Gazeta de Antropología

ISSN 0214-7564
En Internet: ISSN 2340-2792

Gazeta de Antropología, 2023, 39 (1), artículo 09 · Versión HTML

Recibido 17 marzo 2022 | Aceptado 24 octubre 2022 | Publicado 2023-01

Big data y ciencias sociales. Una mirada comparativa a las publicaciones de antropología, sociología y trabajo social

Big Data and Social Sciences. A comparative review at anthropology, sociology, and social work publications

Estrella Gualda Caballero
Catedrática de Sociología, Universidad de Huelva, ESEIS/COIDESO (España)

Alba Taboada Villamarín
Doctoranda del Programa de Ciencias Sociales y Educación, Universidad de Huelva, ESEIS/COIDESO (España)
alba.taboada@alu.uhu.es

Carolina Rebollo Díaz
Contratada Posdoctoral "Margarita Salas", Universidad de Huelva, ESEIS/COIDESO (España)
carolina.rebollo@dstso.uhu.es

40 AÑOS (1982-2022) DE GAZETA DE ANTROPOLOGÍA
MONOGRÁFICO COORDINADO POR ÁNGELES ARJONA Y JOSÉ LUIS SOLANA

RESUMEN
Este artículo revisa la bibliografía internacional sobre big data y explora comparativamente la evolución, características y temáticas de las investigaciones que sobre este tema se encuadran en las áreas de antropología, sociología y trabajo social. Se emplean métodos cuantitativos para la descripción y una estrategia analítica de aprendizaje automático no supervisado al objeto de identificar y agrupar los principales tópicos o temáticas de los artículos. Los resultados confirman que el interés sobre los macrodatos ha llegado antes a la sociología que a la antropología o el trabajo social. Igualmente, se destaca la importancia de las publicaciones inter y multidisciplinares sobre big data en estas disciplinas. Del modelado de temas emergen 13 clústeres, destacando los correspondientes a publicaciones sobre redes sociales, epistemología, metodología e implicaciones del big data, big data y sociedad, salud y machine learning.

ABSTRACT
This article reviews the international bibliography on big data. Comparatively, it explores the evolution, characteristics and themes of the research on this topic that falls within Anthropology, Sociology and Social Work. Quantitative methods are used for the description. Also, we employed an analytical strategy of unsupervised machine learning to identify and group the main themes of the articles. The results confirm that interest in big data has reached sociology before anthropology or social work. Likewise, inter and multidisciplinary publications on big data in these disciplines are highlighted. Also, from the topic modeling analysis, 13 clusters emerged. The most important were those corresponding to publications on social networks, epistemology, methodology and implications of big data, big data and society, health, and machine learning.

PALABRAS CLAVE
macrodatos | Ciencias Sociales | Antropología | Sociología | Trabajo Social | modelado de temas | aprendizaje automático no supervisado
KEYWORDS
big data | Social Sciences | Anthropology | Sociology | Social Work | topic modeling | non-supervised machine learning

1. Introducción

1.1. Big Data y Ciencias Sociales

Uno de los términos que se han introducido en el debate científico global y particularmente en las ciencias sociales en los últimos lustros ha sido el de *big data* o macrodatos, si bien es cierto que la llegada ha sido desigual por áreas científicas o incluso en subcampos de especialización. Cronológicamente, las publicaciones relativas a los macrodatos comienzan antes en ramas de ciencias e ingeniería y arquitectura que en las ciencias sociales (de acuerdo con las publicaciones que se encuentran catalogadas en Web of Science (WoS) y Scopus. Por otra parte, dentro de las ciencias sociales, la incorporación sigue ritmos diferentes, encontrándose la sociología, de acuerdo con el análisis bibliométrico de Edelmann y otros (2020) como una de las que en mayor medida se incorpora a la investigación sobre *big data*.

Para situarnos, por *big data*, macrodatos o datos masivos se entiende algo que va más allá de una gran cantidad o volumen de datos, si bien el volumen es uno de los elementos clave y más claramente consensuados que definen este concepto que alude a conjuntos de datos con rasgos diferentes a los que tradicionalmente se podían encontrar. Detrás del crecimiento explosivo de datos al que hemos asistido, se encuentran cambios tecnológicos de relevancia como el rápido desarrollo de Internet, la Internet de las cosas (IoT) y la computación en la nube (Cloud Computing) (Jin y otros 2015, Hashema y otros 2015).

Se ha popularizado, tanto en la bibliografía técnica y académica como en la divulgativa, la definición de *big data* a través de la llamada "familia de Vs del Big Data" (Patgiri y Ahmed 2016: 17). Algunos de los elementos con los que suele definirse esta área se describen a través de palabras cuya primera letra es precisamente una V (en inglés). La bibliografía, no sin algunas controversias, define el concepto *big data* asociándolo a términos como *volumen*, *velocidad* y *variedad* que nos remiten a aspectos como el incremento de la velocidad y del abanico de fuentes de datos, estructuras y formatos. Con el tiempo, se van sumando otras descripciones de los *big data* relativas a la baja *veracidad* de estos datos (en relación a la desinformación), su alto *valor* empresarial, su *variabilidad*, *volatilidad*, carácter *virtual*, la importancia de su *visualización*, o incluso su complejidad, entre otras descripciones o aspectos que no tenemos espacio para desarrollar aquí pero sobre los que hay abundante bibliografía (Laney 2001, Beyer y Laney 2012, Jin y otros 2015, Del Vecchio y otros 2018, Bartosik-Purgat y Ratajczak-Mrożek 2018, Hashema y otros 2015, Bello-Orgaz y otros 2016, Olshannikova y otros 2017, Bulger y otros 2014). Quizás una de las claves de lo que representan los *big data* para las ciencias sociales hoy es que marcan una clara diferencia respecto al tipo de datos que manejamos en el pasado, así como de los procedimientos metodológicos que necesitamos para abordarlos (Gualda 2022).

En este artículo, a partir de una revisión de la bibliografía internacional existente que recoge investigaciones en torno a los *big data*, nos preguntamos por el interés que los macrodatos han tenido en tres disciplinas de las ciencias sociales que se encuentran muy próximas entre sí, pero que al mismo tiempo tienen sus propias tradiciones y especificidades. Para la elaboración de este artículo nos hemos preguntado por:

1) La evolución y caracterización comparativa de las publicaciones que se han publicado sobre *big data* asociadas a las disciplinas de antropología, sociología y trabajo social. Concretamente, nos ha interesado describir aspectos como el número de publicaciones, revistas y rama científica en las que se han publicado y tipo de publicaciones, desde que se introduce este campo en el panorama internacional.

2) Cómo se agrupan estos trabajos en clústeres o conglomerados, a partir de las temáticas o palabras clave contenidas en sus títulos, resúmenes y palabras clave de cada documento, al efecto de llevar a cabo tanto una descripción como un mapeo visual de los mismos.

3) Si era posible encontrar diferencias entre las publicaciones que han sido clasificadas como de antropología, sociología y trabajo social, o identificar aspectos clave ligados a la tradición, historia o idiosincrasia de cada disciplina, así como desafíos ligados a este campo de estudio.

Nuestra hipótesis de fondo, sin entrar en cuestiones de corte bibliométrico o relativas al impacto de los documentos que exceden este trabajo, es que, más allá del diferente ritmo con el que diferentes ciencias se han ido incorporando a la investigación en el área de *big data*, la tradición disciplinar, así como los métodos científicos que la caracterizan, importa y tiene incidencia en su evolución en relación con el protagonismo atribuido al campo de estudio sobre macrodatos. En este sentido, no es descabellado prever que el interés por llevar a cabo investigaciones en este campo haya sido más habitual en la sociología que en la antropología o el trabajo social, por la extensa tradición cuantitativa de la primera frente a un mayor acento en estudios cualitativos o en la intervención social en la segunda y tercera, respectivamente.

1.2. La aproximación a los *big data* en las áreas de Antropología, Sociología y Trabajo Social

La investigación con *big data* parece haber brindado a las ciencias sociales nuevas oportunidades para obtener un mejor entendimiento de las dinámicas y estructuras sociales en un mundo cada vez más globalizado y dependiente de las nuevas tecnologías. Por ejemplo, los estudios que han utilizado macrodatos a través de una lente antropológica han revelado su capacidad para aportar conocimientos sobre cuestiones tan variadas como las complejas y problemáticas relaciones de la vida pública y privada (Mardeson y otros 2015), la legitimidad y autoridad de los Estados y los desafíos a los mismos (Monroe 2017) o el papel del pasado en la construcción de las identidades políticas actuales (Bonacchi y otros

2018). Objetos de estudio clásicos de la antropología ahora también se desarrollan en contextos o escenarios diferentes como las plataformas de redes sociales, generadoras de una gran cantidad de datos e interacciones sociales, que se han transformado en campos de investigación etnográficos no tradicionales (Bonacchi y otros 2018). La sociología parece haber integrado bastante bien los *big data* para el estudio del comportamiento, percepciones e interacciones sociales. Estos se han utilizado para detectar líderes de comunidades y su influencia (Ahajjam y otros 2018), cuestiones sobre desigualdades de género (Lynn y otros 2019) o la polarización de los debates públicos y de las ideas políticas (Robles y otros 2020). Para el trabajo social, las publicaciones en las redes sociales que generan grandes conjuntos de datos han demostrado tener potencial para comprender mejor fenómenos sociales en profundidad, especialmente los relatos de la vida cotidiana, las experiencias vividas, las percepciones personales y las opiniones que podrían no recogerse a través de los métodos cualitativos tradicionales (Caplan y Purser 2019). Asimismo, se han utilizado grandes conjuntos de datos administrativos con el objetivo de incorporar algoritmos para los sistemas de apoyo a la toma de decisiones y así mejorar el trabajo social con niños y familias (Gillingham 2019), aunque todavía estos sistemas de apoyo se están explorando escasamente (Schneider y Seelmeyer 2019). Esos grandes datos administrativos también han sido utilizados, por ejemplo, para la creación de modelos predictivos basados en el aprendizaje automático para desarrollar sistemas de alerta temprana que sirvan para identificar con antelación a los estudiantes en riesgo de abandono escolar (Chung y Lee 2019) o para predecir la probabilidad de que los niños de acogida de Estados Unidos logren la permanencia legal (Elgin 2018).

A pesar de que el uso de los *big data* para la investigación social ha ido creciendo en los últimos años, la integración de los mismos en las diferentes disciplinas sociales no ha estado exenta de debates. La bibliografía actual parece indicar que los investigadores cualitativos (entre ellos, especialmente los etnógrafos) han sido más reacios a integrar macrodatos y enfoques computacionales en sus trabajos (Abramson y otros 2018). La preocupación radica en que el mayor interés en los grandes datos podría venir a expensas de la etnografía cara a cara (Reyes 2014), un debate importante en áreas como la antropología lingüística, ya que como argumentaba Philips (2013: 93): “si uno está tratando de entender la naturaleza de la relación entre el lenguaje y la cultura, entonces metodológicamente tiene sentido dar prioridad a los datos que provienen de la comunicación en la interacción cara a cara que ocurre socialmente porque ese es el lugar de la constitución de los procesos sociales” (traducción propia). Para otros autores, en cambio, es clave que incorporemos el análisis de Internet para “comprender más profundamente las dinámicas relacionales de la sociedad actual” (Escobar y otros 2022: 91).

Asimismo, las lógicas y prácticas de uso intensivo de datos en diferentes ámbitos de la vida contemporánea han hecho aumentar la “datificación”, es decir, la conversión de aspectos cualitativos de la vida en datos cuantificados (Ruckenstein y Schüll 2017), lo que podría acabar potenciando el privilegio con el que cuenta la investigación cuantitativa para investigar estos ámbitos de la vida social, a expensas de la cualitativa. A esto se le añade que, incorporar enfoques computacionales para el estudio de *big data* es, sin lugar a duda, una tarea compleja para los científicos sociales menos acostumbrados a las técnicas que suelen aplicarse a los mismos (*machine learning, topic modeling*, procesamiento del lenguaje natural y otros tipos de programación) lo que puede haber aumentado también las reticencias para estudiar los fenómenos sociales a través de grandes datos, sea por el hándicap que supone la necesidad de contar bien con conocimientos técnicos de programación y matemáticos o estadísticos, bien con equipos de trabajo inter(trans)disciplinares para abordar estas investigaciones (Gualda y Rebollo 2020). Otros debates tienen que ver con problemas prácticos relacionados con los procesos de generación, recopilación y análisis de datos, así como con la interpretación y aplicación de los resultados (Gillingham y Graham 2017), la generalización de las observaciones al mundo *offline* (Golder y Macy 2014) o el posible sesgo en los datos (McFarland y McFarland 2015).

A estas cuestiones, se le suman otras discusiones de naturaleza epistemológica: por una parte, el escepticismo que manifiestan algunos investigadores sociales respecto a que los datos sean capaces de ofrecer un conocimiento profundo de las personas. Por otra parte, la preocupación sobre el uso ético de los datos, la seguridad y privacidad de estos, el consentimiento informado o la equidad de acceso, ya que, a diferencia de la antropología etnográfica, la analítica de grandes datos no tiene una historia larga y profunda de compromiso con las personas participantes de la investigación ni está totalmente clara la responsabilidad de los proveedores de los datos (Pink y Lanzeni 2018).

No obstante, es cierto que en los últimos años las humanidades digitales han emergido con fuerza suscitándose aspectos de gran interés en el área de *big data*. Por ejemplo, en relación a técnicas

asociadas al manejo de macrodatos o datos masivos como el procesamiento del lenguaje natural para el estudio de aspectos como el lenguaje para comprender a la gente y las culturas. El potencial que tiene la ingente cantidad de datos de los que se dispone actualmente necesita de técnicas automatizadas para ayudar a su manejo y en este sentido, los avances recientes en áreas como el modelado de temas (*topic modeling)*, incrustaciones de palabras (*word embeddings)* y modelos de redes neuronales (*neural networks)* pueden ser de gran ayuda en el futuro en estos campos (Berger y Packard 2021). Por otra parte, el empleo de técnicas de aprendizaje automático (*machine learning)* puede ser de gran utilidad para ayudar en procesos de codificación de una gran cantidad de documentos complejos relativos a aspectos sociales y culturales, en lo que ya existe experiencia (Rona-Tas y otros 2019).

Una manera en la que se han solventado algunos de los desafíos metodológicoses la de aproximarse a los *big data* mediante métodos mixtos, y no exclusivamente cuantitativos, o a través de algoritmos o enfoques computacionales. En la bibliografía reciente encontramos investigaciones que han utilizado, por ejemplo, procesos como el *topic modeling* complementado con un posterior análisis cualitativo en profundidad (Bonacchi y otros 2018), o la extracción y procesamiento de datos a través de lenguajes de programación para obtener una muestra más reducida a la que aplicar filtrados manuales y análisis cualitativos (Brownlie y Shaw 2019). O, incluso, a través de métodos exclusivamente cualitativos, como el análisis temático cualitativo a través de codificación asistida por ordenador (Caplan y Purser 2017). Además, algunos autores han defendido la combinación de la etnografía y las técnicas de minería de grandes conjuntos de textos para enfatizar las fortalezas y abordar las debilidades de cada uno (Abramson y otros 2018). Asimismo, otros autores del campo de las ciencias sociales han defendido las oportunidades que brindan los macrodatos como una manera de motivar a la comunidad científica del campo social hacia su uso: pueden proporcionar técnicas sistemáticas para el análisis de patrones, permitir la obtención de muestras más amplias y nuevas posibilidades de compartir y replicar datos y abordar las preocupaciones fundamentales sobre la validez interna y externa (Abramson y otros 2018). Otros autores añaden que el tratamiento automatizado de los *big data* puede reducir los costes y el tiempo para responder a preguntas antropológicas que, de otro modo, serían difíciles de contestar (Alvard y Carlson 2020). Así como, una nueva perspectiva donde la comprobación de hipótesis no sea lo esencial y se brinde una oportunidad para el descubrimiento teórico a través de la inducción (Goldberg 2015).

2. Métodos

En las páginas siguientes hemos aplicado técnicas propias del aprendizaje automático (*machine learning)* para intentar contestar a nuestras preguntas de investigación. En esta sección comenzamos presentando la metodología seguida para la extracción de las publicaciones científicas que son nuestra fuente de datos secundarios para, posteriormente, explicar los procesamientos y procedimientos analíticos que se han seguido. A lo largo del artículo nos vamos a referir de forma indistinta a los términos“documentos” o “publicaciones” extraídos para aludir a artículos, actas de congreso, libros, capítulos de libro, reseñas bibliográficas, material editorial, notas y cartas.

2.1. Fuente de datos y proceso de selección de documentos

Para la realización de este trabajo se ha llevado a cabo una revisión bibliográfica a partir de las publicaciones sobre *big data* identificadas hasta el año 2020. Por otra parte, se ha efectuado una revisión sistemática de los documentos científicos publicados sobre macrodatos y que a su vez estuvieran enmarcados en antropología, sociología y trabajo social.

Para la selección de los documentos se llevaron a cabo varias búsquedas bibliográficas por separado en las bases de datos de WoS y Scopus. Estas bases de datos fueron elegidas porque, en términos de impacto, contienen artículos de las principales revistas del mundo. A partir de los criterios de búsqueda delimitados de acuerdo con nuestros objetivos (véase en el cuadro 1), y con el objeto de permitir la comparación entre ambas fuentes, se seleccionaron todos los documentos publicados desde el inicio de la serie en estas bases de datos hasta 2020 incluido, sin ninguna otra limitación.

Período de búsqueda	1970- 2020 (50 años)
Sources	Se utilizaron las siguientes bases de datos: 1. Web of Science: https://www.webofscience.com 2. Scopus: https://scopus.com/
Descarga de datos	21/04/2021 (WoS) 28/05/2021 (Scopus)
Términos y sintaxis de búsquedas	Se extraen todos los documentos encontrados en WoS y Scopus que cumplen con la condición de que aparezca la palabra "big data" bien sea en el título, en el resumen o en las palabras clave (tanto las mencionadas por los autores como las referenciadas por la revista que se recogen en WoS y Scopus). Sintaxis de búsqueda en WoS TI=("Big Data") OR AB=("Big Data") OR AK=("Big Data") OR KP=("Big Data") Sintaxis de búsqueda en SCOPUS TITLE-ABS-KEY ("Big Data") AND (EXCLUDE (PUBYEAR, 2022) OR EXCLUDE (PUBYEAR, 2021)
Tipo de fuente	1. Web of Science: Web of Science Core Collection 2. Scopus: Scopus all database
Tipo de documentos	Artículos, libros, capítulos de libros, comunicaciones a congresos, editoriales, notas, cartas y otros materiales.
Submuestra de este artículo	Para generar la base de datos de este artículo, se filtraron los documentos obtenidos con la condición de que en los campos 'Título', 'Resumen', 'Palabras clave del autor' y 'Palabra clave de la revista' apareciera alguno (o varios) de los términos que siguen: 'sociology', 'Sociology', 'SOCIOLOGY', 'SOCIOLOGICAL', 'sociological', 'social work', 'Social Work', 'SOCIAL WORK', 'SOCIAL SERVICES', 'Social work', 'Social services', 'social services', 'Anthropology', 'anthropology', 'ANTHROPOLOGY', 'anthropological', 'ANTHROPOLOGICAL' Antes de aplicar el script se habían identificado en la base de datos global las palabras asociadas a las tres disciplinas.

Cuadro 1. Revisión bibliográfica: Período de búsqueda, fuentes, criterios de búsqueda y submuestra. Fuente: Elaboración propia.

2.2. Artículos extraídos y criterios de inclusión y exclusión para el análisis

Se llevó a cabo una primera extracción de publicaciones en WoS y Scopus que tuvieran la palabra clave *big data* en el título, el resumen o las palabras clave. Esta extracción arrojó un total de 255.846 documentos. A partir de aquí, con ayuda de Python, se filtraron las publicaciones o documentos que junto al término de *big data* contarán con las palabras "antropología", "sociología", "trabajo social" y "servicios sociales", con el objetivo de seleccionar los documentos pertenecientes a las tres disciplinas específicas. Este criterio nos posibilitó encontrar aquellas publicaciones relacionadas con las áreas de estudio sin limitarnos exclusivamente a las publicaciones presentes en las revistas catalogadas como propias de una disciplina, criterio que no era común en WoS y Scopus. La aparición conjunta de alguno de estos términos junto a *big data* opera como criterio de inclusión. Se obtuvieron un total de 1.271 documentos correspondientes a estas áreas. Posteriormente, como se había estado trabajando con dos bases de datos (WoS y Scopus), se unificaron y se hizo una limpieza destinada a depurar los artículos duplicados, operando esta depuración como criterio de exclusión. El proceso de selección seguido se encuentra en el cuadro 2, sintetizado a través de un diagrama de flujos. La muestra final tras aplicar estas operaciones fue de 613 artículos. Los textos completos de las publicaciones que fueron seleccionados finalmente fueron revisados para comprobar que efectivamente cumplían los requisitos fijados.

2.3. Procesamiento y limpieza de datos

La base de datos de los documentos seleccionados (en formato Excel) fue procesada para su análisis con apoyo del lenguaje de programación Python (https://www.python.org/). Para manipular nuestros datos con Python, se empezó adaptándolos a un *dataframe*. Un dataframe es un tipo de estructura de datos que tiene dos dimensiones y que permite guardar los datos en columnas. Se empleó la *biblioteca Pandas* de Python para trabajar con estos datos tabulares. Esta es una biblioteca que cuenta con un conjunto de herramientas (funciones) para llevar a cabo acciones con los datos.

Como se ha avanzado, posteriormente se llevó a cabo una limpieza de datos para eliminar los documentos duplicados al haber fusionado las bases de datos de WoS y Scopus. Se empleó el campo que contiene el DOI a estos efectos, al tratarse de un identificador único de cada publicación. En el caso de los documentos sin DOI (por ejemplo, actas de congresos), se llevó a cabo una revisión manual para detectar y eliminar las publicaciones duplicadas restantes.

El análisis de texto se realizó con la información contenida en los campos: “Título del documento” [Title (Scopus), Article Title (WoS)], “Resumen” (Abstract), “Palabras clave del autor” (Author Keywords) y “Palabras clave de la revista” [Index Keywords (Scopus), Keywords Plus (WoS)]. Para la limpieza de los textos se emplea en Python la librería “Natural Language Toolkit, NLTK” (https://www.nltk.org/), a través de la que se realizaron las siguientes tareas:

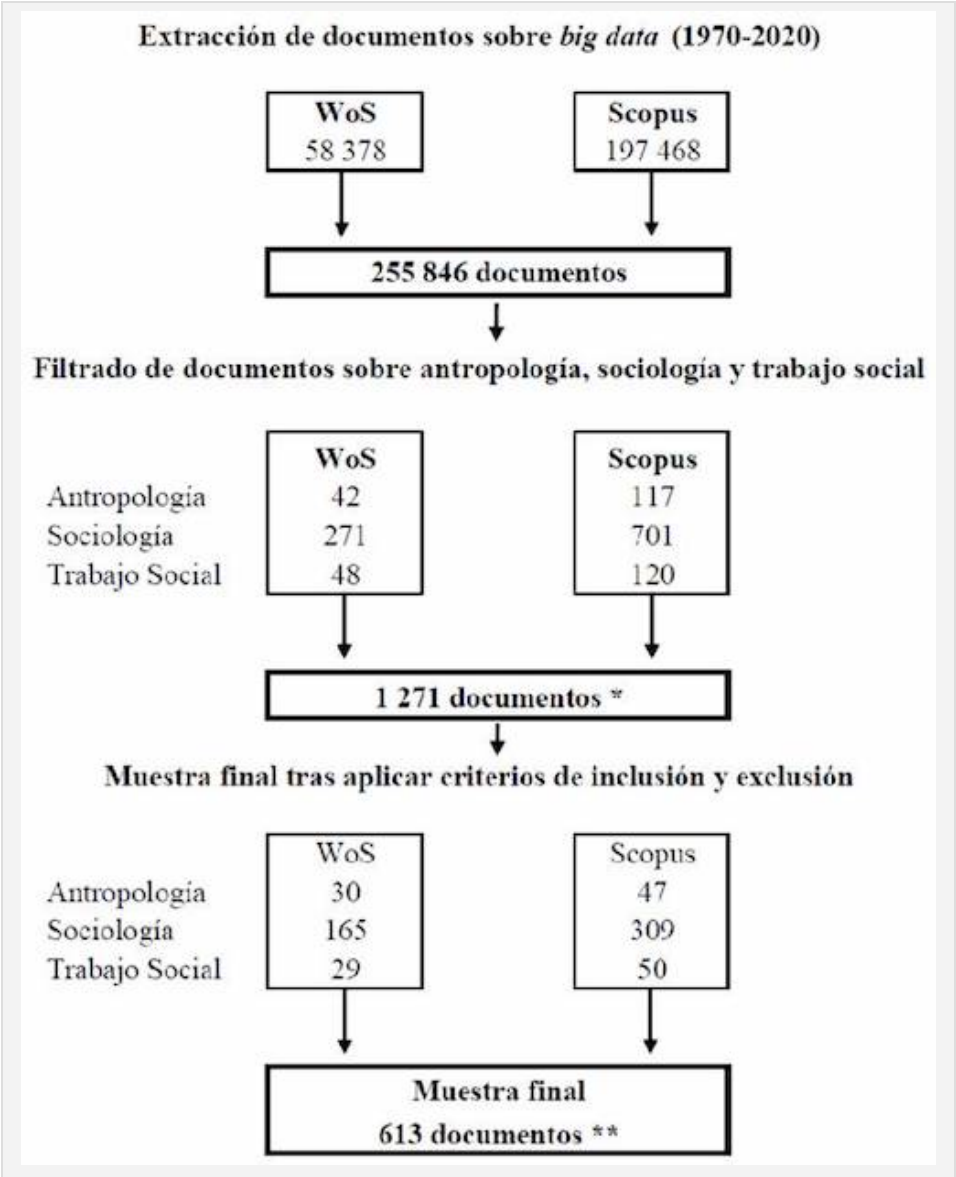
1. Se homogeneizaron todas las palabras pasándolas a minúscula.

2. Se eliminaron todos los signos de puntuación (por ejemplo, exclamaciones, comas, etc.).

3. Se eliminaron las palabras vacías o *stopwords* en inglés, aplicando el listado incluido en la librería NLTK (tales como “and”, “or”, etc.).

4. Se eliminaron los enlaces a páginas webs, si se encontraban.

Por último, dado que se empleó como criterio de búsqueda para la extracción de publicaciones la cadena *big data*, esta se eliminó también en esta fase, al tratarse de una cadena que íbamos a encontrar en todos los casos,con escaso poder discriminatorio.



Cuadro 2. Proceso de selección de documentos. * Se han extraído del cómputo final

los doc. ** Se han extraído del cómputo final los documentos duplicados que se clasifican en más de una disciplina. Fuente: Elaboración propia.

2.4. Análisis de las temáticas de los documentos a partir de una estrategia de aprendizaje automático no supervisado

2.4.1. Aprendizaje automático, supervisado y no supervisado

Una vez que el texto se depuró, con el objetivo de investigar las temáticas que subyacían en los documentos analizados, se aplicaron principalmente dos métodos estadísticos de aprendizaje no supervisado. El aprendizaje no supervisado (*non-supervised machine learning)* es uno de los métodos que se enmarcan en el aprendizaje automático. Entendemos por aprendizaje automático o aprendizaje estadístico (Müller y Guido 2016) al conjunto de herramientas integradas por la estadística, la inteligencia artificial y las ciencias computacionales que permiten el reconocimiento de patrones en los datos. El *machine learning* se puede clasificar como supervisado y no supervisado. En este artículo hemos aplicado algoritmos de aprendizaje no supervisado. Se recomienda el uso de estos algoritmos cuando no se cuenta con etiquetas o respuestas previas asociadas a las observaciones, tal y como sucede en este caso. Esta técnica aplica algoritmos que nos indican la homogeneidad y distancia entre los documentos o publicaciones que estamos analizando. Concretamente, se utiliza la medición de distancias estadísticas para generar conglomerados o clústeres, de forma que se organizan los textos en función de los que son más similares entre ellos y se separan en otros grupos los más distantes (James y otros 2013).

2.4.2. Aprendizaje automático no supervisado: Estrategia de análisis

En primer lugar, se llevó a cabo la vectorialización de los textos. Para procesar automáticamente el lenguaje humano, o lenguaje natural (PLN) (*Natural Language Processing*, NLP), es necesaria una previa transformación de los datos para convertir las “palabras” a un formato capaz de ser traducido por nuestros algoritmos. La vectorización es una de las estrategias con las que contamos para convertir nuestras palabras a formato numérico. En nuestro caso, utilizamos la medida numérica “TF-IDF” que indica la relevancia de una palabra clave en un documento. Cuanto mayor sea el valor TF-IDF, más importante es la palabra clave para el documento. La fórmula está compuesta por dos elementos: la frecuencia de un término (TF), que mide la frecuencia con la que un término específico aparece en un documento. Y la frecuencia inversa del documento (IDF) que mide la frecuencia con la que se utiliza una palabra clave en un conjunto de documentos. TF-IDF es la abreviatura de *Term Frequency Inverse Document Frequency* y es un algoritmo frecuentemente utilizado para transformar el texto en representación numérica (Chaudhary 2020).

En segundo lugar, con los datos vectorizados, se pretendió disminuir la alta dimensionalidad de estos. Nuestros datos presentan una alta dimensionalidad ya que cada palabra es entendida como una variable en cada uno de nuestros documentos (observaciones). Esta alta dimensionalidad se redujo empleando como estrategia un análisis de componentes principales (PCA). El análisis de componentes principales es una herramienta usualmente utilizada en estos casos ya que nos permite resumir un conjunto amplio de variables en un número menor representativo que colectivamente explican la mayoría de la variabilidad en el conjunto original (James y otros 2013). Para esta ocasión se llevó a cabo la reducción manteniendo una varianza del 95%.

En tercer lugar, tras la reducción de la dimensionalidad, aplicamos el algoritmo de aprendizaje no supervisado K-medias. Este es “un enfoque simple y elegante para particionar un conjunto de datos en K grupos distintos, no superpuestos” (James y otros 2013: 386, traducción propia). Usar una estrategia de aprendizaje no supervisado ha implicado la toma de decisión por parte de las investigadoras sobre el número de clústeres en los que se distribuirán los documentos. Para llevar a cabo esta decisión, usamos la distancia euclídea siguiendo la estrategia de Eren y otros (2020). En cuanto a la forma de determinar el número de K (número de conglomerados), se utilizó el método “Elbow” y la función ofrecida por la librería estadítica en Python “kneed” (https://kneed.readthedocs.io/en/stable/) (la más usada para este tipo de casos), resultando en trece el número más óptimo de clústeres para los documentos analizados,de entre las diferentes opciones exploradas.

Una vez decidido el número de conglomerados en los que se clasificarían los documentos, se aplicó el algoritmo K-medias a los datos vectorizados. La aplicación del algoritmo K-medias ha producido como resultado la agrupación de texto, encontrando diferentes patrones en los documentos analizados.

Una vez los documentos fueron clasificados en varios conglomerados, se procedió al modelado de temas (*topic modeling)* en cada uno de ellos. De esta forma, se obtienen las palabras de mayor relevancia en cada grupo de documentos pudiendo identificar más fácilmente el tema o temas del que trata cada conglomerado.El modelado de temas se realizó mediante el algoritmo LDA (*Latent Dirichlet Allocation*). Existen varios algoritmos con el mismo propósito, sin embargo, este es uno de los más usados: “En LDA, cada documento se puede describir mediante una distribución de temas y cada tema se puede describir mediante una distribución de palabras” (Eren y otros 2020, traducción propia). Para este caso específico, se decidió indagar en cada clúster, aplicando un modelo LDA por cada clúster de forma que el algoritmo mostrara las palabras más relevantes de cada uno de los grupos para posteriormente inferir los temas.

Los resultados fueron contrastados con otros procesos de vectorialización como la tokenización por pares o el uso de diferentes grados de n-grams, observando un peor desempeño de los algoritmos LDA directamente aplicados al corpus sin la previa reducción de dimensionalidad o la clusterización de documentos. La comparación de los resultados nos confirmó la necesidad de la reducción previa de la alta dimensionalidad de los textos para la correcta ejecución del modelado de temas.

3. Resultados

3.1. Evolución temporal de documentos y distribución según disciplinas

La producción científica referidaa temas *big data* en su generalidad presenta un interés creciente que año tras año acumula gran cantidad de publicaciones en revistas de impacto. En su mayoría, esta producción está liderada por las áreas de ciencias técnicas, especialmente computación e ingeniería, donde queda contenida la propia naturaleza de los macrodatos. Los primerosartículos que mencionan términos como *large database* o *big data* se pueden encontrar en torno a los años setenta, aunque estos aparecen de forma aislada y discontinua. En ellos se ponende relieve problemas en la gestión y almacenamiento de grandes conjuntos de datos que superan o encuentran problemas para su manejo con los programas informáticos de la época (Wainer y otros 1974).

En nuestra serie de datos solo encontramos 4 artículos entre 1970-1989, cifra que se incrementa a 12 entre 1990-1999 y a 158 ya entre 2000-2009, años a partir de los cuales se incrementa sustancialmente la producción en el campo de los *big data* (gráfico 1), encontrándose publicaciones de manera continuada año a año. Una vez descargadas las publicaciones sobre *big data* en WoS y Scopus, y tras llevar a cabo las tareas de filtrado y limpieza, encontramos que no es hasta 2010 cuando se identifican publicaciones que contienen en su registro palabras como “antropología”, “sociología”, “trabajo social” y “servicios sociales”.

La popularización del término *big data* y su expansión a las diferentes áreas científicas se desarrollará en los años siguientes, sobre todo desde 2012 como se aprecia en el gráfico 1, donde por simplicidad, y dado su carácter anecdótico, no se han recogido los datos de la serie correspondientes a 1970-2009. En la serie global de estos años se produce un incremento de documentos sobre macrodatos sostenido en el tiempo, que es acorde a la evolución de los documentos de sociología, aunque con alguna fluctuación. Un fenómeno que desde 2012 es observable claramente ennuestros datoses el aumento de interés sobre la materia *big data* en todos los campos de conocimiento referidos a las ciencias sociales, que imitan la tendencia de incremento en las publicaciones existentes en el resto de las disciplinas científicas.

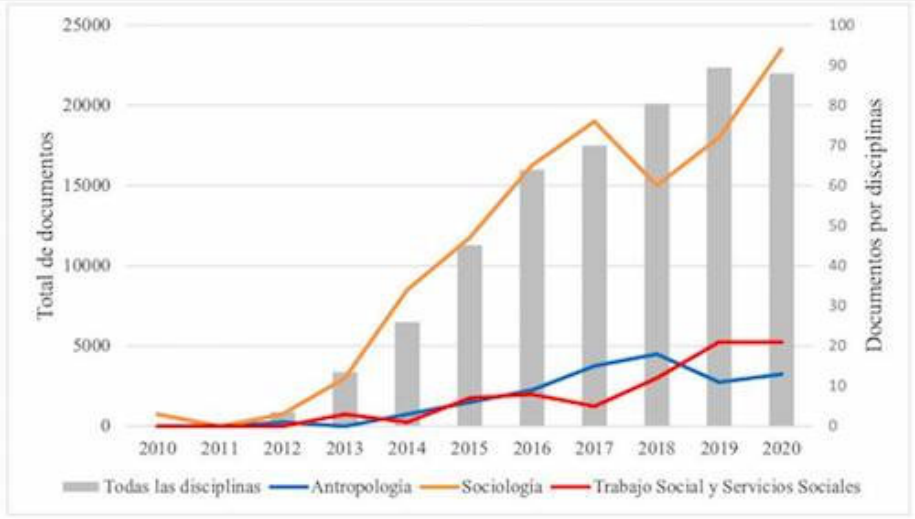


Gráfico 1. Evolución anual de los documentos publicados sobre big data según área científica. La escala de la izquierda representa el total de documentos por año correspondiente a todas las disciplinas (barras). La escala de la derecha representa el total de documentos en las disciplinas de antropología, sociología y trabajo social (líneas). Fuente: Elaboración propia a partir de los datos extraídos de WoS y Scopus (2010-2020).

La aparición de documentos que abordan contenidos relativos a las disciplinas de antropología, sociología y trabajo social y servicios sociales comienza en 2010 para sociología, 2012 para antropología y 2013 para trabajo social. A partir de estas fechas se incrementa de forma sostenida el número de documentos en estas áreas que mencionan en algún momento el término *big data*. Claramente en el caso de sociología el número de documentos es superior frente a las otras dos áreas. Por otra parte, los documentos que estas tres disciplinas aportan a la cifra global relativa al conjunto de documentos es solo del 0,6%, lo que igualmente nos da una pista clara respecto a la dimensión de estas publicaciones en el conjunto de la serie correspondiente a *big data*.

3.2. Documentos sobre *big data*, según las disciplinas de antropología, sociología y trabajo social y servicios sociales

Si bien identificamos inicialmente más de 250.000 documentos sobre *big data* entre 1970-2020 (cuadro 2), tras aplicar el filtrado para extraer aquellos que contuvieran las palabras de antropología, sociología y trabajo social y servicios sociales (apartado 2.2.) y la eliminación de duplicados, solo se obtuvo finalmente un total de 613 documentos, lo cual refleja claramente el escaso peso que aún tienen estas tres disciplinas en el área de estudio de *big data*, lo que se agudiza en el caso de la antropología y el trabajo social y servicios sociales. De los 613 documentos finalmente analizados para este artículo, hay un desequilibrio claro hacia las publicaciones que se relacionan con la sociología (77,4%) frente a aquellas que mencionan trabajo social (12,9%) y antropología (12,6%).

Este resultado no nos parece extraño si tenemos en cuenta la mayor tradición cuantitativa de la sociología, frente a la antropología y el trabajo social, aspecto este que creemos que es importante de cara a una incorporación de los estudios que utilizan macrodatos, por el clarísimo componente estadístico y de programación que hasta ahora ha tenido el análisis de datos o la ciencia de datos.

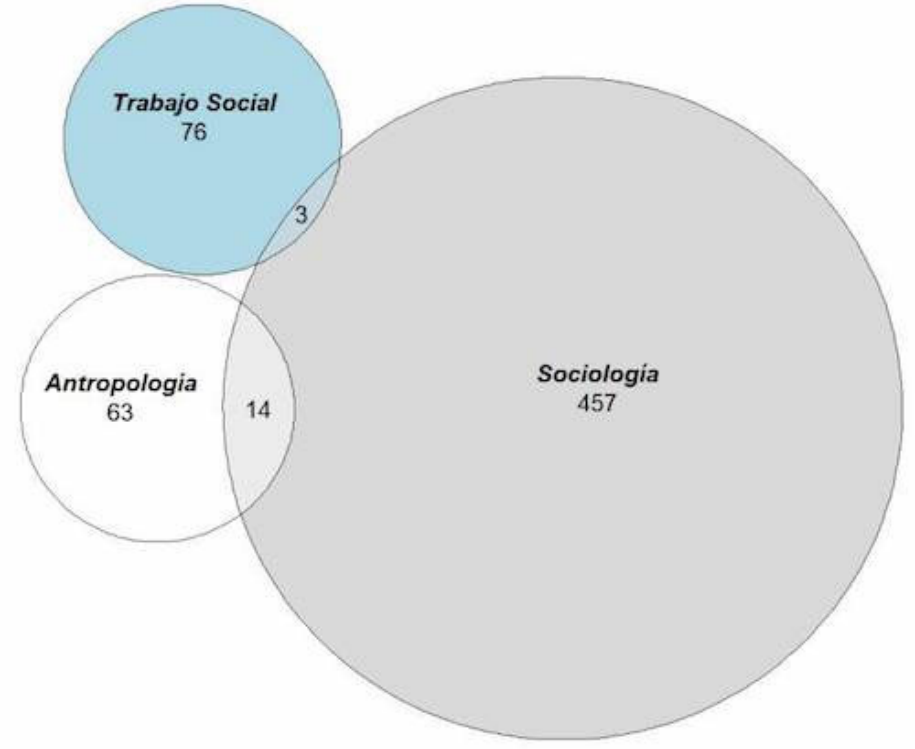


Gráfico 2. Total de documentos según disciplinas. Diagrama de Venn. Número de publicaciones o documentos. En la intersección se encuentran aquellos documentos que fueron clasificados en ambas disciplinas al aplicar el script descrito en 2.2. para seleccionar esta submuestra. Fuente: Elaboración propia con el paquete "eulerr" de R a partir de los datos extraídos de WoS y Scopus (2010-2020).

3.3. Documentos sobre *big data* y antropología, sociología y trabajo social, según su tipo

Por otra parte, de acuerdo con los datos recogidos en el gráfico 3, que contiene la clasificación de nuestros documentos en función de su tipo, según los criterios de WoS y Scopus, predominan dos tipos de documentos: los artículos científicos (66%) y las comunicaciones a congresos (20%). No obstante, se han identificado documentos que versan sobre *big data* y antropología, sociología y trabajo social en otros tipos como, por ejemplo, libros o capítulos de libros (8%). Se apreció igualmente que las publicaciones de artículos son proporcionalmente más frecuentes en antropología y sociología, mientras que comparativamente en trabajo social destacan las comunicaciones en congresos, que tienen mucha menor importancia proporcional en las otras dos áreas. No nos resulta especialmente extraño este resultado, debido a la mayor orientación a la intervención social de esta disciplina.

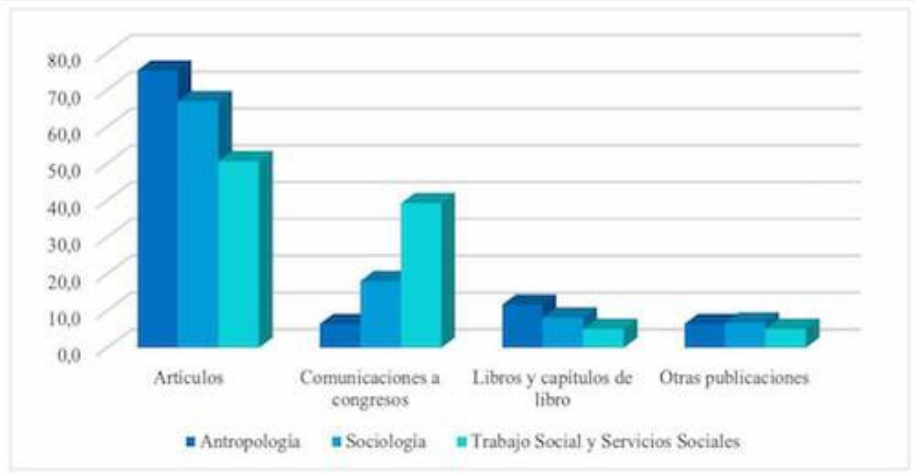


Gráfico 3. Tipo de documentos sobre big data y antropología, sociología y trabajo social (%). En "Otras publicaciones" se agregan notas, material editorial, cartas y reseñas de libros. Fuente: Elaboración propia a partir de los datos extraídos de WoS y Scopus (1970-2020).

3.4. Documentos según categorías científicas

En la tabla 1 se recoge una descripción sintética de las categorías científicas en las que se clasifican los artículos de antropología, sociología y trabajo social y servicios sociales, de forma que pueda compararse en qué categorías encontramos en mayor o menor medida cada disciplina. Los criterios de clasificación de categorías científicas no son los mismos en WoS o Scopus. Para este trabajo, la clasificación de categorías científicas se llevó a cabo según el criterio ofrecido por *Journal Citation Report* (JCR) (https://jcr.clarivate.com/) cuando clasifica a cada revista en categorías diferentes, que a su vez se clasifican en grandes grupos. Se consultaron los títulos de las revistas de cada documento para etiquetar la categoría científica de acuerdo con JCR correspondiente a cada uno de ellos de forma manual. En los casos de las publicaciones referidas a actas de congresos, libros y otros tipos de documentos donde no podía aplicarse como criterio el título de la revista, se hizo una búsqueda de la fuente para clasificarlo bajo las categorías presentes en JCR.

La tabla que sigue, que recoge los grupos principales de categorías científicas según JCR en que se encuentran las publicaciones, revela varios aspectos importantes. Por una parte, tanto en antropología como en sociología son más importantes las publicaciones que se enmarcan dentro la categoría científica correspondiente a sus propias disciplinas (11,7% y 18,8% respectivamente). En cambio, en el trabajo social, aunque un 6,3% de las publicaciones se encuadran en la categoría de trabajo social, destacan las publicaciones que se ubican en categorías clasificadas por JCR en grupos multi e interdisciplinares (hasta un 62,0% en trabajo social frente a un 55,1% en sociología y a 48,1% en antropología). Cuando se aborda la cuestión de los *big data*, la importancia de publicar en revistas clasificadas en categorías interdisciplinares o multidisciplinares es un factor identificativo de las tres disciplinas, lo cual nos presenta un elemento clave de las publicaciones que se producen en el área de *big data.*

	Antropología	Sociología	Trabajo Social y Servicios Sociales	Total
Anthropology	11,7%	0,2%	0,0%	1,5%
Sociology	3,9%	18,8%	2,5%	14,8%
Social Work	0,0%	0,0%	6,3%	0,8%
Multi e Interdisciplinary Sciences	48,1%	55,1%	62,0%	55,5%
Social Sciences	11,7%	7,4%	1,3%	7,0%
Economics & Business	6,5%	4,9%	6,3%	5,2%
Clinical Medicine	6,5%	3,8%	10,1%	4,9%
Computer Science	1,3%	2,5%	2,5%	2,4%
Psychology	5,2%	1,3%	1,3%	1,6%
Engineering	0,0%	1,1%	3,8%	1,3%
Biology & Biochemistry	1,3%	1,5%	0,0%	1,3%
Mathematics	1,3%	0,0%	2,5%	1,1%
Environment/ Ecology	0,0%	1,3%	0,0%	1,0%
Philosophy and Religion	1,3%	0,6%	0,0%	0,7%
Otras	1,3%	1,7%	1,3%	0,8%
Total	100,0%	100,0%	100,0%	100,0%

Tabla 1. Documentos sobre big data y antropología, sociología y trabajo social y servicios sociales, según categorías y grupos en JCR (porcentajes de columna). n=613 documentos. Se han extraído del grupo de “Social Sciences” las categorías de antropología, sociología y trabajo social y servicios sociales para mostrar su peso al inicio de la tabla. Fuente: Elaboración propia a partir de los datos extraídos de WoS y Scopus (1970-2020).

Gran parte de los trabajos que enlazan *big data* con antropología, sociología y trabajo social son de carácter mixto. No obstante, dentro de ese amplio grupo que comprende a categorías inter y multidisciplinares, se destacan algunas categorías correspondientes a revistas clasificadas como: “Interdisciplinary Applications Computer Sciences”, “Interdisciplinary Social Sciences”, “Multidisciplinary Sciences” o “Information Systems Computer Sciences”, siendo de gran relevancia las que tienen que ver con las ciencias computacionales o la ingeniería. Y, con menor relevancia cuantitativa, algunas como: “Education and Educational Research”, “Information and Library Science”, u otras.

Por otra parte, se aprecia igualmente la importancia que tienen las publicaciones en estas disciplinas relativas a *big data* en grupos de categorías científicas como, por ejemplo, las ciencias sociales (sobre todo en antropología y sociología), "clinical medicine" (especialmente en antropología y trabajo social), así como en otros grupos como "economic & business", "psychology" y otros de menor relevancia.

3.5. Clasificación y análisis de temáticas (K-means y LDA)

El segundo objetivo que se proponía este trabajo era el de identificar los temas que destacan en las publicaciones de las disciplinas de antropología, sociología y trabajo social y servicios sociales cuando se asocian o aparecen en investigaciones sobre *big data*. Para ello, se utilizaron técnicas de procesamiento de lenguaje natural (NLP) que permiten encontrar similitudes en las palabras usadas en las publicaciones, de forma que pudiéramos clasificarlas y conocer las temáticas más relevantes que se observan en los textos. Los algoritmos de análisis textual proporcionan este tipo de ventajas (Bird y otros 2009). Las palabras son convertidas a un "lenguaje" comprensible por el algoritmo y a través del cálculo de distancia en la red gramatical, nos proponen agrupaciones en los textos. Como resultado, se extraen las palabras o grupos de palabras más destacadas en cada conglomerado, lo cual ayuda a comprender de qué tratan los documentos o bibliografía que estamos estudiando, así como a identificar patrones. A través del aprendizaje automático no supervisado y el modelado de temas, pudimos identificar que la literatura existente sobre *big data* y las disciplinas aquí estudiadas, se pueden clasificar en 13 tópicos diferentes (más detalles en la sección metodológica 2.4.2.), denominando cada clúster a partir de las palabras o contenidos más identificativos. Se observó también que existen temas transversales que están presentes en todas las agrupaciones. Los 13 temas o tópicos emergentes más destacados en la literatura analizada tras este análisis fueron agrupados, para la exposición que sigue, en cinco grandes bloques por la proximidad de sus temas. La importancia de cada una de estas temáticas en las publicaciones es representada por el número de documentos agrupados en cada uno de los clústeres.

3.5.1. Redes sociales (217 documentos, 35,3% del total)

El clúster 1, "redes sociales", se presenta como el tema más relevante dentro de los estudios de *big data* relacionados con las disciplinas analizadas, con un total de 217 documentos. Un 35,3% de los textos son agrupados bajo esta temática cuando se aplica el modelado de temas. Las publicaciones abordan tres dimensiones principalmente. Por un lado, destacan los artículos altamente críticos con el efecto social de aplicaciones móviles y redes sociales y el funcionamiento de sus algoritmos, así como los problemas de vigilancia y privacidad, transversal a la mayor parte de temáticas. En segundo lugar, se incluyen estudios sobre redes sociales y comunidades desde la óptica de la teoría sociológica y antropológica, donde se promueven mejoras en el análisis de redes y grafos a partir de paquetes de software y modelos computacionales. Estos últimos trabajos se encuentran relacionados además con la tercera tipología de estudios más enfocados en el análisis de opinión, sentimientos y estudios de comunicación en las redes sociales, con la predominancia de Twitter como red social más mencionada.

3.5.2. Big data aplicado a retos y problemáticas de la realidad social (178 documentos, 29% del total)

a) Big data y sociedad (10,8% del total)

El segundo grupo de temáticas (clúster 2) con más relevancia en los textos analizados lo encontramos en relación con el uso de técnicas *big data* durante el proceso de investigación de diversos objetos de estudio de gran importancia para las ciencias sociales. En este grupo, se encuentra una variedad de artículos referidos a casos tan diversos como epidemias, turismo, protección infantil, tendencias culturales, gobernanza medioambiental, agroalimentación, o refugiados, entre otros. Sin embargo, cuatro de los objetos de estudios que destacan, sobresaliendo por sí mismos en los clústeres, son los temas de salud, educación, *smart cities* y urbanismo.

	Clústeres	Nº de artículos	% artículos	Palabras clave (LDA), por asignación latente de Dirichlet	% de artículos de Antropología en el clúster	% de artículos de Sociología en el clúster	% de artículos de Trabajo social en el clúster
Redes sociales	1. Redes sociales	217	35,4%	"social_media", "research", "media", "social", "analysis", "study", "sociology", "digital"	45,5%	35,7%	25,3%
Big data aplicado a retos y problemáticas sociales	2. Big data y sociedad	66	10,8%	"research", "sociology", "analytics", "analysis", "new", "study", "social", "global", "information"	16,8%	11%	5,1%
	3. Salud	52	8,5%	"health","research", "health_care", "care", "medical", "systems", "sociology", "patient", "article", "healthcare	6,5%	7,4%	17,7%
	4. Educación	24	3,9%	"education", "science", "students", "new", "social", "development", "paper", "research", "sociology", "educational"	0%	4%	6,3%
	5. Smart cities	23	3,8%	"smart_city", "cities", "ai", "city", "security", "research", "systems", "privacy", "smart_cities", "technologies"	2,6%	3,2%	10,1%
	6. Urbanismo	13	2,1%	"urban", "new", "city", "research", "urban_computing", "urban_youth", "gentrification", "social", "paper", "sociology"	2,6%	2,3%	0%
Teoría, epistemología e implicaciones de los *big data*	7. Epistemología, metodología e implicaciones del *big data*	95	15,5%	"research", "analysis", "social", "new", "based", "sociology", "also", "science", "networks", "approach"	14,3%	17,1%	5,1%
	8. Sociología digital	10	1,6%	"digital", "sociology", "social", "also", "research","technologies", "studies", "new", "digital_sociology", "authors"	2,6%	1,9%	0%

Cuadro 3. Clasificación y análisis de temáticas (K-means y LDA).
Las palabras claves de cada clúster son el resultado de la aplicación del algoritmo de asignación latente de Dirichlet (LDA). Este algoritmo selecciona estas palabras como las más relevantes de cada conjunto de documentos. Latent Dirichlet Allocation (LDA) = Asignación latente de Dirichlet.

	Clústeres	N° de artículos	% artículos	Palabras clave (LDA), por asignación latente de Dirichlet	% de artículos de Antropología en el clúster	% de artículos de Sociología en el clúster	% de artículos de Trabajo social en el clúster
Ingeniería e internet de las cosas orientadas a toma de decisiones y resolución de problemas	9. Toma de decisiones	26	4,2%	"research", "administrative", "decision_support", "dss", "services", "sociology", "researchers", "new", "science", "systems"	1,3%	2,7%	15,2%
	10. Soluciones de internet de las cosas (IoT)	25	4,1%	"systems", "large_scale", "traffic","based", "sociology", "sets", "new", "planning", "users", "mobile_phone"	1,3%	4,4%	3,8%
	11. Ingeniería	10	1,63%	"higher_education", "phones_tv","extends_virtual", "sciences_published", "branch_belongs", "engineering_devices", "continuous_training", "robotics_mechatronics", "intelligence_co", "electronic_components"	0,0%	1,7%	2,5%
Modelos estadísticos computacionales	12. *Machine learning*	38	6,2%	"algorithms","sociology","applications", "clustering", "prediction", "machine_learning", "time_series", "sampling", "based", "two"	3,9%	6,1%	8,9%
	13. Análisis de texto	14	2,3%	"topic_modeling", "visualization", "black_women", "communities", "interaction", "text_mining", "analyzing_large", "mining", "used", "labor"	2,6%	2,5%	0%
Total		613	100%		100%	100%	100%

Continuación cuadro 3. Clasificación y análisis de temáticas (K-means y LDA)

b) Salud (8,5% del total)

El clúster 3, que hemos denominado “Salud”, reúne al 8,4% del total de los documentos estudiados, lo que expresa claramente la importancia de los trabajos sobre *big data* y salud en las tres disciplinas y, especialmente, en trabajo social (cuadro 1). Estos textos recogen experiencias en la investigación de patologías y problemas de salud específicos en múltiples áreas como la neurociencia o la salud mental, incorporando macrodatos y modelos estadísticos a través de la computación. En muchos casos, se discute sobre la mejora de tecnología relativa a la internet de las cosas (IoT). Si bien, los enfoques que predominan son aquellos que relacionan el estado de la salud en la población y sus dimensiones, retos o problemas sociales.

c) Educación (3,9% del total)

Los documentos relacionados con educación (clúster 4) se definen por la discusión de dos líneas principales. En primer lugar, la crítica al determinismo de los datos en los procesos de gobernanza y políticas referidas a la educación, donde se examinan las plataformas educativas y exigencias estadísticas de resultados. En ocasiones, sin embargo, se valora positivamente el uso de macrodatos como solución eficiente a problemas estructurales propios en la gestión educativa tales como la preparación de planes de estudio o gestión de proyectos. Por otra parte, hay publicaciones que exponen la carencia en los programas educativos de ciencias sociales en la formación de técnicas cuantitativas y habilidades computacionales, lo que en muchos casos dificultan el trabajo interdisciplinar y el avance metodológico de estas disciplinas.

d) *Smart cities* (3,8% del total)

Los artículos clasificados bajo el tópico de ciudades inteligentes (clúster 5) se refieren a un objeto de estudio específico donde los *big data* juegan un papel importante en la planificación territorial, así como en el estudio de la movilidad, el desarrollo de ciudades inteligentes y los efectos tanto positivos como negativos que en ellos podemos encontrar. Los documentos clasificados en el tema *smart cities*, reúnen publicaciones que analizan las ciudades inteligentes desde una óptica amplia. Abordan la dimensión social y sus implicaciones en la distribución del espacio, la vigilancia, mejora del medioambiente y gestión de los datos e internet de las cosas, además de promover soluciones novedosas para el uso de datos enfocado en la mejora de la vida cotidiana en la ciudad a través de la inteligencia artificial.

e) Urbanismo (2,1% del total)

En este caso, aunque la temática se pueda ver relacionada con el concepto de *smart cities* presentado en el clúster anterior, los documentos sobre urbanismo (clúster 6) tienen su especificidad y se diferencian en que mantienen una óptica de sociología urbana como enfoque predominante. En ellos se abordan nuevas metodologías computacionales para el estudio urbano, de regionalización y gentrificación, así como el efecto de las tecnologías relativas a la internet de las cosas en los planeamientos urbanísticos y su implicación en la articulación de procesos participativos.

3.5.3. Teoría, epistemología e implicaciones de los *big data* (105 documentos, 17% del total)

Otra parte de los textos analizados se caracterizan por el enfoque teórico que adquieren los documentos, centrados en discusiones epistemológicas sobre la conceptualización del término *big data*o las implicaciones que tiene el uso de los macrodatos en investigaciones de ciencias sociales. Encontramos dos conglomerados principales bajo este grupo. El clúster 7 es relevante en las tres disciplinas, el clúster 8 en cambio está muy enfocado en la “sociología digital”.

a) Epistemología, metodología e implicaciones del *big data* (15,5% del total)

El clúster 7 agrupa un 15,4% de todas las publicaciones analizadas, siendo la segunda temática más importante cuando relacionamos *big data* con antropología, sociología y trabajo social y servicios sociales. Estos documentos atienden a un enfoque teórico orientado al estudio de cómo se produce el conocimiento científico a través de macrodatos. Se discuten aspectos relacionados con el efecto epistemológico que conlleva el uso de datos masivosen investigaciones de carácter social, indagando en el significado del concepto *big data,* las posibilidades que este tipo de herramientas ofrecen y también, las posibles consecuencias y efecto perverso que puede conllevar su uso aplicado a fenómenos sociales. Igualmente, presentan reflexiones con diversos planteamientos críticos sobre los cambios en la metodología, indagando en el papel del investigador y en las transformaciones epistemológicas y ontológicas que sufrirá tanto el proceso de investigación como la naturaleza del objeto de estudio. Estas discusiones, sin embargo, se plantean de forma transversal en muchos de los documentos analizados, aunque los textos agrupados bajo este clúster tratan estas temáticas de forma central y aparecen como objetivo específico de estudio.

b) Sociología digital (1,6% del total)

El clúster 8 recoge igualmente, documentos de corte fundamentalmente teórico, que exponen de forma crítica aspectos relacionados muy específicamentecon la sociología digital. Destaca la preocupación por el efecto holístico de las sociedades digitales, incurriendo en la brecha digital o en aspectos como el determinismo cuantitativista que supone la traducción de todos los fenómenos sociales en datos y el incremento de la vigilancia en la ciudadanía. Es común en estos documentos la pregunta sobre el significado y descripción del concepto *big data,* así como el efecto metodológico que conlleva su uso por parte de la investigación en sociología de forma similar al clúster anterior.

3.5.4. Ingeniería e internet de las cosas orientadas a toma de decisiones y resolución de problemas (60 documentos, 9,7% del total)

Bajo esta agrupación encontramos investigaciones que indagan sobre el uso de macrodatos como herramienta para la mejora en la toma de decisiones y la gestión o administración de instituciones. Estas publicaciones destacan por ser de carácter técnico, desarrollando tecnologías relacionadas con la ingeniería y el internet de las cosas orientado a problemas de carácter social, posibilitando, además, soluciones novedosas que potencian la interdisciplinariedad entre ciencias técnicas como la ingeniería y otras de carácter social como el trabajo social -esta última, muy presente en estos artículos (el 21,5% de

los documentos referidos a trabajo social y servicios sociales se encuentran en esta agrupación).

a) Toma de decisiones (4,2% del total)

En el conglomerado 9 encontramos documentos relacionados con trabajo social y servicios sociales -acumulando el 15,2% de los documentos que mencionan esta disciplina-. La temática se relaciona con el uso de macrodatos para el desarrollo de sistemas de apoyo en la toma de decisiones en áreas que competen a trabajo social, analizando la intervención de este tipo de herramientas en procesos como, por ejemplo, la identificación de niños vulnerables para introducirlos en programas de prevención del maltrato (Gillingham, 2019). Del mismo modo, se argumenta sobre la relevancia que adquieren los datos administrativos en el entorno de los *big data* para áreas como la justicia penal o la administración pública. Es transversal a los documentos de este clúster además, la reflexión sobre la importancia que han de adquirir losmacrodatos en la toma de decisiones, ya que se plantea la necesidad de la intervención del criterio humano y su predominancia a la hora de determinar una decisión, debido a la posibilidad de sesgos en este tipo de datos.

b) Soluciones de internet de las cosas (4,1% del total)

Los documentos del clúster 10 articulan nuevas propuestas de carácter metodológico y técnico para el estudio de fenómenos digitales, principalmente de movilidad y tráfico, a través de datos producidos por sensores y otros elementos propios de la internet de las cosas. Estas propuestas se presentan como novedosas y exponen formas experimentales de análisis y modelado de datos para llegar a soluciones creativas de problemas sociales. Todos los documentos además presentan una fuerte intención interdisciplinar, llamando a disciplinas como la sociología o el trabajo social para la mejora en la comprensión de estos modelos y la interpretación de los fenómenos estudiados.

c) Ingeniería (1,6% del total)

Aunque reúne pocos documentos, en este conglomerado 11 se resalta la articulación de la ingeniería con las disciplinas de sociología y trabajo social y servicios sociales. Las publicaciones clasificadas en este grupo discuten sobre el potencial de las tecnologías de telecomunicación o ingeniería en el desarrollo social, educacional, sociológico, etc.

3.5.5. Modelos estadísticos computacionales (52 documentos, 8,4% del total)

Se han agrupado aquí dos clústeres que recogen documentos con investigaciones que proponen la experimentación con modelos estadísticos computacionales para su mejora, así como la intención de poder ser aplicados en investigaciones de carácter social. Destaca claramente en este conjunto la sociología, estando presente en el 80% de los documentos agrupados en este clúster.

a) *Machine learning* (6,2% del total)

Los documentos agrupados en el clúster 12 atienden a un enfoque metodológico donde seproponen mejoras en modelos computacionales (algoritmos), principalmente de *machine learning*, destinados a su aplicación en estudios de fenómenos sociales. Este tipo de investigaciones promueven el conocimiento y desarrollo metodológico en áreas más alejadas a las ciencias computacionales o técnicas, permitiendo que científicos sociales incorporen esta tecnología a sus análisis.

b) Análisis de textos (2,3%)

El clúster 13, con poca importancia cuantitativa, solo lo encontramos presente en relación con la antropología y la sociología. Los documentos de este grupo investigan modelos computacionales enfocados al procesamiento de lenguaje natural (lenguaje humano), ensayando mejoras en los algoritmos para el modelado de temas. Además, desarrollan teorías sobre comunicación y análisis cultural que dan soporte al uso de metodologías de aprendizaje automático, permitiendo la comprensión del lenguaje humano de una forma más compleja y completa y, por tanto, la mejora en los algoritmos para su aplicación en investigaciones de carácter social.

4. Discusión y conclusiones

Una de las áreas de investigación que se han introducido con un interés creciente en el debate científico global y particularmente en las ciencias sociales en los últimos lustros ha sido la correspondiente al estudio de los *big data* o macrodatos. La bibliografía sobre *big data* ha adquirido relevancia internacional y se ha expandido en diferentes ramas científicas sobre todo en el siglo XXI, muy especialmente desde 2012 de acuerdo con nuestros datos. Este artículo evidencia cómo la incorporación de este campo de estudio ha sido desigual en diferentes disciplinas, así como, en el caso de las disciplinas en las que nos hemos enfocado, revela que se ha incorporado con mayor intensidad en sociología, que en antropología y trabajo social y servicios sociales. Aunque se trata de tres disciplinas próximas entre sí, sus propias tradiciones y especificidades, pueden ayudarnos a entender esta diversa evolución. Es bastante plausible pensar que la extensa tradición cuantitativa de la primera frente a un mayor acento en estudios cualitativos o en la intervención social en la segunda y tercera, respectivamente, hayan marcado esta diferente evolución. No obstante, las publicaciones que hemos encontrado donde se produce una conexión entre *big data* y estas tres disciplinas ocupan solo el 0,6% del conjunto de publicaciones sobre *big data* en WoS y Scopus (cuadro 2), mientras que otras áreas en las ciencias sociales acumulan hasta ahora más publicaciones. En el conjunto, se trata de las ciencias computacionales o las ingenierías las que más publicaciones tienen en su haber en conexión a los macrodatos. Por otra parte, si bien parece que en el caso de la sociología los estudios asociados a los macrodatos siguen una tendencia creciente, esto no es tan evidente en el caso de la antropología y el trabajo social que siguen un ritmo más discreto que hace dudar de si el enfoque de *big data* vaya a ser en un inmediato futuro un enfoque teórico o metodológico de relevancia para estas disciplinas.

La agrupación de artículos a través de la estrategia de análisis de aprendizaje no supervisado y modelado de temas permitió encontrar una serie de patrones en los documentos que nos ha llevado a una mejor comprensión de los aspectos que más atención han captado en estas disciplinas en conexión con los *big data*. Particularmente importantes son las publicaciones agrupadas en los clústeres denominados “redes sociales” (35,3%), “epistemología, metodología e implicaciones del *big data”* (15,5%), “*big data* y sociedad” (10,8%), “salud” (8,5%) y “*machine learning”* (6,2%). Estos conglomerados agrupan algo más del 75 % de los documentos analizados. Aunque su trabajo no se basa en las mismas fuentes que el nuestro, en la medida en que la fuente de datos es solo Scopus y la perspectiva bibliométrica, nuestra clasificación presenta similitudes con la realizada por Becerra y Ratovicius (2022) sobre literatura en ciencias sociales, psicología y humanidades centrada en big data, coincidiendo en la relevancia de la cuestión metodológica y epistemológica y el estudio de asuntos sociales como las *smart cities*, el urbanismo o la educación. El orden de importancia hemos visto, no obstante, que es desigual según disciplinas (cuadro 3). Quizás entre los aspectos distintivos cabe destacar la gran importancia que el clúster de “redes sociales”, siendo importante para las tres disciplinas, tiene en la antropología, con casi la mitad de sus publicaciones, y la menor importancia de bloques más técnicos o estadísticos. En trabajo social dos conglomerados que destacan frente al resto son, en cambio, los de salud y toma de decisiones, mientras que en sociología la carga que tienen clústeres ligados a aspectos más técnicos y de ingeniería o matemáticos y estadísticos destacan.

Por otra parte, el análisis de la bibliografía sobre *big data* en estos campos permite observar la transversalidad de algunas cuestiones, que aparecen en los diferentes tópicos identificados. De esta forma, aunque los documentos se clasificaron vía LDA en trece clústeres únicos, es común que los textos compartan varias temáticas. Por ejemplo, en la mayor parte de los conjuntos analizados, encontramos presentes preocupaciones de carácter transversal que incurren en las siguientes discusiones: vagueza en la conceptualización del término *big data*, preocupación por la privacidad y protección de datos, e incertidumbre en el efecto epistemológico que conlleva la aplicación de metodologías basadas en los *big data* en los estudios relacionados con ciencias sociales.

Igualmente, nuestros datos sugierenclaramente que cuando se encuentran publicaciones que versan sobre *big data* y alguna de estas tres disciplinas de las ciencias sociales, si bien la tradición de publicar en una revista correspondiente a la propia disciplina se mantiene en ocasiones, globalmente el abanico de revistas donde se publica sobre *big data* es mayor y las fronteras disciplinares se diluyen cuando se trata esta temática, algo por otra parte coherente con el desarrollo de este campo de estudio. La inter y multidisciplinariedad es, probablemente, uno de los elementos más característicos de cualquier estudio que atienda a macrodatos yciencias sociales.

Los retos metodológicos refuerzan la necesidad de buscar alianzas intelectuales y técnicas con otras áreas de conocimiento como las ciencias computacionales. El trabajo con *big data* requiere metodologías sólidas y marcos teóricos de las ciencias sociales, pero, a la vez, una aproximación exclusivamente

antropológica o sociológica puede ser insuficiente. Los macrodatos requieren habilidades y enfoques que, a veces, trascienden las disciplinas (Ruppert 2013), por ello ofrecen una oportunidad para el desarrollo de métodos interdisciplinarios. La asociación entre científicos sociales y científicos de datos y los enfoques híbridos y colaborativos son por tanto una respuesta ante estos retos que muchos investigadores ya están apoyando (Nardulli y otros 2015, Tinati y otros 2014, Gillingham y Graham 2017). Como narra Erikson (2018), durante la crisis sanitaria del Ébola en África, se confió en exceso en los datos de registro de llamadas de teléfonos móviles para detener la propagación de la enfermedad, pero fue el trabajo de los antropólogos el que dio las claves para entender las especificidades culturales que promueven el contagio (obligación social de cuidar de los enfermos o la preparación del cuerpo del fallecido para el entierro). Con esto queremos ejemplificar que, aunque los científicos sociales necesitemos los conocimientos avanzados en técnicas de *big data* de los científicos de datos, también los algoritmos necesitan la visión social de los científicos sociales, de ahí la importancia de un trabajo conjunto entre disciplinas. Junto al establecimiento de alianzas para conseguir equipos interdisciplinares, algunas investigaciones ponen de manifiesto la necesidad de una introducir o mejorar la formación metodológica especializada en *big data* en las carreras de ciencias sociales, tanto en la manipulación de estos grandes conjuntos de datos como en el análisis de estos, incluyendo las herramientas de aprendizaje automático (Mützel 2015). De esta manera, los futuros investigadores podrían contar con conocimientos básicos sobre programación que les permitan tener una visión más clara de los procedimientos para adquirir, procesar y analizar los datos (Halavais 2015, Edelmann y otros 2020, Evans y Foster 2019, Gualda y Rebollo 2020, Gualda 2022).

Se ha mostrado en las páginas precedentes cómo una parte sustancial de la bibliografía existente que conecta *big data* con antropología, sociología y trabajo social se enfoca en estudios de corte cuantitativo, aunque igualmente se encuentran trabajos de reflexión teórica, epistemológica y ética. Igualmente, no escasean trabajos centrados en enfoques mixtos o cualitativos que nos parece aportan una línea muy prometedora de futuro no solo para la sociología, sino especialmente para la antropología y el trabajo social (Nikunen 2021, Andreotta y otros 2019). Especialmente si se quieren abordar tareas clásicas de recolección y análisis e interpretación cualitativa de discursos, textos, palabras, etc. a partir de nuevas fuentes digitales o documentos publicados en Internet (Bonacchi y otros 2018). Son varios los autores que coinciden en las ventajas del uso de técnicas *big data* para llevar a cabo investigaciones de carácter cualitativo, subrayando otros beneficios como en el abaratamiento de los costes y el uso eficaz del tiempo para llevar a cabo estas tareas. De esta forma, se confeccionan nuevas propuestas que puedan dar un soporte teórico y un marco metodológico consistente a este tipo de prácticas (Rose y Lennerholt 2017).

Igualmente, el uso de técnicas de datos masivos es útil si se pretende llevar a cabo una nueva mirada de documentos clásicos, con el apoyo de técnicas de análisis de macrodatos que permiten manejar mayores volúmenes de datos, como las vinculadas al análisis de texto, el aprendizaje automático y otras. Este tipo de enfoques son necesarios para la aproximación a investigaciones que estudien datos originados en Internet o que se enfoquen en el análisis de los *social big data* (Gualda 2022). Son importantes igualmente para dar respuesta a nuevas preguntas de investigación o, incluso, para llevar a cabo una primera selección o filtrado de documentos con criterios sistemáticos cuando es inabordable llevar un análisis exhaustivo de muestras muy amplias. De esta forma, gracias al avance en el procesamiento computacional del lenguaje natural, por ejemplo, grandes corpus de textos pueden ser modelados de forma que se encuentren temáticas predominantes en ellos (Berger y Packard 2021). Igualmente, es posible la automatización de tareas como la clasificación o etiquetación de los datos a través de técnicas de *machine learning*, permitiendo abarcar una mayor cantidad de datos que si se realizara de forma manual (Gui y otros 2020), como se ha llevado a cabo en este artículo.

Si bien nuestra búsqueda y análisis se ha limitado a extraer las publicaciones que contenían la cadena exacta “big data” y a partir de aquí se han extraído una submuestra de documentos que contenían los términos de “antropología”, “sociología”, “trabajo social” y “servicios sociales”, dejando fuera otras posibles palabras clave asociadas con estas tres disciplinas, el grado de innovación tan importante que se encuentra en algunos de los textos analizados de cara a las investigaciones en Internet y a abrir escenarios donde se pueda profundizar en investigaciones inter y multidisciplinares, así como la combinación, son un incentivo para seguir avanzando en este camino.

Notas

Esta publicación ha sido posible gracias al apoyo del Ministerio de Universidades que ha beneficiado a Carolina Rebollo con una Ayuda Margarita Salas para la formación de jóvenes doctores incluidas en las Ayudas para la recualificación del Sistema Universitario Español 2021-2023, financiadas por la Unión Europea- NextGenerationEU. Agradecemos también el apoyo del grupo de investigación "Estudios Sociales e Intervención Social" (ESEIS) y del centro de investigación de "Pensamiento Contemporáneo e Innovación para el Desarrollo Social" (COIDESO) para la realización de este artículo.

Bibliografía

Abramson, Corey M. (y otros)
2018 "The promises of computational ethnography: improving transparency, replicability, and validity for realist approaches to ethnographic analysis", *Ethnography*, nº 19 (2): 254-284. https://doi.org/10.1177/1466138117725340

Ahajjam, Sara (y otros)
2018 "A new scalable leader-community detection approach for community detection in social networks", *Social Networks*, nº 54: 41-49. https://doi.org/10.1016/j.socnet.2017.11.004

Alvard, Michael (y David Carlson)
2020 "Identifying patch types using movement data from artisanal fishers from the commonwealth of dominica", *Current Anthropology*, nº 61 (3): 380-387.
https://www.journals.uchicago.edu/doi/abs/10.1086/708720?journalCode=ca

Andreotta, Matthew (y otros)
2019 "Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis", *Behavior Research Methods*, nº 51 (4): 1766-1781.
https://doi.org/10.3758/s13428-019-01202-8

Bartosik-Purgat, Malgorzata (y Milena Ratajczak-Mrożek)
2018 "Big data analysis as a source of companies' competitive advantage: A review", *Entrepreneurial Business and Economics Review,* nº 6 (4): 197. https://doi.org/10.15678/EBER.2018.060411

Becerra, Gastón (y Cristian Rotovicius)
2022 "Social Sciences and Humanities on Big Data: a Bibliometric Analysis", *Journal of Information Systems and Technology Management*, nº 19: e202219011. https://doi.org/10.4301/s1807-177520221901

Bello-Orgaz, Gema (y otros)
2016 "Social big data: Recent achievements and new challenges", *Information Fusion*, nº 28: 45-59. https://doi.org/10.1016/j.inffus.2015.08.005

Berger, Jonah (y Grant Packard)
2021 "Using Natural Language Processing to Understand People and Culture", *American Psychologist*, Advance online publication, nº 1: 13. http://dx.doi.org/10.1037/amp0000882

Beyer, Mark (y Douglas Laney)
2012 "The Importance of "Big Data": A Definition", *Gartner Research*.
https://www.gartner.com/en/documents/2057415/the-importance-of-big-data-a-definition

Bird, Steven (y otros)
2009 *Natural language processing with Python*. O'Reilly.

Bonacchi, Chiara (y otros)
2018 "The heritage of Brexit: Roles of the past in the construction of political identities through social media", *Journal of Social Archaeology*, nº 18 (2): 174-192. https://doi.org/10.1177/1469605318759713

Brownlie, Julie (y Frances Shaw)
2019 "Empathy rituals: Small conversations about emotional distress on Twitter", *Sociology*, nº 53 (1):

104-122. https://doi.org/10.1177/0038038518767075

Bulger, Monica (y otros)
2014 *Engaging Complexity: Challenges and Opportunities of Big Data*. London, NEMDOE.

Caplan, Mary A. (y Gregory Purser)
2019 "Qualitative inquiry using social media: A field-tested example", *Qualitative Social Work*, nº 18 (3): 417-435. https://doi.org/10.1177/1473325017725802

Chaudhary, Mukesh
2020 "TF-IDF Vectorizer scikit-learn". https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a

Chung, Jae Young (y Sunbok Lee)
2019 "Dropout early warning systems for high school students using machine learning", *Children and Youth Services Review*, nº 96: 346-353. https://doi.org/10.1016/j.childyouth.2018.11.030

Del Vecchio, Pasquale (y otros)
2018 "Creating value from Social Big Data: Implications for Smart Tourism Destinations", *Information Processing & Management*, nº 54 (5): 847-860. https://doi.org/10.1016/j.ipm.2017.10.006

Edelmann, Achim (y otros)
2020 "Computational Social Science and Sociology", *Annual Review of Sociology*, nº 46. https://doi.org/10.1146/annurev-soc-121919-054621

Elgin, Dallas J.
2018 "Utilizing predictive modeling to enhance policy and practice through improved identification of at-risk clients: Predicting permanency for foster children", *Children and Youth Services Review*, nº 91: 156-167. https://doi.org/10.1016/j.childyouth.2018.05.030

Erikson, Susan L.
2018 "Cell phones≠ self and other problems with big data detection and containment during epidemics", *Medical Anthropology Quarterly*, nº 32 (3): 315-339. https://doi.org/10.1111/maq.12440

Escobar, Modesto (y otros)
2022 "Análisis de la dinámica, la estructura y el contenido de los mensajes de Twitter: violencia sexual en #Cuéntalo", *Empiria. Revista de metodología de Ciencias Sociales*, nº 53: 89-119. https://doi.org/10.5944/empiria.53.2022.32614

Evans, James (y Jacob G. Foster)
2019 "Computation and the Sociological Imagination", *Contexts*, nº 18 (4): 10-15. https://doi:10.1177/1536504219883850

Gillingham, Philip
2019 "Can predictive algorithms assist decision-making in social work with children and families?" *Child abuse review*, nº 28 (2): 114-126. https://doi.org/10.1002/car.2547

Gillingham, Philip (y Timothy Graham)
2017 "Big data in social welfare: The development of a critical perspective on social work's latest 'electronic turn'", *Australian Social Work*, nº 70 (2): 135-147. https://doi.org/10.1080/0312407X.2015.1134606

Goldberg, Amir
2015 "In defense of forensic social science", *Big Data & Society*, nº 2 (2): 2053951715601145. https://doi.org/10.1177/2053951715601145

Golder, Scott A. (y Michael W. Macy)
2014 "Digital footprints: Opportunities and challenges for online social research", *Annual Review of Sociology*, nº 40: 129-152. https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-071913-043145

Gualda, Estrella
2022 "Social big data, sociología y ciencias sociales computacionales", *Empiria. Revista de metodología*

de Ciencias Sociales, nº 53: 147-177. https://doi.org/10.5944/empiria.53.2022.32631

Gualda, Estrella (y Carolina Rebollo)
2020, "Big data y Twitter para el estudio de procesos migratorios: Métodos, técnicas de investigación y software", *Empiria. Revista de metodología de Ciencias Sociales*, nº 46: 147-177. https://doi.org/10.5944/empiria.46.2020.26970

Gui, Yong (y otros)
2020 "Three faces of the online leftists: An exploratory study based on case observations and big-data analysis", *Chinese Journal of Sociology,*nº 6 (1): 67-101. https://doi.org/10.1177/2057150X19896537

Halavais, Alexander
2015 "Bigger sociological imaginations: Framing big social data theory and methods", *Information, Communication & Society*, nº 18 (5): 583-594. https://doi.org/10.1080/1369118X.2015.1008543

Hashema, Ibrahim Abaker Targio (y otros)
2015 "The rise of big data on cloud computing: review and open research issues", *Information Systems*, nº 47: 98-115. https://doi.org/10.1016/j.is.2014.07.006

James, Gareth (y otros)
2013 *An Introduction to Statistical Learning* (vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jin, Xiaolong (y otros)
2015 "Significance and Challenges of Big Data Research", *Big Data Research*, nº 2: 59-64. https://doi.org/10.1016/j.bdr.2015.01.006

Laney, Doug
2001 "3D Data Management: Controlling Data Volume, Velocity, and Variety", *Gartner*, 6 febrero, nº 949. https://idoc.pub/documents/3d-data-management-controlling-data-volume-velocity-and-variety-546g5mg3ywn8

Lynn, Freda B. (y otros)
2019 "A rare case of gender parity in academia", *Social Forces*, nº 98 (2): 518-547. https://doi.org/10.1093/sf/soy126

Manderson, Lenore (y otros)
2015 "On secrecy, disclosure, the public, and the private in anthropology: an introduction to supplement 12", *Current Anthropology*, nº 56 (S12): S183-S190. https://www.journals.uchicago.edu/doi/full/10.1086/683302

McFarland, Daniel A. (H. Richard McFarland)
2015 "Big data and the danger of being precisely inaccurate", *Big Data & Society*, nº 2 (2): 2053951715602495. https://doi.org/10.1177/2053951715602495

Monroe, Kristin V.
2017 "Tweets of surveillance: Traffic, Twitter, and securitization in Beirut, Lebanon", *Anthropological theory*, nº 17 (3): 322-337. https://doi.org/10.1177/1463499617729296

Müller, Andreas C. (y Sarah Guido)
2016 *Introduction to machine learning with Python: A guide for data scientists*. Sebastopol, CA, O'Reilly Media, Inc., 1ª edición.

Mützel, Sophie
2015 "Facing big data: Making sociology relevant", *Big Data & Society*, nº 2 (2): 2053951715599179. https://doi.org/10.1177/2053951715599179

Nardulli, Peter F. (y otros)
2015 "A progressive supervised-learning approach to generating rich civil strife data", *Sociological methodology*, nº 45 (1): 148-183. https://doi.org/10.1177/0081175015581378

Nikunen, Kaarina

2021 “Ghosts of white methods? The challenges of Big Data research in exploring racism in digital context”, *Big Data & Society*, nº 8 (2): 205395172110489. https://doi.org/10.1177/20539517211048964

Olshannikova, Ekaterina (y otros)
2017 “Conceptualizing Big Social Data”, *Journal of Big Data,* nº 4 (3). https://doi.org/10.1186/s40537-017-0063-x

Patgiri, Ripon (y Arif Ahmed)
2016 “Big Data: The V’s of the Game Changer Paradigm”. *18th IEEE High Performance Computing and Communications, Sydney*: 17-24. https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0014

Philips, Susan
2013 “Method in anthropological discourse analysis: The comparison of units of interaction”, *Journal of linguistic anthropology*, nº 23 (1): 82-95. http://www.jstor.org/stable/43103109

Pink, Sara (y Debora Lanzeni)
2018 “Future anthropology ethics and datafication: Temporality and responsibility in research”, *Social Media+ Society*, nº 4 (2): 2056305118768298. https://doi.org/10.1177/2056305118768298

Reyes, Ángela
2014 “Linguistic anthropology in 2013: Super-new-big”, *American Anthropologist*, nº 116 (2): 366-378. https://doi.org/10.1111/aman.12109

Robles, Jose Manuel (y otros)
2020 “The polarization of ‘La Manada’: the public debate in Spain and the risks of digital political communication”, *Tempo Social*, nº 31: 193-216. https://doi.org/10.11606/0103-2070.TS.2019.159680

Rona-Tas, Akos (y otros)
2019 “Enlisting Supervised Machine Learning in Mapping Scientific Uncertainty Expressed in Food Risk Analysis”, *Sociological Methods & Research*, nº 48 (3): 608-641. https://doi:10.1177/0049124117729701

Rose, Jeremy (y Christian Lennerholt)
2017 “Low cost text mining as a strategy for qualitative researchers”, *Electronic Journal on Business Research Methods*, nº 15 (1): 2-16. https://academic-publishing.org/index.php/ejbrm/article/view/1352

Ruckenstein, Minna (y Natasha Dow Schüll)
2017 “The datafication of health”, *Annual Review of Anthropology*, nº 46: 261-278.
https://doi.org/10.1146/annurev-anthro-102116-041244

Ruppert, Evelyn
2013 “Rethinking empirical social sciences”, *Dialogues in Human Geography*, nº 3 (3): 268-273. https://doi.org/10.1177/2043820613514321

Schneider, Diana (y Udo Seelmeyer)
2019 “Challenges in using big data to develop decision support systems for social work in Germany”, *Journal of Technology in Human Services*, nº 37 (2-3): 113-128.
https://doi.org/10.1080/15228835.2019.1614513

Tinati, Ramine (y otros)
2014 “Big data: methodological challenges and approaches for sociological analysis”, *Sociology*, nº 48 (4): 663-681. https://doi.org/10.1177/0038038513511561

Wainer, Howard (y otros)
1974 “Trebig: A 360/75 Fortran program for three-mode factor analysts designed for big data sets”, *Behavior Research Methods & Instrumentation*, nº 6 (1): 53-54.
https://doi.org/10.3758/BF03200290