ruidosas en la correlación.

El objetivo de este experimento no es encontrar qué método produce mayores correlaciones, sino más bien estudiar si los resultados de BiasPMI son similares a los de BiasWE, de uso más difundido en este tipo de análisis.

| Sesgo                | Correlación   | PMI  | SGNS | FastText | GloVe |
|----------------------|---------------|------|------|----------|-------|
| Sesgo de género      | $r$           | 0.51 | 0.49 | 0.47     | 0.46  |
|                      | $r$ ponderado | 0.45 | 0.62 | 0.63     | 0.69  |
| Sesgo de sentimiento | $r$           | 0.43 | 0.59 | 0.59     | 0.58  |
|                      | $r$ ponderado | 0.34 | 0.66 | 0.66     | 0.64  |
| Sesgo étnico         | $r$           | 0.14 | 0.44 | 0.36     | 0.30  |
|                      | $r$ ponderado | 0.15 | 0.51 | 0.20     | 0.43  |

Tabla 4.2: Coeficientes de correlación de Pearson entre el sesgo de acuerdo al juicio humano y el sesgo textual medido con cada métrica. Los coeficientes son significativos con un nivel de confianza superior 99 %, exceptuando el  $r$  y el  $r$  ponderado de GloVe para el sesgo étnico, y el  $r$  de PMI para el sesgo étnico, los cuales tienen p-valores  $> 0,10$ .

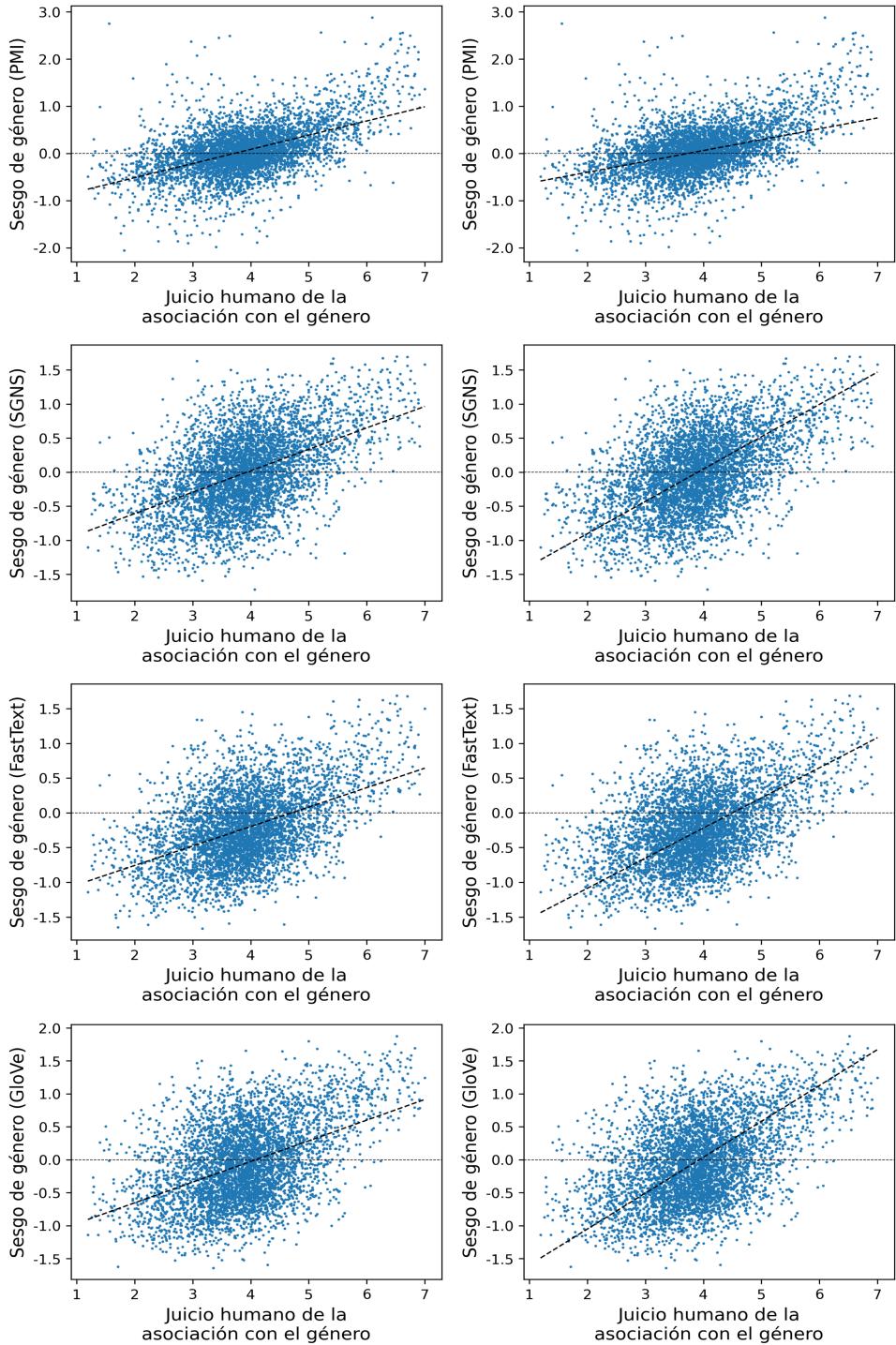


Figura 4.3: Relación entre el sesgo de género textual y de acuerdo al juicio humano en las palabras de Lewis y Lupyán [2020]. Cada punto representa una palabra objetivo. Las rectas representan un ajuste lineal de los datos. En la segunda fila el ajuste pondera cada punto por el error estándar de la estimación de sesgo textual (los intervalos de confianza asociados no se muestran para mayor claridad).

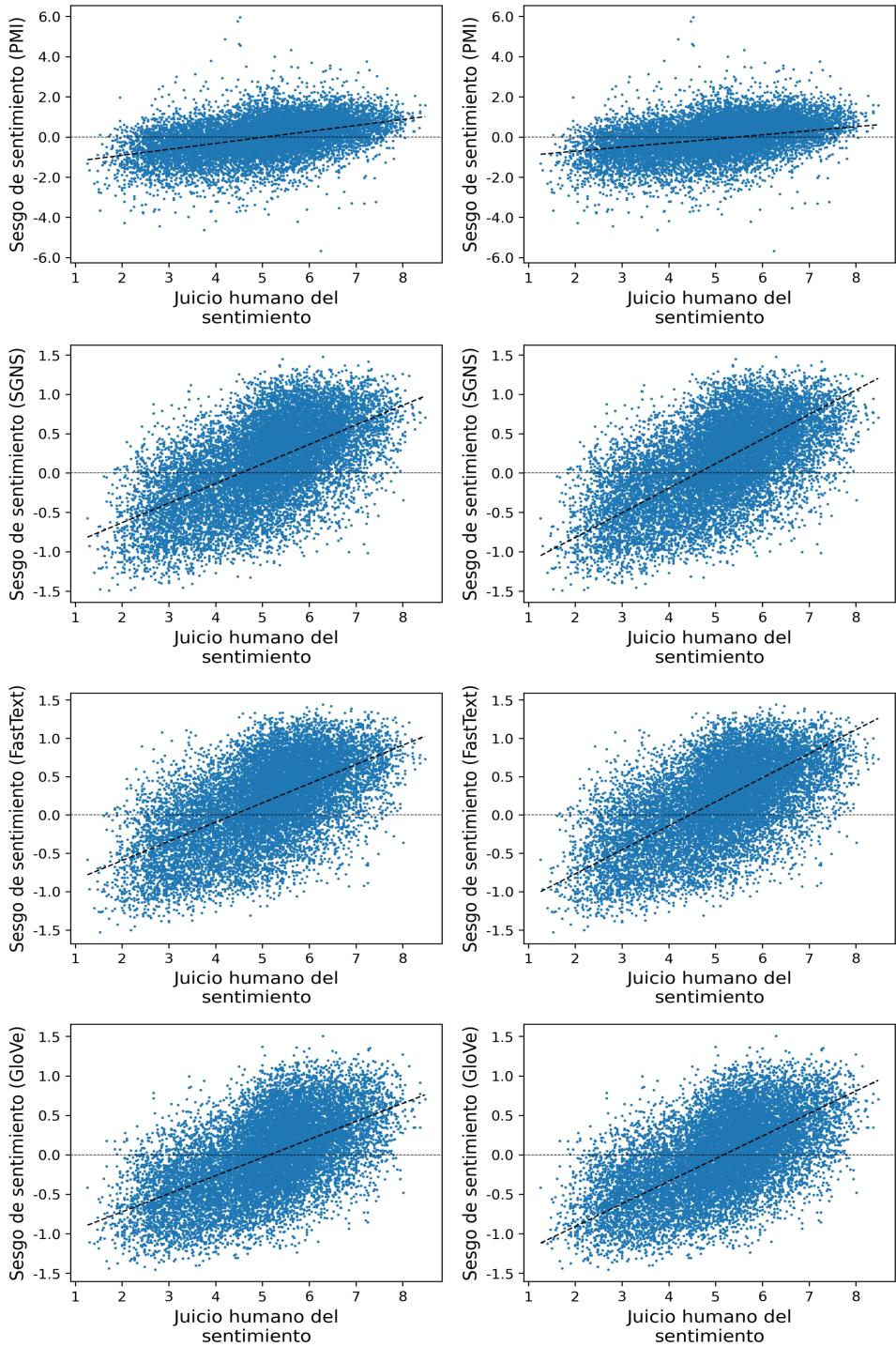


Figura 4.4: Relación entre el sesgo de sentimiento textual y de acuerdo al juicio humano en las palabras de Toney y Caliskan [2021]. Cada punto representa una palabra objetivo. Las rectas representan un ajuste lineal de los datos. En la segunda fila el ajuste pondera cada punto por el error estándar de la estimación de sesgo textual (los intervalos de confianza asociados no se muestran para mayor claridad).

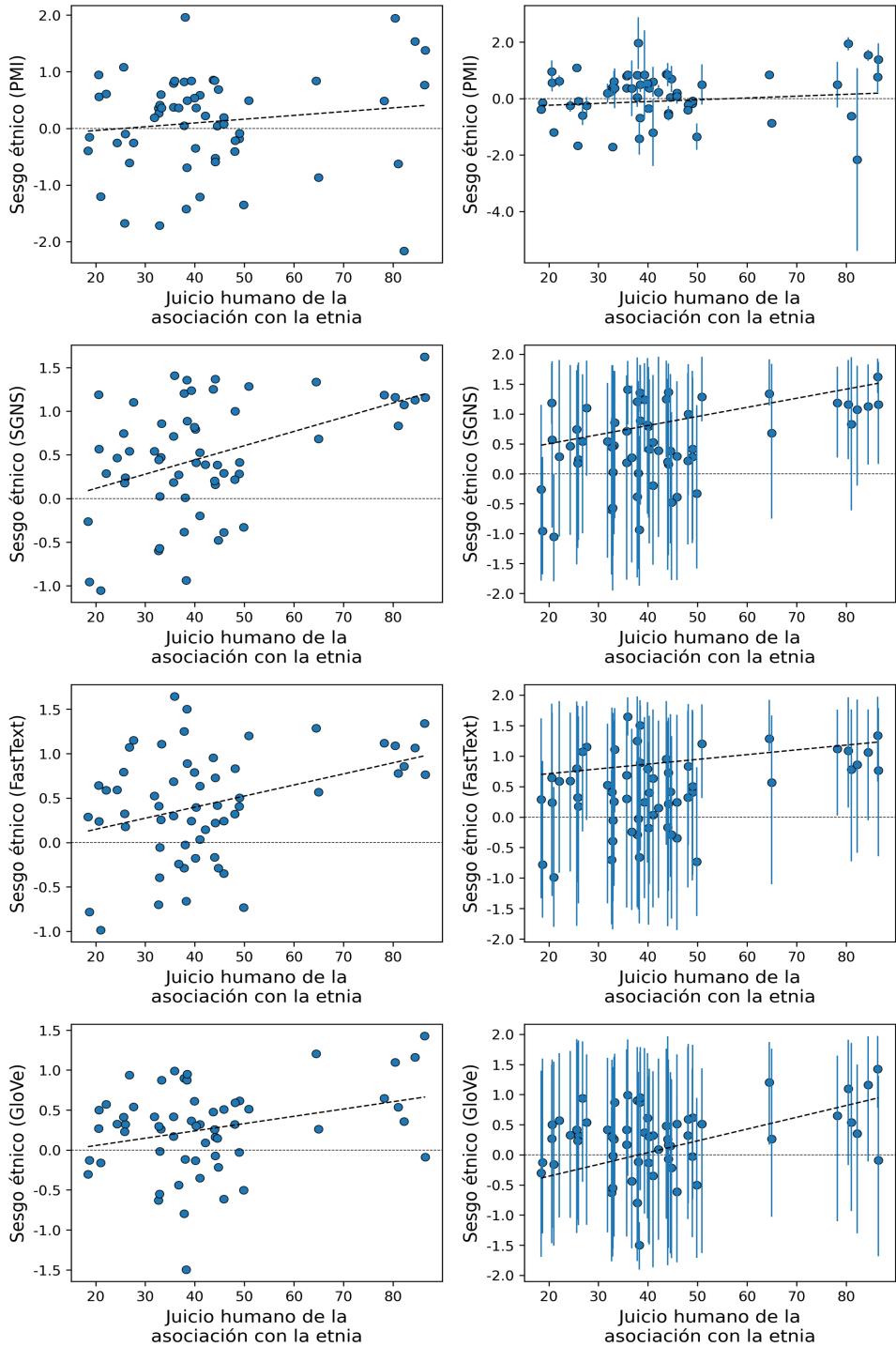


Figura 4.5: Relación entre el sesgo étnico textual y de acuerdo al juicio humano en las palabras de Kozlowski *et al.* [2019]. Cada punto representa una palabra objetivo. Las rectas representan un ajuste lineal de los datos. En la segunda fila el ajuste pondera cada punto por el error estándar de la estimación de sesgo textual (los intervalos de confianza se indican con barras de error).

Encontramos **correlaciones positivas** entre todas las métricas de

**sesgo textual y los sesgos de acuerdo al juicio humano** (ver la Tabla 4.2 con los coeficientes de correlación; los diagramas de dispersión asociados se encuentran en las Figuras 4.3, 4.4 y 4.5). Esto es consistente con los hallazgos previos de los estudios en los cuales están basados los tres experimentos [Kozlowski *et al.*, 2019, Lewis y Lupyán, 2020, Toney y Caliskan, 2021], los cuales usan métricas de sesgo textual basadas en *word embeddings*.

También hallamos que, en general, el uso de ponderadores para calcular la correlación aumenta los coeficientes de BiasWE. Esto implica que reducir la ponderación de estimaciones más ruidosas (donde la diferencia de similitudes con respecto a las listas de palabras *A* y *B* no es tan sistemática) tiende a incrementar la correlación del sesgo de este *corpus*. En cambio, los coeficientes de correlación de BiasPMI tienden a caer o permanecer constantes cuando se usan ponderadores, lo cual indica que los pesos tienden a ser mayores en palabras donde el sesgo textual estimado no coincide con el sesgo de acuerdo al juicio humano. No tenemos ninguna hipótesis sobre por qué el efecto de usar ponderadores es distinto para BiasWE y BiasPMI.

Remarcamos que esto no significa que para cada estimación individual de sesgo los errores estándar de cada método sean mutuamente intercambiables o igualmente útiles, dado que capturan tipos de variabilidad distinta (ver secciones 3.3 y 4.2). Esto se ve reflejado en que los intervalos de confianza de BiasWE tienden a ser relativamente más anchos en BiasWE que en BiasPMI, como se observa en la Figura 4.5.

Tomando como referencia el coeficiente de correlación más alto de cada tipo de métrica (BiasPMI vs. BiasWE) en cada sesgo, se obtienen correlaciones más altas con BiasWE que con BiasPMI. La diferencia es particularmente alta en el caso del sesgo étnico ( $\text{máx}(r_{\text{BiasWE}}) = 0.51$  vs.  $\text{máx}(r_{\text{BiasPMI}}) = 0.15$ ), donde además el grado de correlación de BiasPMI es especialmente débil.

Una hipótesis que puede explicar este resultado es que en este estereotipo las **asociaciones de segundo orden** juegan un rol importante, y éstas no son capturadas por BiasPMI. Determinados consumos culturales como el *basketball* pueden estar más asociados a *black* que a *white* no porque *basketball* aparezca relativamente más en el contexto de palabras que explícitamente refieren a *black* (las palabras de contexto), sino porque comparten vecinos en común, como pueden ser los nombres de las personas que juegan al *basketball*.

Este tipo de asociación puede ser capturada por los puntajes de juicio

humano y por BiasWE, pero no necesariamente por BiasPMI. Una manera de capturar esto con BiasPMI puede ser extendiendo las listas de palabras de contexto, por ejemplo con nombres propios o apellidos típicos de cada etnia (como se realiza, por ejemplo, en Garg *et al.* [2018]). En la sección 4.4 discutimos con mayor detalle las diferencias entre BiasPMI y BiasWE en términos de los tipos de asociaciones semánticas y aspectos del *corpus* que capturan.

Enfatizamos que no consideramos que los coeficientes de correlación con el juicio humano (Tabla 4.2) sean una medida de la bondad global de una métrica de sesgo. El objetivo de nuestro análisis es analizar en qué medida los resultados que se obtienen con los dos tipos de métricas son similares. Con otras listas de palabras probablemente se obtengan otros resultados. Asimismo, es probable que Wikipedia no sea necesariamente representativa de los estereotipos sociales codificados en los puntajes de las encuestas que se usan en cada experimento. Incorporar *corpora* de otros dominios podría dar lugar a correlaciones mayores o menores.

Por otra parte, hasta donde sabemos, no existen *corpora* anotados con los sesgos que contienen. Disponer de ese *ground truth* sería valioso para la tarea de medir sesgos en textos porque permitiría evaluar las métricas de manera más directa, comparando el sesgo textual estimado con el sesgo anotado. Sin embargo, no es obvio cómo anotar la cantidad de sesgo de una palabra en un *corpus*, particularmente si el *corpus* es grande. En los tres experimentos que estudiamos, las “anotaciones” son promedios de valoraciones de humanos sobre la semántica de las palabras, y no tienen en cuenta ningún *corpus* en particular. La falta de estas etiquetas hace que no esté claro cómo sacar conclusiones sobre si un método es globalmente mejor que otro a la hora de medir el sesgo en palabras específicas de un texto.

## 4.4 Interpretación de las estimaciones

En esta sección, analizamos las ventajas y desventajas de medir los sesgos con *word embeddings* en comparación a una métrica de asociación de primer orden como BiasPMI. En particular, usaremos ejemplos concretos de mediciones de sesgo de género binario para **comparar la interpretación de las estimaciones con cada método**.

Como hemos visto en la sección 3.2, BiasPMI puede expresarse intrínsecamente en términos de probabilidades condicionales (ecuación 3.6). El sesgo se interpreta en este caso como el logaritmo de cuánto más pro-

bable es encontrar palabras de  $C$  en el contexto de palabras en  $A$  que en el contexto de palabras en  $B$ . Dado que Bias<sub>PMI</sub> **puede interpretarse de forma transparente en términos de coocurrencias de primer orden**, podemos decir que Bias<sub>PMI</sub> es una métrica absoluta que no requiere comparaciones con otros sesgos para ser interpretada.

Podemos ilustrar esta propiedad con un ejemplo individual. Cuando medimos el sesgo de género binario (femenino/masculino) en el *corpus* de Wikipedia, obtenemos que Bias<sub>PMI</sub>(*nurse*)  $\approx 1,3172$ . Por lo tanto,

$$\frac{P(\text{nurse}|A)}{P(\text{nurse}|B)} \approx e^{1,3172} \approx 3,7330.$$

Esto significa que es aproximadamente un 273,30 % más probable encontrar la palabra *nurse* en el contexto de palabras femeninas ( $A$ ) que en el contexto de palabras masculinas ( $B$ ).

En el caso de BiasWE, si bien existen estudios que buscan interpretar cómo se forman los espacios vectoriales de los *embeddings* [Ethayarajh *et al.*, 2019, Levy y Goldberg, 2014, Levy *et al.*, 2015] o que analizan los patrones semánticos que codifican [Bolukbasi *et al.*, 2016, Gonen y Goldberg, 2019, Zhao *et al.*, 2017], no existe una interpretación transparente de las métricas de sesgo basadas en *embeddings* en términos de las coocurrencias de palabras en los textos.

Esta falta de interpretabilidad de BiasWE se debe a que la similitud entre *embeddings* como SGNS, GloVe y FastText puede captar asociaciones entre palabras tanto de primer orden (sintagmáticas) como de segundo orden (paradigmáticas) o superior [Altszyler *et al.*, 2018, Schlechtweg *et al.*, 2019]. Entonces, cuando usamos *embeddings* para medir sesgos, no es posible saber si los resultados se deben a coocurrencias de primer orden generalizadas o se derivan de coocurrencias de orden superior poco transparentes [Brunet *et al.*, 2019, Rekabsaz *et al.*, 2021]. En definitiva, mientras **la relación entre BiasWE y la distribución de palabras no es transparente**, en el caso de Bias<sub>PMI</sub> sí lo es porque PMI es estrictamente una métrica de asociación de primer orden.

Ejemplificamos esta diferencia en interpretabilidad con el caso de la palabra *evil*. En Wikipedia, el Bias<sub>PMI</sub> de género de *evil* es igual a  $-0,25$ , lo que indica una mayor probabilidad de aparecer en el contexto de palabras de contexto masculino ( $B$ ) en comparación con las femeninas ( $A$ ). Por el contrario, BiasWE =  $0,23$  con SGNS. Aunque sabemos que esto se interpreta como un sesgo femenino, es difícil comprender el origen exacto de este resultado porque está influido por coocurrencias de segundo

orden o superior, y no podemos medir el peso que tiene cada uno de estos factores.

A continuación mostramos otros ejemplos que ilustran cómo el orden de asociación puede influir en las estimaciones de sesgo de género:

- Hay palabras no incluidas en el grupo *A* que intrínsecamente tienen un sesgo femenino, como *women*, *wife*, *ladies* o *girls*. Sin embargo, estas palabras aparecen con  $\text{BiasPMI} < 0$  (sesgo masculino) porque tienden a aparecer relativamente más en el contexto de palabras masculinas (grupo *B*) que femeninas (grupo *A*) en Wikipedia (por ejemplo, *his wife*). Probablemente la asociación de estas palabras con las del grupo *A* sea de segundo orden o paradigmática (tienden a tener vecinos similares), y este aspecto no es capturado por  $\text{BiasPMI}$ .
- Ocupaciones como *assistant* y *secretary* están estereotípicamente asociadas al género femenino [Caliskan *et al.*, 2017]. Sin embargo, obtenemos valores de  $\text{BiasPMI}$  de género de  $-0,19$  y  $-0,24$ , respectivamente (ambos con  $p$ -valor  $< 10^{-5}$ ). Esto indica una asociación de primer orden más fuerte con las palabras de contexto masculinas (*B*) en comparación con las femeninas (*A*); por ejemplo, por la presencia de estructuras como *his assistant*. Al usar  $\text{BiasWE}$  con SGNS obtenemos valores de  $0,70$  y  $0,43$ , respectivamente: el sesgo invierte su signo. Esto podría deberse al efecto de la asociación de segundo orden: estas palabras probablemente tienen vecinos similares a las palabras del grupo *A*, es decir, aparecen en contextos similares (por ejemplo, contextos relacionados a la oficina o el cuidado).
- Palabras como *harbor*, *navy*, *dock*, *port*, *fleet*, *steam*, *pier*, *sailor*, *ship*, y *sail* tienen valores positivos de  $\text{BiasPMI}$  de género en Wikipedia. Esto puede deberse a que en el idioma inglés, los barcos se suelen denominar con pronombres femeninos como *she*. Los valores de  $\text{BiasWE}$  con SGNS para estas palabras son, en cambio, inferiores a  $0$ , posiblemente porque estas palabras están más relacionadas con el género masculino a través de coocurrencias de segundo orden (e.g. los barcos y los varones pueden aparecer en contextos similares relacionados a la marina o la guerra).

Otro aspecto interesante a considerar es que las métricas de sesgo que capturan asociaciones de segundo orden tienen la ventaja de **gestionar la raleza de los datos (*data sparsity*)**. Dado que todo *corpus* es limitado, algunas de las muchas formas que pueden tomar los conceptos objetivo y de contexto pueden aparecer raramente en el texto. En este escenario en el que las coocurrencias son ralas (*sparse*), cuando

usamos BiasWE, puede que no sea necesario incluir todas las palabras relacionadas con los conceptos objetivo y de contexto para medir sesgos exitosamente, dado que los *embeddings* pueden captar sinonimia. En el caso de BiasPMI, este problema debe abordarse aumentando las listas de palabras con sinónimos y formas de las palabras de interés.

Para ejemplificar esto, consideremos el caso de los sinónimos *nourish* y *nurture*, que tienen frecuencias diferentes en el corpus de Wikipedia (700 y 3,000, respectivamente). Con BiasPMI, obtenemos un sesgo de 0,33 para *nurture* ( $p$ -valor  $< 10^{-4}$ ). Sin embargo, si hubiéramos utilizado en su lugar su sinónimo menos frecuente *nourish*, el BiasPMI habría sido  $-0,10$  y no habría sido estadísticamente significativo ( $p$ -valor  $\approx 0,66$ ). En este caso no habríamos podido determinar si realmente no hay sesgo o si la cantidad de datos es insuficiente para determinar esto.

Esto demuestra que, en general, es aconsejable incluir todos los sinónimos y variaciones pertinentes del término cuyo sesgo intentamos medir cuando usamos BiasPMI. Por otra parte, cuando se usa BiasWE podríamos confiar en el hecho de que los *embeddings* pueden capturar sinonimia; en este ejemplo particular BiasWE con SGNS arroja valores positivos para ambas palabras.

Por último, en investigaciones recientes hemos demostrado que BiasWE, **a diferencia de BiasPMI, puede producir resultados engañosos porque captura inadvertidamente las disparidades en las frecuencias de las palabras de contexto** [Valentini *et al.*, 2022].

Para dar cuenta de este problema en el contexto de la interpretabilidad, analizamos las estimaciones de sesgo de *stopwords*, es decir, palabras muy frecuentes con poco contenido semántico (ver Tabla 4.3). Mientras que BiasPMI no tiende a detectar sesgo de género sistemáticamente en alguna dirección en las *stopwords*, BiasWE con cualquiera de los tres métodos de generación de *embeddings* tiende a producir estimaciones de sesgo siempre negativas i.e. con sesgo masculino. Esto se debe a que las palabras de contexto masculinas (grupo *B*) son más frecuentes que las femeninas (grupo *A*) en el corpus de Wikipedia, y a que los *embeddings* tienen la capacidad de capturar la frecuencia de las palabras [Valentini *et al.*, 2022].

En consecuencia, no podemos interpretar qué aspectos del corpus dan lugar a las estimaciones de sesgo que observamos. Cuando usamos BiasWE, en definitiva, no solo es difícil determinar el tipo de asociación que da origen a los resultados, sino también si el sesgo semántico detectado es genuino o si sólo se debe a la disparidad en las frecuencias de las palabras

| Palabra | PMI   | SGNS  | FastText | GloVe |
|---------|-------|-------|----------|-------|
| which   | -0.08 | -0.50 | -0.59    | -0.51 |
| first   | 0.03  | -0.15 | -0.28    | -0.50 |
| after   | -0.06 | -0.60 | -0.73    | -0.62 |
| have    | 0.13  | -0.42 | -0.35    | -0.45 |
| other   | 0.07  | -0.57 | -0.45    | -0.44 |
| all     | -0.05 | -0.55 | -0.58    | -0.64 |
| over    | -0.20 | -0.80 | -0.87    | -0.76 |
| only    | 0.02  | -0.46 | -0.56    | -0.60 |
| most    | -0.13 | -0.53 | -0.56    | -0.62 |
| up      | 0.11  | -0.63 | -0.67    | -0.64 |
| used    | -0.13 | -0.72 | -0.38    | -0.68 |
| under   | -0.16 | -0.96 | -1.12    | -1.15 |
| part    | 0.07  | -0.31 | -0.41    | -0.40 |
| many    | -0.15 | -0.42 | -0.60    | -0.64 |
| well    | 0.04  | -0.29 | -0.44    | -0.52 |
| name    | 0.13  | -0.26 | -0.28    | -0.50 |
| several | -0.09 | -0.36 | -0.53    | -0.53 |
| same    | 0.05  | -0.06 | -0.40    | -0.54 |
| former  | -0.21 | -0.60 | -0.59    | -0.59 |
| system  | -0.55 | -1.01 | -0.77    | -0.93 |

Tabla 4.3: Sesgo de género de las 20 *stopwords* más frecuentes de las palabras del experimento de Glasgow [Scott *et al.*, 2019].

de contexto.

# Capítulo 5

## Conclusiones

En los últimos años, la temática de los sesgos en los modelos de aprendizaje automático ha suscitado una gran atención. Aunque numerosos estudios han explorado los sesgos presentes en los modelos, siguen siendo escasas las herramientas diseñadas para medir y analizar directamente los sesgos en los textos. Además, estas herramientas suelen carecer de interpretabilidad. En esta tesis, abordamos estas deficiencias introduciendo y analizando una métrica basada en *Pointwise Mutual Information* (PMI) para medir sesgos en *corpora*. A través de nuestra investigación, hemos destacado las diferencias del enfoque basado en PMI sobre las métricas tradicionales basadas en *word embeddings* estáticos, como SGNS, GloVe y FastText, poniendo especial énfasis en las ventajas.

Una de las principales contribuciones de nuestro trabajo es la introducción de la métrica basada en PMI como **método sencillo, interpretable y computacionalmente eficiente de medir sesgos textuales**. A diferencia de las métricas basadas en *embeddings*, que carecen de transparencia e interpretabilidad, nuestro enfoque ofrece una interpretación clara en términos de coocurrencias de primer orden, y por lo tanto, una comprensión más intuitiva de los sesgos subyacentes presentes en los textos.

Además, introducimos una **técnica paramétrica para estimar la incertidumbre asociada a las estimaciones** de la métrica basada en PMI. Esta manera de medir la variabilidad permite a los investigadores determinar hasta qué punto los valores medidos pueden atribuirse a fluctuaciones estadísticas. A diferencia de las metodologías tradicionales basadas en el remuestreo, como el *bootstrapping* y las pruebas de permutación utilizadas habitualmente con los *embeddings*, nuestro enfoque paramétrico brinda una estimación más informativa de la verdadera variabilidad en las mediciones de sesgo.

Mediante una serie de experimentos, presentamos **evidencia empírica que respalda las ventajas del método basado en PMI** frente a las métricas basadas en *embeddings*. Estos resultados demuestran, asimismo, que la métrica basada en PMI muestra asociaciones similares

con el juicio humano que las métricas basadas en *embeddings* en determinados escenarios, como los sesgos de género; mientras que en otros casos, como los estereotipos étnicos, PMI y *embeddings* arrojan resultados divergentes. Esta distinción subraya las diferencias fundamentales en las asociaciones semánticas capturadas por los *embeddings* y PMI, respectivamente.

La disponibilidad de herramientas para medir sesgos en los textos es limitada, y la interpretabilidad de estas herramientas es aún más acotada. Nuestro trabajo aborda esta necesidad proporcionando un método transparente e interpretable. Creemos que los estudios centrados en la interpretabilidad y las propiedades estadísticas de las métricas son de suma importancia para el NLP. Al facilitar análisis más transparentes e interpretables, nuestro enfoque puede ayudar a los investigadores del área a hacer análisis más rigurosos de los sesgos potencialmente presentes en textos. En última instancia, nuestro trabajo puede contribuir a mejorar la equidad y la imparcialidad de los sistemas de Inteligencia Artificial, pues permite también **estudiar y controlar los sesgos de los datos de entrenamiento**.

Además, nuestra contribución es valiosa para los estudios de ciencias sociales computacionales. La métrica basada en PMI puede funcionar como una **herramienta cuantitativa que complementa a los análisis lingüísticos y sociológicos existentes, más cualitativos, de los sesgos culturales**, mejorando así el conjunto de herramientas analíticas a disposición de los investigadores sociales. Para los estudios en ciencias sociales es particularmente importante no sólo disponer de una métrica transparente para cuantificar estereotipos, sino también de pruebas estadísticas e intervalos de confianza que capten la variabilidad relevante.

Subrayamos la importancia de la interpretabilidad y el rigor estadístico en el desarrollo de herramientas de medición de sesgos en general, y animamos a seguir explorando y perfeccionando estos métodos.

# Bibliografía

- Agresti, A. (2003). *Categorical data analysis*, volumen 482. John Wiley & Sons.
- Aka, O., Burke, K., Bauerle, A., Greer, C., y Mitchell, M. (2021). Measuring model biases in the absence of ground truth. En *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.
- Altszyler, E., Sigman, M., y Fernández Slezak, D. (2018). Corpus specificity in LSA and word2vec: The role of out-of-domain documents. En *Proceedings of The Third Workshop on Representation Learning for NLP*, pp. 1–10, Melbourne, Australia. Association for Computational Linguistics.
- Benjamini, Y. y Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1):289–300.
- Blodgett, S. L., Barocas, S., Daumé III, H., y Wallach, H. (2020). Language (technology) is power: A critical survey of “bias” in NLP. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online. Association for Computational Linguistics.
- Bojanowski, P., Grave, E., Joulin, A., y Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., y Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. En Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., y Garnett, R., editores, *Advances in Neural Information Processing Systems*, volumen 29. Curran Associates, Inc.
- Bordia, S. y Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.
- Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., y Zemel, R. (2019).

- Understanding the origins of bias in word embeddings. En Chaudhuri, K. y Salakhutdinov, R., editores, *Proceedings of the 36th International Conference on Machine Learning*, volumen 97 de *Proceedings of Machine Learning Research*, pp. 803–811. PMLR.
- Caliskan, A., Bryson, J. J., y Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.
- Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., y Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240.
- Church, K. W. y Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
- Davison, A. C. y Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- DeFranza, D., Mishra, H., y Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology*, 119(1):7–22.
- Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Duchi, J., Hazan, E., y Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.
- Ethayarajh, K., Duvenaud, D., y Hirst, G. (2019). Understanding undesirable word embedding associations. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1696–1705, Florence, Italy. Association for Computational Linguistics.
- Gálvez, R. H., Tiffenberg, V., y Altsyler, E. (2019). Half a century of stereotyping associations between gender and intellectual ability in films. *Sex Roles*, 81(9):643–654.

Garg, N., Schiebinger, L., Jurafsky, D., y Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Gonen, H. y Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., y Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Hoyle, A. M., Wolf-Sonkin, L., Wallach, H., Augenstein, I., y Cotterell, R. (2019). Unsupervised discovery of gendered language through latent-variable modeling. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1706–1716, Florence, Italy. Association for Computational Linguistics.

Jones, J. J., Amin, M. R., Kim, J., y Skiena, S. (2020). Stereotypical gender associations in language have decreased over time. *Sociological Science*, 7(1):1–35.

Jurafsky, D. y Martin, J. H. (2009). *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Kiritchenko, S. y Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. En *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Kozlowski, A. C., Taddy, M., y Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Levy, O. y Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. En Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., y Weinberger, K. Q., editores, *Advances in Neural Information Processing Systems*, volumen 27. Curran Associates, Inc.

- Levy, O., Goldberg, Y., y Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.
- Lewis, M. y Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, 4(10):1021–1028.
- Lu, K., Mardziel, P., Wu, F., Amancharla, P., y Datta, A. (2020). Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pp. 189–202.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., y Dean, J. (2013). Distributed representations of words and phrases and their compositionality. En Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., y Weinberger, K. Q., editores, *Advances in Neural Information Processing Systems*, volumen 26. Curran Associates, Inc.
- North, B. V., Curtis, D., y Sham, P. C. (2002). A note on the calculation of empirical p values from monte carlo procedures. *The American Journal of Human Genetics*, 71(2):439–441.
- Pennington, J., Socher, R., y Manning, C. D. (2014). Glove: Global vectors for word representation. En *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.
- Řehůřek, R. y Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. En *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta. ELRA. <http://is.muni.cz/publication/884893/en>.
- Rekabsaz, N., West, R., Henderson, J., y Hanbury, A. (2021). Measuring societal biases from text corpora with smoothed first-order co-occurrence. *Computing Research Repository*, arXiv:1812.10424.
- Schlechtweg, D., Oguz, C., y Schulte im Walde, S. (2019). Second-order co-occurrence sensitivity of skip-gram with negative sampling. En *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 24–30, Florence, Italy. Association for Computational Linguistics.
- Scott, G., Keitel, A., Becirspahic, M., Yao, B., y Sereno, S. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51:1258–1270.
- Toney, A. y Caliskan, A. (2021). ValNorm quantifies semantics to re-

- veal consistent valence biases across languages and over centuries. En *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7203–7218, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Turney, P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Valentini, F., Rosati, G., Blasi, D., Fernandez Slezak, D., y Altszyler, E. (2023). On the Interpretability and Significance of Bias Metrics in Texts: a PMI-based Approach. En *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, Canada. Association for Computational Linguistics.
- Valentini, F., Rosati, G., Slezak, D. F., y Altszyler, E. (2022). The undesirable dependence on frequency of gender bias metrics based on word embeddings. En *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics.
- Warriner, A. B., Kuperman, V., y Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207.
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., y Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. En *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., y Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. En *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, New Orleans, Louisiana. Association for Computational Linguistics.