Sesgos algorítmicos en las redes sociales

ANDRÉS ABELIUK

Profesor Asistente del Departamento de Ciencias de la Computación de la Universidad de Chile e investigador del Centro Nacional de Inteligencia Artificial (CENIA). Ph.D en Ciencias de la Computación por la Universidad de Melbourne, Australia. Líneas de investigación: computación social e inteligencia colectiva, análisis de redes sociales e impacto de la inteligencia artificial en la sociedad.

aabeliuk@dcc.uchile.cl

RESUMEN. Los sistemas de recomendación en las redes sociales pueden mejorar la toma de decisiones y reducir la carga cognitiva, pero también pueden reforzar los sesgos existentes y distorsionar las percepciones de las personas. Los orígenes de los sesgos pueden ser tanto algorítmicos como humanos, y pueden interactuar entre sí amplificándose mutuamente. Por lo tanto, es importante comprender cómo los humanos interactúan con los algoritmos y cómo estos sistemas pueden influir en nuestras percepciones y decisiones en línea.
Este artículo examina investigaciones recientes sobre sesgos en redes sociales, destacando dos estudios realizados por el autor y sus colegas del centro USC Information Sciences Institute. El primer estudio explora cómo las decisiones de a quién seguir en las redes sociales pueden distorsionar la forma en que las personas se comparan con su círculo cercano, haciendo que un rasgo parezca más popular a nivel local entre los amigos de lo que realmente es a nivel global. El segundo estudio analiza el impacto de los algoritmos de selección de contenido en Twitter, indicando que los sistemas de recomendación pueden acentuar la desigualdad de atención, amplificando desproporcionadamente la voz de unos pocos usuarios.

Sistemas de recomendación en redes sociales

Las plataformas de medios sociales han reducido las barreras para publicar, permitiendo que cada vez más personas compartan información en línea y participen en el discurso público. Observamos a nuestros pares para aprender normas sociales, evaluar riesgos e informarnos de diversos temas. Sin embargo, estas observaciones pueden estar sistemáticamente sesgadas, distorsionando nuestra visión del mundo.

En el contexto digital, los datos reflejan nuestros propios sesgos cognitivos y sociales, y pueden manifestarse de muchas maneras con importantes consecuencias para la forma en que interactuamos y nos comunicamos en línea. Por ejemplo, los sesgos de actividad pueden influir en cómo se difunden las noticias y la información en las redes sociales: durante el 2011 el 2% de los usuarios de Twitter produjeron el 50% de las publicaciones. Si sólo un pequeño porcentaje de usuarios genera la mayoría de los contenidos, esto puede limitar la diversidad de perspectivas y voces que se escuchan en línea. Los sesgos de datos también son una preocupación importante en el ámbito digital. La mayoría de los sitios web están en inglés: más del 50% están en ese idioma, mientras que el porcentaje de angloparlantes en el mundo es aproximadamente del 13%. Esto puede sesgar la forma en que se accede a la información y limitar la exposición a diferentes perspectivas y culturas [1].

La producción constante de contenidos ilimitados provoca una sobrecarga de información en los usuarios, lo que ha llevado a las plataformas de contenido a adoptar algoritmos de recomendación para ayudar a los usuarios a interactuar con el contenido disponible de forma más eficaz. Estos algoritmos están diseñados para organizar la información para los usuarios y recomendarles un subconjunto manejable de contenidos en función de sus preferencias y comportamientos anteriores. Sin embargo, los algoritmos pueden amplificar los sesgos existentes e introducir otros nuevos en el sistema. Por ejemplo, los algoritmos de recomendación pueden amplificar la popularidad de productos en la plataforma presentando a los usuarios contenidos que ya son populares [2], lo que puede llevar a la amplificación de ciertos estereotipos y prejuicios. Por lo tanto, es importante que las empresas que desarrollan estos algoritmos sean conscientes de los posibles sesgos y trabajen para mitigarlos.

Redes sociales como Twitter, Facebook, Instagram y LinkedIn crean un "feed social" o "línea de tiempo" personalizado a partir de los contenidos generados por las personas a las que los usuarios siguen. Esta línea de tiempo sirve como mecanismo algorítmico para curar la exposición de información de un usuario usando algoritmos de selección que destacan determinados contenidos de determinadas personas a las que un individuo sigue. En esencia, los sistemas de recomendación son una forma automatizada de predecir nuestras preferencias basándose en nuestras interacciones con el sistema, por ejemplo, los "me gusta", las películas que vemos, etc. De este modo "aprenden" patrones de preferencias a través de nuestro comportamiento individual y colectivo. Las recomendaciones son el resultado de patrones descubiertos en los datos de muchas personas diferentes que utilizan la misma plataforma de contenidos. A través de la inteligencia colectiva, estos sistemas tienen un enorme potencial para mejorar la toma de decisiones individuales y disminuir la carga cognitiva en la búsqueda de información. Las recomendaciones son el resultado de patrones descubiertos en los datos de muchas y diversas personas que utilizan una misma plataforma de contenidos.

Sesgos en redes sociales

Aunque las recomendaciones algorítmicas son útiles para mitigar la sobrecarga de información, la selección algorítmica de contenidos también presenta inconvenientes. Los algoritmos pueden atrapar a los usuarios en grupos

homogéneos o "burbujas de filtros" [3]. Esto se debe a un proceso de retroalimentación entre el sesgo de confirmación, inherente a la psicología humana, y a la exposición selectiva a la información inducida por la tecnología (ver Figura 1). Al sobreexponer a los usuarios a contenidos que confirman ideas, actitudes y creencias preexistentes, se limita la diversidad de información a la que están expuestos [4]. Por otro lado, la selección algorítmica de contenidos puede utilizarse para exponer a las personas a puntos de vista más diversos. Para mitigar los sesgos y posibles consecuencias negativas en los medios digitales, es importante comprender la interacción y los mecanismos que existen entre los individuos y la información (filtrada algorítmicamente) a la que están expuestos.

Facebook realizó un estudio en el que un grupo de investigadores analizó cómo la red social y los algoritmos influyen en la exposición mediática de los usuarios [5]. Examinaron las interacciones de 10 millones de usuarios estadounidenses con las noticias compartidas en su sección de noticias y cuantificaron la diversidad de contenidos a los que estaban expuestos. Los resultados mostraron que muchos usuarios partidistas estaban expuestos a contenidos ideológicamente transversales, aunque de manera asimétrica (los liberales están menos expuestos que los conservadores a ideas opuestas). Además, se determinó que la selección individual de contenidos por parte de los usuarios influye más en la reducción de la exposición a contenidos opuestos que la selección algorítmica de contenidos.

Una limitación del estudio de Facebook es que se basa en una muestra estática a lo largo del tiempo, lo que no permite evaluar los efectos cumulativos y a largo plazo que los algoritmos de recomendación pueden tener en el comportamiento social (ver Figura 1). Estudios anteriores han demostrado que las recomendaciones algorítmicas, cuando se combinan con decisiones individuales, pueden alterar el ecosistema de la información. Por ejemplo, la introducción de la función de recomendación de amigos "a quiénes seguir" en Twitter aceleró el crecimiento de cuentas ya populares [6]; los resultados de búsqueda en Google para consultas políticas difieren significativamente en función del historial de navegación previo de los usuarios [7]; y las noticias seleccionadas automáticamente por Apple News proceden de fuentes menos diversas que las seleccionadas por humanos [8].

Durante el 2011, el 2% de los usuarios de Twitter produjeron el 50% de las publicaciones.

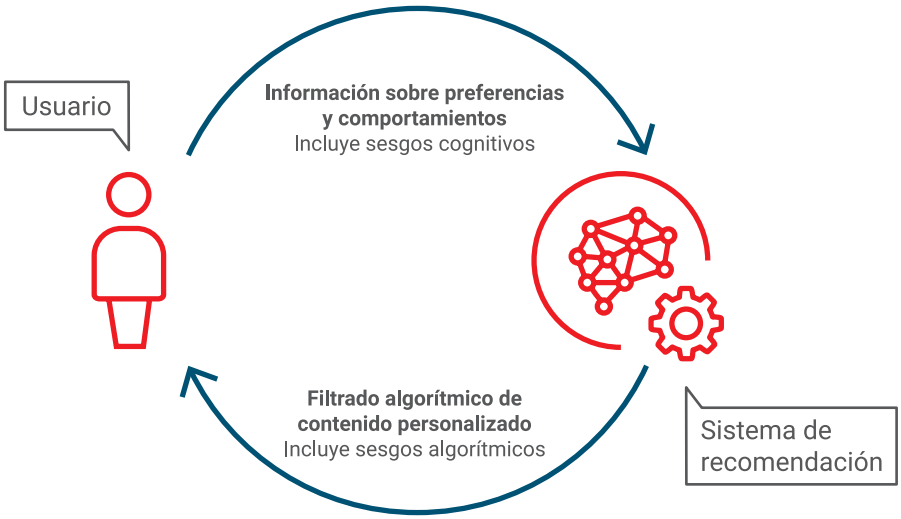
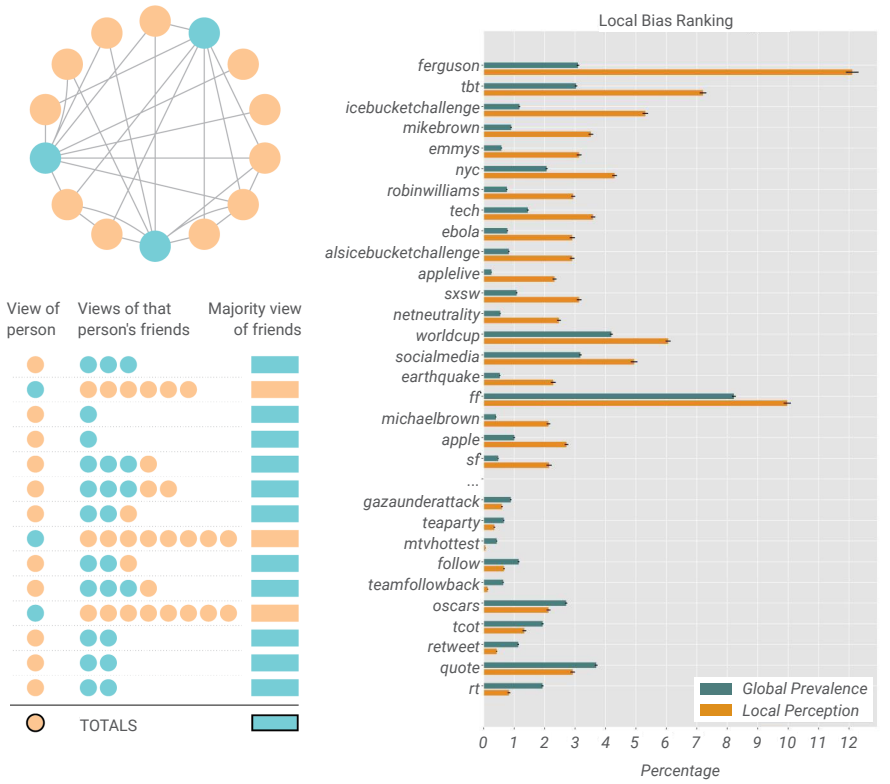


Figura 1. Ciclo de retroalimentación humano-algorítmica en sistemas de recomendación. Los sesgos cognitivos que interactúan con los sistemas de recomendación pueden incluir el sesgo de confirmación, la exposición selectiva, la homofilia, entre otros. Estos sesgos interactúan con los algoritmos de recomendación y pueden contribuir a la creación y amplificación de sesgos en las redes sociales.

La ilusión de la mayoría

Siguiendo la línea de investigaciones anteriores, nuestro trabajo consiste en identificar y caracterizar condiciones que puedan distorsionar las percepciones de las personas en las redes sociales en el corto y largo plazo. Una importante fuente de sesgo en las redes sociales es la "paradoja de la amistad", que afirma que las personas son, en promedio, menos populares que sus amigos. Está afirmación estadística tiene como consecuencia que puede distorsionar la forma en que nos comparamos con nuestro círculo íntimo, ya que cualquier rasgo correlacionado con la popularidad es susceptible de ser percibido erróneamente. Por ejemplo, esto puede explicar por qué los adolescentes suelen sobreestimar el consumo de alcohol o las conductas de riesgo de sus compañeros [9].

La percepción que un individuo tiene sobre la prevalencia de ciertos rasgos,



Fuente: https://www.washingtonpost.com/graphics/business/wonkblog/majority-illusion/.

Figura 2. Izquierda: Ejemplo de una red social donde se observa el fenómeno conocido como "ilusión de la mayoría", en el que la mayoría de los nodos pueden tener la falsa percepción de estar en minoría. **Derecha**: La clasificación de hashtags populares de Twitter basada en sesgo local. En la clasificación se incluyen los 20 superiores y los 10 inferiores. Los hashtags pueden parecer mucho más populares de lo que son en realidad (por ejemplo, #ferguson) o pueden parecer menos populares (por ejemplo, #oscars) debido al sesgo de percepción local.

como el género, la afiliación política o el uso de un hashtag específico, viene determinada por su prevalencia dentro de su círculo social más cercano. En nuestro estudio, hemos identificado una nueva paradoja en las redes sociales dirigidas, donde existe una asimetría entre seguidores y seguidos. Esta paradoja hace que un rasgo parezca más popular localmente entre los amigos de un individuo que globalmente entre todas las personas de la red [10]. En la Figura 2 (izquierda) se muestra una red en la que la mayoría de los nodos tienen una opinión (representada por círculos naranjas), pero tienen una mayoría de amigos con una opinión diferente (representada por círculos azules). En esta red, se puede observar un fenómeno conocido como "ilusión de la mayoría" [11], en el que la mayoría de los nodos pueden tener la falsa percepción de estar en minoría.

Existen dos condiciones que refuerzan este sesgo de percepción: en primer lugar, una correlación positiva entre los atributos de los individuos y su popularidad, y en segundo lugar, una correlación positiva entre los atributos de los individuos y la atención de sus seguidores. La primera condición sugiere que existe un sesgo cuando las personas populares tienen ciertos atributos. La segunda condición indica que la influencia se amplifica cuando estas personas populares son seguidas por "buenos oyentes", es decir, aquellos que siguen a menos personas y, por tanto, pueden prestar más atención a los influyentes. Estas condiciones pueden ser el resultado de sesgos en las preferencias durante la formación de la red, impulsados, por ejemplo, por la homofilia.

En nuestro estudio empírico, confirmamos las afirmaciones anteriores analizando datos de la red social Twitter. En concreto, medimos la popularidad percibida de los hashtags. Los hashtags cumplen múltiples funciones, desde organizar contenidos a expresar opiniones o conectar temas y personas. Calculamos la prevalencia global de un hashtag como la proporción de personas que lo utilizan, y su popularidad percibida como la proporción de amigos que lo utilizan. Descubrimos que algunos hashtags parecían ser mucho más populares de lo que realmente eran debido al sesgo de percepción local. Estos hashtags estaban relacionados con movimientos sociales, memes y acontecimientos de actualidad. Nuestros datos datos se recopilaron en 2014, y algunos de los hashtags más sesgados estaban asociados con el popular "Ice Bucket Challenge" y las protestas que comenzaron en Ferguson (Missouri), tras el tiroteo mortal del afroamericano Michael Brown (ver Figura 2). Es posible que el sesgo de percepción amplificara la difusión de estos temas.

Nuestro trabajo sugiere que una forma de mitigar el sesgo de percepción es cambiar las conexiones de la red para permitir que llegue más información a los usuarios desatentos. Esto abre nuevas vías de investigación para explorar cómo sistemas de recomendación pueden sugerir a quién seguir para mitigar el sesgo de percepción en las redes

Aunque las recomendaciones algorítmicas son útiles para mitigar la sobrecarga de información [...], pueden atrapar a los usuarios en "burbujas de filtros".

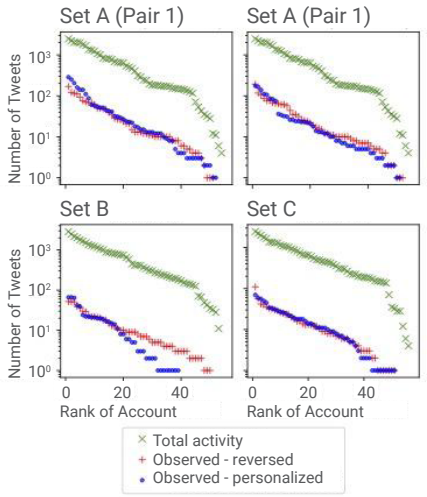


Figura 3: Distribución de frecuencia de rango de los tweets observados y actividad real por conjunto. El verde es la actividad real del usuario, el rojo es la actividad observada del usuario en la condición invertida y el azul es la actividad observada del usuario en la condición personalizada. Cada panel representa lo observado por par de bots.

sociales. Adicionalmente, proponemos un algoritmo que explota la paradoja de la amistad en las redes sociales para estimar la prevalencia real de un atributo con menos error que otros métodos. Esencialmente, la idea que subyace al algoritmo es que las percepciones de seguidores aleatorios deberían tener menos varianza en comparación con las percepciones de individuos aleatorios, porque los seguidores aleatorios están más informados que las personas aleatorias, ya que según la paradoja de la amistad tienden a tener más amigos.

La selección algorítmica amplifica la desigualdad

En un estudio posterior, utilizamos una metodología para medir el impacto de la personalización algorítmica en la exposición de contenidos en las redes sociales [12]. En nuestros experimentos, medimos el impacto de los algoritmos de selección de Twitter en el contenido que los usuarios ven, evaluando las diferencias entre la línea de tiempo personalizada y la línea de tiempo cronológica inversa, también conocidas como "For You" y "Following" en Twitter, respectivamente. Para ello, creamos cuatro pares de cuentas bot que se conectan simultáneamente varias veces al día. Cada par de bots es idéntico, salvo que uno está configurado para ver los tweets en una línea de tiempo personalizada, mientras que el otro está configurado para ver los tweets en orden cronológico. Estos bots sólo observan los tweets y no realizan ninguna acción. Las cuentas seguidas por cada par de bots se seleccionaron aleatoriamente de una lista de las 200 cuentas de Twitter anticientíficas y procientíficas más populares relacionadas con la pandemia de COVID-19, recopilada entre el 21 de enero de 2020 y el 1 de mayo de 2020 [13]. Estas cuentas tienen un amplio rango de popularidad, con un número de seguidores que oscila entre 1.000 y 10 millones. Utilizando esta metodología, recopilamos 14.213 tweets durante junio de 2020.

La Figura 3 muestra el número de tweets por cuenta en orden descendente de frecuencia, donde la cuenta más activa tiene un rango de 0 y así sucesivamente. Cada panel representa un conjunto de cuentas seguidas por nuestros bots, y los tweets observados en cada línea temporal algorítmica se compararon con la actividad total producida por las cuentas. De estas comparaciones se desprenden dos observaciones principales: en primer lugar, la distribución observada de tweets es "heavy-tailed", lo que significa que una minoría de las cuentas son muy activas y la mayoría tiene valores pequeños en la cola. En segundo lugar, los tweets observados representan menos del 10% del contenido total creado por las cuentas seguidas durante el periodo de estudio.

Es importante señalar la heterogeneidad de la actividad entre los amigos en las redes sociales. Como muestra la Figura 3, algunas de las cuentas seguidas por nuestros bots de auditoría son mucho más activas que otras. Para medir esta desigualdad en la actividad, utilizamos el coeficiente de Gini sobre el número de tweets de cada cuenta. El coeficiente de Gini es una medida de desigualdad utilizada habitualmente en diversos ámbitos y varía entre 0 (igualdad perfecta) y 1 (desigualdad máxima). Por ejemplo, el Gini de desigualdad de ingresos en Chile, que se considera alto, ronda alrededor de 0,45 en los últimos año.

En nuestra investigación, descubrimos que el coeficiente de Gini de la actividad de los amigos en Twitter es elevado (mayor a 0,5), lo que refleja una desigualdad significativa en la actividad de las cuentas que seguimos. Esta desigualdad puede tener implicaciones importantes para la exposición a contenidos en las redes sociales, ya que los algoritmos de recomendación pueden sesgar aún más la exposición a contenidos populares o producidos por cuentas muy activas. Comparando la línea de tiempo personalizada con la presentada en orden cronológico (inverso) y, además, con el conjunto no filtrado de mensajes generados por las cuentas que siguen los bots, podemos cuantificar los sesgos de selección algorítmica en Twitter. El algoritmo de selección de Twitter modifica sutilmente los tweets de los seguidores observados por el

Hemos identificado una nueva paradoja [...] que hace que un rasgo parezca significativamente más popular localmente, entre los amigos de un individuo, que globalmente, entre todas las personas de la red.

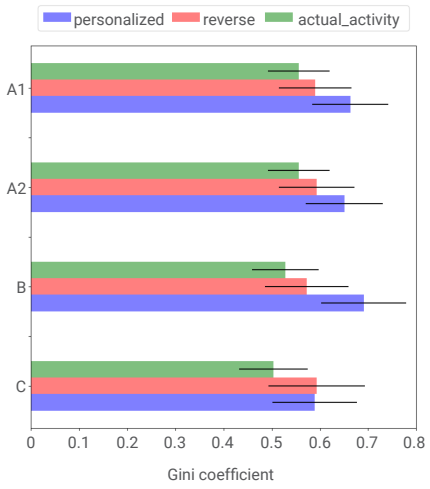


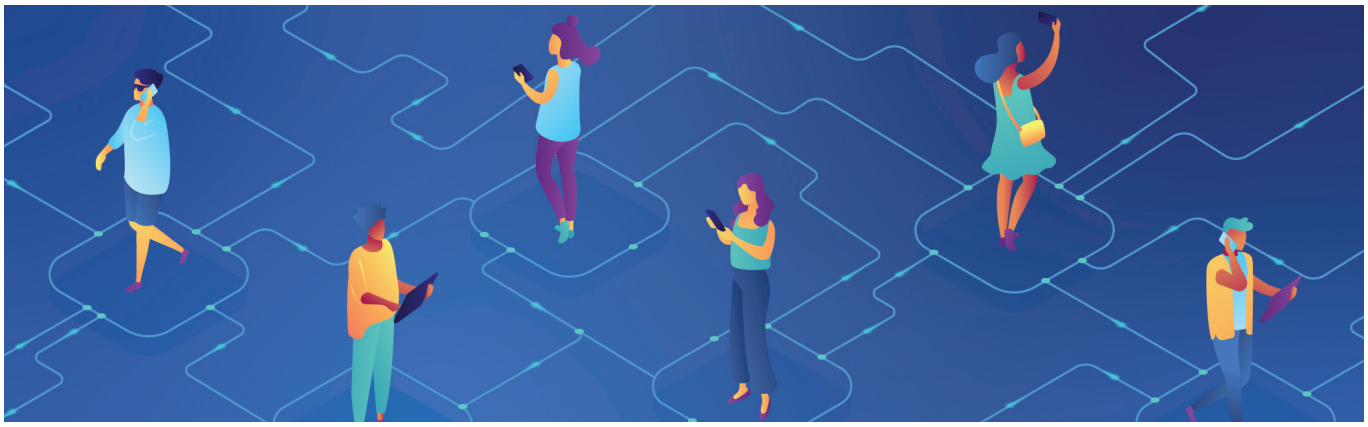
Figura 4: Desigualdad de exposición de las cuentas en Twitter. El coeficiente de Gini de la actividad de las cuentas muestra una mayor desigualdad en la actividad observada en comparación con la actividad real. Además, la desigualdad de la exposición en la cronología personalizada es mayor que en la cronología inversa. Las barras de error son intervalos de confianza del 95% calculados a partir de 1.000 muestras *bootstrap* entre sesiones.

usuario. Identificamos tres tipos de sesgo: popularidad, recencia y exposición. Como resultado de la selección, el usuario ve una distribución diferente de la actividad real, con unos pocos usuarios que reciben una parte desproporcionada de la atención total. Sorprendentemente, la línea de tiempo cronológica también distorsiona la actividad observada, aunque no tanto.

En la Figura 4 se muestra una comparación del Gini entre la actividad total de las cuentas (barras verdes) y la actividad de las cuentas considerando sólo los mensajes que aparecieron en la línea de tiempo personalizada (barras azules) o cronológica (barras rojas) de cada bot. Se observa que la actividad observada de las cuentas en la línea de tiempo personalizada (barras azules) es más sesgada que su actividad real. El coeficiente de Gini de la actividad observada de los amigos en la línea de tiempo personalizada oscila entre 0,585 (conjunto C) y 0,703 (conjunto B). Resulta curioso que la actividad de las cuentas observada en la línea de tiempo cronológica (barras rojas) también está más sesgada que su actividad real, y oscila entre 0,565 (conjunto B) y 0,608 (conjunto A). Es decir, todas las formas de ranking algorítmico, incluso el ranking cronológico que uno esperaría que sea neutro, amplifican los sesgos en la actividad de las redes sociales.

Conclusión

El papel de los sistemas algorítmicos sigue creciendo, desplazando a las formas tradicionales de moderación de contenidos. Los efectos emergentes que amplifican sesgos en las redes sociales no sólo están relacionados con los algoritmos, sino también con la interacción de estos con distintos componentes del sistema sociotécnico. Por tanto, es importante comprender las estructuras de incentivos de las plataformas, así como las interacciones entre humanos y algoritmos. Una mayor transparencia por parte de las empresas de plataformas sociales mejoraría nuestra comprensión de estos aspectos fundamentales. Más allá de las redes sociales, los algoritmos de recomendación se han abierto camino en casi todos los campos, como el comercio electrónico, el cine, la música, el arte, la salud, la alimentación, el derecho y las finanzas. Esto abre la puerta a colaboraciones interdisciplinares entre la industria y el mundo académico con un gran potencial de impacto y retos importantes en cuanto al impacto de estas tecnologías en la sociedad. ■



REFERENCIAS

[1] Baeza-Yates, R. 2018. “Bias on the Web”. Communications of the ACM 61 (6): 54–61.

[2] Abeliuk, A., G. Berbeglia, M. Cebrián, y P. Van Hentenryck. 2015. “The Benefits of Social Influence in Optimized Cultural Markets”. PLoS ONE 10 (4).

[3] Pariser, E. 2011. The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. Penguin.

[4] Chong, Su., y A. Abeliuk. 2019. “Quantifying the Effects of Recommendation Systems”. In 2019 IEEE International Conference on Big Data (Big Data), 3008–15. IEEE.

[5] Bakshy, E., S. Messing, y L. A. Adamic. 2015. “Exposure to Ideologically Diverse News y Opinion on Facebook”. Science 348 (6239): 1130–32.

[6] Su, J., A. Sharma, y S. Goel. 2016. “The Effect of Recommendations on Network Structure.” In Proceedings of the 25th International Conference on World Wide Web, 1157–67. WWW ’16. Republic y Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

[7] Robertson, R. E., D. Lazer, y C. Wilson. 2018. “Auditing the Personalization y Composition of Politically-Related Search Engine Results Pages”. In Proceedings of the 2018 World Wide Web Conference, 955–65. WWW ’18. Republic y Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.

[8] Bandy, J., y N. Diakopoulos. 2020. “Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News”. Proceedings of the International AAAI Conference on Web and Social Media 14 (May): 36–47.

[9] Perkins, H. 2002. “Social Norms and the Prevention of Alcohol Misuse in Collegiate Contexts”. Journal of Studies on Alcohol, Supplement, no. s14 (March): 164–72.

[10] Alipourfard, N., B. Nettasinghe, A. Abeliuk, V. Krishnamurthy, y K. Lerman. 2020. “Friendship Paradox Biases Perceptions in Directed Networks”. Nature Communications 11 (1).

[11] Lerman, K., X. Yan, and X. Wu. 2016. “The ‘Majority Illusion’ in Social Networks”. PLOS ONE 11 (2): e0147617.

[12] Bartley, N., A. Abeliuk, E. Ferrara, and K. Lerman. 2021. “Auditing Algorithmic Bias on Twitter”. In 13th ACM Web Science Conference 2021, 65–73. WebSci ’21. New York, NY, USA: Association for Computing Machinery.

[13] Rao, A., F. Morstatter, M. Hu, E. Chen, K. Burghardt, E. Ferrara, y K. Lerman. 2020. “Political Partisanship and Anti-Science Attitudes in Online Discussions about Covid-19”. arXiv.