Social big data y sociología y ciencias sociales computacionales

1. INTRODUCCIÓN

Este artículo se enmarca globalmente, como indica el título, en el campo transdisciplinar emergente de las Ciencias Sociales Computacionales y, más específicamente, en lo que podríamos llamar Sociología Computacional (Edelmann et al. 2020; Evans y Foster 2019). Si a veces cuesta trabajo, dada la complejidad de nuestros objetos de estudio, trazar una línea clara entre las disciplinas asociadas a las Ciencias Sociales (por sus múltiples hibridaciones e intercambios), las áreas de investigación cuyo objeto requiere necesariamente observar fenómenos que transcurren en Internet, presentan un extra de complejidad, en la medida en que investigar sobre lo que ocurre en Internet frecuentemente nos aproxima a investigaciones que se sitúan, necesariamente, en la frontera del conocimiento, yendo mucho más allá de las Ciencias Sociales. Este sería el caso sobre el que queremos reflexionar en este artículo, al hilo del desarrollo actual de la investigación que toma como referencia datos sociales masivos, macrodatos sociales o Social Big Data[1]. Si bien el paradigma de los datos masivos parece haberse incorporado primero a las "ciencias duras" antes que a otras (Burgess y Bruns 2012), la situación inicial ha ido cambiando, provocando nuevos desafíos para la Sociología.

El campo de las Ciencias Sociales Computacionales, entre ellas la Sociología Computacional, está en expansión ofreciendo una variedad de direcciones de trabajo interesantes en algunos subcampos. Más allá de lo descriptivo, los sociólogos están contribuyendo al desarrollo de nuevas teorías y a la construcción de metodos híbridos en las Ciencias Sociales Computacionales que combinan métodos clásicos y modernos de la Sociología (Edelmann et al. 2020). Motiva el artículo, principalmente, el interés por identificar en el campo de las Ciencias Sociales algunas de las implicaciones metodológicas de este tipo específico de investigación en la cual la materia prima procede de Internet. Advertimos que a lo largo de estas páginas algunos de los dilemas que se van a presentar y que conciernen al método y a las técnicas, son comunes en las Ciencias Sociales y, si cabe, habituales también en las llamadas Humanidades Digitales. Igualmente, se comparten dilemas en algunos casos con otras Ciencias, en la medida en que la investigación en Internet presenta algunas limitaciones y ventajas ligadas al manejo de la información que son comunes independientemente de la disciplina desde la que se trabaja.

Las páginas que siguen introducen la especificidad conceptual de los Social Big Data frente al término de Big Data. A partir de aquí, exploramos qué suponen los procesos de investigación en este campo, frente a la investigación tradicional en las Ciencias Sociales, enfatizando dilemas emergentes relativos a los métodos y técnicas que se aplican, sus ventajas y limitaciones. La investiga-

[1] Por la amplia difusión que tienen en este campo científico términos como Big Data u otros, hemos optado a lo largo del artículo por usar terminología técnica tanto en inglés como en español. En ocasiones, para lectores no familiarizados, hemos incorporado entre corchetes otros términos técnicos frecuentes en inglés.

ción en el área de Big Data ha estado asociada desde sus orígenes a la llamada Ciencia de Datos y a la Inteligencia de Negocios [Data Science y Businness Intelligence] y ha enfatizado el manejo de algoritmos y estadística avanzada. Progresivamente ha ido profundizando también en áreas de la inteligencia artificial como el aprendizaje automático, las redes neuronales artificiales o el aprendizaje profundo, entre otros, alineándose a veces con unas ciencias sociales que tienden a la predicción. Argumentamos a continuación sobre la pertinencia para la Sociología, así como para otras ciencias, de avanzar igualmente en un enfoque basado en métodos mixtos en el campo de los Social Big Data, repensando el vínculo micro – macro en este campo de estudio. A través de un estudio de caso que intercalamos, ilustramos igualmente algunas potencialidades y limitaciones de este tipo de investigación, lo que nos permitirá trazar algunos de los desafíos metodológicos que pueden incorporarse en una agenda de investigación en esta área. Un área de investigación en la que se puede progresar de manera excepcional en el desarrollo de la transdisciplinariedad y la hibridación en las Ciencias, enriqueciéndolas.

2. SOCIAL BIG DATA: LA CONFLUENCIA DE LOS MEDIOS SOCIALES, EL ANÁLISIS DE DATOS Y LOS DATOS MASIVOS EN LA INVESTIGACIÓN

De acuerdo con Bello-Orgaz, Jung y Camacho (2016), en el campo de los Social Big Data se produce la confluencia de tres grandes áreas: los medios sociales, el análisis de datos y los datos masivos, que se conforman como un área interdisciplinar donde los medios sociales destacan como fuente principal de datos. En este artículo, al hilo de la relevancia que alcanzan las plataformas de redes sociales, por la importante generación de contenidos que se produce en ellas, centramos nuestra atención en el área específica de Social Big Data, sin menospreciar la importancia que tiene la generación de otro tipo de datos masivos. De hecho, no son infrecuentes las investigaciones cuyos análisis se basan en fuentes relacionadas tanto con medios sociales como con otros medios de obtención de la información (sensores, estadística pública, GPS, cartografía, etc.) (Bartosik-Purgat y Ratajczak-Mrożek, 2018; Kumar y Jaiswal, 2019; He y Xiong, 2018; Piccialli y Jung, 2017; Olshannikova et al., 2017).

El área de Social Big Data se nutre principalmente de contenidos que se producen en Internet, en los medios sociales, como plataformas de comunicación en línea, con un protagonismo importante de algunas redes sociales. Algunas fuentes típicas de este tipo de datos son desde redes sociales como Facebook, a blogs, pasando por microblogs como Twitter. Noticias sociales (como Reddit), medios para el etiquetaje o marcadores sociales (como lo fuera Delicious), medios para el intercambio de archivos de fotos o videos (Instagram, YouTube, TikTok), páginas wikis (Wikipedia, Wikihow), sitios basados en preguntas y respuestas de los usuarios (tales a Yahoo! Answers o Ask.com) y otros que basan su actividad

en la formulación de reseñas o críticas sobre servicios y establecimientos (como Yelp o TripAdvisor) (Jin et al, 2015).

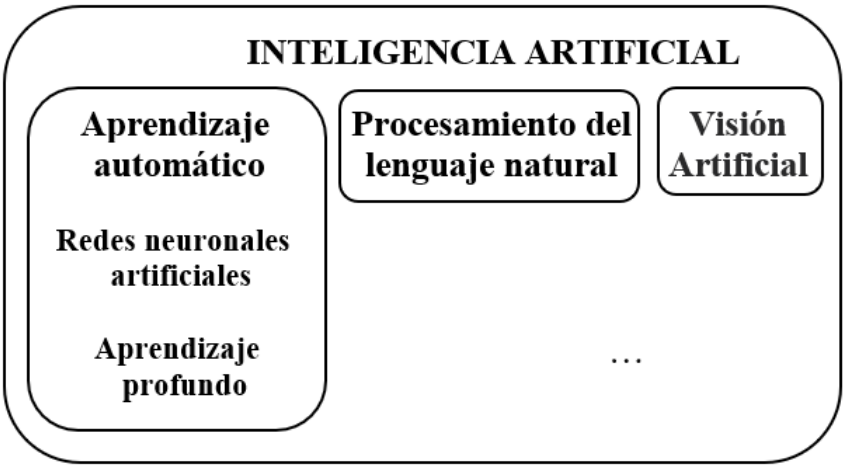
Para el análisis de los datos sociales masivos que se generan en estos u otros medios sociales se desarrollan métodos y técnicas de análisis que progresivamente han ido conformando las especialidades en Ciencia de Datos e Inteligencia de Negocios, con el objeto de avanzar en el manejo de amplios volúmenes de información, su procesamiento y análisis con el fin de generar conocimiento útil (Del Vecchio, 2018). La intersección entre las tres parcelas (social media, big data y data analysis) va asociada a continuos y novedosos desarrollos técnicos para el análisis de datos. En las secciones que siguen, exponemos algunos aspectos del trabajo metodológico con datos sociales masivos, con la idea de mostrar algunos elementos claves de los procesos de investigación en las Ciencias Sociales Computacionales con este tipo de datos, así como algunas de sus ventajas y limitaciones. Se exponen algunas orientaciones que apuntan hacia el fortalecimiento del paradigma predictivo en la Sociología (Chen, Wu, Hu, A. *et al.* 2021), al tiempo que destacamos el amplio margen de avance si se profundiza en los métodos mixtos y la sociología cualitativa si se incorporan al análisis de datos sociales masivos.

El tipo de retos que se plantean con los datos sociales masivos se ven condicionados por algunas de las características propias de los datos masivos, en la medida en que son datos con rasgos muy diferentes a los que manejábamos antaño en aspectos como el volumen, velocidad, variedad, veracidad, valor, validez, variabilidad, volatilidad, virtual, visualización, visibilidad (para más detalles sobre la diversidad de Vs con que se han descrito estos datos y su evolución, véase en Bello-Orgaz, Jung y Camacho 2016; Patgiri y Ahmed 2016; Laney 2001; Beyer y Laney, 2012; Bulger, Taylor y Schroeder, 2014; Hashema et al., 2015).

3. INTELIGENCIA ARTIFICIAL, APRENDIZAJE AUTOMÁTICO, REDES NEURONALES ARTIFICIALES Y APRENDIZAJE PROFUNDO: REPERCUSIONES EN LAS CIENCIAS SOCIALES Y LA SOCIOLOGÍA

Si bien la inteligencia artificial tiene un amplio recorrido, los avances en los últimos años, con el desarrollo del aprendizaje automático, las redes neuronales artificiales y el aprendizaje profundo han hecho que algunos expertos sugieran que en este campo se está produciendo un cambio de paradigma (Chen, Wu, Hu, A. *et al.* 2021), con repercusiones entre otras, en las Ciencias Sociales y la Sociología. Los términos que encabezan el título de esta sección, aunque no son sinónimos, es común encontrarlos en el mismo escenario. Introducimos brevemente algunos elementos clave en estos campos, como ejemplos muy representativos de avances actuales, al objeto de reflexionar sobre los efectos que tiene para las Ciencias Sociales y la Sociología la introducción de estas áreas, y específicamente para el campo de datos sociales masivos en el que se centra este artículo.

Cuadro 1. Inteligencia artificial y algunas ramas de utilidad para el análisis de datos sociales masivos



Fuente: Elaboración propia. El esquema tiene como objetivo simplemente enmarcar algunos de los términos que se usan a continuación.

3.1. Inteligencia artificial o artificial inteligence [AI]

Con la metáfora "inteligencia artificial" [AI] se define todo un campo de conocimiento que explora la capacidad que tienen las computadoras para mostrar un comportamiento "inteligente", por ejemplo, a través del desarrollo de funciones cognitivas (la parte relativa a la inteligencia) que llevan a cabo máquinas (la parte artificial). Este comportamiento abarca un amplio espectro de acciones, como pueda ser la de resolver problemas. Si en sus inicios la inteligencia artificial funcionaba a base de crear o configurar una serie de reglas que decían a las computadoras lo que debían hacer, con la aparición del aprendizaje automático se produce un cambio cualitativo y las máquinas realizan funciones que van más allá del seguimiento de reglas.

De esta forma, hoy la inteligencia artificial se refiere al estudio, al desarrollo y a la aplicación de técnicas que permiten a las computadoras adquirir ciertas habilidades propias de la inteligencia humana como, por ejemplo, el reconocimiento de imágenes, la comprensión del lenguaje natural, estudiar y resolver problemas, entender los contextos, aprender a hacer tareas nuevas y otras. Se recogerían en este campo técnicas que permiten a los ordenadores imitar la inteligencia humana usando la lógica, reglas tipo *si-entonces* (*if-then rule*) que establecen condiciones que deben pasar (*si*) para que se pueda afirmar una segunda parte (*entonces*), árboles de decisión y aprendizaje automático, incluyendo aquí redes neuronales artificiales y aprendizaje profundo.

3.2. Aprendizaje automático o *machine learning* [ML]

Cuando en el ámbito de la inteligencia artificial aparece el aprendizaje automático, este aporta a las computadoras una capacidad real de aprendizaje, que se asemeja más a la idea recogida en la metáfora o concepto de "inteligencia artificial". El aprendizaje automático se configura, por tanto, como una de las ramas dentro de la inteligencia artificial que se especializa en los aspectos relativos al aprendizaje. Sin ser las únicas, algunas ramas relevantes para la Sociología y el análisis de datos sociales masivos son, por ejemplo, el procesamiento del lenguaje natural [PLN][2] que, al permitir que las máquinas entiendan el lenguaje humano, posibilita avanzar en aplicaciones no solo para tareas como la traducción automática, sino también en otras como el procesado y análisis de datos sociológicos; la visión artificial, que enseña a los ordenadores a ver y entender las imágenes digitales[3], el reconocimiento automático del habla; las redes neuronales artificiales o el aprendizaje profundo[4].

El aprendizaje automático crea inteligencia artificial, lo que hace normalmente a través del uso de técnicas estadísticas para programar algoritmos que puedan aprender por su cuenta a realizar tareas y mejorar en su ejecución a través de la experiencia. Frente a las reglas fijas de antaño, el machine learning es una de técnica de entrenamiento que se utiliza para crear y mejorar dicho comportamiento (clasificar datos automáticamente, predecir, etc.) (Chen, Wu, Hu *et al.* 2021). Con la llegada de los datos masivos, por otra parte, se facilitan estos procesos de aprendizaje, por cuanto los ordenadores puedan contar con una cantidad y diversidad de datos para ser procesados, analizados e incorporados en el proceso de aprendizaje, siguiendo un esquema en el que proporcionando conjuntos amplios de datos como *inputs* se obtiene una mejora en el rendimiento al solucionar problemas (Robles *et al.* 2020).

Lo característico del aprendizaje automático es la introducción de procesos de aprendizaje, entre los que están la capacidad de aprender, razonar y mejorarse por sí mismos a través de diferentes técnicas o modelos (Molina y Garip 2019; Di Franco y Santurro 2021). Como subrayan Molina y Garip (2019) y Di Franco y Santurro (2021), el aprendizaje automático, incluyendo las redes neuronales artificiales, es resultado de la intersección de varias disciplinas (estadística, matemáticas, informática, neurociencia) que usan algoritmos para extraer información y conocimiento de los datos masivos y heterogéneos, con aplicaciones que alcanzan a ciencias sociales como la economía, las ciencias políticas y la sociología (Di Franco y Santurro 2021).

---

[2] Natural Language Processing (NLP).
[3] Computer visión (CV).
[4] Deep Learning (DL).

3.3. Redes neuronales artificiales o artificial neural networks [ANN]

Por otra parte, entre las técnicas o modelos que se incorporan al machine learning, están las redes neuronales artificiales, caracterizadas por imitar a las redes neuronales biológicas, en la medida en que encuentran su inspiración en cómo se comporta el cerebro humano (las neuronas, sus conexiones y la transmisión de información). Las redes neuronales artificiales son capaces de extraer patrones y detectar aspectos que el ser humano o el empleo de otras técnicas computacionales no pueden (Plebe y Grasso 2019). A partir de aquí, se crean modelos para intentar resolver problemas complejos a través de técnicas algorítmicas. Es clave el entrenamiento de la red, que trata de encontrar la combinación que mejor se ajusta buscando la mayor precisión del algoritmo en un proceso normalmente iterativo que se detiene cuando se alcanza el grado de error establecido por el investigador. Una red ya entrenada se puede usar para otras aplicaciones, por ejemplo, para hacer predicciones o clasificaciones. Las redes neuronales virtuales están diseñadas para analizar grandes volúmenes de datos, a partir de lo que extraen enseñanzas que aplican para la realización de tareas. Algunas de estas tareas, de aplicabilidad en las Ciencias Sociales y la Sociología, son el análisis de imágenes, la traducción de textos, el reconocimiento del habla y otras.

Dado que las redes neuronales analizan un conjunto muy ingente de datos haciendo durante todos los procesos innumerables tests, en realidad los investigadores no pueden saber con certeza qué factores fueron los que aportaron más capacidad de aprendizaje y mejora, aunque es viable corroborar el conocimiento aportado. Existen muchas vías por las que se puede llevar a cabo el proceso de aprendizaje en las redes neuronales (ejemplo de ello son aprendizaje supervisado, aprendizaje no supervisado, aprendizaje de refuerzo, aprendizaje fuera de línea, aprendizaje en línea), así como diferentes arquitecturas para las redes neuronales que se diferencian en cómo fluye la información. Aunque una descripción de estos aspectos excede del objetivo de este artículo, pueden consultarse un desarrollo más extenso con aplicaciones para las ciencias sociales y la sociología (Plebe y Grasso 2019; Molina y Garip 2019; Robles et al. 2020; Franco y Santurro 2021).

Un ejemplo de este enfoque con aplicación sociológica es el estudio reciente de Gabdrakhmanova y Pilgun (2021) que presenta un caso en el que las redes neuronales se han desarrollado para resolver aspectos diversos como la gestión eficaz de los sistemas urbanos y la resolución de conflictos urbanísticos. Para ello se crea un algoritmo a partir de fuentes diversas de datos que recopilan huellas digitales de redes sociales, microblogging, blogs, mensajería instantánea, foros, reseñas y videos dedicados a la construcción del Noroeste Chord (NEC) en Moscú. Se trata de un caso donde se usan diferentes estrategias y modelos matemáticos con el fin de detectar, prevenir y abordar tempranamente conflictos en la planificación urbana para ganar en eficiencia. Son muchas las aplicaciones en curso o potenciales de las redes neuronales artificiales que se desarrollan hoy en campos muy diversos, tanto asociados con los datos sociales masivos como con otro tipo de datos.

3.4. Aprendizaje profundo o deep learning [DL]

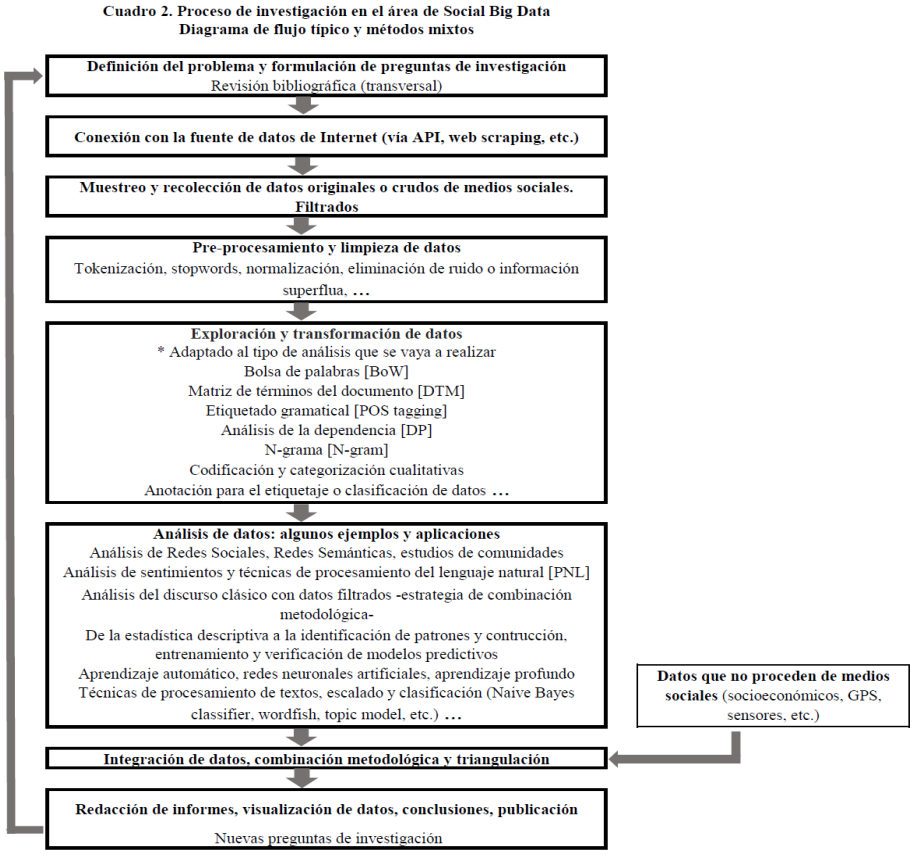
Las redes neuronales artificiales trabajan con capas de neuronas. Conforme hay más capas, más profunda es la red, y se alcanza mayor capacidad de aprendizaje y de procesamiento. Cuando la red neuronal artificial alcanza niveles muy profundos, se crea lo que se conoce como aprendizaje profundo o deep learning, que es un área de la inteligencia artificial que ha crecido mucho en poco tiempo (Plebe y Grasso 2019). Cuando se alcanza este nivel de profundidad, la novedad, en comparación, es que no solo se aprenden conceptos, sino también se pueden comprender contextos complejos. Esto hace que si el aprendizaje automático se caracteriza por la capacidad que algunos sistemas de inteligencia artificial tienen para auto-aprender y corregir errores basados en sus acciones previas, el aprendizaje profundo es capaz también de tomar decisiones a partir de los datos. Un ejemplo clásico de las aplicaciones fuera del campo de los datos sociales masivos es el diseño de los vehículos automáticos actuales. Otros ejemplos de utilidad para la investigación sociológica son los de los algoritmos que permiten que el software se entrene a sí mismo para realizar tareas como reconocimiento de voz e imágenes al exponer a las redes neuronales de múltiples capas a una gran cantidad de datos.

4. LOS PROCESOS DE INVESTIGACIÓN EN EL ÁREA DE SOCIAL BIG DATA Y MÉTODOS MIXTOS

Del mismo modo que ocurre en otras ciencias, una de las características sobresalientes de los datos masivos es que la magnitud y la complejidad de los datos que se producen y se quieren analizar es tan grande que afecta a diferentes momentos del proceso de investigación, siendo inviable llevar a cabo diversas operaciones de recolección, procesado y análisis con métodos y técnicas tradicionales. En el Cuadro 2 destacamos algunos hitos del proceso de investigación en el área de Social Big Data, incorporando diversas posibilidades analíticas que, más allá de los habituales propósitos predictivos comunes al área de Big Data, profundizan en el vínculo micro – macro y los métodos mixtos, propios de la Sociología y las Ciencias Sociales, en el contexto del procesamiento y análisis de datos masivos. Hemos incorporado en el cuadro técnicas como el análisis de redes sociales o el análisis del discurso, que conectan con un ejemplo que mostramos posteriormente. La idea de fondo es que, si bien la tónica más habitual al pensar en Big Data ha sido trabajar desde la perspectiva mayoritariamente cuantitativa, así como en la construcción de modelos y el avance de una ciencia social predictiva, para la Sociología y otras Ciencias Sociales, la recolección de datos masivos puede ir ligada igualmente a una mayor diversificación de los procesamientos y análisis de datos, desde una concepción en la que el objeto de conocimiento transita por el continuum macro – meso – micro. Este foco, además de producir igualmente conocimiento útil para diferentes fines, es también acorde a las metas y métodos propios de las Ciencias Sociales. Ciencias que ya, de hecho,

han indagado suficientemente en el ámbito de los medios sociales, aunque no habitualmente desde una aproximación tan centrada en los datos masivos.

Cuadro 2. Proceso de investigación en el área de Social Big Data. Diagrama de flujo típico y métodos mixtos



Fuente: Elaboración propia. El flujo de trabajo y las acciones descritas no son exhaustivas ni rígidas.

5. INTERNET COMO FUENTE: REPENSANDO EL MUESTREO PARA ACCEDER A DATOS SOCIALES MASIVOS

El estudio sobre los contenidos producidos en los medios sociales es un tipo de investigación basada en documentos que proceden de Internet. Como otros trabajos de este tipo, se trata de una investigación fundamentada en fuentes secundarias que maneja documentos de índole diverso (textuales, visuales, sonoros, audiovisuales) recopilados en medios sociales. La recogida de la información o la producción de datos es una de las tareas claves en la investigación, e implica definir algún tipo de criterio muestral, sea basado en criterios de representación estadística, teórica, u otros, siempre en función de los objetivos que tengamos. En el caso concreto de los medios sociales, estas tareas son quizás más complejas de emprender para los científicos sociales y aquí -en el establecimiento de criterios muestrales y la extracción de datos- se encuentra uno de los cambios y desafíos principales en las investigaciones basadas en Social Big Data, si las comparamos con otro tipo de investigaciones más asentadas en las Ciencias Sociales.

Junto a la dificultad recurrente de establecer un claro marco muestral, la investigación en medios sociales cuenta con la dificultad añadida de la recolección de datos, en la medida en que dicha recolección requiere en ocasiones el manejo de técnicas ligadas a conocimientos informáticos (principalmente programación), sobre todo cuando los objetivos de nuestra investigación nos conducen a la necesidad de obtener datos masivos o macrodatos. El establecimiento de criterios muestrales para acceder a datos de medios sociales en Internet supone un cambio de gran relevancia en el proceso de investigación. Por ejemplo, al estar los medios sociales dominados por empresas, es frecuente no poder contar con datos suficientemente desagregados y fiables sobre las poblaciones de referencia, lo cual es un hándicap para investigaciones cuantitativas, pues el desconocimiento del tamaño y características del universo afecta directamente tanto a las posibilidades del diseño muestral, como a las de la inferencia.

Otro ejemplo de las dificultades que entraña la investigación en este campo sería el de la volatilidad de los datos frente a otro tipo de investigaciones sociales. El que los datos sean perecederos hace más difícil (o imposible) la posibilidad de investigaciones de carácter retrospectivo, como ocurre cuando son las propias empresas (Twitter, Facebook, etc.) las que borran algunas de sus cuentas (y con ello sus mensajes), a pesar de que ya han podido tener un importante impacto social en aspectos como la propagación de discursos de odio (Avaaz, 2019). Aunque algunos de los mensajes borrados hayan podido ser captados y almacenados previamente por investigadores, o incluso aunque algo de esta información haya podido quedar almacenada para la posteridad en el repositorio Internet Archive (https://archive.org/), la indefinición del universo a la que aludíamos y la volatilidad de los datos son importantes hándicaps para la investigación. Un ejemplo claro sería la suspensión de la cuenta de Twitter de Donald Trump el 9 de enero de 2021, tras el asalto al Capitolio.

Como consecuencia, la investigación basada en la obtención de datos de medios sociales adopta nuevos procedimientos técnicos. Por ejemplo, durante el muestreo, al contactar con las APIs es habitual establecer criterios de búsqueda para realizar extracciones automatizadas de datos. Esta estrategia difiere claramente del clásico establecimiento de un universo de partida a partir del que se diseñan muestras con un criterio de representatividad estadística para, por ejemplo, la realización de una encuesta (en el enfoque clásico cuantitativo). Frente a muestras probabilísticas, que permiten que todos los elementos de la población tengan la probabilidad de ser seleccionados y con ello se puedan construir parámetros y realizar inferencias, en el área de Big data difícilmente se puede contar con muestras de este tipo. Como señalaran Burgess y Bruns (2012), uno de los problemas cuando este tipo de datos son almacenados por empresas es que no hay forma de saber cuán incompletos están los datos que descargamos. Este tipo de limitaciones, más que llevarnos a descartar este tipo de investigaciones, plantean diversos tipos de retos tales a los de la reducción de los sesgos de selección y ponderación, así como otros que han sido ya discutidos (Keiding & Louis 2016; Elliott & Valliant 2017; Franke et al. 2016; Morstatter et al. 2013).

6. LIMPIEZA Y PROCESAMIENTO: DE LA MINERÍA AL ANÁLISIS DE DATOS Y LOS RETOS DERIVADOS DE LA MAGNITUD Y LA COMPLEJIDAD

Sea más o menos acertado el término Big Data, es evidente que las Ciencias se encuentran tras el desarrollo de Internet y los avances tecnológicos que corren en paralelo, con el desafío de manejar la ingente, compleja y continua cantidad de datos que circulan o están accesibles en la red. A veces estos datos proceden de la naturaleza y son medidos, por ejemplo, a través de sensores, mientras que en otras ocasiones puede tratarse de datos que se originan en las redes sociales. En el campo de Social Big Data el foco principal lo tienen aquellos datos que proceden de medios sociales, si bien es cierto que muchas investigaciones, encuadradas de alguna forma en enfoques de métodos mixtos, dan valor también a datos procedentes de otras fuentes -como, por ejemplo, datos censales, encuestas, etc.- para la explicación de fenómenos sociales (Rahman et al., 2020; Flores, 2017).

Aunque en este artículo nos enfocamos sobre todo en los desafíos para las Ciencias Sociales, y prestamos más atención a la Sociología, la preocupación por el manejo de macrodatos alcanza cada vez más a las ciencias en general (sociología, ciencia política, ciencia de la información, periodismo, economía, psicología, geografía, astronomía, ecología, física, matemáticas, etc.). Esto se debe principalmente a que los métodos y técnicas de investigación clásicos no suelen ser suficientes para el manejo de los volúmenes y complejidad de los datos que se generan hoy en día en el mundo natural y social y a los que, cada vez más, la tecnología que se desarrolla permite aproximarnos.

Con el crecimiento de los Big Data, emergen diversidad de técnicas y procedimientos destinados a la realización de análisis de datos sobre todo de carácter textual, auditivos y visuales. Las características descritas arriba sobre los Big Data obligan necesariamente a que se produzcan cambios sustanciales en cuanto a la captura, procesamiento y análisis de datos si nuestro objeto es desarrollar investigaciones basadas en Internet desde esta perspectiva. Existe mucha bibliografía, sobre todo técnica, que describe, sistematiza y profundiza en estas técnicas analíticas, su estado actual y retos presentes y futuros (una buena revisión para datos sociales masivos se encuentra en Bello-Orgaz, Jung, Camacho, 2016).

En lo que sigue, no obstante, nos referimos solo a algunas técnicas, sin ningún ánimo de exhaustividad, con el objeto de hacer visibles algunos puentes de interés entre métodos clásicos de las Ciencias Sociales, y algunas de las técnicas comunes en las Ciencias Sociales Computacionales actuales. Uno de los aspectos que nos interesa es, igualmente, sintetizar alguno de los cambios sustantivos en los procesos de investigación social que supone el incorporarse al estudio de Social Big Data, así como reflexionar sobre las aportaciones a la discusión sobre el vínculo micro – macro que este nuevo campo abre. Si bien el procesamiento y análisis de datos en el campo de Social Big Data puede abarcar el estudio tanto de datos textuales, auditivos o visuales, dada la limitación de espacio, vamos a centrarnos sobre todo en algunos cambios metodológicos de relevancia respecto al análisis de textos, cuando estos son de gran magnitud.

6.1. Un cambio de escala en el procesamiento de datos textuales

La ingente cantidad de datos que se genera en los medios sociales conduce al desarrollo de técnicas de extracción de la información que permitan obtener datos estructurados para abordar el análisis de datos masivos que frecuentemente son no estructurados. Diversos procedimientos, que se ayudan de algoritmos para el procesamiento y el limpiado de información superflua para el análisis, facilitan el manejo posterior de los grandes datos que se han recolectado. Las tareas de pre-procesamiento de datos textuales previas al análisis siempre deben hacerse de acuerdo con los objetivos de investigación que hayamos formulado, si bien, como exponemos a continuación, hay una serie de pasos que son habituales para trabajar con grandes volúmenes de información textual. Recordamos algunas de las tareas típicas basándonos, entre otros, en el artículo de Welbers, van Atteveld y Benoit (2017), que recoge perfectamente el proceso habitual que suele hacerse en el análisis de textos, lo apliquemos al manejo de grandes o pequeños corpus de datos.

La investigación empírica en el campo de los Social Big Data arranca a partir de la obtención de datos procedentes de Internet. Una vez tenemos estos datos, hemos de leerlos con el lenguaje de programación o programa que vayamos a emplear para su procesamiento y análisis. Aunque son muchos los recursos que se pueden usar hoy en día, el uso de R como lenguaje de programación de software libre para el análisis de textos destaca cada vez más en Ciencias Socia-

les. Sobresale igualmente, para llevar a cabo otro tipo de análisis e incluso para recolectar datos de diferentes medios sociales, con aplicaciones cuantitativas y cualitativas.

Lenguajes como Python o R, así como otros, permiten llevar a cabo operaciones habituales relativas a la lectura, procesamiento y limpieza de datos. En el caso de las indicaciones respecto al análisis de textos basadas R, Welbers, van Atteveld y Benoit (2017), podemos ver un ejemplo de cómo se articula el paquete Quanteda en R de forma coherente con otros paquetes complementarios en R para el análisis de textos (Benoit et al., 2018), posibilitando su uso para el estudio de datos sociales masivos.

En el proceso básico de trabajo, una vez que leemos los datos, contaríamos ya con un corpus a partir del que podríamos empezar a limpiar y preparar los datos o incluso realizar algunos análisis exploratorios [EDA, Exploratory Data Analysis]. Para emprender algunas de las operaciones más interesantes y potentes que conlleva el análisis de datos sociales masivos (en este caso textuales), lo habitual es preparar un corpus que permita trabajar con toda esa cantidad de información de manera más rápida y efectiva. Aparecen aquí operaciones básicas del procesamiento de textos tales a la lectura o importación de los textos originales y su conversión o división en tokens. Los tokens son unidades significativas de texto (palabras -lo más habitual-, n-grams, frases, párrafos), llamándose *tokenización* al proceso de dividir el texto en tokens (Silge y Robinson, 2021).

Otras tareas básicas en las primeras fases buscan la normalización, con el objetivo de generar textos uniformes, de forma que el análisis sea más eficiente y se facilite. Entran en juego aquí transformaciones del estilo a homogeneizar el texto *de mayúscula a minúscula* o, por ejemplo, eliminar las *tildes y ñs*, si esto fuera de utilidad para contestar a nuestras preguntas de investigación. Otra operación frecuente es la de eliminar del corpus las *stopwords*, o palabras vacías de contenido informativo. Para ello se compara el corpus con una lista de stopwords y se aplican instrucciones a través de un script (u otro procedimiento) para que se eliminen estas palabras superfluas antes del análisis. Palabras del estilo a "lo, la, los, las, y, de, que…". Esta operación reduce el tamaño del corpus y mejora el trabajo computacional. Otra de las tareas típicas es eliminar *los signos de puntuación*, o los *números* (si no son necesarios en nuestro análisis), y tratándose de redes sociales, entre las rutinas típicas se pueden borrar espacios vacíos entre palabras, caracteres especiales tales a @, #, \\, los trazos de urls o incluso los emojis. Muchas operaciones se ayudan de expresiones regulares que simplifican los procesos de búsqueda de la información que se quiere modificar o limpiar. Cualquiera de estas tareas, aunque son típicas, vendrán condicionadas por nuestros objetivos de investigación, así como por los procedimientos analíticos que posteriormente se van a aplicar. Hay que tener en cuenta que, en la medida en que siempre conservamos los textos originales antes de procesar, es posible usar el corpus original para contestar unas preguntas de investigación, y el texto limpio o procesado para hacer otras operaciones que no requieran trabajar con todos los datos.

Además de estas tareas elementales de limpieza de textos, es habitual llevar a cabo otras de *radicación* [stemming] que consisten en normalizar o convertir las palabras en sus raíces para facilitar el recuento de términos. Las raíces son la parte de las palabras que no varían, no necesariamente las palabras en sí mismas. Esta operación convierte palabras conjugadas en raíces [stems], lo que de nuevo posibilita reducir el número de elementos que forman nuestros textos. Otra operación posible, a veces usada en vez de la radicación, es la *lematización* [lemmatization], que es una técnica que convierte palabras flexionadas o derivadas en sus lemas. El lema está formado por una serie de caracteres que forman una unidad semántica con significante y significado. El procedimiento normalmente emplea un diccionario para reemplazar las palabras con sus lemas. Estas estrategias se benefician de los avances en el área del procesamiento del lenguaje natural [NLP, Natural Language Processing] (Martí 2003; Vajjala et al. 2020). Actualmente son operaciones que se pueden realizar con apoyo de diversos paquetes de software, así como, por ejemplo, a través de librerías o bibliotecas que funcionan en lenguajes como Python (NLTK, Spacy) o R (quanteda, tm, tidytext, OpenNLP), citando algunas muy usadas. Tras estos procesos se suele reducir el tamaño de datos a procesar, lo que mejora la capacidad de cálculo, lo cual es esencial si trabajamos con datos sociales masivos.

6.2. De los textos pre-procesados a la matriz de términos [Document-term matrix, DTM]

Una vez que el corpus se ha procesado, se puede crear una matriz documento-término o matriz de términos del documento [Document-term matrix, DTM], donde las filas corresponden a los documentos de la colección y en las columnas se encuentran los términos. En las celdas encontramos la frecuencia con la que ocurre cada término (Benoit et al. 2018). A partir de aquí, se puede calcular la frecuencia del término [*term frequency, tf*], o frecuencia en que una palabra ocurre en un documento y la frecuencia inversa del documento [*inverse document frequency*, idf], o frecuencia en que el término ocurre en una colección de documentos (Silge y Robinson 2021). Las estadísticas tf-idf ayudan a medir la importancia de un término en una colección de documentos.

Según el enfoque de minería de textos [*Data mining*] empleado, hay diversas vías para el análisis. Por ejemplo, una aproximación está basada en el modelo de "bolsa de palabras" [BoW, bag-of-words], de acuerdo con el cálculo de las frecuencias de palabras que hay en un documento, sin tener en cuenta la posición u orden de cada palabra en el texto, aunque el contexto en muchas ocasiones puede ser clave. En el modelo de BoW se analizarían las palabras como simples token, sobre la base de diccionarios. Otro enfoque es aquel en el que se lleva a cabo un análisis sintáctico - semántico [semantic parsing] en el que se tienen en cuenta el orden en el que están las palabras en el documento, así como el tipo de palabras. Es importante el contexto en este caso.

Las operaciones previas, sin ser las únicas, explicadas muy someramente, se realizan con frecuencia en investigaciones basadas en grandes datos sociales o Social Big Data, en ocasiones al objeto de reducir el tamaño de los corpus con los que operar o hacerlos más amigables y eficientes para la computación (Gualda y Rebollo, 2020).

7. MÁS ALLÁ DE LOS ALGORITMOS: REPENSANDO EL VÍNCULO MICRO Y MACRO EN EL CONTEXTO DE LOS SOCIAL BIG DATA

Hasta ahora la Ciencia de Datos ha puesto un gran énfasis en asociar los datos masivos al desarrollo de técnicas estadísticas avanzadas, en gran medida al servicio de potenciar la capacidad predictiva y aportar valor en diferentes ámbitos empresariales. En este apartado presentamos un breve ejemplo de investigación al objeto de mostrar algunas potencialidades analíticas de los Social Big Data, en la línea de reforzar un enfoque de métodos mixtos, de gran utilidad para las Ciencias Sociales y la Sociología. Al mismo tiempo, nos paramos brevemente en dar unas pinceladas sobre varias técnicas de análisis para mostrar la versatilidad que propicia el desarrollo de enfoques mixtos y la combinación metodológica en la Sociología. Hace ya un tiempo Alexander y Giesen (1987) plantearon que la lealtad a puntos de partida limita el éxito en los intentos de integración micro – macro y que era necesario cambiar radicalmente de punto de arranque teórico para lograr un vínculo inclusivo. Más que una estrategia de combinación se planteaba cambiar de modelo de partida. Münch y Smelser (1987) destacaron el error de priorizar unos niveles sobre otros, enfatizando las interrelaciones entre los niveles micro y macro y la necesidad de caracterizar los procesos transicionales y emergentes que se mueven en ambas direcciones como agenda para los próximos años. En esta sección, aplicado al caso de los datos sociales masivos, exploramos este desafío, que continua con una larga tradición sociológica.

7.1. De los procesos clásicos de codificación y categorización al aprendizaje automático y la inteligencia artificial

Si en la investigación clásica se llevan a cabo procedimientos de clasificación de la información (codificación y categorización) básicamente forma manual, en el trabajo actual en el ámbito de los Social Big Data, es habitual implementar otros procedimientos. Uno de los empleados en investigaciones sociales y políticas es el método de codificación semiautomática con diccionario. En este, a través de varias etapas que pueden incluir la codificación manual de muestras aleatorias se puede ir mejorando la fiabilidad al codificar (Casas, Davesa y Con-

gosto, 2016.; Gallego, Gualda y Rebollo, 2017; Arcila, Blanco y Valdez, 2020; Arcila et al. 2021).

La automatización, para llevar a cabo la clásica tarea de "codificar" en Ciencias Sociales (preguntas abiertas, discursos, etc.), es necesaria ante la ingente cantidad de datos que se manejan. Conlleva habitualmente desarrollar actividades de programación, sea a través de lenguajes como Python, R u otros, o a través de paquetes que incorporan algunas de estas potencialidades, como es el caso de Tableau o Power BI, que permiten, en un entorno quizás más amigable para científicos sociales no acostumbrados a programar, usar algoritmos para el procesamiento de los datos, expresiones regulares o incluso conectar con estos lenguajes de programación. Otras opciones las aportan, por ejemplo, paquetes tales a Atlas ti o NVivo -amen de otros similares- que incorporan rutinas de codificación automática o de limpieza de stopwords en un entorno que no requiere la elaboración de scripts complejos).

En tareas clásicas como la codificación, incorporar procesos de codificación semiautomática (con momentos de revisión de códigos manualmente como verificación) suele ser muy útil para aplicar los códigos resultantes de forma automática a la serie de datos que se está procesando y analizando, lo que permite alcanzar datos masivos, algo que manualmente sería inviable. A veces es necesario llevar a cabo, desde cero, el proceso de elaboración de libros de códigos o diccionarios iniciales, si bien en ocasiones se pueden encontrar diccionarios o lexicones ya preparados en investigaciones similares, incluso en otras lenguas, por lo que podrían adaptarse y mejorarse.

Los procesos de codificación cualitativa (manuales o semiautomáticos), que ya de por sí son complejos, se nutren a veces ante los Social Big Data de estrategias derivadas de la inteligencia artificial donde se aplican técnicas de aprendizaje automático [machine learning], sea este supervisado, no supervisado o por refuerzo [supervised, unsupervised, reinforcement machine learning] que buscan el reconocimiento de patrones en los datos, por ejemplo, para ayudar en el proceso de clasificación de textos, a partir del entrenamiento de modelos que sin ser programados específicamente, pueden ayudar a resolver problemas computacionales a través del ensayo y error [reinforcement learning], del etiquetaje previo de algunos datos con los que se entrena a las máquinas [supervised learning] o de la localización de similitudes [non supervised learning]. Estas estrategias son útiles en el proceso de clasificación de ingentes cantidades de datos. El mismo aprendizaje automático cuenta con diferentes elementos de utilidad para la sociología si quiere aproximarse al análisis de este tipo de datos (Molina y Garip 2019). Por otra parte, estas técnicas de aprendizaje automático, u otras citadas menos sofisticadas, encuentran un uso cada vez mayor para el análisis de datos documentales de Internet, en el estudio de fenómenos como el bullying (Bellmore et al. 2015); tensiones sociales y análisis de sentimientos en comunidades on line (Burnap et al. 2016; Shu et al. 2017; Bello-Orgaz, Hernandez-Castro y Camacho 2017), discursos de odio y teorías de la conspiración (Arcila, Blanco y Valdez 2020; Arcila et al. 2021; Calderón, De la Vega y Herrero 2020; Gualda 2020) y una variedad de otros fenómenos sociales.

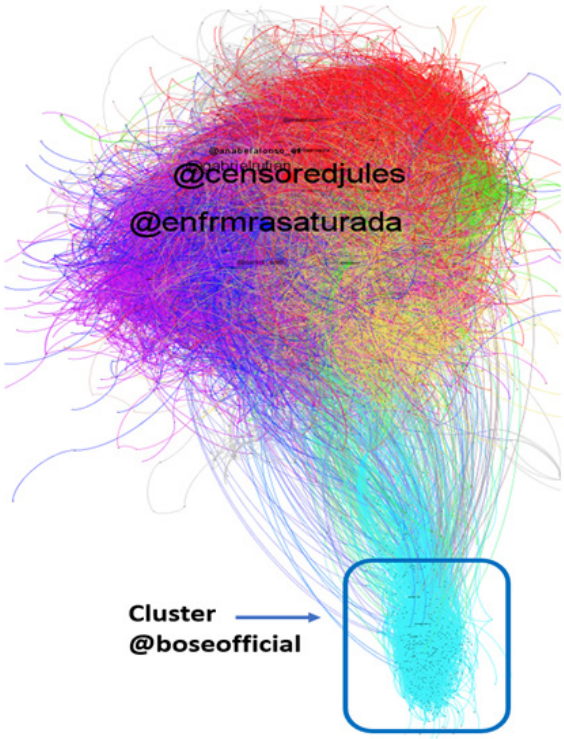
7.2. Análisis de redes sociales y micro discursos

Siguiendo con el ejemplo del análisis textual en el marco de los Social Big Data, una de las líneas que nos parece más sugerente para las Ciencias Sociales Computacionales es la que conecta el análisis de redes sociales con los datos sociales masivos, encontrando aquí diversidad de posibilidades de explotación de datos aún no suficientemente exploradas, a veces por las propias limitaciones técnicas. Si nos detenemos por un segundo en un mensaje típico en Twitter -similar en otros medios sociales-, en el contexto de un tuit se pueden encontrar varias etiquetas [#hashtags] que, consideradas conjuntamente, muestran, aunque sea de forma simbólica, por lo abreviado, un micro discurso. A partir de aquí, con procedimientos de filtrado, pueden elaborarse análisis cualitativos destinados a interpretar aspectos como la simbología o condensación de significados que una sola etiqueta incorpora (llamada a la movilización, crítica social, solidaridad, identidad, etc.) (Gualda 2016). Al mismo tiempo, se pueden estudiar desde la perspectiva de las redes sociales analizando relaciones entre co-hashtags, o co-ocurrencias de etiquetas en el mismo espacio de un tuit, a los efectos de encontrar patrones significativos que a simple vista no se aprecian ante el amplio volumen de datos. De igual modo, podría hacerse respecto a las co-palabras, si pensamos en el texto completo de un tuit, en las palabras que se encuentran en un blog, etc., lo que nos puede permitir no solo conocer elementos clave del discurso, sino también sus conexiones, para profundizar en otras facetas.

En nuestro estudio sobre el entramado de actores y mensajes que se difundieron en Twitter con motivo de la manifestación de Madrid del movimiento negacionista en agosto de 2020, se extrajo la información a partir de la etiqueta que se usó para la difusión de esta movilización. De esta forma, como criterio de búsqueda para la descarga de datos a partir de la API Twitter se empleó la cadena: #madrid16a. El criterio establecido en la plataforma que usamos para descargar los datos (*t-hoarder_kit*, de Congosto, 2016; Congosto, Basanta y Sánchez, 2017) fue simplemente que se descargaran los tuits que contenían dicha etiqueta. Una estrategia de muestreo muy diferente a las clásicas. A partir de los datos descargados se extrajo una red de retuits y llevamos a cabo un análisis de redes sociales, con el objetivo de identificar las diferentes comunidades en Twitter que estaban conversando este día apoyando o no la manifestación.

Con ayuda de Gephi, un paquete de análisis de redes sociales, se representó visualmente la red de retuits y se calcularon estadísticas de modularidad y centralidad de los actores de dicha red. El análisis visual y los datos de modularidad permitió identificar una comunidad que se encontraba a gran distancia del resto por su comportamiento relacional. Esta contaba con algunos actores conocidos, promotores del negacionismo, como fue el caso de la cuenta de usuario de Miguel Bosé (@boseofficial, cuenta que se suspendió por Twitter), que ocupaba una gran centralidad de grado e intermediación en ese cluster. Otras centrales fueron las cuentas de @paseamosjuntos, @gatunaguerrera o @cisnenegrojmal, también con importantes grados de entrada, reflejo ello de que sus mensajes fueron altamente retuiteados (Figura 1).

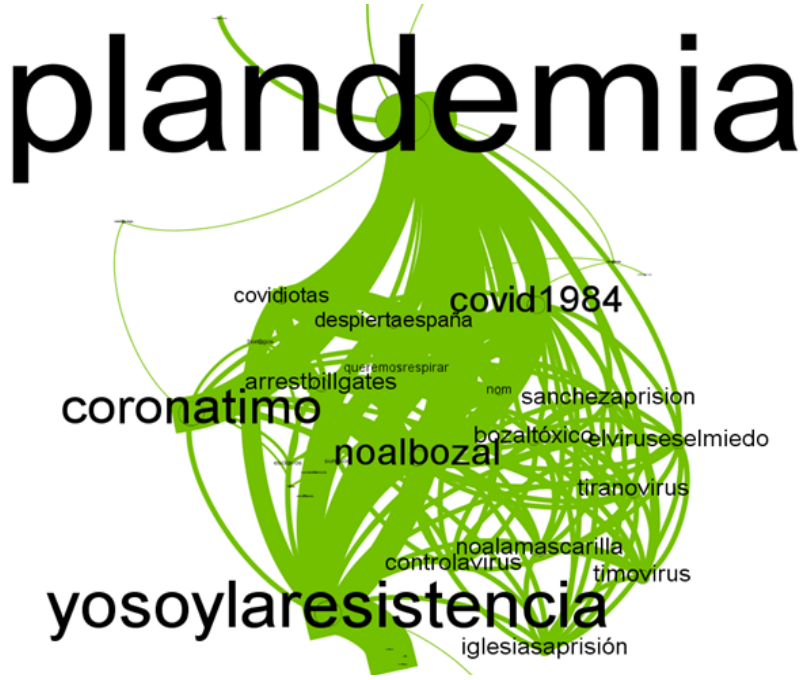
Figura 1. Red de retuits de #madrid16a



Fuente: Elaboración propia a partir de Gephi. Se representa la red de retuits según el grado de entrada, algoritmo de distribución Force Atlas 2. Dataset de #madrid16a (270.708 tuits). Tuits extraídos vía API streaming y API rest con la herramienta t-hoarder_kit (Congosto, 2016).

Para acercarnos a los contenidos que se estaban difundiendo por los actores de esa comunidad o cluster, ya en R, vinculamos a partir de la identificación de cada @user la base de datos completa de tuits con los datos de modularidad y centralidad obtenidos en Gephi para cada actor. De esta forma, pudimos aislar una submuestra de tuits publicados por los actores que integraban el cluster negacionista, y estudiar su comportamiento a partir de los mensajes publicados. El segundo grafo (Figura 2) muestra la red de principales co-hashtags del cluster donde se encontraba la cuenta @boseofficial. El grafo mide las relaciones entre hashtags que comparten el mismo espacio de un tuit. Mostramos solo los hashtags que forman la principal comunidad de co-hashtags (representados en verde). El discurso tiene un claro tinte negacionista, como igualmente se aprecia a partir de la elaboración de una clasificación de hashtags más frecuentes (Figura 3).

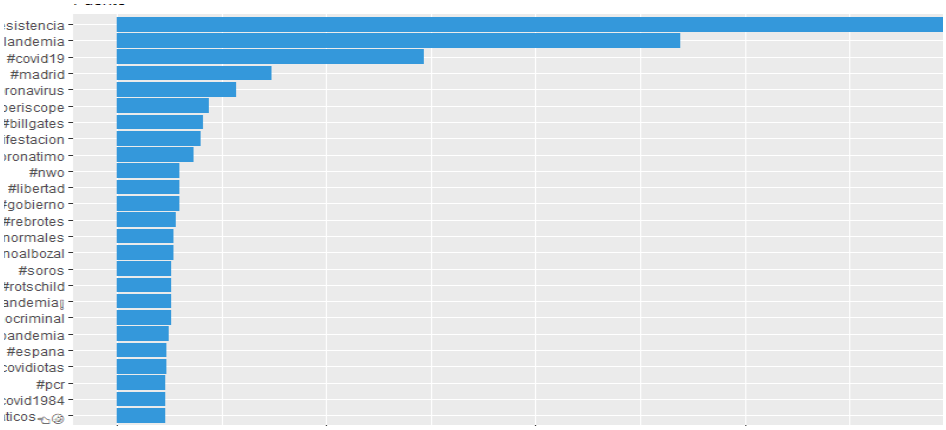
Figura 2. Red de co-hashtags, cluster @boseofficial



Fuente: Elaboración propia a partir de Gephi. Se representa la principal subred de co-hashtags extraída de los tuits del cluster @boseofficial.

No parece casualidad la proximidad encontrada en este cluster entre hashtags tales a #plandemia, #yosoylaresistencia, #coronatimo, #bozaltóxico, #tiranovirus, #controlavirus, #covid1984, #elviruseselmiedo, #noalbozal, #noalamascarilla, #timovirus, etc. El discurso negacionista y conspiracionista se manifiesta en el mismo espacio de crítica y ataque tanto al gobierno como a actores claves internacionales: #arrestbillgates, #sanchezaprision, #iglesiasaprisión, así como de referencias a la privación de libertad: #queremosrespirar, #covidesnazismo, etc. Algunoss hashtags como palabras clave que comparten espacio con otros desarrollan un micro discurso plagado de metáforas, con una importante carga simbólica (#noalbozal). La exploración sobre las pautas que adquiere el discurso en el contexto de microblogs tiene un gran interés desde la perspectiva de los métodos cualitativos.

Figura 3. TOP 25, hashtags más frecuentes en el cluster @boseofficial



Fuente: Elaboración propia a partir de R. Clasificación elaborada a partir de las etiquetas más frecuentes en el cluster @boseofficial.

7.3. Análisis de sentimientos, discursos de odio y polarización

En un contexto donde la polarización, los discursos de odio o las noticias falsas son frecuentemente estudiados en procesos de comunicación, a pesar de su dificultad técnica (Shu, 2017; Arcila et al., 2020; Arcila-Calderón, et al., 2021; Sánchez y Arcila, 2020; MacAvaney et al., 2019), no es extraño que el análisis de sentimientos suscite interés en el marco del análisis de textos para recoger las opiniones que se vuelcan en las redes (Rahman et al. 2020; Liu, 2012).

El interés en el análisis de sentimientos conecta igualmente directamente con algunos desarrollos que, basados más en el análisis de redes sociales, teorizan sobre comunidades que conversan dentro de sí mismas en cámaras de eco o especie de "burbujas de filtro" (Cinelli et al. 2020; Brugnoli, Cinelli, Quattrociocchi, *et al.* 2019; Stout, Coulter y Edwards, 2017). Más allá de cuestiones teóricas, hay importantes desafíos técnicos en este campo (y aún más complejidad en el caso del análisis de videos o fotos). Uno muy evidente tiene que ver con las dificultades para clasificar contenidos procedentes de medios sociales que, como ocurre con otros textos, siempre cuentan con problemas tales a la ambigüedad, los dobles sentidos, y otras peculiaridades del lenguaje (MacAvaney et al., 2019). Adicionalmente, hay una gran laguna aún respecto a las capacidades para el procesamiento y el análisis en el mismo proceso de investigación de texto, imágenes, vídeos o audios con la idea de estudiar las conexiones entre sus contenidos, o para observar el componente emocional de algunos datos publicados en medios sociales. Por ejemplo, aunque se ha identificado el componente emocional que incorporan los robots sociales a campañas políticas (Ferrara, 2017), cuando se analizan datos masivos para la Sociología es de gran interés

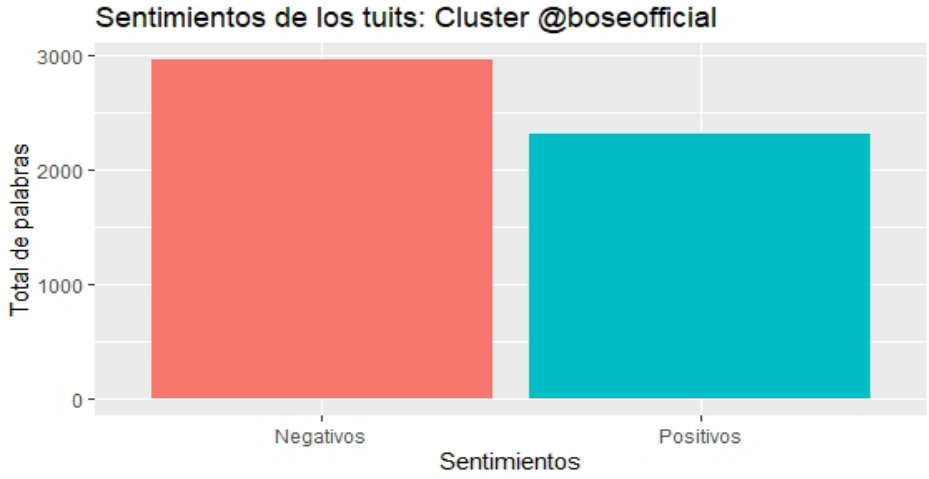
poder saber si en un mismo mensaje (por ejemplo, un tuit), el discurso de odio se encuentra tanto en el texto como en la imagen o vídeos adjuntos, o incluso en las URLs que se recomiendan.

Trabajar con algunos conjuntos de datos (de Twitter, por ejemplo) requiere especificar el idioma a través del que se van a clasificar los sentimientos. En ocasiones hay complejidades añadidas porque no es infrecuente que ante determinadas temáticas se escriban textos que incorporan palabras de varias lenguas (hashtags, por ejemplo). A veces, la resolución de este problema va de la mano de un algoritmo que reconoce el idioma principal a partir de umbrales que se delimitan, y esto a la vez es una ganancia, porque este tipo de estrategias posibilita el manejo de datos masivos, pero al mismo tiempo es una pérdida, en la medida en que siempre hay márgenes de error que asumir.

Muchos paquetes de análisis trabajan con diccionarios y lexicones en inglés, pero es posible encontrar una serie de ellos (adaptados a lenguajes diversos como R o Python) de carácter multilingüe. Basándonos en la librería Syuzhet sobre análisis de sentimientos que trabaja en R, hemos explorado emociones y sentimientos de los textos correspondientes al cluster de @boseofficial. Para ello, se ha usado el lexicón en español de este paquete (Mohammad, 2015). En él, a partir de los textos de los tuits, tras el preprocesado y limpieza de textos, se proporciona un puntaje de sentimientos para cada tuit a partir del uso del diccionario NRC de sentimientos en español. Se aportan dos tipos de puntajes. Uno evalúa si los tuits, de acuerdo con los sentimientos que expresa su vocabulario, son positivos o negativos. Otro, respecto a las emociones: Ira, expectación, disgusto, miedo, alegría, tristeza, sorpresa, confianza, donde se evalúa la intensidad en que están asociadas las palabras a diferentes emociones. Una de las dificultades para este este tipo de análisis es precisamente que no siempre es viable encontrar lexicones en todas las lenguas. Aparte de las ambigüedades, dobles sentidos, etc.

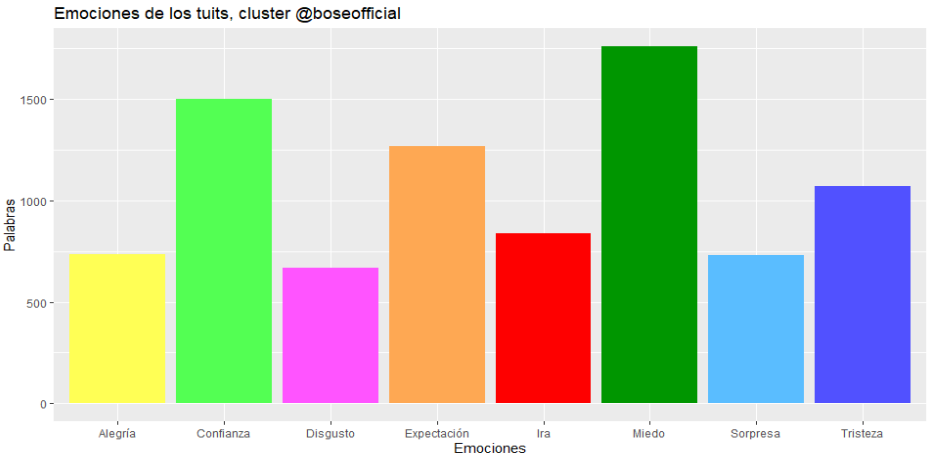
Pese a que habitualmente la bibliografía alude a que los contenidos que se vuelcan en Twitter son más positivos que negativos, en el caso de los tuits asociados al cluster negacionista, hemos encontrado mayor carga de sentimientos negativos (Figura 4) o de emociones negativas como Ira, etc. (Figura 5). Un paso más en el análisis podría ser delimitar las palabras clave que caracterizan cada grupo de emociones, y compararlas con el conjunto de datos total, a los efectos de identificar algunos aspectos que pueden estar sobrecargando el discurso que se encuentra en las redes y que conviene explorar con más atención con ayuda de técnicas clásicas de corte cualitativo.

Figura 4. Sentimientos asociados a los tuits del cluster @boseofficial



Fuente: Elaboración propia en R. Para la clasificación de sentimientos en los tuits se empleó la librería Syuzhet.

Figura 5. Emociones asociadas a los tuits del cluster @boseofficial



Fuente: Elaboración propia en R a partir de la clasificación de emociones de los tuits con la librería Syuzhet.

La Figura 6 recoge las principales palabras de los tuits asociados a cada emoción, a modo de ejemplo. Aunque permite una primera visión, habría que seguir profundizando para conectar estas palabras con el contenido completo de los tuits donde aparecen. Otras estrategias son posibles, más sencillas o complejas, permitiendo la articulación metodológica. Desde identificar los mensajes concre-

tos donde aparecen los principales términos asociados a cada emoción para estudiar el contexto en el que aparecen, hasta profundizar en aspectos como puedan ser la descripción del tipo de hashtags con los que algunas palabras clave pueden compartir espacio en el microdiscurso que se elabora en un tuit. Otros objetivos podrían ser conocer cuáles son las URLs que se recomiendan, o analizar, con criterios visuales, si los videos y fotos que acompañan a los tuits refuerzan o no las emociones que parecen expresar globalmente los tuits. Además de diversidad de aspectos complementarios que podrían explorarse de forma cuantitativa si se comparan entre sí las comunidades de una serie de datos que previamente ha sido sometida a un análisis de redes sociales. De esta forma, podría afinarse en la identificación de los discursos principales expresados por los usuarios de los diferentes subgrupos de una red social.

8. AVENIDAS HACIA UNA SOCIOLOGÍA COMPUTACIONAL, MÉTODOS MIXTOS Y PREDICCIÓN SOCIAL

8.1. Predicción social y métodos mixtos

Recientemente, Chen, Wu, Hu *et al.* (2021) sugieren que la predicción social basada en el aprendizaje automático tiene gran importancia en el horizonte de la sociología y las ciencias sociales, en aspectos tales a la obtención de indicadores latentes de interés para la sociología, la generación de hipótesis teóricas, ayudar a llevar a cabo inferencias causales, la estimación de valores perdidos o incompletos a través del aprendizaje automático y promover la innovación. Apuntan a un cambio de paradigma que se abre en las ciencias sociales a raíz de la introducción del aprendizaje automático. Este radica en el desarrollo de tres orientaciones: la cuantitativa (basada en la correlación y la causalidad), la cualitativa y la predicción cuantitativa, esta última a raíz del rápido desarrollo del aprendizaje automático y de factores como la posibilidad de contar con datos masivos representativos del contexto histórico que puedan nutrir el aprendizaje automático de más información que facilite la "predicción social" (Chen, Wu, Hu *et al.*, 2021). Por otra parte, en el campo de los datos sociales masivos, en conexión con los avances que se introducen en la inteligencia artificial, se observa una diversidad de aplicaciones orientadas hacia la predicción, se trate de investigaciones que buscan pronosticar la personalidad basándose en datos relativos a la interacción en redes sociales (Jang, 2021), el uso de algoritmos de machine learning en marketing para predecir el comportamiento del consumidor (Baptiste, 2020), el establecimiento de predicciones que ayudan a tomar decisiones política u otros usos en ciencias sociales (véase en Molina y Garip, 2019).

A pesar de la intensidad con que parece revitalizarse la predicción en la investigación social de la mano del machine learning, y aunque se subrayan diversidad de aspectos positivos como, entre otros, la habilidad de las redes neuronales artificiales para adaptarse a diversos tipos de datos (Di Franco y Santurro, 2021), se mantienen muchas de las incógnitas y limitaciones de antaño respecto

a la capacidad que tenemos para predecir fenómenos sociales. Por otra parte, dado que las redes neuronales y especialmente el deep learning trabajan sobre la base de capas ocultas donde se realizan operaciones en los datos, esto genera mecanismos de "caja negra" que suponen igualmente una limitación y una falta de transparencia para la investigación social, en la medida en que la manera de trabajar estas redes hace imposible ver el proceso completo que genera los resultados de aprendizaje, si bien las predicciones puedan ser o parecer robustas. Esto dificulta conocer aspectos clave para la sociología cómo la manera en que se relacionan las variables entre sí (Di Franco y Santurro, 2021).

Un aspecto complementario que, entre otros, dificulta la predicción tiene que ver con que algunos modelos predictivos se asientan en procesos de preprocesado o transformación de datos donde los algoritmos manejan umbral es que pueden ser una fuente de error. Un caso típico que puede citarse es respecto al etiquetaje automático de palabras sobre la base de algoritmos basados en la distancia de Levenshtein o algoritmos de distancia fonética, o cuando se aplican procesamientos de normalización léxica u otros relativos al procesamiento del lenguaje natural destinados minimizar el efecto de las faltas ortográficas, omisiones y errores gramaticales que frecuentemente se encuentra en textos publicados en los medios sociales (Ahmed, 2015). Por no decir de los intentos de desambiguación aplicados automáticamente pueden fallar. Otras fuentes de error se encuentran en estrategias o análisis concretos como cuando se evalúan los mensajes asignando un puntaje con relación a la carga emocional detectada, que a veces no detectan la sutilidad o los dobles sentidos de los mensajes, o se necesita, para una adecuada comprensión, un buen conocimiento del contexto que difícilmente un lexicón estandarizado puede recoger si no se aplican procesos semiautomatizados.

El umbral establecido por los algoritmos importa. Estrategias de clasificación o reducción de datos, que pueden ser algunas de las formas en las que se hace viable la investigación basada en datos sociales masivos, puede tener, en este sentido, sus pros y sus contras. Una de las ventajas es que todas las tareas de preprocesado y transformación de datos facilitan aplicar una serie de técnicas que con los textos originales serían menos eficientes (o incluso imposibles) ante el ingente volumen y complejidad de los datos. Pero, dado que cualquier transformación de los datos crudos desvirtúa el original, y puede hacer que se pierdan algunos sentidos interpretativos o matices, es preciso valorar siempre este contraste entre las pérdidas y las ganancias antes de ejecutar cada paso al investigar.

Las bondades atribuidas en la historia a la predicción en las ciencias sociales no son nuevas. No obstante, sin rechazar la posibilidad de esbozar horizontes o escenarios a partir del análisis de datos sociales masivos, nos parece importante para las Ciencias Sociales mantener un sano escepticismo respecto a nuestra capacidad predictiva, sobre todo cuando se entiende esta como causalidad, a pesar de que un buen diagnóstico con el apoyo de datos sociales masivos puede llevar a la toma de decisiones mejor informadas. En este sentido, la posibilidad de aplicar modelos avanzados de estadística destinados a la predicción social no deben considerarse como una panacea, sino más bien como la oportunidad de

poder aproximarnos a la ingente cantidad de datos complejos que se encuentran en los medios sociales y que serían inaccesibles de otra forma.

Por otra parte, la constatación histórica de la existencia de factores imprevisibles que afectan a la sociedad hace que los modelos, que suelen ser útiles para esbozar escenarios, adelantarse a posibles problemas y ayudar a la toma de decisiones, haya que considerarlos siempre con cierta reserva, huyendo del espejismo de la predictibilidad (la pandemia de COVID-19 es un buen ejemplo de nuestra falta de control de los potenciales imprevistos o elementos que invalidan en la práctica muchas predicciones).

El trabajo con los datos (sociales) masivos ha estado fuertemente influido hasta ahora por disciplinas como la informática, la ingeniería, la estadística o las matemáticas. No obstante, desde la perspectiva sociológica, del mismo modo que desde las Ciencias Sociales Computacionales y las Humanidades Digitales, otras miradas complementarias son posibles, entre las que nos parecen especialmente productivas las que se encuentran enmarcadas en los métodos mixtos, la hibridación y el pluralismo teórico y metodológico, al hilo de los que emergen nuevas preguntas de investigación. En las páginas precedentes hemos mostrado un ejemplo a este respecto.

Algunas de las limitaciones señaladas arriba no invalidan el estudio de datos sociales masivos, si bien plantean la enorme importancia de ser concientes de hasta dónde pueden llegar en sus conclusiones, lo que puede ser variable según las ciencias y el propósito de cada investigación. Nuestra invitación es quizás más a la prudencia. En el campo de los datos sociales masivos otros elementos que invitan a la cautela son, por ejemplo, que estamos a expensas de datos proporcionados por empresas, datos que normalmente no sabemos cuán completos o sesgados están pero son la base que nutre nuestro análisis. En este sentido, una mirada crítica y la orientación al esbozo de escenarios de cara a la prevención pueden ser más útiles que predicciones estrictas.

8.2. Avenidas de la Sociología computacional en el área de Social Big Data

Diferentes tipos de desafíos y/o hándicaps aparecen en el escenario de las nuevas Ciencias Sociales Computacionales cuando estos se aplican al área de los datos sociales masivos. Desde todo lo que comporta ser capaz de manejar grandes volúmenes de información en un tiempo razonable (inviable con estrategias clásicas), hasta avanzar en aspectos como la profundización y mejora del estudio de los datos generados en streaming, con lo que ello implica tanto de manejo de gran cantidad de información en tiempo real, como de formulación de analíticas e indicadores sintéticos de utilidad. Igualmente, el avance en la identificación y el estudio de las ambigüedades del lenguaje, los matices o la sutilidad del discurso en datos masivos.

La mejora en el análisis conjunto de texto, fotos, audio y video, que permita obtener una visión más integral del contenido de cada mensaje es igualmente una

línea muy prometedora y compleja para la Sociología, que entraña la necesidad de avances tanto en la línea de almacenaje como de procesamiento y analítica de fotos, audios y videos, con gran dependencia, por ejemplo, de los avances en el reconocimiento de audio e imagen, o en la mejora con tecnologías de big data en procesos como los que llevan a poder revisar a fondo los vídeos, indexarlos o trabajar con metadata para una búsqueda y recuperación más fácil y certera de contenidos, aspectos estos en que algunas áreas de la inteligencia artificial avanzan. Conectados a los anteriores, se encuentran otros retos complejos relacionados con la ética y la confidencialidad al trabajar con datos sensibles a veces. Junto al desafío de poder articular los planos micro y macro con diferentes tipos de datos, está también el avance en el manejo de la complejidad computacional para el procesamiento de grandes volúmentes de información estructurada y no estructurada.

Otro aspecto importante es preguntarse por cambios venideros en la Sociología y en otras Ciencias Sociales cuando se incorpora esta línea de trabajo. Además de la posibilidad de transferir algunos métodos y técnicas desarrollados para el manejo de datos masivos, que pueden ser adaptados para otro tipo de datos, observamos varios caminos complementarios para la disciplina. Por una parte, la apertura de una nueva especialización de trabajo en el campo de la Sociología Computacional requiere necesariamente que nuevas generaciones de Sociólogos reciban formación para ser capaces de abordar los nuevos desafíos al conocimiento que plantea la gran cantidad de información publicada en los medios sociales. Esto, en la práctica, implica incorporar nuevas materias de conocimiento en planes de estudio, tales al aprendizaje de Lenguajes de Programación, para superar hándicaps técnicos y adquirir habilidades para trabajar escribiendo código para recolectar, procesar y analizar datos sociales (Edelmann, Wolff, Montagne y Bail 2020; Evans y Foster 2019; Gualda y Rebollo 2020). Otra vía es reforzar la presencia de los equipos inter y transdisciplinares con profesionales de ramas más técnicas, especialmente para el manejo de aspectos técnicamente más complejos. Una tercera aproximación guarda relación con la profundización en el manejo de paquetes de software más amigables o que requieren menor capacitación técnica, si bien, son limitados a veces para el diseño y realización de investigaciones con datos sociales masivos.

La Sociología Computacional saca ventaja de nuevas herramientas y fuentes de datos para amplificar su alcance y escala, abriendo a su vez nuevos espacios en diferentes especialidades de la Sociología. “Y, sobre todo, amplia la imaginación sociológica” (Evans y Foster 2019, traducción propia). Entre los aspectos más sugerentes que se potencian en el ámbito de la Sociología Computacional, siendo una de sus facetas el análisis de datos sociales masivos, se encuentra el uso de nuevos tipos de datos para revisitar preguntas sociológicas antiguas (en los ámbitos micro-macro) o que se desarrollen creativamente métodos híbridos que contribuyan a profundizar en el espacio entre niveles de análisis si se ligan teorías de corte macro con procesos a un nivel micro (Edelmann et al. 2020). Apreciamos, por tanto, el gran potencial que aporta para la Sociología el estudio de datos sociales masivos y el enriquecimiento que produce la articulación

de métodos y técnicas convencionales y modernas, así como la fertilidad de la hibridación entre diferentes ciencias en la Sociología y Ciencias Sociales Computacionales en construcción.