The challenges of Big Data in social research: gender biases and situated knowledge

Gazcón Nuñez, Mariana

Mariana Gazcón Nuñez
mariana.gazcon@gmail.com
Asociación Progreso para México, México

Analéctica
Arkho Ediciones, Argentina
ISSN-e: 2591-5894
Periodicidad: Bimestral
vol. 9, núm. 59, 2023
revista@analectica.org

Recepción: 24 Mayo 2023
Aprobación: 30 Junio 2023

URL: http://portal.amelica.org/ameli/journal/251/2514741003/

DOI: https://doi.org/10.5281/zenodo.10475495

Resumen: El Big Data supone la confluencia de múltiples tendencias tecnológicas que, si bien venían madurando durante la primera década del siglo XXI, es a partir de 2011 cuando explosionan e irrumpen a gran escala, derivado del uso masivo de redes sociales, la reducción de los costes de internet, mayor velocidad de ancho de banda, mejor infraestructura de conectividad, internet de las cosas, geolocalización y datos en la nube (Joyanes, 2013). El término refiere no solo a las grandes cantidades de datos que se generan de manera continua, provenientes de personas, organismos, comportamientos o sociedades en conjunto (Lesnichevsky, 2021); sino también a arquitecturas diseñadas para capturar y analizar el volumen de datos a altas velocidades de procesamiento y recepción-distribución de dicha información (Joyanes, 2013).

Palabras clave: big data, investigación social, género.

Abstract: Big Data represents the confluence of multiple technological trends that, although they were maturing during the first decade of the 21st century, it is from 2011 onwards that they explode and break out on a large scale, derived from the massive use of social networks, the reduction of costs of internet, higher bandwidth speed, better connectivity infrastructure, internet of things, geolocation and data in the cloud (Joyanes, 2013). The term refers not only to the large amounts of data that are generated continuously, coming from people, organisms, behaviors or societies as a whole (Lesnichevsky, 2021); but also to architectures designed to capture and analyze the volume of data at high speeds of processing and reception-distribution of said information (Joyanes, 2013).

Keywords: big data, social research, gender.

El Big Data supone la confluencia de múltiples tendencias tecnológicas que, si bien venían madurando durante la primera década del siglo XXI, es a partir de 2011 cuando explosionan e irrumpen a gran escala, derivado del uso masivo de redes sociales, la reducción de los costes de internet, mayor velocidad de ancho de banda, mejor infraestructura de conectividad, internet de las cosas, geolocalización y datos en la nube (Joyanes, 2013). El término refiere no solo a las grandes cantidades de datos que se generan de manera continua, provenientes de personas, organismos, comportamientos o sociedades

en conjunto (Lesnichevsky, 2021); sino también a arquitecturas diseñadas para capturar y analizar el volumen de datos a altas velocidades de procesamiento y recepción-distribución de dicha información (Joyanes, 2013).

La cantidad de información generada por segundo es incalculable, según la infografía "The Internet in every minute 2023", publicada por eDiscovery Today (2023) y Legal Tech Media Group (LMTG!) se estima que cada minuto se envían 241.2 millones de correos electrónicos en todo el mundo, 18.8 millones de mensajes de texto, 10.4 millones de vistas en Instagram, 6.94 millones de emojis enviados, 6.3 millones de conferencias en Zoom, 2.4 millones de búsquedas en Google, entre otros datos generados en sesenta segundos; calculando que en 2021, la cantidad de datos producida en el mundo rondó los 79 zettabytes[1], esperando se duplique para 2015 (Djuraskovic, 2023).

Los Big Data representan fuentes de datos no tradicionales (bases de datos relacionales con esquemas fijos, como hojas de cálculo); contienen datos semiestructurados (aquellos sin formatos fijos mediante etiquetado o marcadores, por ejemplo, etiquetas de XML o HTML) y datos no estructurados (datos sin campos fijos como multimedia, mensajería o correos), siendo sus tres grandes dimensiones el volumen, la velocidad y la variedad (3V), considerando adicionalmente la veracidad y el valor, como características de los Big Data (Joyanes, 2013).

El modelo de las tres V considera el volumen como las cantidades masivas de datos que se crean y almacenan diariamente, pasando a la era del zettabyte en un corto periodo de tiempo y calculando que dicha información se multiplicará significativamente en los próximos años; la velocidad de los datos es una de las características más importantes del Big Data, esto implica un procesamiento en tiempo real de la información generada para la toma de decisiones oportuna, cinco minutos pueden ser tardíos para un análisis en la actualidad y, finalmente, la variedad de las fuentes de datos en el Big Data, desde aquellas estructuradas hasta los datos no estructurados que se generan en las redes sociales (Joyanes, 2013).

Esto versa en el mito de que "todos los aspectos de la vida pueden y deben convertirse en datos" (Becerra & Castorina, 2023, p. 2) y, supone una creencia de naturaleza epistémica respecto a lo que es el conocimiento, cómo se construye, a qué objetivos sirve, y con qué criterios y valores lo evaluamos socialmente (Becerra & Castorina, 2023). Actualmente la producción del conocimiento científico está influenciada por los avances tecnológicos que implican la sofisticación de métodos de recolección y análisis de datos. Los datos están cargos de poder, y "la potencia aumenta porque la lógica del mundo es en red de atracción: los datos llaman a más datos, a la concentración" (Benítez-Eyzaguirre, 2019).

Bajo este fundamento surgen discusiones relacionadas a las marcos epistémicos y ontológicos en el campo del Big Data, focalizándose en aspectos que expresan la relación conocimiento-sociedad, donde se discute "la naturaleza del dato y la pretendida objetividad de su análisis" (dato crudo/interpretado) y "la manera de recortar su objeto de referencia" (disputa entre conducta/sujeto) (Becerra & Castorina, 2023); afirmando que quienes diseñan dichas estructuras de datos poseen "conocimiento del dominio, es decir saben cuáles son los objetos y procesos importantes, cómo se pueden definir y cómo se relacionan en términos de estructura y causalidad" (de Vos, 2014, como se citó en Suárez, 2021, p. 323),

siguiendo a Gitelman y Jackson (como se citó en Suárez, 2021, p. 320), "los datos probablemente no tengan ninguna relación con la verdad o la realidad más allá de la realidad que ellos mismos nos ayudan a construir".

A partir de esto, se trazan reflexiones respecto a los sesgos heredados en la producción del conocimiento científico, como la preocupación en la existencia de deficiencias en la calidad de los datos masivos que se encuentran disponibles, particularmente sesgos de género, lo que puede conducir a conclusiones erróneas a partir de los mismos, poniendo de manifiesto que quienes analizan los datos "deben ser conscientes de la necesidad de que el nivel de agregación de los datos no menoscabe su calidad" (Sáinz et al. 2020); recalcando en esto a lo que Haraway enuncia con su propuesta de conocimiento situado y parcial, en el que sostiene que se tiene

> una visión parcial de la realidad y que es indispensable reconocer que el lugar (físico, epistémico y simbólico) desde donde miramos influenciará el conocimiento que producimos. La inclusión de sujetos y grupos sociales marginalizados en los espacios de producción de conocimiento sería, por lo tanto, indispensable, pero no porque éstos ocupen un lugar epistémico privilegiado, sino porque aportarán un conocimiento diferenciado de la realidad que, juntamente con los demás saberes, nos permitirá tener una visión más completa (Martínez et al. 2014).

Conceptualizando estos sesgos estadísticos que hacen "referencia a errores sistemáticos que distorsionan los datos o los análisis efectuados sobre ellos" (Bercovich, 2020, como se citó en Lesnichevsky, 2021, p. 28), siendo no siempre conscientes por quien les genera, pero asumiendo supuestos como verdades a partir de la neutralidad, desembocando esto en dos problemas: "la apariencia de la objetividad y el desarrollo de la ciencia basada en universalidades donde los sesgos pasan por naturales" (Lesnichevsky, 2021).

Retomando a Haraway (1995) señala que "el «género» fue desarrollado como una categoría para explorar lo que suele entenderse por «mujer», para problematizar lo que había sido tomado como regla inamovible", añadiendo que las teorías feministas apuestan por un proyecto de "ciencia del sucesor" que oferta una visión más adecuada del mundo, con vistas de vivir bien y "en relación crítica y reflexiva con nuestras prácticas de dominación y con las de otros y con las partes desiguales de privilegio y opresión que configuran todas las posiciones" (Haraway, 1995).

Volviendo con esto a la potencialidad del Big Data, de mejorar la calidad de vida de las sociedades partiendo de políticas dirigidas y descubrimientos de relevancia científica; todo esto, "en la medida que sus premisas, metodologías y usos estén orientadas por premisas éticas y epistemológicas y políticas orientadas al bien común y la justicia social" (Lesnichevsky, 2021), permitiendo revelar prácticas ocultadas o invisibilizadas bajo el conocimiento hegemónico. Implicando que la perspectiva de género y feminista "refiere la necesidad de involucrarse con el contexto de los datos para acercarse de la mejor manera posible a la realidad" (Lesnichevsky, 2021), de igual forma posibilita rastrear la complejidad social y la multiplicidad de factores implicados (Suárez, 2021).

El «feminismo de datos», se modela, como una forma de pensar los datos, en su uso y límites, partiendo de la distribución inequitativa del poder en el mundo y teniendo como base la experiencia directa y el compromiso con la co-liberación, en esta "idea de que los sistemas de poder opresivos nos perjudican a todas las

personas, que socavan la calidad y la validez de nuestro trabajo, y que nos impiden crear un impacto social verdadero y duradero con la ciencia de datos" (D'Ignazio & Klein, 2020), también ayuda a recordarnos que antes de datos, hay personas; "personas que ofrecen su experiencia para ser contadas y analizadas, personas que realizan ese conteo y análisis, personas que visualizan los datos y promueven los hallazgos de cualquier proyecto en particular y las personas que usan el producto al final" y, al mismo tiempo, personas que no son contadas y problemas que no se puede abordar sólo con los datos.

Finalmente, el feminismo de datos reconoce siete principios bajo los cuáles se fundamenta el

1. examinar el poder partiendo del análisis de cómo opera en el mundo,
2. desafiar al poder y trabajar por la justicia,
3. elevar la emoción y la encarnación que proviene de las personas y cuerpos sensibles en el mundo,
4. repensar el binarismo y las jerarquías, junto con otros sistemas de clasificación y conteo que perpetúan la opresión,
5. reconocer el pluralismo, priorizando conocimientos locales, indígenas y experienciales,
6. considerar el contexto partiendo de que los datos no son neutros ni objetivos sino producto de relaciones desiguales y
7. hacer el trabajo visible, colaborativo, que se construye de la mano de muchas (D'Ignazio & Klein, 2020).

Concluyendo con que, el feminismo de datos propone reformular sobre el pensar y hacer datos, ampliando el horizonte que lleve a la construcción de un conocimiento menos excluyente y más amplio, a fin de aportar en la mejora de las condiciones de las personas; comprehendiendo, desde el diseño y narrativa de los mismos, los fenómenos sociales en todas las aristas de la complejidad.

Referencias

Becerra, G. & Castorina, J.A. (2023). Hacia un análisis de los marcos epistémicos del big data. *Cinta de moebio*, (76), 50-63. https://dx.doi.org/10.4067/s0717-554x2023000100050

Benítez-Eyzaguirre, L. (2019). Ética y transparencia para la detección de sesgos algorítmicos de género. Estudios sobre el Mensaje Periodístico, 25 (3), 1307-1320. https://dx.doi.org/10.5209/esmp.66989

D'Ignazio, C. & Klein, L.F. (2020). *Data Feminism.* The MIT Press. Traducción al español 2023. https://data-feminism.mitpress.mit.edu/bienvenida

Djuraskovic, O. (4 de octubre de 2023). *Estadísticas de Big Data 2023: ¿Cuántos datos hay en el mundo?* First Site Guide. https://firstsiteguide.com/es/big-data-stats/

eDiscovery Today. (20 de abril de 2023). *2023 Internet Minute Infographic, by eDiscovery Today and LTMG!: eDiscovery Trends*.https://ediscoverytoday.com/2023/04/20/2023-internet-minute-infographic-by-ediscovery-today-and-ltmg-ediscovery-trends/

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

Joyanes Aguilar, L. (2013). *Big Data, Análisis de grandes volúmenes de datos en organizaciones.* Alfaomega.

Lesnichevsky Boronat, M. (2021). *Aportes feministas al futuro de la ciencia de datos.* [Tesis de maestría]. Universitat de Barcelona. URI http://hdl.handle.net/2445/184186

Martínez Martínez, L. M., Biglia, B., Luxán, M., Fernández, C., Azpiazu Carballo, J., Bonet Martí, J. (2014). Experiencias de investigación feminista: propuestas y reflexiones metodológicas. *Athenea digital: revista de pensamiento e investigación social*, Vol. 14 Núm. 4 (diciembre 2014), p. 3-16. DOI 10.5565/rev/athenea.1513

Sáinz, J., Arroyo, L. y Castaño, C. (2020). *Mujeres y digitalización: de las brechas a los algoritmos*. Madrid: Instituto de la Mujer y para la Igualdad de Oportunidades, 130 p. Mujeres, Tecnología y Sociedad Digital. https://cendocps.carm.es/documentacion/2020_Mujeres_digitalizacion.pdf

Suárez Val, H. (2021). Marcos de Datos de Feminicidio. *Informatio. Revista Del Instituto De Información De La Facultad De Información Y Comunicación*, 26(1), 313-346. https://doi.org/10.35643/Info.26.1.15

Notas

1 Un zettabyte equivale a 1 billón de gigabytes, comparando esto con almacenar 12,288 millones de videos 4k.