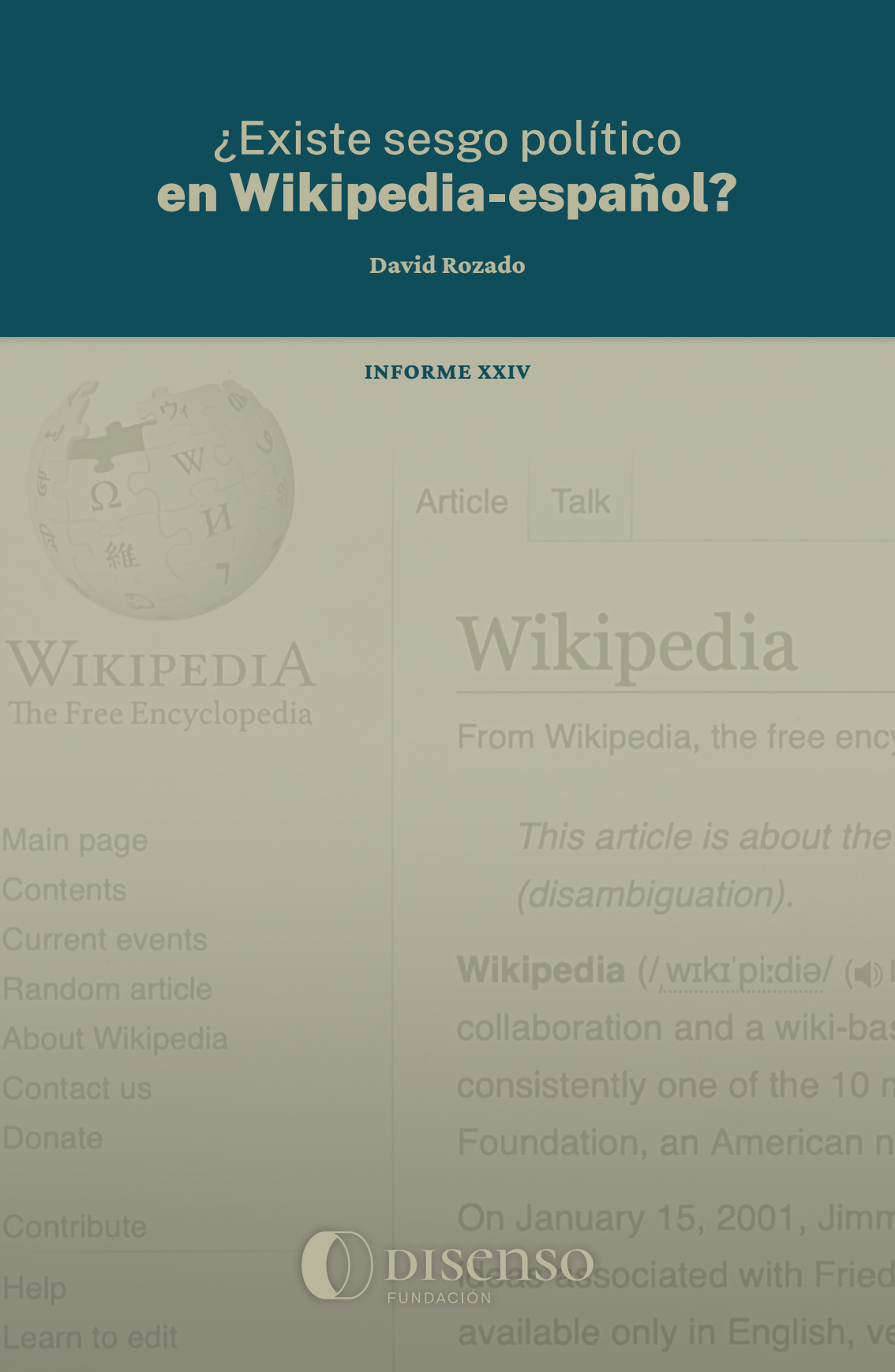
¿Existe sesgo político en Wikipedia-español?

David Rozado

INFORME XXIV



DISENSO
FUNDACIÓN

FUNDACIÓN DISENSO

Pº. del General Martinez Campos 21, 1ºA.
28010, Madrid
info@fundaciondisenso.org
prensa@fundaciondisenso.org

Índice

David Rozado es doctor en Filosofía y Ciencia Informática por la Universidad Autónoma de Madrid. En la actualidad ejerce de profesor titular de Informática en Te Pūkenga – The New Zealand Institute of Skills and Technology neozelandés. Su campo de estudio es el aprendizaje automático (AI) y el análisis de datos.

1. Introducción

La enciclopedia gratuita colaborativa Wikipedia fue puesta en línea el 15 de enero de 2001 por sus creadores Jimmy Wales y Larry Sanger. En los más de 20 años desde su creación, Wikipedia se ha convertido gradualmente en un recurso informativo indispensable para millones de usuarios. Un análisis publicado en el año 2005 en la revista *Nature* reveló que el conjunto de contenidos de Wikipedia-inglés, que en la actualidad cuenta con más de 6 millones de artículos redactados y modificados por colaboradores voluntarios no remunerados, es en promedio de una calidad similar a los de Enciclopedia Británica.[1]

La popularidad del proyecto Wikipedia es indudable. El dominio web Wikipedia.org ocupa la posición número 9 en el ranking de dominios web más visitados del mundo, recibiendo cada mes 4.300 millones de visitas. Así, la posibilidad de que el contenido de Wikipedia contenga sesgo político es preocupante debido a la enorme capacidad de la página web para influir las percepciones de un amplio porcentaje de la humanidad.

Desde su fundación, Wikipedia ha aspirado a presentar artículos que carezcan de sesgos. Uno de los principios, que todos los artículos de Wikipedia aspiran a lograr, es un "punto de vista neutral" (NPOV, por sus siglas en inglés), junto a la "verificabilidad". Si un artículo refleja NPOV, las opiniones en conflicto se presentan una al lado de la otra, con todos los puntos de vista significativos representados. Esta aspiración parece bastante factible en algunos contextos, ya que no debería ser difícil de lograr cuando los artículos cubren temas poco controvertidos, cargados de información objetiva y que se pueden verificar contra muchas fuentes. Ese entorno caracteriza a la gran mayoría de los artículos de Wikipedia sobre temas científicos. Sin embargo, algunos temas con connotaciones políticas carecen de estas características ideales. Es, por lo tanto, importante esclarecer qué potenciales sesgos podrían existir en temas donde parte de la información es controvertida, subjetiva o inverificable.

La posibilidad de que exista sesgo político en el contenido de Wikipedia ha sido investigada con anterioridad. Un estudio publicado en el año 2012 fue el primero en reportar la existencia de un sesgo a la izquierda del centro político

1 J. Giles, "Internet encyclopaedias go head to head," Nature, vol. 438, no. 7070, Art. no. 7070, Dec. 2005, doi: 10.1038/438900a.

en el contenido de Wikipedia-inglés.[2] Este estudio fue replicado más recientemente en el año 2018 con unos resultados muy similares al estudio original, y que confirmaron la existencia de un sesgo político a favor de puntos de vista alineados con el partido Demócrata de los Estados Unidos.[3] Otros estudios también han documentado sesgo en los juicios sobre fuentes y sesgo en la aplicación del arbitraje en artículos con disputas editoriales que mostraban favoritismo institucional hacia puntos de vista a la izquierda del centro.[4]

Incluso uno de los creadores de Wikipedia, Larry Sanger, ha manifestado en público repetidamente su opinión de que Wikipedia contiene un alto grado de sesgo político. Sanger ha sido explícito en su dictamen de que los artículos de Wikipedia presentan una perspectiva izquierdista y liberal o *"del establishment"*. Debido a estos sesgos, Sanger ha acusado a Wikipedia de abandonar su política de neutralidad (punto de vista neutral) y, por consiguiente, considera a Wikipedia como poco fiable.[5]

Una limitación de los estudios previos sobre sesgos políticos en el contenido de Wikipedia es que la mayoría de dichos análisis se han llevado a cabo sobre el contenido de Wikipedia en inglés. Existe por tanto la necesidad de caracterizar la existencia o ausencia de sesgos políticos en los artículos de Wikipedia escritos en idioma español. El objetivo de este informe es cubrir este hueco en la literatura académica y determinar si existe sesgo político en el contenido de Wikipedia-español.

---

2 S. Greenstein and F. Zhu, "Is Wikipedia Biased?," American Economic Review, vol. 102, May 2012, doi: 10.1257/aer.102.3.343.

3 S. Greenstein and F. Zhu, "Do experts or crowd-based models produce more bias? evidence from encyclopedia britannica and wikipedia," MIS Q., vol. 42, no. 3, pp. 945–960, Sep. 2018, doi: 10.25300/MISQ/2018/14084.

4 "The left-wing bias of Wikipedia | Shuichi Tezuka and Linda A. Ashtear," The Critic Magazine, Oct. 22, 2020. https://thecritic.co.uk/the-left-wing-bias-of-wikipedia/ (accessed Mar. 14, 2023).

5 "Larry Sanger," Wikipedia. Feb. 25, 2023. Accessed: Mar. 18, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Larry_Sanger&oldid=1141551296

2. Metodología

En este trabajo derivamos modelos de embeddings del contenido de Wikipedia-español para cuantificar la frecuencia con la que, en los artículos de Wikipedia español, palabras clasificadas como positivas o negativas por léxicos de sentimiento son utilizadas en las inmediaciones de palabras con connotaciones políticas, tales como nombres de líderes políticos, ideologías o nombres de partidos políticos (ver Apéndice I para una descripción extendida de la metodología de análisis empleada en este informe y en trabajos previos de la literatura académica).

En nuestros experimentos utilizamos el léxico de sentimiento AFINN[6] traducido al español y que contiene un total de 2930 palabras anotadas por su polaridad de sentimiento positivo o negativo, pero resultados similares se obtienen utilizando otros léxicos de sentimiento tales como NRC[7] o iSOL[8]. La distribución de palabras en el léxico AFINN no es simétrica (el número de palabras anotadas como de sentimiento negativo supera al número de palabras anotadas como de sentimiento positivo). Para obtener una distribución simétrica de anotaciones de sentimiento realizamos un podado aleatorio de palabras negativas en el léxico AFINN con un resultado final de un léxico simétrico de 2066 palabras (1033 positivas y 1033 negativas).

En la Figura 1 se puede observar cómo este método captura en el contenido de Wikipedia español asociaciones intuitivas, utilizando palabras a modo de ilustración y con connotaciones ampliamente aceptadas como positivas: *vida*, *paz*, *salud* o *democracia*, tienden a asociarse de forma preferencial en el contenido de Wikipedia-español con un amplio número de palabras clasificadas como positivas (en verde) por el léxico de sentimiento AFINN. Sin embargo, sus antónimos: *muerte*, *guerra*, *enfermedad* o *dictadura,* tienden a asociarse de manera preferencial en el contenido de Wikipedia-español con palabras clasificadas como de sentimiento negativo (en rojo) por el léxico AFINN. En la Figura 1, las líneas verticales discontinuas indican la media estadística de

6 "AFINN Sentiment Lexicon — sentiment_afinn • corpus." http://corpustext.com/reference/sentiment_afinn.html (accessed Mar. 18, 2023).

7 "NRC Emotion Lexicon." https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (accessed Mar. 18, 2023).

8 "iSOL | Red TIMM." http://timm.ujaen.es/recursos/isol/ (accessed Mar. 18, 2023).

intensidad de asociación de cada distribución de palabras (positivas: en verde, negativas: en rojo). Estos resultados, a pesar de ser poco sorprendentes, son presentados aquí para ilustrar la capacidad de los modelos de embeddings para capturar asociaciones latentes en un corpus de texto. (En el apéndice de este informe se pueden encontrar más detalles sobre esta metodología).

Distribución de asociaciones en artículos de Wikipedia-español entre términos ilustrativos y un léxico de sentimiento con palabras positivas y negativas (N=2066)

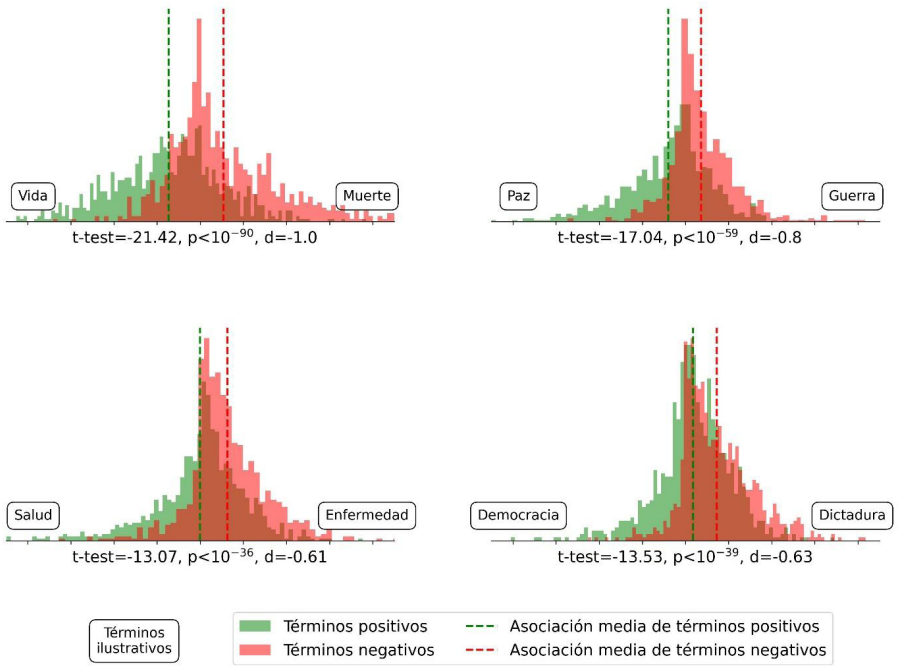


Figura 1. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y una serie de términos ilustrativos con connotaciones ampliamente aceptadas como positivas en el polo izquierdo de cada eje horizontal ('vida', 'paz', 'salud' y 'democracia') y negativas ('muerte', 'guerra', 'enfermedad' y 'dictadura') en el polo derecho de cada eje horizontal.

En nuestros análisis experimentales, realizamos primordialmente una serie de comparaciones entre términos con connotaciones políticas en el eje izquierda–derecha. La hipótesis nula (***null hypothesis***) es que no existe una diferencia en el contenido de Wikipedia-español a la hora de asociar palabras positivas o negativas con cualquiera de los polos (izquierda/derecha) de dicho eje con connotaciones políticas.

En cada experimento, llevamos a cabo un test estadístico para evaluar si existe una diferencia entre los dos grupos con carga política, que están siendo comparadas y representadas cada una en un polo del eje horizontal, y las asociaciones positivas y negativas predominantes en los artículos de Wikipedia. Generalmente, situamos a las palabras con connotaciones de izquierda en el polo izquierdo y las palabras con connotaciones de derechas en el polo derecho.

El test empleado para evaluar si existe una diferencia estadísticamente significativa entre las asociaciones de sentimiento con las palabras representadas en los polos del eje ideológico es un *t-test* independiente, del cual reportamos también su *p-value* para determinar si la diferencia entre las asociaciones positivas y negativas con los términos políticos analizados es estadísticamente significativa. Debido al alto volumen de palabras en el léxico AFINN simétrico (N=2066), los resultados tienden a ser estadísticamente significativos ya que, cuando existe un efecto, incluso sí este es excepcionalmente leve, los *p-values* son proporcionales a N. Esta limitación es una de las críticas que normalmente recibe el uso de *p-values* en el análisis estadístico de diferencias entre grupos. Para neutralizar esta crítica metodológica, en nuestro análisis reportamos también una medida de tamaño de efecto (*effect size*): *Cohen's d,* que, en el contexto de nuestro análisis, es probablemente más informativa.[9] Una interpretación comúnmente utilizada es referirse a los tamaños del efecto *Cohen d* como pequeños (d = 0.2), medianos (d = 0.5) y grandes (d = 0.8), basándose en los puntos de referencia sugeridos por el propio Cohen.[10]

En el siguiente análisis, valores de *Cohen d* negativos indican una tendencia a asociar palabras clasificadas por el léxico de sentimiento AFINN como positivas con palabras en el polo izquierdo del eje horizontal que, en nuestros análisis, generalmente denotan orientación ideológica a la izquierda del centro político. Correspondientemente, las asociaciones preferenciales de palabras negativas se asocian con términos en el polo derecho del eje horizontal, que denotan orientación ideológica a la derecha del centro político. Valores de *Cohen d* positivos sugieren asociaciones opuestas: una tendencia a asociar palabras positivas con términos que referencian la derecha política y palabras negativas con términos con connotaciones de izquierda ideológica. La literatura académica previa ha caracterizado que estas asociaciones preferenciales, en un gran

9 D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs," Front Psychol, vol. 4, p. 863, Nov. 2013, doi: 10.3389/fpsyg.2013.00863.

10 J. Cohen, Statistical Power Analysis for the Behavioral Sciences. Academic Press, 2013.

volumen de texto, correlacionan de forma sustancial con el sesgo político de dicho texto[11] y, por lo tanto, en este estudio, las utilizamos como un indicador que cuantifica el sesgo político en el contenido en español de Wikipedia.

Para concluir, en este trabajo derivamos modelos de embedding de palabras a partir de artículos de Wikipedia-español y analizamos las asociaciones latentes prevalentes en dichos artículos entre términos con connotaciones políticas y palabras clasificadas como de sentimiento *positivo* o *negativo* por un léxico de sentimiento. En cada análisis, animamos al lector a prestar atención al valor *Cohen's d*, que cuantifica la intensidad del sesgo ideológico incrustado en el contenido de Wikipedia español para cada experimento. En los experimentos donde el polo izquierdo del eje horizontal contiene palabras con connotaciones de izquierda ideológica y el polo derecho contiene palabras con connotaciones de derecha ideológica, los valores de *Cohen's d* negativos sugieren sesgo ideológico a favor de la izquierda. Por el contrario, los valores de *Cohen's d* positivos sugieren sesgo ideológico a favor de la derecha (En el Apéndice de este informe el lector puede encontrar detalles adicionales sobre la metodología de embeddings aplicada al análisis de grandes Corpus de texto).

11 Rozado, D., & Al-Gharbi, M. (2021). Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets. Journal of Computational Social Science, 1-22.

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

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

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

1. Resultados

La Figura 2 muestra las asociaciones de sentimiento más comunes en el contenido de Wikipedia-español entre palabras con connotaciones positivas (en verde) y negativas (en rojo) y nombres de presidentes de los Estados Unidos. Los resultados sugieren que, en el contenido de Wikipedia en español, existe una mayor tendencia a asociar palabras positivas con presidentes de Estados Unidos pertenecientes al Partido Demócrata que con presidentes de Estados Unidos pertenecientes al Partido Republicano (observar los valores de *Cohen's d* negativos en cada experimento). Estos resultados son similares a resultados previos que documentaron sesgo ideológico a favor del Partido Demócrata en el contenido de Wikipedia en inglés.[12]

Distribución de asociaciones en artículos de Wikipedia-español entre nombres de presidentes de EEUU y un léxico de sentimiento con palabras positivas y negativas (N=2066)

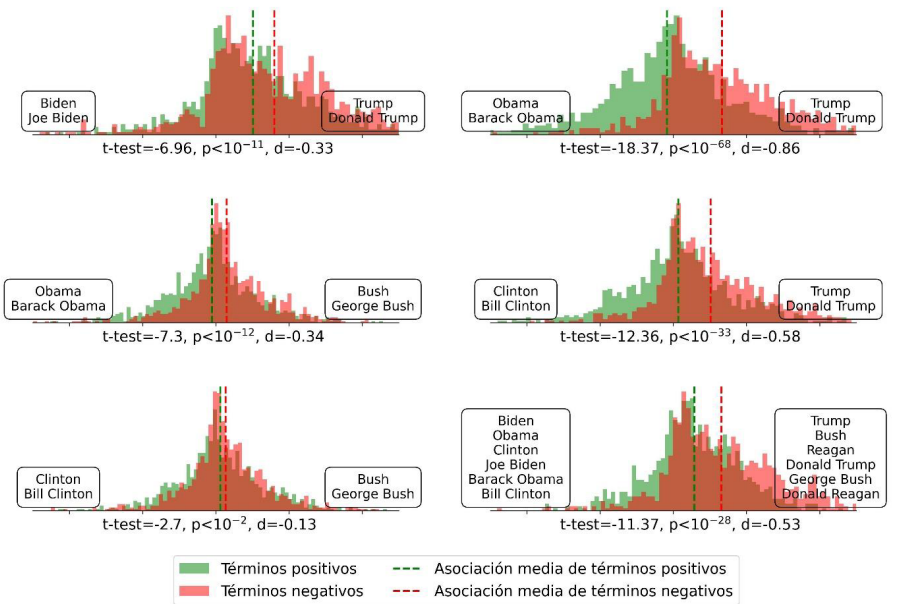


Figura 2. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y nombres de presidentes de Estados Unidos del Partido Demócrata (polo izquierdo de cada eje

12 S. Greenstein and F. Zhu, "Do experts or crowd-based models produce more bias? evidence from encyclopedia britannica and wikipedia," MIS Q., vol. 42, no. 3, pp. 945–960, Sep. 2018, doi: 10.25300/MISQ/2018/14084.

horizontal) y del Partido Republicano (polo derecho de cada eje horizontal).

En nuestro siguiente análisis utilizamos varios términos que denotan orientación ideológica. Los resultados en esta ocasión son heterogéneos. En el contenido de Wikipedia español, las palabras con connotaciones positivas tienden a asociarse más a menudo con términos tales como *izquierda* o *progresismo* que con términos tales como *derecha* o *conservadurismo*. Sin embargo, las palabras positivas tienden a asociarse más a menudo con términos que denotan *centro derecha* qué *centro izquierda*. En nuestro análisis no hemos encontrado diferencias significativas en las asociaciones de sentimiento entre términos tales como *socialismo* y *capitalismo*. Tampoco existen diferencias sustanciales en el contenido de Wikipedia-español a la hora de asociar palabras positivas o negativas con palabras que denotan extremismo político tales como *extrema izquierda o comunismo y extrema derecha o fascismo*.

Distribución de asociaciones en artículos de Wikipedia-español entre términos que denotan orientación política y un léxico de sentimiento con palabras positivas y negativas (N=2066)

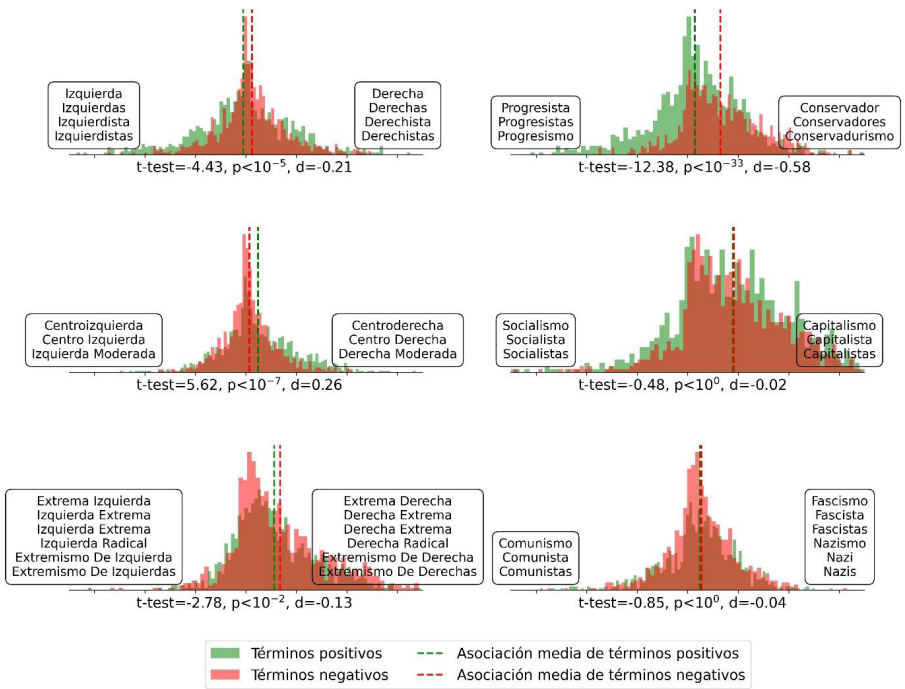


Figura 3. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y términos que denotan orientación ideológica de izquierdas (polo izquierdo de cada eje horizontal)

y de derechas (polo derecho de cada eje horizontal).

Un análisis de los términos que referencian a los principales partidos políticos en España muestra una tendencia en los artículos de Wikipedia-español a asociar palabras positivas con los partidos a la izquierda del centro político tales como PSOE, Izquierda Unida y Podemos sobre partidos a la derecha del centro político, tales como el PP o Vox. Incluso un partido político de izquierdas asociado con la defensa y justificación del terrorismo durante muchos años, Bildu, muestra asociaciones de sentimiento muy similares a las de un partido de derechas democrático que nunca ha justificado la violencia como Vox.

Distribución de asociaciones en artículos de Wikipedia-español entre nombres de partidos políticos y un léxico de sentimiento con palabras positivas y negativas (N=2066)

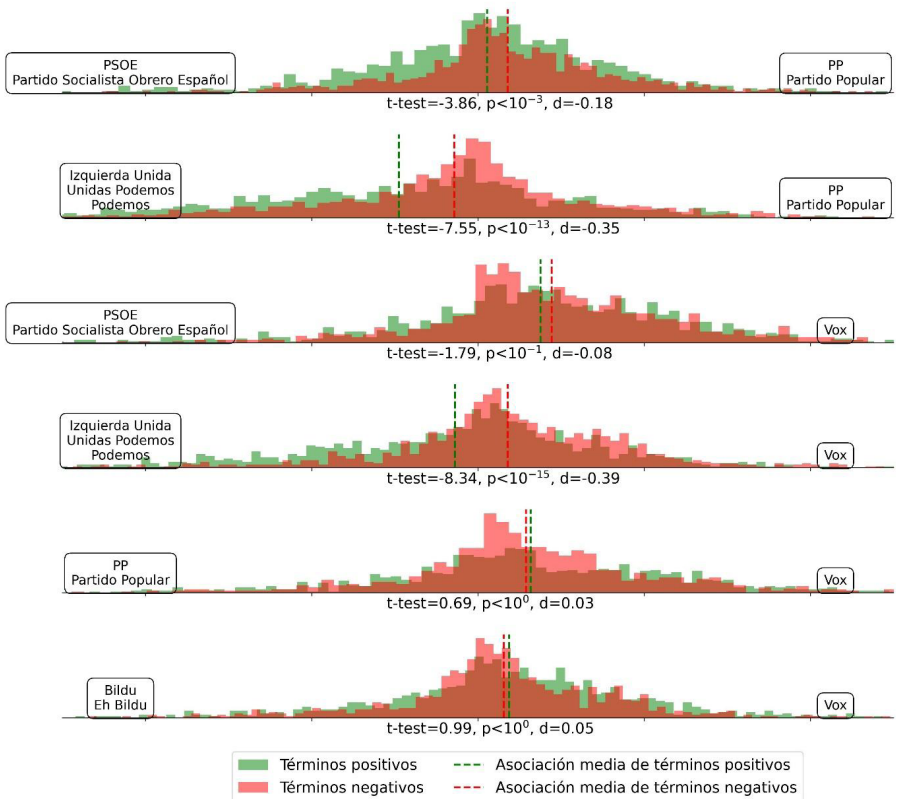


Figura 4. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y de los principales partidos políticos españoles de izquierdas (polo izquierdo de cada eje horizontal)

y de derechas (polo derecho de cada eje horizontal).

Un análisis de todos los presidentes y vicepresidentes de Gobierno de España pertenecientes al PSOE y al PP también muestra una tendencia en los artículos de Wikipedia-español a asociar más a menudo palabras negativas con los presidentes y vicepresidentes del PP y palabras positivas con los presidentes y vicepresidentes del PSOE.

Distribución de asociaciones en artículos de Wikipedia-español entre nombres de figuras políticas y un léxico de sentimiento con palabras positivas y negativas (N=2066)

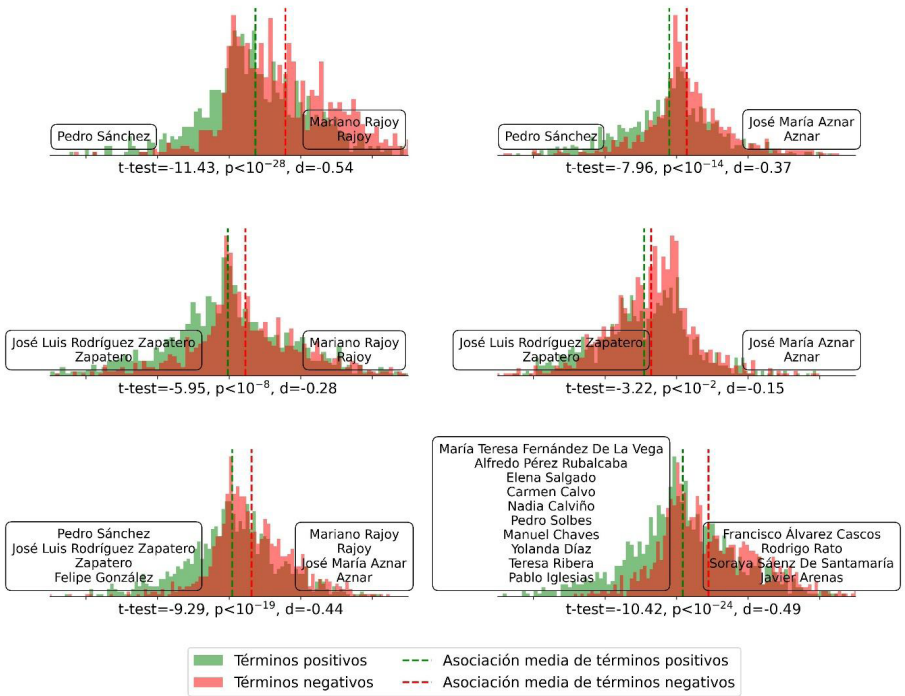


Figura 5. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y nombres de presidentes y vicepresidentes del gobierno de España pertenecientes al PSOE y Podemos (polo izquierdo de cada eje horizontal) y al PP (polo derecho de cada eje horizontal).

Un análisis comparativo entre varios líderes políticos españoles y el líder de Vox, Santiago Abascal, muestra una tendencia en los artículos de Wikipedia-español a asociar palabras positivas con Pedro Sánchez y Alberto Núñez Feijóo con respecto a Santiago Abascal, pero una ausencia de sesgo cuando se compara a Santiago Abascal con Isabel Díaz Ayuso, Pablo Iglesias o Yolanda Díaz.

En una comparación puntual entre Alberto Núñez Feijóo e Isabel Díaz Ayuso, el sesgo político en el contenido de Wikipedia es favorable a Alberto Núñez Feijóo.

Distribución de asociaciones en artículos de Wikipedia-español entre nombres de figuras políticas y un léxico de sentimiento con palabras positivas y negativas (N=2066)

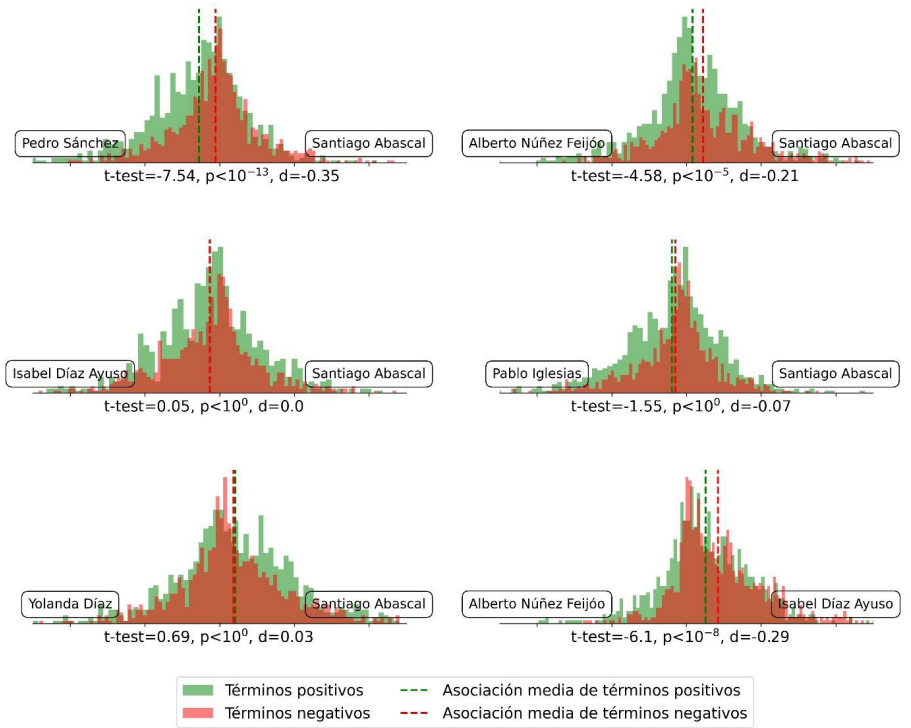


Figura 6. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y nombres de líderes políticos españoles pertenecientes a partidos políticos de izquierdas (polo izquierdo de cada eje horizontal) y de derechas (polo derecho de cada eje horizontal).

La siguiente figura visualiza las asociaciones preferenciales en el contenido de Wikipedia-español entre términos positivos y negativos y los presidentes de las 10 comunidades autónomas más pobladas de España, que han tenido presidentes pertenecientes al PP o al PSOE desde el inicio de la democracia. El análisis muestra una clara tendencia en el contenido de Wikipedia-español a asociar más a menudo palabras negativas con los presidentes de comunidades autónomas pertenecientes al PP y palabras positivas con los presidentes de

comunidades autónomas pertenecientes al PSOE.

Distribución de asociaciones en artículos de Wikipedia-español entre presidentes de comunidades autónomas del PSOE y del PP y un léxico de sentimiento con palabras positivas y negativas (N=2066)

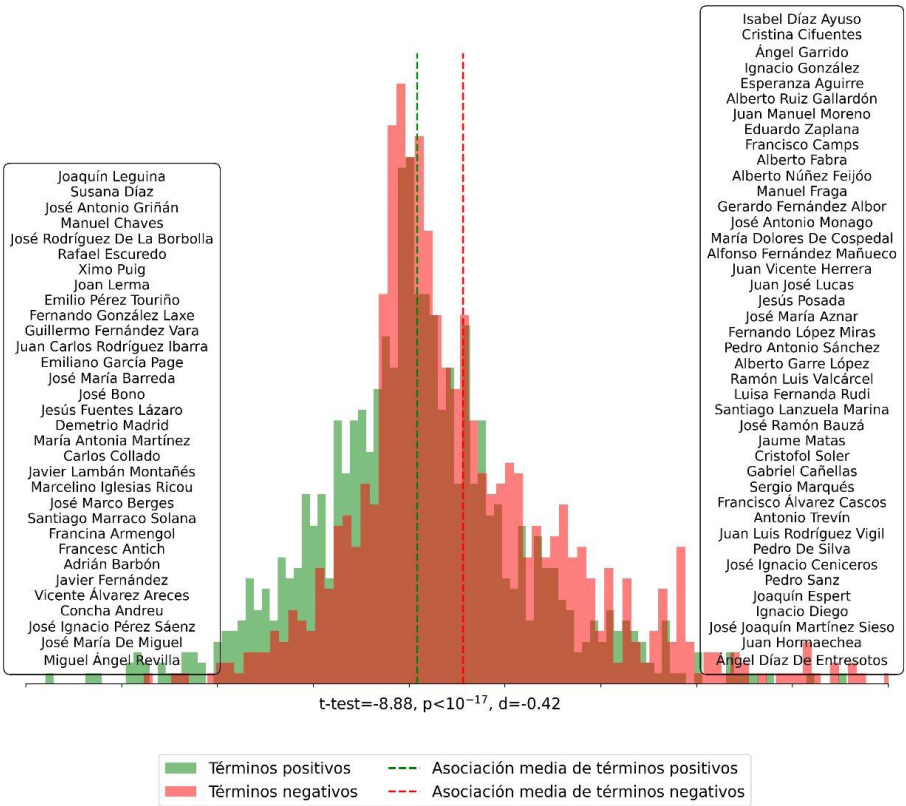


Figura 7. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y nombres de presidentes autonómicos españoles pertenecientes al PSOE (polo izquierdo de cada eje horizontal) y al PP (polo derecho de cada eje horizontal).

Es importante señalar que el análisis anterior agrega a un amplio número de presidentes de comunidades autónomas. Un análisis más granular de comunidades específicas muestra un cierto grado de heterogeneidad de asociaciones. En la mayoría de las comunidades autónomas, existe un sesgo político favorable al PSOE en los artículos de Wikipedia en español. Sin embargo, en el análisis de los presidentes de la Junta de Andalucía, la mayoría de las asociaciones

negativas en el contenido de Wikipedia son con los presidentes de la Junta de Andalucía pertenecientes al PSOE.

Distribución de asociaciones en artículos de Wikipedia-español entre presidentes de CCAA y un léxico de sentimiento con palabras positivas y negativas (N=2066)

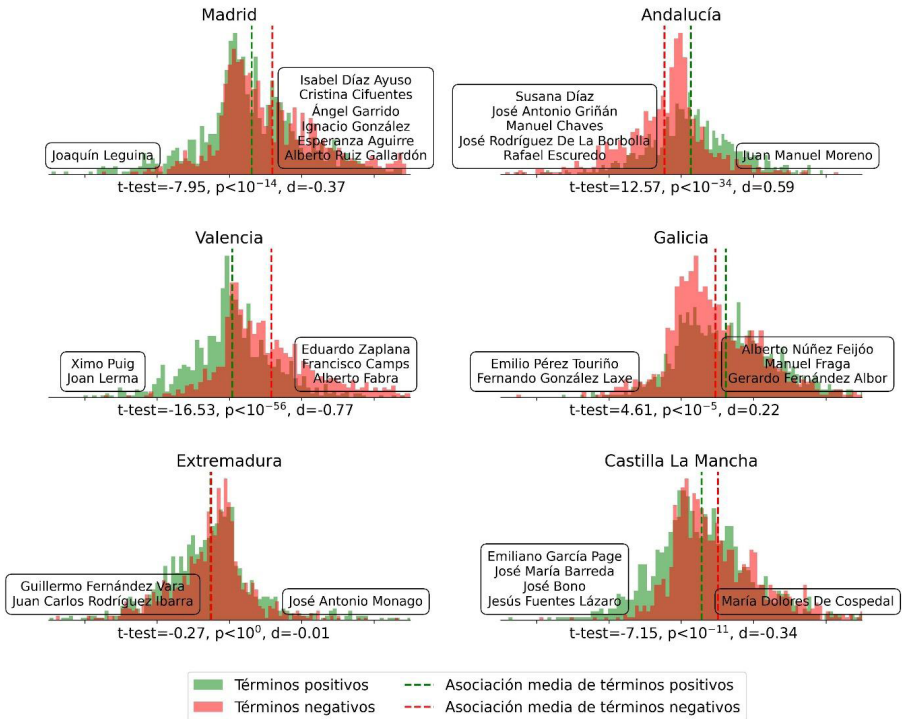


Figura 8. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y nombres de presidentes autonómicos españoles pertenecientes al PSOE (polo izquierdo de cada eje horizontal) y al PP (polo derecho de cada eje horizontal).

Un análisis adicional de 50 políticos influyentes listados por la Fundación Marques de Oliva[13] y ordenados por su pertenencia a la izquierda o a la derecha del centro político y excluyendo a políticos de partidos de centro y nacionalistas (Ciudadanos, UPyD, PNV, CIU, ERC, etc), también muestra un

13 "Los 50 políticos más influyentes en España en 2020 – Fundación Marqués de Oliva." https://fundacionmarquesdeoliva.com/estudio-de-los-500-espanoles-mas-influyentes-de-2020/los-50-politicos-mas-influyentes-en-espana-en-2020/ (accessed Mar. 19, 2023).

sesgo similar al que hemos documentado en las figuras anteriores y que es favorable al PSOE.

Distribución de asociaciones en artículos de Wikipedia-español entre personalidades políticas del PP y PSOE y un léxico de sentimiento con palabras positivas y negativas (N=2066)

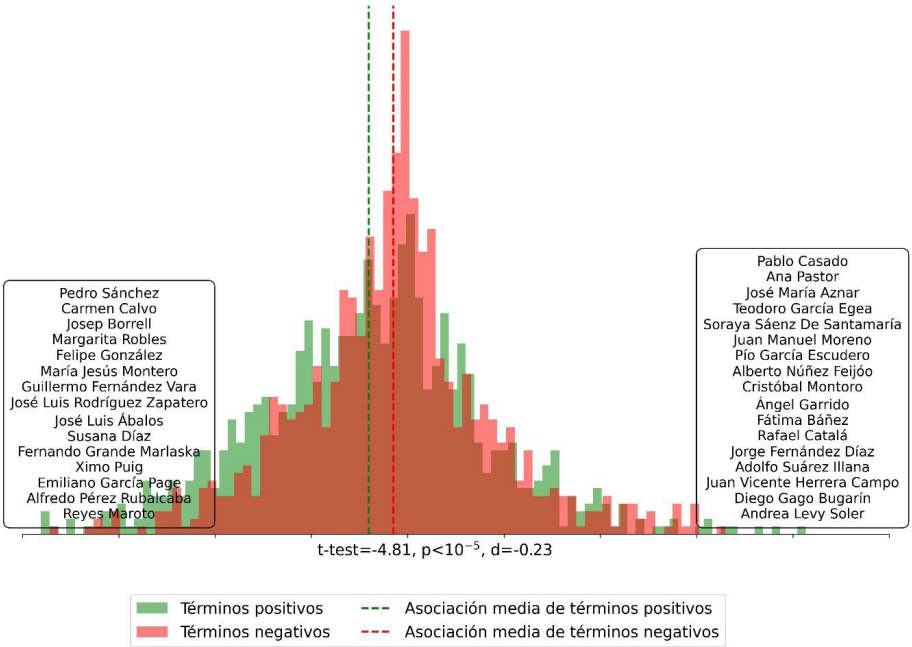


Figura 9. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y nombres de líderes políticos españoles a la izquierda del centro político (polo izquierdo de cada eje horizontal) y a la derecha del centro político (polo derecho de cada eje horizontal).

Una comparación entre algunas de las principales figuras políticas dentro de los partidos políticos nacionales en España muestra un favoritismo en el contenido de Wikipedia español favorable al PSOE (pero no a Podemos), con respecto al PP y a Vox.

Distribución de asociaciones en artículos de Wikipedia-español entre nombres de figuras políticas y un léxico de sentimiento con palabras positivas y negativas (N=2066)

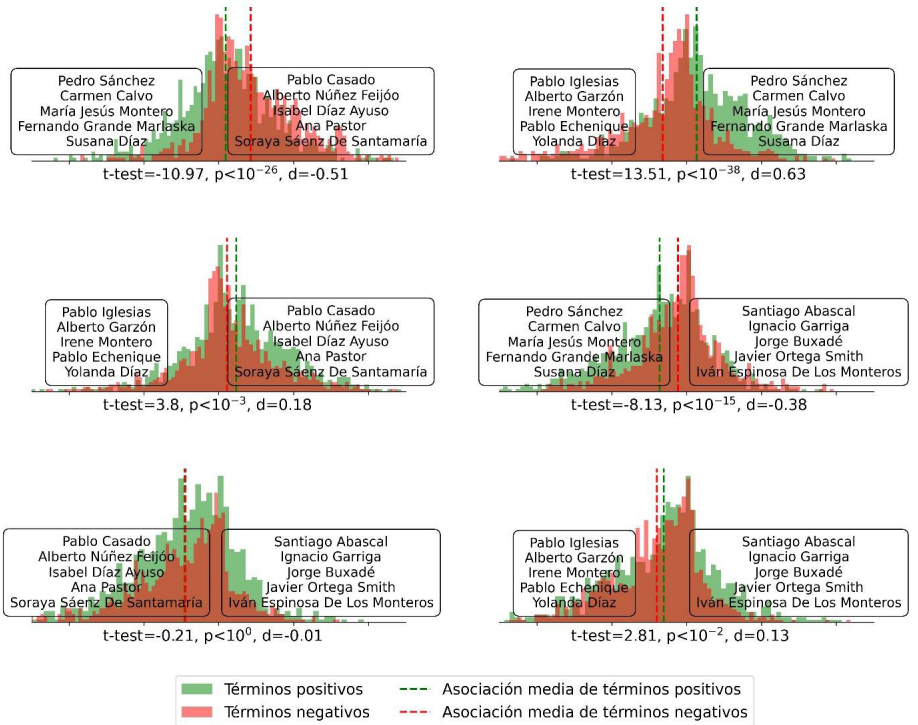


Figura 10. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y nombres de líderes políticos españoles pertenecientes a partidos de izquierdas (polo izquierdo de cada eje horizontal) y de derechas (polo derecho de cada eje horizontal).

Un breve análisis de líderes políticos latinoamericanos revela un sesgo en Wikipedia a favor de políticos a la izquierda del centro político en Colombia y Brasil. En el caso de Venezuela, las asociaciones positivas tienden a realizarse con la oposición venezolana. Los artículos de Wikipedia también muestran favoritismo por un dictador de extrema izquierda, Fidel Castro, sobre un dictador de extrema derecha, Augusto Pinochet.

Distribución de asociaciones en artículos de Wikipedia-español entre nombres de figuras políticas latinoamericanas y un léxico de sentimiento con palabras positivas y negativas (N=2066)

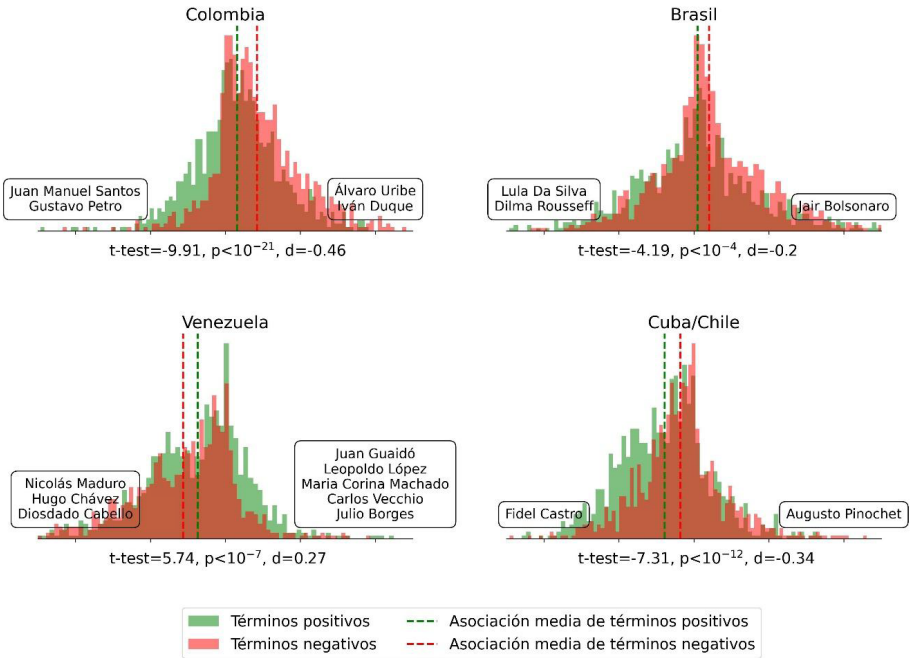


Figura 11. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y nombres de líderes políticos latinoamericanos a la izquierda del centro ideológico (polo izquierdo de cada eje horizontal) y a la derecha del centro ideológico (polo derecho de cada eje horizontal).

Para concluir nuestro análisis, llevamos a cabo un breve estudio de asociaciones preferenciales en el contenido de Wikipedia español entre términos positivos y negativos, y una serie de medios de comunicación y periodistas españoles considerados a la izquierda y a la derecha del centro político. Los resultados de dicho análisis también muestran un sesgo político a favor de la izquierda en el contenido de los artículos de Wikipedia-español. Esto sugiere que el sesgo político en Wikipedia-español no se circunscribe tan solo a términos explícitamente políticos, sino que se extiende también a otros ámbitos o figuras públicas con alineaciones políticas implícitas.

Distribución de asociaciones en artículos de Wikipedia-español entre medios y periodistas por orientación ideológica y un léxico de sentimiento con palabras positivas y negativas (N=2066)

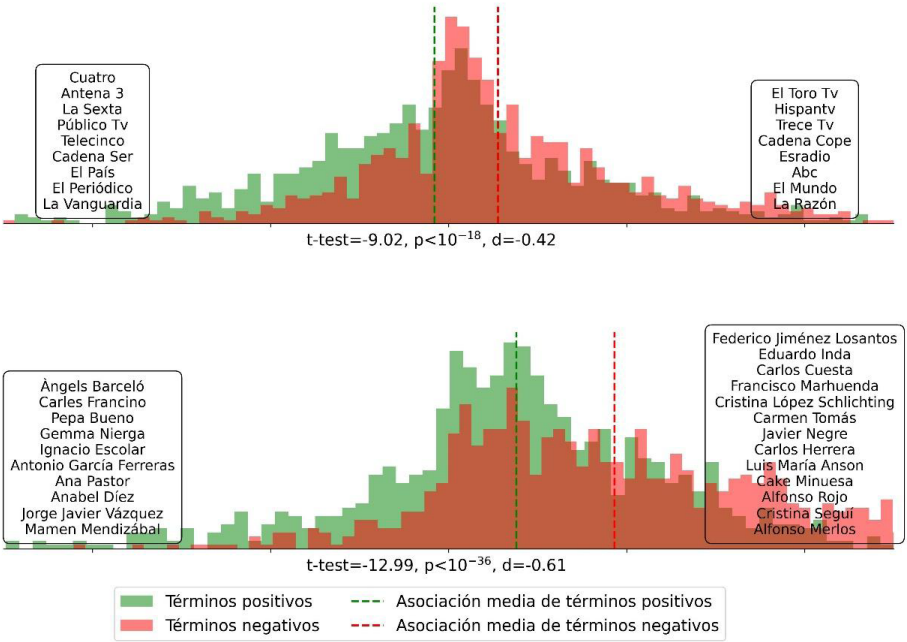


Figura 12. Distribución de asociaciones en el contenido de Wikipedia español entre palabras clasificadas como positivas (en verde) o negativas (en rojo) por el léxico de sentimiento AFINN y medios de comunicación y periodistas alineados ideológicamente a la izquierda del centro político (polo izquierdo de cada eje horizontal) y a la derecha (polo derecho de cada eje horizontal).

2. Conclusiones

Este estudio ha puesto de manifiesto la presencia de un sesgo político favorable a la izquierda ideológica en el contenido de Wikipedia en español. Sin embargo, es importante subrayar que la mayoría de los tamaños de efecto obtenidos en el análisis abarcan un rango de leves a moderados.

A pesar de los supuestos esfuerzos de la plataforma Wikipedia por mantener la neutralidad en su contenido, nuestro análisis ha identificado ahí tendencias que favorecen a políticos, partidos políticos e ideologías a la izquierda del centro político y desfavorecen a partidos políticos, ideologías o políticos alineados con la derecha ideológica, lo que pone en tela de juicio la imparcialidad de Wikipedia-español.

Es necesario enfatizar que el sesgo ideológico que hemos documentado se manifiesta primordialmente en favoritismo hacia el PSOE, sus líderes políticos y la izquierda ideológica convencional y no se extiende a manifestaciones más extremistas de la izquierda, como el partido político Podemos, sus líderes políticos o ideologías de extrema izquierda como el comunismo.

El efecto que ejerce sobre los usuarios hispanohablantes el sesgo político favorable a la izquierda ideológica, en el contenido de Wikipedia en español, debería ser analizado en profundidad en investigaciones futuras.

Es fundamental que Wikipedia y sus colaboradores trabajen proactivamente para neutralizar la existencia de sesgos políticos en su contenido, garantizando así que la enciclopedia en línea más consultada del mundo sea un espacio de conocimiento objetivo y confiable. La transparencia, el compromiso con la diversidad de puntos de vista y la educación en materia de sesgos cognitivos son herramientas esenciales para abordar este desafío y proteger la integridad de la información contenida en Wikipedia.

3. Apéndice – Metodología extendida

Descripción general

Los métodos para analizar el contenido de textos mediante codificadores humanos, como el análisis interpretativo de textos o la codificación cualitativa, han sido útiles en el análisis sociológico y cultural de cantidades de texto relativamente pequeñas. Sin embargo, estos métodos están limitados por su incapacidad para escalar a grandes volúmenes de texto y por la baja consistencia de anotaciones entre evaluadores al examinar temas con connotaciones sutiles. Las técnicas computacionales de análisis de texto, como el análisis de redes semánticas y la modelización de tópicos, evitan algunas de las limitaciones impuestas por el uso de evaluadores humanos en el análisis de contenido textual.[14]

Los algoritmos de embeddings de palabras son una técnica de modelado de lenguaje relativamente reciente que captura, en representaciones vectoriales, las asociaciones semánticas prevalentes dentro de un gran volumen de texto.[15] Se ha demostrado previamente que los espacios de embeddings destilan en su configuración geométrica cantidades sustanciales de información factual y semántica contenida en los textos utilizados para entrenar dichos embeddings.[16] Varios estudios previos han demostrado que los embeddings de pa-

14 A. C. Kozlowski, M. Taddy, and J. A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," Am Sociol Rev, vol. 84, no. 5, pp. 905–949, Oct. 2019, doi: 10.1177/0003122419877135.

15 T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119. Accessed: May 11, 2017. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

16 T. Mikolov, W. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations," in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, 2013, pp. 746–751. Accessed: Mar. 19, 2019. [Online]. Available: http://aclweb.org/anthology/N13-1090 A. C. Kozlowski, M. Taddy, and J. A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class throu-

labras también absorben muchos de los sesgos predominantes dentro de un sistema cultural, como resultado de que dichos modelos de embeddings son entrenados en artefactos lingüísticos culturales, tales como repositorios de libros contemporáneos, que presumiblemente contienen sesgos culturales.[17]

Varios estudios previos han utilizado embeddings de palabras entrenados en los Ngramas de libros de Google, un corpus diacrónico de libros que abarca un período de 100 años, para rastrear la evolución de un conjunto reducido de asociaciones y estereotipos culturales en torno a las dimensiones de género, etnia[18] y clase social [19]. La metodología utilizada en dichos trabajos consiste en realizar un seguimiento de la evolución temporal de las palabras en los espacios vectoriales de embeddings derivados de dicho corpus. También se ha observado que las asociaciones descubiertas en embeddings de palabras diacrónicas se correlacionan significativamente con las asociaciones conscientes e inconscientes expresadas por individuos tanto en encuestas históricas como contemporáneas.[20]

---

gh Word Embeddings," Am Sociol Rev, vol. 84, no. 5, pp. 905–949, Oct. 2019, doi: 10.1177/0003122419877135. D. Rozado, "Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types," PLOS ONE, vol. 15, no. 4, p. e0231189, Apr. 2020, doi: 10.1371/journal.pone.0231189. Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), E3635-E3644. D. Rozado and M. al-Gharbi, "Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets," J Comput Soc Sc, 2021, doi: 10.1007/s42001-021-00130-y.

17 A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," Science, vol. 356, no. 6334, pp. 183–186, Apr. 2017, doi: 10.1126/science.aal4230.

18 N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," PNAS, vol. 115, no. 16, pp. E3635–E3644, Apr. 2018, doi: 10.1073/pnas.1720347115.

19 A. C. Kozlowski, M. Taddy, and J. A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," Am Sociol Rev, vol. 84, no. 5, pp. 905–949, Oct. 2019, doi: 10.1177/0003122419877135.

20 A. C. Kozlowski, M. Taddy, and J. A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," Am Sociol Rev, vol. 84, no. 5, pp. 905–949, Oct. 2019, doi: 10.1177/0003122419877135.

En resumen, está ampliamente demostrado en la literatura académica la capacidad de la técnica de embeddings de palabras derivadas de un corpus de texto para capturar asociaciones latentes en dichos textos, similares a las asociaciones conscientes e inconscientes que muestran los seres humanos en las culturas. Por lo tanto, un modelo de embeddings de palabras, entrenado en un corpus específico dentro de un sistema cultural dado, puede servir como un indicador útil para elucidar las asociaciones idiosincrasia utilizadas por el grupo que produjo el corpus de texto.

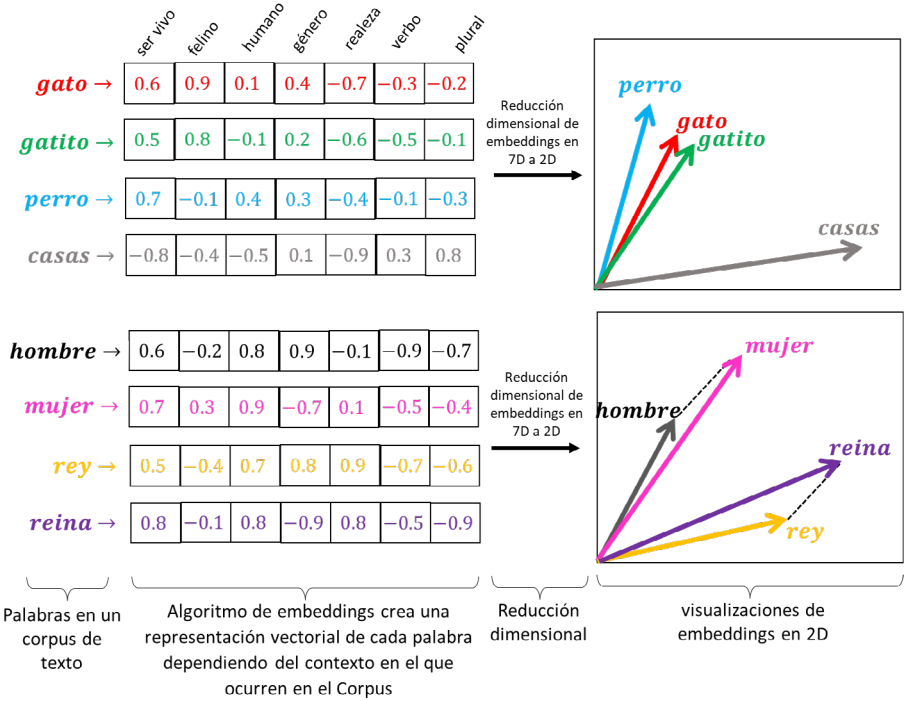
Descripción técnica

Los embeddings de palabras (también conocidos como vectores de palabras) son representaciones numéricas de los roles sintácticos y semánticos de las palabras en un corpus de texto. El éxito de los embeddings de palabras en el modelado del lenguaje se debe a su capacidad para mapear la coocurrencia estadística de las palabras y sus contextos en un corpus de texto a posiciones en el espacio vectorial que representan los significados con los que se usan las palabras en el corpus. Tras entrenar un modelo de embeddings de palabras en un gran volumen de texto, el espacio de embeddings resultante contiene una estructura espacial semánticamente significativa, de tal manera que las palabras con significados similares, como *gato* y *gatito*, tienden a ubicarse en regiones adyacentes del espacio vectorial.

La cercanía en el espacio vectorial no solo captura la similitud semántica, sino también las asociaciones entre pares de palabras. Es decir, las palabras que tienden a coocurrir en contextos similares, incluso si no son intercambiables semánticamente, también ocupan regiones cercanas en el espacio vectorial. Por ejemplo, pares de palabras relacionadas como *coche* y *combustible*, aunque no son intercambiables semánticamente, tienden a ser adyacentes en los modelos de embeddings debido a su coocurrencia contextual en textos escritos. Esta característica ha demostrado ser útil para el análisis sociológico de las asociaciones culturales contenidas en grandes corpus de texto. La siguiente figura proporciona una visión conceptual de los embeddings de palabras.

En este trabajo derivamos 5 modelos de embeddings del contenido de Wikipedia en español (eswiki-20220501-pages-articles-multistream). La razón de utilizar varios modelos es que cada estimación de un modelo de embeddings se inicia con parámetros estocásticos y es, por lo tanto, no determinista. Utilizando la media de asociaciones en varios modelos entrenados en el mismo

Corpus (Wikipedia-español) se reduce el ruido estadístico en la estimación. Para generar cada modelo de embeddings de palabras, se utilizó la implementación de word2vec en la librería gensim. La arquitectura de bolsa de palabras continuas (CBOW) tuvo un rendimiento ligeramente mejor que la arquitectura Skip-Gram en las métricas de validación comúnmente utilizadas para evaluar



la calidad de los modelos de embeddings, por lo que esta fue la versión del algoritmo utilizada en todos nuestros análisis posteriores.

Para entrenar el modelo de embeddings de palabras, se utilizaron los siguientes parámetros: dimensiones de los vectores=300, tamaño de ventana contextual=10, muestreo negativo=10, muestreo de palabras frecuentes reducidas=0.001, frecuencia mínima de 10 (solo se incluyen en el vocabulario del modelo de embeddings términos que aparecen más de 10 veces en el corpus), número de iteraciones de entrenamiento (épocas) sobre él corpus=5. El exponente utilizado para dar forma a la distribución de muestreo negativo fue el valor predeterminado de 0.75.

La precisión de los modelos de embedding entrenados en el contenido de

Wikipedia español fue comparable a varios modelos de embedding populares entrenados en corpora como Twitter o Google Books en tareas de similitud, asociación y analogía de palabras.

En este informe, la cuantificación de sesgo político en el contenido de Wikipedia se lleva a cabo en el espacio vectorial de los embeddings derivados del contenido de Wikipedia-español. Es decir, cada término/palabra en nuestro análisis está representada por un vector numérico (embedding) derivado del uso de dicho término en el contenido de Wikipedia-español.

Proyecciones de palabras en ejes culturales y relaciones factuales en el mundo empírico

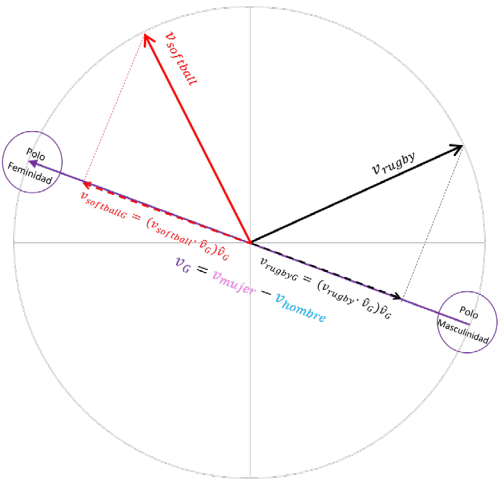
Como se ha demostrado en trabajos previos, se pueden derivar dimensiones semánticamente significativas del espacio vectorial de modelos de embeddings usando operaciones de álgebra vectorial. Esta técnica explota la estructura semántica inherente en los espacios de embeddings para extraer espectros culturales tales como la etnia, el género o el estatus socioeconómico. Por ejemplo, la diferencia vectorial correspondiente a la resta de la representación vectorial de la palabra *hombre* de la representación vectorial de la palabra *mujer* en el espacio de embeddings. El resultado de esta resta, , puede interpretarse como un eje cultural/demográfico (es decir, un vector) en el espacio de embeddings que apunta desde la masculinidad hacia la feminidad.

Varios trabajos previos[21] han demostrado que las regularidades estructurales culturalmente significativas en el espacio vectorial pueden usarse para resolver tareas de razonamiento analógico como *'hombre es a mujer como rey es a...?'* La resolución de esta analogía en el espacio vectorial se logra realizando la operación de álgebra vectorial . Es decir, agregando el vector de género que apunta desde la masculinidad hacia la feminidad al vector representativo de la palabra *rey*, que da como resultado una ubicación en el espacio vectorial muy cercana a la solución de la analogía, la representación vectorial de la palabra *reina*, . Por lo tanto, el significado semántico en los modelos de embeddings de palabras no solo está contenido en la dirección de vectores de palabras específicas, sino también en la dirección de nuevos vectores resultantes de operaciones algebraicas entre vectores representativos de las palabras existentes en el vocabulario del corpus donde fue entrenado el modelo de embeddings.

Varios trabajos previos han demostrado también que las connotaciones culturales de una palabra en un corpus pueden estimarse calculando la proyección ortogonal de su representación vectorial sobre dimensiones culturales

21 T. Mikolov, W. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations," in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, 2013, pp. 746–751. Accessed: Mar. 19, 2019. [Online]. Available: http://aclweb.org/anthology/N13-1090

de interés. Es decir, el coseno del ángulo entre el vector representativo de la palabra *rugby* () y un vector/eje cultural como el género , proporciona una estimación de la asociación en el corpus de la palabra *rugby* con los polos del eje cultural: *hombres* y *mujeres*. Para la interpretación apropiada de la siguiente figura, se notifica al lector de qué en el deporte universitario de Estados Unidos el softball es un deporte practicado principalmente por mujeres, de ahí que la proyección del vector representativo de la palabra *softball*, , esté cercana al polo femenino del eje de género .



En la literatura académica se han utilizado varios métodos para medir la frecuencia de las asociaciones contenidas en el espacio geométrico de las palabras en los modelos de embeddings, pero todas ellas tienden a generar resultados similares ya que suelen ser algebraicamente similares. Aquí, siguiendo la metodología pionera de trabajos previos[22], creamos ejes que denotan categorías políticas como izquierda–derecha o Rajoy–Zapatero en modelos de

22 A. C. Kozlowski, M. Taddy, and J. A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," Am Sociol Rev, vol. 84, no. 5, pp. 905–949, Oct. 2019, doi: 10.1177/0003122419877135.

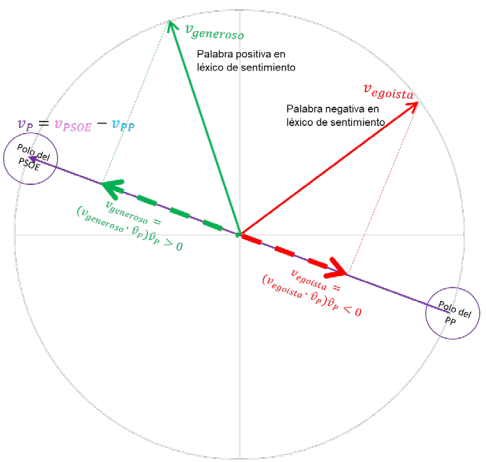
Rozado, D. (2020). Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. PloS one, 15(4), e0231189.

Rozado, D., & Al-Gharbi, M. (2021). Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets. Journal of Computational Social Science, 1-22.

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

embeddings de palabras. Es decir, construimos ejes culturales que capturan el espectro de orientación ideológica con términos como *conservadores* o *derechas* formando el polo conservador del eje, y términos como *progresistas* o *izquierdas* formando el polo izquierdista del eje. De forma similar se puede crear un eje entre presidentes del Gobierno de España como Rajoy y Aznar en el polo conservador y Zapatero, Pedro Sánchez o Felipe González en el eje representativo de la izquierda.

A continuación, proyectamos sobre esos ejes con connotaciones políticas palabras clasificadas como positivas o negativas en un léxico de sentimiento. Si existe una diferencia significativa entre las proyecciones medias de las palabras positivas y negativas en el eje de orientación política, esto indica una preferencia en el corpus a asociar palabras positivas/negativas con cada polo opuesto del eje de orientación política.



Para validar que la metodología utilizada en este trabajo captura muchas de las asociaciones de sentimiento intuitivas mantenidas por los humanos, en la sección de Metodología del informe creamos cuatro ejes culturales ilustrativos en el modelo de embeddings de palabras, derivado de los artículos de Wikipedia-español. Los cuatro ejes culturales son Vida–Muerte, Paz–Guerra, Salud–Enfermedad y Democracia–Dictadura. Proyectando un léxico de sentimiento con palabras positivas y negativas sobre cualquiera de estos ejes, se producen proyecciones que tienden a asociar las palabras clasificadas como de sentimiento positivo con la *vida*, la *paz*, la *salud* y la *democracia* y las palabras clasificadas como negativas con la *muerte*, la *guerra*, la *enfermedad* y la *dictadura*.

Estos resultados sugieren que los modelos de embeddings de palabras entrenados en el corpus de artículos de Wikipedia-español capturan muchas de las asociaciones intuitivas de sentimientos positivos y negativos que son prevalentes en el contexto cultural. Por lo tanto, es razonable utilizar esta técnica para medir las asociaciones de sentimiento prevalentes con respecto a la orientación política en modelos de embeddings derivados del contenido textual de los artículos de Wikipedia.

DISENSO
FUNDACIÓN

Actividad subvencionada por el Ministerio de Cultura

fundaciondisenso.org