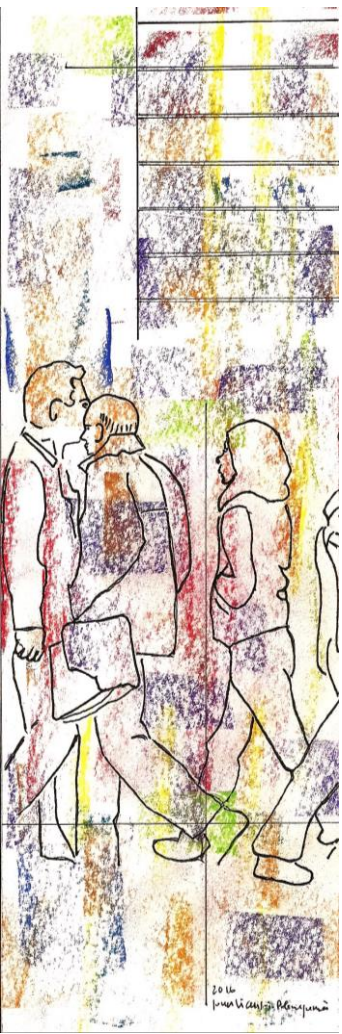
INCASI
Working Paper Series

2025, No. 17



INCASI *International Network for Comparative Analysis of Social Inequalities*

Ciencias sociales, métodos computacionales e inteligencias generativas

Juan Barri
Sandra Fachelli

Funded by the European Union

Horizon Europe – INCASI2 Project
Marie Skłodowska-Curie Actions (MSCA)
Staff Exchanges (GA-101130456)

Ciencias sociales, métodos computacionales e inteligencias generativas

Juan Barri[1]
Sandra Fachelli[2]

[1] Escuela de Filosofía, Facultad de Filosofía y Humanidades,
UNC, Córdoba, Argentina.
juan.barri@unc.edu.ar

[2] Departamento de Sociología, Universidad Pablo de Olavide, Sevilla
España,
sfachelli@upo.es

Funded by the European Union

Horizon Europe – INCASI2 Project
Marie Skłodowska-Curie Actions (MSCA)
Staff Exchanges (GA-101130456)

This paper was elaborated in the context of the INCASI2 project, *A New Measure of Socioeconomic Inequalities for International Comparison*, that has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Sklodowska-Curie, Staff Exchanges, grant agreement No 101130456 (https://incasi.uab.es). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Dipòsit Digital de Documents
Bellaterra, Cerdanyola del Vallès (Barcelona), Spain
Universitat Autònoma de Barcelona

Ciencias sociales, métodos computacionales e inteligencias generativas

Juan Barri
Sandra Fachelli

Abstract

El presente artículo busca indagar en el potencial de la aplicación de un conjunto de métodos computacionales en la investigación en ciencias sociales, en el contexto de emergencia de las llamadas inteligencias generativas y los modelos fundacionales. Para ello recorreremos una serie de dimensiones que consideramos relevantes en el cruce entre la reflexión sociológica y la aplicación de una serie de técnicas computacionales que se utilizan regularmente en contextos de producción y gestión, pero que son poco o subutilizadas en el campo de las ciencias sociales. Los temas que abordaremos son: a. ¿existen las ciencias sociales computacionales? b. Lenguajes computacionales y ciencias sociales. c. el acceso a la información en la era del big data y la huella digital. d. análisis exploratorio y estadísticas descriptivas. e. Maching learning y ciencias sociales. f. Procesamiento de Lenguaje Natural. g. Grafos y h. Análisis de trayectoria. Somos conscientes de que queda un amplio abanico de recursos sin mencionar, pero en rigor de un tratamiento adecuado preferimos centrarnos en los enumerados.[1]

Palabras clave
Ciencias Sociales, Inteligencias Generativas, Métodos Computacionales, Big Data

Índice

Introducción

Para comenzar es necesario señalar que desde el lanzamiento de ChatGPT a fines de 2022 vienen sucediendo a nivel global una serie de cambios abruptos y significativos en el ámbito de la economía, la producción, el modelo científico tecnológico y la política.[2] Asistimos a la

[1] Este artículo se inició en la estancia de Juan Barri en la UPO y en el contexto del Seminario "Ciencias sociales, Métodos computacionales e Inteligencias generativas", impartido en la Facultad de Ciencias Sociales de la UPO el 7 de noviembre de 2024 por el primer autor.

[2] Un ejemplo de ello fue el "terremoto" de la bolsa norteamericana acaecido el 27 de enero de 2025, que provocó pérdidas inéditas a las empresas tecnológicas más grandes de la economía mundial, producido por el lanzamiento del modelo generativo DeepSeek de origen chino, explicita el lugar central que tienen los modelos de inteligencia generativa en el capitalismo global. CNN: "A shocking Chinese AI advancement called DeepSeek is..." https://edition.cnn.com/2025/01/27/tech/deepseek-stocks-ai-china/index.html

consolidación de un paradigma socio productivo que parece ser la secuela del modelo tecnoliberal del que habla E. Sadín (2018) al referirse al proceso de *siliconización del mundo*.

Esta transformación se inscribe en un cambio estructural más amplio, que algunos autores comienzan a caracterizar como el advenimiento de una "revolución de la IA" con profundas consecuencias sociales. Por ejemplo, Xie y Avila (2025) analizan cómo la inteligencia artificial generativa puede reconfigurar la desigualdad global, favoreciendo a las grandes potencias tecnológicas y alterando las estructuras ocupacionales hacia una posible "sociedad post-conocimiento", escenarios que las ciencias sociales computacionales deberán ayudar a comprender y criticar.

Es probable que la transición hacia la utilización regular de los modelos generativos en las ciencias sociales lleve tiempo y mucho debate, y que el paso de una primera etapa de aproximación heurística y fragmentada a su uso generalizado, consistente y científicamente validado sea una tarea compleja y desafiante. Puesto en la clave analítica de Bourdieu su impacto alcanzará suficiente madurez una vez que estos recursos pasen a formar parte de las reglas sociales y epistémicas que regulan los intercambios y la producción del conocimiento en el campo de las ciencias sociales.

Consideramos que se vuelve perentorio reflexionar sobre el cambio de contexto y avanzar desde las ciencias sociales en caracterizar estos nuevos escenarios. Lejos de una postura ludista utópica, consideramos que profundizar en el análisis de las técnicas y metodologías computacionales en el contexto del Big Data ayudará de manera decisiva a ampliar el arsenal de recursos con los que los cientistas sociales pueden mejorar los diagnósticos, repensar la teoría y abordar las tendencias en curso. Además, resulta de interés ofrecer a la sociedad civil interpretaciones científicas validadas sobre los cambios en curso, brindando insumos para disputar las lecturas tanto catastrofistas como las del positivismo ingenuo.

El propósito de este trabajo es aportar un marco integrador que articule la reflexión sociológica con la incorporación crítica de métodos computacionales en un contexto marcado por el auge de las inteligencias generativas. Para lograrlo, empleamos una estrategia de revisión conceptual con sistematización exploratoria, que combina el mapeo de nociones y debates clave con el análisis de casos y herramientas representativas (NLP, grafos, aprendizaje automático, minería de procesos) a fin de delimitar alcances, límites y oportunidades para las ciencias sociales. Esta aproximación, de carácter transversal y no exhaustiva, prioriza la claridad didáctica y la pertinencia aplicada, y se orienta a abrir un espacio de investigación y formación que pueda ser replicable y ampliado en distintos contextos institucionales.

1. Las ciencias sociales computacionales

Lazer et al. (2009) definen a la ciencia social computacional como una disciplina emergente que aprovecha la recopilación y análisis masivo de datos digitales para estudiar el comportamiento humano y social. Los autores argumentan que, aunque disciplinas como la biología y la física han avanzado gracias al análisis de grandes volúmenes de datos, las ciencias sociales aún tienen barreras institucionales, metodológicas y éticas que dificultan su desarrollo. Los autores destacan los beneficios potenciales de analizar interacciones sociales a gran escala mediante registros de llamadas, redes sociales en línea y rastreo de movimientos humanos, lo que podría

IoT Analytics: "DeepSeek implications: Generative AI value chain winners & losers" https://iot-analytics.com/winners-losers-generative-ai-value-chain/.
CNBC: "Stock market news for Jan. 26, 2025: DeepSeek, NVDA spark sell off" https://www.cnbc.com/2025/01/26/stock-market-news-for-jan-26-2025.html

Reuters: "DeepSeek sparks AI stock selloff; Nvidia posts record market-cap loss" https://www.reuters.com/technology/chinas-deepseek-sets-off-ai-market-rout-2025-01-27/
https://www.reuters.com/markets/us/nasdaq-futures-tumble-chinas-ai-push-rattles-big-tech-2025-01-27/

transformar la comprensión de la comunicación, la productividad y la salud pública.

Según Oliveira y Cardoso Sampaio (2023), las Ciencias Sociales Computacionales emergen como un campo hibrido formado por la intersección de las ciencias sociales y las ciencias de la computación, en un contexto de crecimiento exponencial de los datos y los métodos computacionales. En la actualidad encontramos poca bibliografía científica que aborde el tema y una oferta reducida de cursos académicos en Ciencias Sociales Computacionales[3]. Vale la pena destacar el trabajo realizado por el equipo Computational Human Dynamic Lab, dirigido por el profesor Marton Karsai, cuyas investigaciones combinan métodos de análisis de datos, modelado matemático y técnicas computacionales, para estudiar dinámicas sociales y comportamientos humanos[4].

Conviene tener en cuenta que la idea de unas Ciencias Sociales Computacionales empieza a circular incluso antes que exista un campo como tal. Si bien hay registros del uso del concepto y expresiones sobre el potencial de estas metodologías (Lazer et al. (2009) y Conte et al (2012), consideramos que aciertan Rosati et al (2013) al señalar que expresa más bien, al menos en sus comienzos, una propuesta de giro metodológico.

Este giro metodológico va acentuando su alcance en la medida que las tendencias que lo originaron se profundizan, al punto que en la actualidad nos encontramos con tres condiciones cruciales a tener en cuenta: 1) La transformación de las relaciones sociales de producción a partir del desarrollo de las fuerzas productivas que impulsan las IA generativas, que está provocando un cambio sin precedentes las formaciones sociales. 2) Con estos cambios se transforman también las relaciones sociales y técnicas que, como decía Bourdieu (2003), regulan ciertos procesos y mecanismos de la comunidad científica, y producen conocimiento a través de la circulación y el intercambio del conocimiento 3) Estamos, como señala Sadin, asistiendo a un cambio en las condiciones socio cognitivas que ordenan las prácticas del sujeto que realiza la objetivación. En este punto el renovado ejercicio de la vigilancia epistemológica será condición imprescindible para construir un horizonte supervisado en relación a las herramientas epistémicas y técnicas que se construyan en las mediaciones con las inteligencias generativas.

También es cierto que en el abordaje de este tema existen riesgos relacionados con la privacidad, la propiedad de los datos y la necesidad de modelos de colaboración entre la academia y la industria. Lazar et al (2023) enfatizan la urgencia de establecer una infraestructura académica abierta y regulaciones adecuadas para fomentar esta nueva área del conocimiento

Respecto del primer factor es necesario abordar las investigaciones que desde las ciencias sociales y otras disciplinas afines avanzan sobre estas transformaciones, ya que es importante contar con información precisa y suficiente sobre las múltiples dimensiones de la transición. Por mencionar algunos textos en esta línea, el conjunto de artículos presentados en la publicación *OK Pandora* (2024), el ya

[3] Entre los cuales destaca el ofrecido por el profesor Martin Hibert y un nutrido grupo de especialistas en el tema dictado de manera online por Universidad de California Davies (USA). En ese curso se presta especial atención a los métodos computacionales y al crecimiento exponencial de los datos en la era del Big Data. Con un enfoque al que clasificaríamos entre empirista e instrumental, la reflexión se concentra en el uso adecuado de las técnicas y como estas pueden ofrecer nuevas formas de acceso a la realidad social. El curso ofrece una interesante propuesta en tanto invita a conocer y reconocer que los cambios tecnológicos en materia de tecnologías digitales impactan de manera decisiva sobre la producción de conocimiento y, por ello, las ciencias sociales computacionales no pueden ser ajenas a estas transformaciones.

[4] Entre las temáticas investigadas por el equipo están: grafos, patrones de movilidad humana, difusión de la información, Epidemiología computacional, desigualdades sociales. Para más información ver: https://www.martonkarsai.com

mencionado texto de Sadin y otros del mismo autor, y el trabajo de Berti (2022) son aportes valiosos a la temática. Actualmente se observa el crecimiento de investigaciones en esta materia.

Respecto del impacto de los cambios sobre las ciencias sociales, resulta indispensable pensar a las Ciencias Sociales Computacionales desde un enfoque crítico que problematice las relaciones entre tecnología, poder y desigualdad. Como plantean autores de la sociología digital (Couldry y Mejías, 2019), los procesos de datificación no son neutrales, sino que se inscriben en dinámicas de acumulación y control que redefinen los vínculos sociales y las formas de subjetivación. Así, la posibilidad de analizar comportamientos a gran escala debe articularse con una reflexión sobre quién produce los datos, bajo qué intereses circulan y de qué manera su uso puede reproducir o, en cambio, desafiar las estructuras de dominación existentes. La tarea de la sociología, entonces, no se limita a incorporar herramientas computacionales, sino también a reponer categorías clásicas —como clase, género, etnicidad o territorio— en escenarios donde los algoritmos y las plataformas median crecientemente la vida social. Parece así oportuno abordar un enfoque realista y relacional como el que propone Bourdieu (1999) que ayude a caracterizar las condiciones específicas del campo de las ciencias sociales, sus fronteras y el modo en que los nuevos avances técnicos repercuten, o no, en su interior.

En efecto, reconocer las dinámicas propias del campo científico implica atender a las luchas simbólicas y materiales que lo atraviesan. Como sugiere la sociología de la ciencia, la incorporación de nuevas herramientas técnicas no es un proceso lineal, sino que se ve mediado por relaciones de poder, jerarquías disciplinares y capitales específicos (académicos, políticos, tecnológicos). Desde esta perspectiva, el ingreso de metodologías computacionales al campo de las ciencias sociales abre disputas por la legitimidad de ciertos saberes, la distribución de recursos y la definición de agendas de investigación. Por ello, además de un reto teórico y de política institucional, se trata de un desafío práctico en torno a cómo se configuran las posiciones dominantes y dominadas dentro del campo y de qué manera estas posiciones condicionan las posibilidades de innovación, colaboración y apertura interdisciplinaria.

Por último, y en estrecha relación con lo recién señalado, en el nivel de las prácticas es necesario abordar estas transformaciones a nivel de su curricularización en la formación de grado y posgrado, en las instancias de formación de los equipos de investigación e impulsando las mismas desde las agencias de ciencia y técnica, como está empezando a ocurrir en algunas universidades de Argentina y España ante la creciente demanda de información sobre estos avances.[5]

Lo decisivo en relación a los esfuerzos institucionales y colectivos es considerar el impacto del desarrollo de los métodos computacionales y el crecimiento exponencial de los datos (producidos, capturables y procesables) sobre el campo de las ciencias sociales. Entre los métodos que se utilizan podemos señalar: desarrollo del análisis factorial, las técnicas de procesamiento de lenguajes naturales (NLP, en inglés), extracción automática de contenido, sistemas de información geográfica y social, análisis de redes sociales, modelado computacional, aprendizaje supervisado y no supervisado y análisis de trayectorias, entre otros. En todos los casos estamos hablando de técnicas que se aplican en diversos campos de la investigación y producción y que exceden con

[5] En esa dirección, el Observatorio de Empleo de la UPO se constituye como un ejemplo concreto de institucionalización de estas transformaciones, ya que ha logrado articular producción de conocimiento, formación de recursos humanos y transferencia de información estratégica hacia la sociedad. El seminario impartido en la Universidad Pablo de Olavide en el seno del proyecto INCASI2 en este contexto constituye, asimismo, una muestra del modo en que estas iniciativas pueden convertirse en espacios de experimentación pedagógica e innovación metodológica. Allí se conjugan el trabajo académico, la reflexión crítica y la apropiación de nuevas herramientas computacionales, avanzando en la necesaria actualización de los planes de estudio y en la consolidación de una cultura investigadora capaz de responder a los retos emergentes que plantea la sociedad digital.

creces el campo de las ciencias sociales computacionales[6].

Finalmente, y en dirección a las tareas de vigilancia epistemológica, es importante reconocer que, si en este proceso se le da preponderancia al enfoque metodológico sobre los postulados teóricos, asociando la formulación teórica con el modelado experimental, se corre el riesgo de no observar el principio metodológico que plantean Bourdieu, Passeron y Chamboredon (1975) acerca de que el vector va de lo racional a lo real, y que los hechos sociales por si solos no nos dicen nada. Un enfoque empirista que se subordine a enfatizar las correlaciones, y una visión formalista y abstracta de la teoría, dejan en un segundo plano a las teorías sociales y, con ello, al valioso aporte que estos marcos de referencia teórica hacen a la investigación en ciencias sociales. Se trata, más bien, de re-inteligir los procesos de construcción del objeto a la luz de los nuevos aportes y desde una perspectiva multidimensional, más que refundar el campo. Grandes desafíos esperan respuestas colectivas y nuevas polémicas de la razón epistemológica.

2. Lenguajes computacionales y ciencias sociales

Los cambios en la matriz tecnológica conllevan también una trasformación orientada hacia una presencia cada día más decisiva de las tecnologías digitales. En este sentido los desarrollos en materia de software y hardware son también cruciales para entender las tendencias actuales. El abandono de las formas "analógicas" de producción hacia modelos digitales es un rasgo de época que se expresa en un abanico de recursos computacionales al servicio del desarrollo industrial, comercial y logístico. La base técnica de este desarrollo sostenido son modelos informáticos formulados en una variedad de lenguajes computacionales (Python, Java, R, SQL, C++, etc.) que se renuevan periódicamente y amplían sus posibilidades de uso.

Paralelamente, estos recursos computacionales pasan a integrar de manera cada vez más decisiva los procesos de investigación científica, en la medida en que hacen posible un tipo de tratamiento de datos y modelado que habilita nuevos horizontes a las investigaciones. Si bien la aplicación de modelos estadísticos y representaciones gráficas estructuradas no es nueva en disciplinas de las llamadas ciencias físico naturales, y en menor medida en las ciencias sociales, los actuales modelos impulsados en el área de la inteligencia generativa y el aprendizaje automático han impactado en extensión y en profundidad en la investigación aplicada. En el caso de las ciencias sociales estos recursos están más focalizados a un grupo de técnicas estadísticas que permiten trabajar con grandes volúmenes de datos y analizar relaciones. Consideramos que el desarrollo de las inteligencias generativas permite, sin lugar a duda, ampliar el horizonte de recursos técnicos a los que recurrir en materia de investigaciones en ciencias sociales, al reducir el costo de acceso a los modelos computacionales, en tanto y en cuanto el conocimiento experto de esos lenguajes no opera ya como una barrera para su utilización.

2.1 Python y R

La elección del lenguaje de trabajo pasa a ser, entonces, una decisión teórica y técnica que impacta sobre las posibilidades de incorporación de los recursos computacionales al proceso de construcción del objeto. En tal sentido conviene distinguir dos lenguajes utilizados de manera recurrente por la comunidad académica: Python y R.

[6] El análisis factorial permite reducir una multiplicidad de variables a un número reducido de factores subyacente, el NLP utiliza algoritmos para entender y generar respuestas en lenguaje humano, la extracción automática de contenido facilita el acceso a múltiples fuentes de información, los sistemas de información geográfica y social nos permiten trabajar con datos geolocalizados, etc. Veremos con más detalle algunas de estas técnicas en los próximos apartados.

El primero es un lenguaje de propósito general, y es el lenguaje de programación más utilizado en la actualidad, en especial en el área de la ciencia de datos, el aprendizaje automático y los Modelos Grandes de Lenguaje. Es un lenguaje de sintaxis sencilla con una amplia colección de librerías para el análisis y el modelado. Encontramos, además, herramientas como los jupyter notebooks que facilitan la visualización de los resultados y la documentación del código, y que están cada día más presentes en los procesos de producción de conocimiento científico.

Desde una perspectiva sociológica, resulta interesante observar cómo la elección de Python no sólo responde a cuestiones técnicas, sino también a dinámicas comunitarias y culturales que configuran el campo científico. La extensa comunidad internacional de usuarios y desarrolladores de Python genera un capital social específico, materializado en foros, repositorios de código abierto y proyectos colaborativos, que permite una circulación más democrática del conocimiento técnico. Este carácter abierto y cooperativo, en el que se comparte y se discute colectivamente, favorece la apropiación de recursos computacionales por parte de grupos de investigación con menos acceso a infraestructuras costosas, ampliando así las posibilidades de producción de conocimiento. De este modo, la difusión de Python en las ciencias sociales computacionales no sólo tiene implicaciones metodológicas, sino también políticas y epistemológicas, ya que redefine las fronteras entre expertos técnicos y académicos, y entre centros con mayores o menores recursos. Por ejemplo, Brooker (2019) en Programming with Python for Social Scientists muestra cómo investigadores con poca experiencia pueden incorporar Python en el diseño de investigación social, lo que ilustra el papel formativo del lenguaje.

Por otro lado, R es un lenguaje de programación de código abierto y se utiliza para cálculos estadísticos, análisis de datos y aprendizaje automático. Es el lenguaje de programación más utilizado por la comunidad científica en el tratamiento de la información en los procesos de investigación, en particular en análisis estadístico, la generación de gráficos y el aprendizaje automático. Ofrece una gran cantidad de paquetes y bibliotecas específicas para análisis estadístico, lo que lo convierte en un valioso recurso para quienes se dedican a la investigación científica. Tiene también sus propios cuadernos en formato Markdown y es quizá, el lenguaje computación de investigación de mayor presencia y desarrollo dentro de la comunidad de cientistas sociales.

Desde la óptica sociológica, el uso de R también se asocia con la consolidación de comunidades académicas transnacionales que comparten código, desarrollan paquetes específicos y establecen estándares metodológicos en torno al análisis de datos sociales. La lógica colaborativa del software libre fortalece una cultura de apertura y transparencia que ha sido reclamada históricamente en las ciencias sociales, en tanto posibilita la replicabilidad de los estudios y la comparación entre contextos. Asimismo, la amplia tradición de R en la investigación social ha favorecido la integración de técnicas clásicas —como los modelos de regresión o el análisis de correspondencias— con enfoques más recientes, como el análisis de redes sociales o el aprendizaje automático, lo que permite mantener un puente entre herencias metodológicas y nuevos desafíos computacionales. Kahmann, Niekler y Wiedemann (2021) presentan el Leipzig Corpus Miner, herramienta basada en R, que permite combinar minería de contenido y análisis cuantitativo y cualitativo, ejemplificando la potencia de R en estudios sociales contemporáneos. Asimismo, Vissoci et al. (2013) proponen un marco para la investigación interactiva y reproducible en ciencias sociales usando R, subrayando la importancia de la transparencia, la apertura del código y la documentación como condiciones para una investigación sólida y ética. En consecuencia, R no sólo constituye una herramienta técnica potente, sino también un dispositivo institucional que fomenta la cooperación, la reflexividad crítica y la democratización del acceso a metodologías avanzadas en la disciplina.

Los dos lenguajes mencionados permiten, entre otras cosas, el desarrollo de procesos complejos

de análisis de datos y ofrecen una amplia gama de recursos para el tratamiento de la información de diverso tipo y formato. Además, el impulso de las inteligencias generativas ha permitido algo que hasta hace muy poco implicaba un arduo y tedioso trabajo: la traducción de un bloque de código expresado en un tipo de lenguaje a otro. Este es un punto no menor en tanto que permite potenciar los recursos (expresados en paquetes y librerías) en aquellos casos que los que han sido construidos por la comunidad científica para fines específicos en un lenguaje, y circunstancialmente no encuentran equivalentes en el otro. Una de las tareas con la que nos encontraremos en breve será una clasificación exhaustiva de estos recursos por dominio de conocimiento específico, su posibilidad de traducción y la combinación con nuevos métodos y estrategias computacionales.

2.2 NodeXL, Dedoose, Elicit

Asistimos asimismo a la emergencia de una serie, cada día más amplia y diversa, de aplicaciones desarrolladas específicamente para la investigación en el campo de las ciencias sociales y la integración de los distintos tipos de datos. Podemos decir que representan una segunda generación de soportes informáticos respecto de softwares clásicos como SPSS, Stata o Atlas.ti. Entra las más conocidas están: NodeXL, Dedoose y cuentan con sus propios manuales de procedimiento y espacios de intercambio. También han aparecido plataformas impulsadas por inteligencias generativas orientadas a la investigación académica y la consulta de fuentes bibliográficas como Elicit o Dimensions AI.

Más allá de su funcionalidad instrumental, estas plataformas expresan un cambio en el modo en que se organizan y legitiman las prácticas de investigación en ciencias sociales. NodeXL, por ejemplo, ha permitido que técnicas de análisis de redes sociales —antes restringidas a entornos estadísticos complejos— se vuelvan accesibles a un público académico más amplio gracias a su integración con Excel y a la posibilidad de visualizar datos de manera inmediata. Dedoose, por su parte, constituye un avance significativo en la investigación mixta, ya que combina el análisis cualitativo con el cuantitativo en entornos colaborativos en línea, favoreciendo dinámicas de trabajo en equipo distribuidos. Finalmente, Elicit y Dimensions AI, en tanto asistentes potenciados por inteligencia artificial, abren un nuevo horizonte al automatizar parte de la revisión de literatura y la sistematización de información bibliográfica, aunque también plantea desafíos éticos vinculados a la opacidad algorítmica y a la dependencia de infraestructuras privadas.

Desde una perspectiva sociológica, la emergencia de estas herramientas puede leerse como parte de una segunda ola de informatización de la investigación social, en la que el software ya no solo procesa datos, sino que incide directamente en la definición de problemas, en la elección de técnicas y en la organización del trabajo académico. De este modo, los debates metodológicos se ven entrelazados con discusiones sobre autonomía científica, acceso desigual a los recursos y vigilancia epistemológica, recordándonos que cada soporte técnico no es neutro, sino que se inserta en campos de poder y en luchas por la legitimidad disciplinar. Investigaciones recientes muestran que asistentes de inteligencia artificial como Elicit pueden servir como herramientas complementarias valiosas en la elaboración de revisiones sistemáticas, particularmente para etapas como la búsqueda bibliográfica y la selección de artículos, mejorando la eficiencia sin reemplazar completamente los métodos tradicionales (Bolaños, Salatino, Osborne & Motta, 2024). Este tipo de hallazgos refuerzan la idea de que plataformas como Elicit, NodeXL o Dedoose no solo ofrecen facilidades técnicas, sino que también moldean prácticas de investigación, alianzas institucionales y debates metodológicos sobre rigor, transparencia y autoridad académica.

2.3 Lenguajes y modelos

Para cerrar este breve apartado queremos señalar nuevamente la decidida importancia que tiene el desarrollo de los llamados Modelos Fundacionales (ChatGPT, Claude, Meta, DeepSeek, etc.) en la transición asistida de los cientistas sociales hacia la generación del código

y los modelos a los que haremos referencia en los siguientes apartados.

En este contexto, no se trata únicamente de reconocer la potencia técnica de los Modelos Fundacionales, sino de analizar cómo su irrupción redefine el lugar de los cientistas sociales en el proceso de producción de conocimiento. Estas herramientas no sólo facilitan la transición hacia el aprendizaje del código y la experimentación con modelos, sino que también generan un cambio en la división del trabajo académico, abriendo la posibilidad de que investigadores sin formación computacional avanzada accedan a recursos de análisis sofisticados. Al mismo tiempo, emergen nuevos desafíos vinculados a la dependencia de infraestructuras privadas, a la transparencia de los procesos algorítmicos y a la necesidad de fortalecer competencias críticas que permitan usar estas tecnologías sin perder de vista las mediaciones sociales, culturales y políticas que atraviesan todo proceso de investigación. Por último, la llamada *delegación cognitiva* aparece también como un problema a analizar en profundidad ante estos cambios.

3. El acceso a la información en la era del Big Data

Uno elemento distintivo de los cambios en el actual contexto socio histórico es el crecimiento exponencial de los datos, y la importancia que tienen los mismos en el desarrollo de la economía, al punto que se los denomina como el "petróleo del siglo XXI". Datos y meta datos, en su carácter de commodities, son fuente fundamental de un nuevo tipo de extractivismo y se generan en una escala sin precedentes. La producción, mensura, captura y organización de los mismos es el pilar fundamental de las principales empresas a nivel global, que comandan los bloques del poder en el capitalismo actual. Este proceso es caracterizado como un cambio ontológico implicado en la expansión de las tecnologías digitales a todas las esferas del quehacer humano (Berti, 2022).

Además, hay evidencia empírica y técnica que respalda los roles que esas empresas desempeñan en la infraestructura informática global. Por ejemplo, un estudio sobre Amazon DynamoDB (Elhemlali et al., 2022) describe cómo este servicio NoSQL de Amazon Web Services maneja peticiones a gran escala: durante eventos de alta demanda como el "Prime Day", el sistema alcanzó picos de 89.2 millones de solicitudes por segundo, manteniendo disponibilidad alta y latencias muy bajas, lo que demuestra cómo el control de metadatos, almacenamiento distribuido y respuesta en tiempo real son componentes centrales de la operación de Amazon.
En otro caso, investigaciones recientes sobre el consumo energético de nodos acelerados por GPU de Nvidia (Latif et al., 2024) muestran las exigencias físicas implicadas en entrenar modelos grandes. Un nodo con 8 GPUs modelo H100 de Nvidia durante entrenamiento mostró una demanda energética instantánea de ~8.4 kW, incluso con utilización alta, y se observó que al escalar el tamaño de lote (batch size) se reducen los costos energéticos totales de entrenamiento, lo que evidencia también que la infraestructura —hardware, conexión, enfriamiento, eficiencia— no es solo un soporte pasivo sino un factor clave en la factibilidad material de los Modelos Fundacionales.
Asimismo, acerca de Meta (2024), se ha documentado que la empresa realiza iniciativas de acceso de datos para investigación académica, como es el caso del Ad Library Dataset, que proporciona datos agregados sobre anuncios electorales, sobre asuntos de interés público y métricas de interacciones, con el propósito declarado de permitir estudios de impacto social y transparencia en plataformas de redes sociales.

Estas contribuciones científicas muestran que no se trata solo de capacidad técnica o discursiva, sino de que hay componentes medibles — rendimiento, infraestructura, políticas de datos, acuerdos de disponibilidad académica — que respaldan la afirmación de que empresas como Amazon, Nvidia y Meta configuran efectivamente los marcos en los que se produce y se regula el acceso a la información.

3.1 Minería de datos y Webscraping

Este crecimiento exponencial de los datos impacta, además, en la disponibilidad y

diversidad de fuentes de información a la que podemos acceder como investigadores de las ciencias sociales, como en los recursos técnicos que permiten tales procesos de captura. Un término que permite describir este proceso de extracción continua de la información es la llamada minería de datos, un proceso que busca capturar datos para descubrir patrones de información valiosa en grandes conjuntos de datos. Y en este proceso un recurso muy utilizado en el área de producción es lo que se conoce como web scraping o extracción de datos web mediante el uso de programas informáticos o bloques de código producidos ad hoc. Esta es una metodología que viene ganando terreno en el campo de las ciencias y que, al mismo tiempo, nos obliga tanto a reconocer su complejidad técnica, los límites éticos de su aplicación, como a pensar e instrumentar protocolos de accesos seguros y confiable de captura de la información.

Investigaciones como las de Luscombe et al. (2022) muestran que web scraping permite superar barreras en los datos oficiales, acceder a fuentes diversas, y fomentar valores como la apertura y transparencia en las ciencias sociales. Además, hay límites éticos y legales: consentimiento tácito o explícito de los propietarios del sitio, términos de servicio (Terms of Service), privacidad de los datos, riesgo de sobrecarga del servidor, uso indebido, anonimización, impactos no deseados en sujetos humanos. Ejemplos como el marco legal y ético recientemente propuesto en "Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations" (Brown et al., 2024) ofrecen recomendaciones concretas para operar estos métodos de manera responsable. En consecuencia, es imprescindible que los proyectos de investigación que usan web scraping cuenten con protocolos claros de acceso seguro, transparencia en los criterios de extracción, aprobación ética institucional (comités de ética), y criterios de verificación de datos para asegurar confiabilidad, reproducibilidad y responsabilidad científica.

Ya en un texto del 2013, Marres y Weltevrede ofrecen un enfoque esclarecedor sobre cómo el scraping no es solamente una herramienta técnica sino un "dispositivo" que transforma la investigación social al introducir formas específicas de captura de datos propias de los medios digitales. En Scraping the Social? señalan que el scraping tiene la capacidad de reestructurar la investigación en al menos dos dimensiones: primero, al importar supuestos analíticos externos al campo social (por ejemplo, una preocupación por lo nuevo o lo instantáneo —freshness); segundo, al hacer disponibles categorías ya formateadas por prácticas mediáticas (como marcas de fecha, enlaces, formatos de metadatos) que, lejos de ser contaminantes, pueden constituir materias primas analíticas valiosas si se las reconoce como tales. En su estudio de perfilado de temas como "austeridad" en Twitter y Google, los autores muestran cómo los formatos y ciclos de vida propios de los datos en línea moldean el objeto mismo de análisis, convirtiendo al web scraping en un medio que no solo extrae datos, sino que participa activamente en la formación del objeto investigativo.

3.2 Las API

Una alternativa al scraping, que busca mitigar los problemas generados por este tipo de metodologías de captura de la información en la web, es lo que se conoce como conexiones API (Interfaz de Programación de Aplicaciones), y que tiene un enorme desarrollo en la actualidad. Las API son un conjunto de definiciones y protocolos que se utilizan para diseñar e integrar el software de las aplicaciones. Son recursos de software que hacen las veces de intermediarios cuya función es facilitar la integración, el intercambio de datos y la colaboración entre diferentes programas informáticos. Son recursos valiosos para la investigación científica en tanto facilitan la integración y el intercambio de datos, pueden automatizar tareas repetitivas e iterar interacciones, permitiendo a los investigadores dedicar más tiempo al análisis y la interpretación de datos antes que a la compleja tarea de recopilación y limpieza.

Existen en la actualidad numerosos sitios gubernamentales y del sector privado que eligen sistematizar y regular el acceso a la información bajo estos protocolos permitiendo una gestión más eficiente y segura de intercambio de datos.

En este contexto la documentación que explicita los protocolos de intercambio, y que suele estar disponible en las plataformas que disponen de los datos, es un recurso clave para instrumentar los pedidos y el acceso a los datos.

3.3 Los repositorios de bases de datos

Otro fenómeno emergente es el surgimiento de diferentes sitios que, además de estimular e impulsar la ciencia de datos, cuentan con una gran cantidad de bases de datos públicas, como el caso de Kagle.com que en febrero de 2025 contaba con más de 430 mil bases de datos de libre acceso. Sumado a ello se han facilitado los mecanismos de búsqueda de la información con la implementación de herramientas como Google data search (https://datasetsearch.research.google.com) creadas para visualizar las bases de datos presentes en la web y sus respectivos sitios de acceso.

Un fenómeno que no es nuevo, pero que también se ha visto impactado por los cambios en el contexto global, son las bases de datos públicas, que han mejorado o actualizado sus canales de acceso. Un ejemplo de ello es el portal de datos abiertos del gobierno de los Estados Unidos (https://data.gov) donde se pueden encontrar muchísimos datos abiertos recopilados por el gobierno del citado país. También hay universidades como la de Harvard y la de Míchigan que han creado sitios para poner a disposición bases de datos académicos: https://dataverse.harvard.edu, https://www.icpsr.umich.edu. En el caso de Argentina el portal de datos abiertos que tienen disponibles más de 1200 datasets públicos de diverso tipo (https://datos.gob.ar/)

3.4 Plataformas combinadas

Entre los nuevos tipos de instrumentos generados para la gestión de datos en el mundo académico aparecen una serie de iniciativas que combinan artículos académicos, bases de datos, y modelos computacionales. Estos sitios permiten analizar la articulación concreta entre conocimiento científico, bases de datos y algoritmos computacionales. Un ejemplo es la plataforma Zenodo (https://zenodo.org), de acceso abierto y creada por CERN (Organización Europea para la Investigación Nuclear), que permite a investigadores, académicos y profesionales compartir y almacenar todo tipo de productos de investigación, incluidos datasets, publicaciones, software, videos y más. Otra iniciativa en esta línea es Papers with code (https://paperswithcode.com/), una plataforma en la que encontramos artículos científicos y el código que se utilizó en esas producciones, además de algunos datasets. En este repositorio vamos a ver vinculados papers académicos con su código fuente y la metodología.

3.5 Huella Digital

Para finalizar este apartado quisiéramos hacer referencia a un concepto novedoso que nos permite entender la naturaleza de los cambios en la producción y acceso a la información y cómo pueden impactar en las ciencias sociales. Y nos referimos al concepto de huella digital, que hace referencia a cada interacción en el mundo digital que da lugar a un registro de las personas. Cuando hacemos algo en una computadora o en un teléfono celular digitalmente, esa información queda archivada en formato digital, dejando una huella de nuestro comportamiento en diferentes medios digitales. Es un efecto de la digitalización de muchos procesos y procedimientos. Es un tipo de información que se produce en tiempo real, que es utilizada sistemáticamente en el sector industrial y comercial para tomar decisiones y que permite fusionar datos muy variados y complejos. Una dimensión clave de este emergente es el no muestreo, la idea de que potencialmente trabajamos con todo lo que hay, de que registramos toda la población.

Hay una serie de trabajos y experiencias que explicitan el enorme potencial de trabajar con huellas digitales. Por cuestiones de extensión solo haremos mención al pionero e innovador artículo de Blumenstock, Cadamuro y On (2015) que mediante un proceso de ingeniería inversa despliegan una investigación en la que combinan información obtenida mediante una encuesta, con los metadatos de todos los usuarios de telefonía de ese país -provisto por la empresa que

brinda los servicios telefónicos en Ruanda- para relacionar modelos supervisados y no supervisados, construyendo un modelo que les permiten predecir la posición socioeconómica de los ruandeses con un alto grado de exactitud.

4. Explorar y transformar datos en entornos digitales

Una de las fases más significativas de los procesos de investigación en ciencias sociales está relacionada al tratamiento exploratorio de los datos y su preparación en el proceso de construcción del objeto, en la que un trabajo riguroso con la información es condición de posibilidad de alcanzar resultados confiables en términos científicos.

Los cambios tecnológicos a los que venimos haciendo referencia permiten reestructurar parte de este trabajo y llevarlos a entornos de ejecución como los cuadernos jupyter que facilitan un trabajo más dinámico con diversas fuentes de datos. Entre las tareas específicas a desarrollar podemos señalar: resumen de características principales, identificación de patrones y relaciones entre variables, detección de anomalías, verificación de hipótesis, eliminación de duplicados, detección de valores faltante, normalización, imputación, etc.

Este proceso es aún más complejo cuando las fuentes de datos ya no se reducen a las clásicas y confiables bases de datos estructuradas y preprocesadas. En estos casos la preparación de los datos para su posterior análisis demanda una atención especial. Desde una simple adecuación de formato en los datos hasta el proceso de recodificación de variables categóricas para su tratamiento en un modelo de análisis estadístico, requieren atención y supervisión por parte de quienes llevan adelante las tareas de manipulación. Esto se conoce en el ámbito de la ciencia de datos como "resolver los problemas de calidad de los datos".

Un punto importante del uso de los lenguajes computacionales en estas tareas tiene que ver con la enorme capacidad de recursos que facilitan y aceleran el análisis estadístico de los datos y permiten escalar. Dos de las bibliotecas de Python, Pandas y Numpy más utilizadas, ofrecen muchos recursos funcionales en este sentido, a lo que se suman librerías específicas para el tratamiento estadístico, que facilitan y mejoran la tarea interpretativa de la información, al ofrecer una descripción de sus características clara y concisa. En cuanto a la representación gráfica, es importante mencionar librerías de uso extendido como Matplotlib, Seaborn y SciPy, que se han convertido en recursos fundamentales para el análisis de datos. Matplotlib es una de las bibliotecas más antiguas y versátiles de Python, utilizada para generar gráficos estáticos, animados e interactivos, lo que permite construir visualizaciones detalladas y personalizadas. Seaborn, que se construye sobre Matplotlib, facilita la creación de gráficos estadísticos más complejos con menor cantidad de código y con estilos estéticos predeterminados, lo que resulta muy útil en investigaciones sociales que requieren comunicar patrones y relaciones de manera clara. Por su parte, SciPy amplía el ecosistema científico de Python, integrando funciones de estadística, álgebra lineal, optimización y procesamiento de señales, con lo cual no solo se limita a la visualización, sino que aporta herramientas analíticas avanzadas que pueden ser representadas gráficamente en combinación con Matplotlib o Seaborn. En conjunto, estas bibliotecas permiten a los investigadores sociales traducir grandes volúmenes de datos en representaciones visuales comprensibles, facilitando la interpretación y comunicación de los hallazgos.

Procedimientos habituales en el análisis estadístico de los datos, como el estudio de las medidas de tendencia central o las medidas de dispersión, pueden ejecutarse de manera sencilla y ordenada en estos entornos, en un flujo de trabajo estructurado. Asimismo, las estadísticas descriptivas y generación de histogramas, gráficos de barra o diagramas de caja facilitan la interpretación visual de la distribución y el comportamiento de los datos. Por último, el análisis univariado de los datos y la representación estadística bivariada son, también, recursos útiles para avanzar en una

primera representación exhaustiva de la información disponible.

5. Machine Learning y Ciencias Sociales

El aprendizaje automático (machine learning) es un subcampo de la inteligencia artificial que se ocupa de cómo construir programas informáticos (algoritmos y modelos computacionales) que mejoren automáticamente con la experiencia (Mitchell, 1997). La clave está en que los algoritmos "aprenden" a detectar la estructura y los patrones en los datos para mejorar su desempeño y tomar decisiones basadas en datos. En las ciencias sociales, este potencial resulta particularmente atractivo, ya que el acceso a volúmenes masivos de información —encuestas, datos administrativos, redes sociales, sensores— abre nuevas posibilidades para comprender fenómenos complejos. Sin embargo, la calidad y el volumen de los datos son condiciones críticas: modelos entrenados con información sesgada o insuficiente no solo reducen la fiabilidad, sino que pueden amplificar desigualdades ya existentes (Lazer & Radford, 2017). En este terreno se distinguen varias modalidades de aprendizaje. El aprendizaje supervisado emplea datos etiquetados para predecir variables de interés, siendo útil para estimar resultados educativos o laborales a partir de características sociodemográficas. El aprendizaje no supervisado busca identificar estructuras latentes sin etiquetas previas, lo que permite descubrir comunidades en redes sociales o segmentar grupos poblacionales en encuestas. Por su parte, el aprendizaje por refuerzo, más asociado a la economía computacional y la robótica, también empieza a utilizarse en simulaciones sociales para analizar interacciones entre agentes en contextos de cooperación y competencia (Kitchin, 2014). De este modo, cada tipo de ML ofrece ventajas específicas para abordar distintos problemas de investigación.

Las aplicaciones en ciencias sociales son diversas. En el análisis de texto y discurso, los algoritmos de clasificación permiten detectar narrativas dominantes, identificar discursos de odio en plataformas digitales o analizar la polarización política. En el estudio de la movilidad social, los modelos predictivos ayudan a proyectar la probabilidad de alcanzar ciertos logros educativos o posiciones laborales a partir de trayectorias longitudinales. En la investigación sobre comportamiento electoral, se aplican técnicas de minería de datos y ML para prever patrones de voto y medir la influencia de campañas en redes sociales. Incluso en la sociología urbana, los datos de geolocalización permiten estudiar procesos de segregación espacial y movilidad cotidiana, generando información antes inaccesible con métodos tradicionales (Nelson, 2020).

Cuando indagamos en el modo en que los algoritmos de aprendizaje automático se aplican a un determinado contexto, es importante señalar que para su correcta ejecución se vuelve imprescindible conocer el dominio del problema, es decir el ámbito de conocimiento en el que el modelo va a ser desarrollado, y el tipo de datos con el que trabajaremos.

La incorporación de ML en ciencias sociales plantea retos epistemológicos y éticos. Uno de ellos es la opacidad algorítmica: mientras los modelos complejos como las redes neuronales profundas producen predicciones muy precisas, su falta de interpretabilidad dificulta comprender los mecanismos sociales subyacentes. Otro es el problema de los sesgos de datos, que reproducen y potencian desigualdades al reflejar estructuras históricas de discriminación. Finalmente, cuestiones vinculadas a la privacidad y consentimiento obligan a reflexionar sobre el uso legítimo de datos personales y sobre la necesidad de marcos regulatorios claros (Bourdieu, 2003).

Ante ello, el desafío metodológico radica en articular el potencial predictivo del ML con la tradición crítica de las ciencias sociales. Como propone Nelson (2020) en su enfoque de computational grounded theory, se trata de combinar la minería de datos con el desarrollo teórico reflexivo, integrando la potencia de los algoritmos con la capacidad interpretativa y crítica del investigador social. De este modo, la vigilancia epistemológica que reclamaba Bourdieu se convierte en condición indispensable para evitar el determinismo

técnico y garantizar que el uso del aprendizaje automático contribuya a la construcción de un conocimiento social riguroso, reflexivo y socialmente situado.

5.1 Algoritmos supervisados y no supervisados

Los modelos de aprendizaje automático más utilizados pueden ser clasificados en dos grupos: supervisados y no supervisados. Se llama algoritmos supervisados a aquellos modelos de machine learning que son entrenados con un conjunto de datos etiquetados, lo que significa que para cada entrada de un conjunto de datos conocemos la etiqueta de salida. El objetivo de estos modelos es poder predecir la etiqueta o un valor de los datos nuevos basándose en los datos de entrenamiento. Entre los algoritmos más utilizados en las ciencias sociales tenemos los modelos de regresión lineal simple y múltiple, la regresión logística, y en menor medidas los árboles de decisión, los bosques aleatorios y las Máquinas de Soporte Vectorial (SVM). Los modelos de regresión lineal simple y múltiple, permiten analizar la relación entre una variable dependiente y una o varias variables independientes, constituyendo una de las herramientas clásicas para medir efectos e inferir tendencias. La regresión logística resulta especialmente útil cuando el fenómeno a explicar es de carácter categórico o binario, como puede ser la participación electoral, la inserción laboral o la pertenencia a determinados grupos sociales. Junto a estas técnicas más tradicionales, en los últimos años ha ido creciendo el uso de algoritmos propios del aprendizaje automático, como los árboles de decisión, que facilitan la interpretación de procesos de clasificación a través de reglas jerárquicas; los bosques aleatorios, que mejoran la precisión de los árboles al combinar múltiples modelos; y las máquinas de soporte vectorial, que se utilizan para encontrar fronteras de separación entre clases de datos de forma óptima. Aunque su empleo es aún menos extendido que el de las regresiones, estas herramientas representan un paso importante hacia la incorporación de enfoques más complejos y predictivos en la investigación social.

Las tareas que llevan adelante estos modelos pueden ser clasificadas como de clasificación y de regresión. Los algoritmos de clasificación son aquellos cuyo objetivo es asignar etiquetas o categorías a los registros basándose en sus propiedades y características. En cambio, en los algoritmos de regresión la tarea del modelo es predecir un valor continuo, basándose en las características de los datos. Un ejemplo puede ser la predicción del nivel de ingresos de una persona basándose en el nivel educativo, sus años de experiencia y el sector laboral en el que trabaja.

Los algoritmos no supervisados son aquellos modelos computacionales que trabajan con datos que no tienen asignada una etiqueta de salida. El objetivo de esos modelos es identificar patrones y estructuras subyacentes. Este tipo de herramientas son útiles para tareas de clusterización y reducción de dimensionalidad. En el caso de los algoritmos de clusterización, como K-means (un algoritmo basado en la partición de los datos), los algoritmos de agrupamiento jerárquico o DBSCAN (Density-Based Spatial Clustering of Applications with Noise, que identifica grupos basados en la densidad de puntos), el objetivo es agrupar los datos basándose en las similitudes de sus características, de forma que las instancias que forman parte de un grupo sean más similares entre sí que con las de los otros grupos. Es una herramienta muy útil para la segmentación, el análisis de comunidades y la identificación de tipologías.

Respecto de las técnicas de reducción de dimensionalidad hay que señalar que el objetivo central de estos instrumentos es reducir la cantidad de variables a considerar, al aplicar complejos procedimientos que van extrayendo las características más importantes de los datos que capturan la mayor variabilidad de los mismos. Las técnicas más utilizadas en las ciencias sociales son: el Análisis de Componentes Principales, el Análisis de Correspondencias Múltiples y el Análisis de Factores Latentes.

Para finalizar el apartado quisiéramos mencionar el trabajo de Grimer et al (2021) que analiza

cómo el crecimiento exponencial de los datos favorece la aplicación de los modelos de machine learning en las ciencias sociales, y proponen un modelo agnóstico sin hipótesis rígidas para abordar los problemas de investigación y un enfoque que propone una mirada "adaptada" de las técnicas del aprendizaje automático al contexto teórico y técnico de las ciencias sociales.

6. Procesamiento de Lenguaje Natural

El procesamiento del lenguaje natural (NLP, en inglés) es también una rama de la inteligencia artificial, que busca otorgar a las computadoras la capacidad de entender el lenguaje humano. NLP articula la lingüística computacional con modelos estadísticos, Machine Learning y de Deep Learning. Detrás de estas técnicas está la idea de que las computadoras puedan "comprender" el significado de las expresiones humanas, en particular la intensión y el sentimiento de quien interactúa con ellas (por texto o voz). Entre las principales tareas y algoritmos de NLP podemos mencionar: Análisis de sentimiento, Name Entity Recognition (NER), Stemming y Lemmantization, Bag of Words, Wordclouds, detección de postura (stance detection), extracción de relaciones (relation extraction) y clasificación de temas (topic modeling).

El Análisis de sentimiento es una de las técnicas más utilizadas en NLP, e implica seleccionar un fragmento de un texto y determinar si los datos tienen un sentido positivo, negativo o neutral. Se señala que a partir de la aplicación de esta metodología se podría inferir el sentido que expresa el texto (por ejemplo, un comentario, una reseña o un documento). Por ejemplo, recientemente se han realizado revisiones sistemáticas que sitúan al análisis de sentimiento como uno de los dominios de mayor crecimiento, destacando cómo los modelos basados en deep learning y transformadores (transformers) superan los enfoques lexicográficos tradicionales en precisión y generalización (Mao et al., 2024).

Se llama Reconocimiento de entidades (NER, en inglés) a una técnica que se usa para ubicar y clasificar entidades nombradas en texto (categorías). Estas entidades pueden ser: personas, organizaciones, instituciones, lugares, fechas, cantidades, valores monetarios, etc. Estudios recientes como el de Keraghel et al (2024) resaltan cómo los métodos basados en modelos de transformers y aprendizaje profundo han incrementado significativamente la precisión de NER, permitiendo reconocer entidades en dominios específicos, con escasos datos anotados, y adaptar los modelos a diferentes tipos de texto (doméstico, redes sociales, científicos). Además, Hu et al., (2024) documentan que, junto a los métodos tradicionales basados en reglas y aprendizaje supervisado clásico, los modelos modernos incluyen además embeddings contextuales, redes neuronales recurrentes y modelos híbridos, los cuales mejoran la capacidad de capturar ambigüedad lingüística y variaciones de estilo en distintos dialectos o idiomas. Estas mejoras técnicas tienen implicaciones sociales: al mejorar la detección automática en idiomas menos representados o contextos locales, se abre la posibilidad de reducir sesgos geográficos y culturales, siempre que los conjuntos de datos de entrenamiento reflejen la diversidad de usos lingüísticos.

Stemming y Lematización son técnicas similares: el steaming es una técnica que relaciona las palabras con su raíz, permitiendo agrupar palabras que tienen una raíz común, aunque puede en algunos casos generar palabras difíciles de reconocer. Por otro lado, lematización incluye un paso adicional de análisis lingüístico: reconoce la categoría gramatical (parte de la oración: verbo, sustantivo, adjetivo, etc.) y consulta un léxico o realiza un análisis morfológico para transformar la palabra en su forma canónica o lema. Esto permite que "running", "runs", "ran" se lematice como "run", y que palabras irregulares como "better" se relacionen con "good", si el contexto lo requiere.

El valor de estas distinciones cobra relevancia práctica en ciencias sociales: por ejemplo, en tareas de clasificación de textos, temas o sentimientos, donde la precisión semántica puede afectar interpretaciones de grupos sociales

o patrones discursivos; también en recuperación de información, donde el balance entre velocidad (stemmers más rápidos) y exactitud lingüística (lematizadores más precisos) importa según el objetivo investigativo. Estudios recientes muestran que modelos lematizadores modernos, especialmente aquellos que integran análisis de contexto y modelos estadísticos o de deep learning, logran errores significativamente menores que métodos basados solamente en reglas o heurísticas simples (Karwatowski & Pietron, 2022; Muller, Cotterell & Fraser, 2024).

El modelo Bag of Words (BoW), es una herramienta que trabaja sobre una representación que convierte el texto en vectores de longitud fija. Esto permite representar texto en una estructura numérica de manera que se pueda procesar en modelos de aprendizaje automático. Este enfoque permite que modelos de aprendizaje automático procesen texto de forma cuantitativa, olvidando el orden de las palabras, la sintaxis, o relaciones semánticas más finas, pero sacando provecho de la frecuencia como indicador de importancia dentro del documento. Varios estudios han explorado las ventajas y limitaciones del BoW, por ejemplo Salim & Mustafa (2022) lo señala como uno de los métodos clásicos más simples y rápidos de implementar, útil especialmente en tareas exploratorias o como baseline, pero indica que su simplicidad conlleva debilidades: vectores extremadamente dispersos (sparse vectors), escalabilidad dificultada al crecer el vocabulario, y pérdida de contexto semántico, lo que afecta la interpretación en ámbitos sensibles como género, cultura o dialecto. Otro trabajo reciente de Graff et al. (2025) demuestra que, para ciertos problemas de clasificación de texto (sentimientos, spam, etc.), los modelos BoW siguen siendo competitivos frente a representaciones más modernas cuando los datos son abundantes y bien etiquetados, siempre que se complementen con técnicas como TF-IDF o selección de características (feature selection). También se ha trabajado en mejoras al BoW clásico para mitigar esos problemas: uso de n-grams (pares o tríos de palabras contiguas), reducción de dimensionalidad (por ejemplo, con LSA, PCA), y enriquecimiento con embeddings que agregan semántica, todo ello para preservar significado contextual sin perder la capacidad de escalar y de interpretarse.

Wordcloud o nube de palabra es una técnica muy utilizada en las ciencias sociales que permite identificar y representar las palabras clave en un texto. En la nube de palabras, las palabras más frecuentes tienen una fuente más grande, mientras que las palabras menos frecuentes tienen una fuente más pequeña o más delgada. Schubert et al. (2017), propone mejoras metodológicas como incorporar un corpus de fondo para normalizar la frecuencia de palabras, usar algoritmos como t-SNE para posicionar palabras relacionadas semánticamente de modo que las palabras co-ocurrentes se visualicen próximas entre sí, lo cual enriquece la interpretación al reducir la arbitrariedad del layout aleatorio que, de otro modo, domina muchas aplicaciones de nubes de palabras. Sin embargo, estos estudios también subrayan que las nubes de palabras tienen limitaciones importantes: no capturan el contexto en el que aparece la palabra (lo cual puede llevar a errores de interpretación), no distinguen entre formas distintas de una misma raíz o formas con matices semánticos diferentes, y favorecen la atención visual sobre la lectura crítica —es decir, el que una palabra aparezca grande puede llamar mucho la atención, incluso si no es conceptualmente central para el argumento cualitativo. Estudios de casos en Brasil (Vilela et al., 2020) han aplicado nubes de palabras como herramienta de apoyo en análisis de contenido cualitativo, observando que la técnica sirve para estimular la interpretación y destacar temas emergentes, pero requiere que los investigadores acompañen la nube con una interpretación rigurosa y conocimiento del contexto.

Para cerrar quisiéramos referirnos a otra investigación en el campo de las ciencias sociales que trabaja con técnicas de NLP. En el artículo de Pérez et al. (2022) los autores realizan un interesante trabajo en la que muestran cómo la inclusión de información de contexto en las técnicas de NLP mejora la detección de los discursos de odio en las redes sociales. En particular, desarrollan un nuevo corpus rioplatense recogido en Twitter en el contexto de

la pandemia de COVID-19, donde los mensajes se toman junto con las respuestas de otros usuarios a noticias, lo que permite captar no solo el contenido del mensaje aislado, sino la conversación en que se inserta. Los experimentos muestran que los modelos que utilizan este contexto insertado (por ejemplo, identificando cómo las respuestas condicionan si un post es catalogado como discurso de odio o no, y considerando el tópico de la noticia que motiva la conversación) presentan rendimiento estadístico superior al de los modelos convencionales que analizan los mensajes de forma aislada.

Este estudio es relevante por varias razones metodológicas y sociológicas: primero, porque evidencia que los discursos de odio no son manifestaciones desconectadas, sino que adquieren sentido dentro de interacciones sociales, lo que refuerza la idea de que para comprender fenómenos discursivos digitales se necesita capturar no solo el "qué se dice", sino el "cómo", "dónde" y "quién responde". Segundo, porque pone en evidencia la importancia de construir corpus culturales/locales (en este caso, variedad dialectal rioplatense) que reflejen los usos del lenguaje en contextos particulares, lingüísticos y políticos, evitando sesgos que privilegien al inglés u otros dominios más atendidos. Finalmente, este enfoque con contexto ayuda a mejorar no solo métricas técnicas (precisión, recall, F1), sino también la validez interpretativa de los resultados socialmente significativos: por ejemplo, permite distinguir mejor cuando un mensaje aparentemente ofensivo se convierte en discurso de odio dependiendo del contexto conversacional, lo que tiene implicaciones éticas para cómo diseñamos sistemas de moderación o políticas públicas. Los autores señalan que, aunque agregar contexto mejora los resultados, también aumenta la complejidad del modelo, exige mayores recursos computacionales, y puede generar nuevas fuentes de sesgo (por ejemplo, si solo se analiza contexto para ciertos idiomas o grupos con alta representación en redes). Además, está el desafío de cómo definir los límites del "contexto relevante" (cuántas respuestas, de qué tipo, cuánta historia previa de conversación) para que el modelo sea manejable sin perder interpretabilidad.

En suma, esta investigación de Pérez et al. reafirma muchos de los puntos tratados en este apartado: que técnicas NLP como análisis de sentimiento, NER, Bag of Words, etc., se vuelven más efectivas cuando se complementan con datos contextuales; que la calidad, diversidad y origen del corpus importan críticamente; que los modelos técnicos no funcionan independientemente del marco teórico y de las condiciones sociales bajo las cuales operan; y que la incorporación de estos enfoques debe acompañarse de reflexión ética, institucional y metodológica.

7. Grafos

Hablar de grafos es hacer referencia a redes. Los grafos son estructuras matemáticas compuestas por nodos (o vértices) y aristas (también llamados enlaces). Los nodos son entidades que tienen propiedades, y pueden ser personas, instituciones, lugares, etc. Las aristas dan cuenta de las relaciones que existen entre los nodos, representando sus relaciones o conexiones. En ciencias sociales, los grafos son una herramienta muy útil para modelar y analizar redes de relaciones e interacciones (sociales, económicas o políticas). Podemos pensar en los nodos, por ejemplo, como los agentes que forman parte de una red de relaciones, estudiantes de una carrera universitaria, trabajadores de una fábrica, etc. Al pensar las relaciones sociales que se establecen entre individuos podemos caracterizarlas como los vínculos que los ligan, sea que pensemos a estos como relaciones simétricas o asimétricas[7].

Existen diferentes tipos de grafos, los grafos dirigidos con conexiones en una sola dirección, grafos no dirigidos en los que las conexiones o

[7] Es interesante notar que el enfoque bourdesiano el acento está puesto en analizar las redes desde las estructuras de las relaciones y no en términos de interacciones más o menos espontáneas o regulares. Ello llama a pensar la dimensión relacional e identificar las condiciones objetivas que hacen posibles las interacciones o estructuras vinculares. Ver Gutiérrez (2008).

relaciones no tienen una dirección univoca, y los grafos ponderados en el que las aristas tienen un peso o valor asociado, por ejemplo, la fuerza de una relación puede ser mayor en una dirección que en la otra.

La cantidad y variedad de procesos y relaciones que pueden ser analizadas mediante grafos en ciencias sociales es enorme: análisis de las redes sociales, físicas y virtuales entre individuos o grupos y la influencia de los agentes en las mismas, análisis de redes comerciales, flujos financieros y cadenas de distribución. Se puede abordar las relaciones familiares, las estructuras de poder (ej., relaciones de clase) u otro tipo de relaciones sociales menos estructuradas. También se pueden analizar procesos políticos, los grupos de poder y conflictos sociales.

En el análisis de redes sociales virtuales, este recurso permite definir algunas métricas claves como la centralidad de un grafo, que nos indica la influencia de los nodos. También se puede analizar la cantidad de conexiones que forman parte de un nodo y la densidad de conexiones de un grupo o subgrupo.

Existen diferentes métricas y valores para analizar un grafo, así como una variedad de recursos para representarlos gráficamente. También encontramos distintos tipos de softwares y librerías para trabajar con el análisis de grafos. Dos de las más reconocidas son Neo4j y Gephi. Neo4j es software de base de datos orientado a grafos creado en 2007 por E. Eifrem, P. Neubauer y J. Svensson, que permite gestionar grandes volúmenes de datos conectados y ejecutar consultas complejas de forma eficiente, muy útil en investigaciones que trabajan con datos masivos o no estructurados. Gephi se ha consolidado como una de las plataformas más utilizadas para la exploración visual de redes, permitiendo aplicar algoritmos de detección de comunidades y layouts que facilitan la interpretación gráfica (Bastian et al., 2009).

En Python una de las librerías más utilizadas es NetworkX, creada en 2005 por A. Hagberg, que ofrece diversas funciones para la creación, manipulación y análisis de grafos. Incluye funciones para calcular métricas de centralidad, encontrar cliques o subgrafos, aplicar algoritmos de flujo y generar visualizaciones básicas, integrándose además con librerías como Matplotlib o Plotly para gráficos más complejos (Hagberg et al., 2008). En paralelo, librerías como igraph (disponible en Python y R) ofrecen mayor eficiencia para redes de gran tamaño, mientras que Graph-tool se destaca por su velocidad al trabajar con algoritmos avanzados.

En España hay trabajos pioneros como el de Morillas (1995) que nos muestran que existe un conjunto de investigaciones de referencia que es oportuno recuperar para profundizar en el análisis de los casos de uso de la teoría de los grafos en ciencias sociales. Este tipo de estudios tempranos abrió un campo de exploración que más tarde se vería ampliado con investigaciones aplicadas a fenómenos concretos, como el análisis de redes de colaboración científica, la dinámica de movimientos sociales o la estructura de las élites políticas. Autores como Lozares (1996), Molina (2001) o Lozarez et al. (2003) profundizaron en la aplicación de métricas de centralidad, densidad y cohesión para describir relaciones sociales en contextos organizativos y comunitarios, sentando las bases de un enfoque metodológico que hoy se ha consolidado bajo el paraguas del análisis de redes sociales (ARS). Este proceso permitió que la teoría de grafos dejara de ser solo un recurso matemático abstracto y pasara a convertirse en un lenguaje común en la sociología empírica española y latinoamericana. En este sentido, resulta relevante destacar que la recuperación y actualización de estos aportes no solo permite reconocer la trayectoria del campo, sino también incorporar perspectivas históricas y comparativas en el análisis de problemas sociales contemporáneos.

Finalmente, comentar que si bien hay un enorme potencial en desarrollar el análisis de grafos en ciencias sociales, los estudios sobre grafos, al igual que los desarrollos sobre redes neuronales no son nuevos, llevan al menos unas décadas de trabajo en el campo de las ciencias sociales, y de la economía en particular (Wasserman & Faust, 1994; Goyal, 2009). Sin embargo, lo que aparece como novedoso y al alcance de las

investigaciones de los no expertos en estadísticas o matemáticas son los recursos de software que facilitan su aplicación a diversos campos y procesos (Bastian et al., 2009; y Hagberg et al., 2008).

8. Análisis de Trayectoria y Minería de Procesos

El análisis de trayectorias en las ciencias sociales tiene una larga y heterogénea tradición. Autores como Pierre Bourdieu (1997 y 2011), Guitérrez et al (2021), han dado cuenta del potencial analítico de la categoría. También hay trabajos como los de Roberti (2017) que analizan los usos, significados y potencialidades de la categoría y las metodologías concurrentes, contribuyendo a una comprensión más amplia de los itinerarios vitales, profesionales o colectivos.

En esta línea, podemos encontrar investigaciones que se centran en trayectorias individuales y familiares, así como estudios que exploran trayectorias colectivas (ya sean de clase social, de comunidades u organizaciones). Estas trayectorias pueden ser analizadas mediante la construcción de tipologías comparativas, que permiten agrupar recorridos según características comunes, o a partir de análisis de casos en profundidad, que resaltan la singularidad de los procesos y sus contextos.

En este artículo nos detendremos a analizar la metodología longitudinal, tomando como referencia el artículo de Yepes-Cayuela (2018). La autora busca hacer operativo el concepto de trayectoria laboral a partir de una biblioteca de R que se llama TraMineR, desarrollada en la Universidad de Ginebra, que integra el análisis de secuencias y la metodología Optimal Matching Analysis (OMA). Esta metodología sirve para trabajar con datos estructurados secuenciados a través del tiempo y que mantienen un orden cronológico, permitiendo construir tipologías de trayectorias en base a la "homología" de las secuencias de las trayectorias.

Lo interesante de este tipo de análisis es que no trabaja con eventos individuales de manera aislada, sino que permite comparar trayectorias a lo largo del tiempo, a partir del análisis de la similitud estructural de esos recorridos de los individuos analizados.

En este tipo de procedimientos hay que realizar una adaptación de los datos a los modelos de análisis, lo que permite utilizar posteriormente herramientas como la matriz de costos que define la regla de transformación entre trayectorias y que es necesaria para construir la matriz de distancias, a partir de las cuales es factible medir similitudes y distancias de las trayectorias, facilitando además la identificación de clusters o grupos.

En el caso del análisis de trayectorias en Python existe la posibilidad de medir las distancias entre las trayectorias a través de otras metodologías como Dynamic Time Warping (DTW) que no requiere elaborar matrices de costo. La elección de la metodología debe estar asociada al tipo de análisis que nos interese a utilizar. Además, en este campo se han desarrollado enfoques complementarios como las distancias de Levenshtein, utilizadas para comparar secuencias discretas y medir su similitud a partir del número mínimo de transformaciones necesarias, o los modelos de secuencias ocultas de Markov (HMM), que permiten capturar la dinámica probabilística de los estados en una trayectoria. Estas técnicas han demostrado ser particularmente útiles en ciencias sociales para analizar patrones de movilidad laboral, itinerarios educativos o trayectorias vitales, ya que ofrecen flexibilidad frente a la variabilidad temporal de los eventos y facilitan la comparación de casos individuales con tipologías colectivas (Abbott, 1995; Studer & Ritschard, 2016).

El análisis de trayectorias en base a datos secuenciados en modelos computacionales dinámicos, favorece, además, la interpretación de los datos mediante gráficas como el análisis de las secuencias transversales de las trayectorias, los gráficos como heatmap que representa una matriz de distancia o un gráfico de barras que nos permite visualizar el análisis del tiempo de permanencia promedio de cada estado. A estas representaciones se suman otros recursos como

los diagramas de Sankey, que muestran los flujos y transiciones entre estados de manera intuitiva, o los gráficos de alineamiento de secuencias, que permiten comparar visualmente varios itinerarios individuales en paralelo. Asimismo, las técnicas de clúster aplicadas a secuencias ofrecen la posibilidad de agrupar trayectorias con patrones semejantes, lo cual facilita la construcción de tipologías empíricas y la interpretación de tendencias colectivas. Estos procedimientos han demostrado ser particularmente útiles en el estudio de trayectorias laborales, educativas y vitales, pues permiten captar no solo la distribución de eventos en el tiempo, sino también la intensidad, la recurrencia y la variabilidad entre individuos y grupos. Como señalan Abbott (1995) y Studer y Ritschard (2016), la visualización y comparación de trayectorias mediante representaciones gráficas no solo enriquecen la exploración descriptiva, sino que también constituyen un paso fundamental en la construcción de explicaciones más sólidas acerca de la desigualdad y la movilidad social.

Por último, un área con un potencial interesante en el análisis de trayectorias para datos estructurados es lo que se conoce como Process Mining (Minería de Procesos). Se denomina Minería de Procesos a un conjunto de técnicas y procedimientos orientados a la extracción de información de eventos y procesos, modelar procesos y predecir patrones en los flujos. Estas técnicas, formuladas por el investigador holandés Wil van der Alast (2016) tienen como fin mejorar procesos y sistemas, a través de la realización de diagnósticos precisos sobre procesos en curso y relaciones dinámicas.

Además, process mining integra múltiples perspectivas, no solo la estructura secuencial de las tareas, sino también la dimensión temporal, la dimensión de los recursos (quiénes realizan qué actividad) y otras dimensiones como el costo, los retrasos o cuellos de botella. Por ejemplo, técnicas como process discovery construyen modelos a partir de registros históricos, conformance checking evalúa desviaciones entre lo esperado y lo observado, y desviación de proceso (bottleneck analysis) identifica los pasos que toman tiempos excesivos o que se usan frecuentemente, lo cual es muy útil para políticas públicas, diseño organizativo o evaluación de trayectorias institucionales. Otra ventaja es que estas herramientas permiten visualizar procesos complejos mediante modelos gráficos (BPMN, redes de Petri), dashboards interactivos y análisis predictivo, lo que hace más accesible la exploración para investigadores que no son expertos en estadística avanzada.

No obstante, como toda técnica con ambiciones predictivas y de diagnóstico, la minería de procesos conlleva retos: la calidad y la integridad de los registros de eventos (event logs) es crucial, ya que datos incompletos, timestamps erróneos o recursos mal etiquetados pueden distorsionar los modelos; además, la interpretabilidad de los flujos encontrados requiere interpretación teórica para no caer en determinismos técnicos. En ese sentido, la incorporación del conocimiento estructural (por ejemplo, la teoría organizativa o la sociología institucional) es indispensable para que el análisis no solo describa procesos, sino que los explique.

Si bien estas metodologías fueron desarrolladas para el control, la gestión y el modelado de procesos en el ámbito productivo y comercial, consideramos que la utilización de los algoritmos basados en las Redes Petri, principal insumo de la minería de procesos, puede ser un valioso complemento en el análisis de trayectoria en ciencias sociales. Esto no solo por la dinámica secuencial del análisis, sino porque permite ver interrelaciones entre los recorridos, los puntos de convergencia y bifurcación de las trayectorias, y analizar los momentos en que el flujo de las trayectorias se ralentiza o se interrumpe.

La combinación de Process Mining con técnicas como Optimal Matching Analysis (OMA) o Dynamic Time Warping (DTW) abre nuevas posibilidades en la investigación social aplicada. Mientras que la minería de procesos permite modelar de manera explícita los flujos de eventos, detectar desviaciones y comparar la ejecución real con los modelos normativos, OMA y DTW ofrecen herramientas potentes para medir similitudes y distancias entre trayectorias individuales o colectivas. La integración de estas metodologías incrementa la

capacidad operativa del análisis, ya que no solo se logra una reconstrucción más fiel de las secuencias de prácticas y eventos, sino que también se favorece la identificación de patrones comunes y divergentes entre distintos grupos sociales.

Este cruce metodológico resulta especialmente valioso en estudios de trayectorias educativas, laborales o de movilidad social, donde la dimensión temporal y la identificación de bifurcaciones son claves para comprender la reproducción de desigualdades o la emergencia de nuevos itinerarios. Además, el uso combinado de estas técnicas posibilita una mayor triangulación analítica: por un lado, con la minería de procesos se accede a una visión sistémica de los flujos; por otro, con OMA o DTW se capturan matices en las secuencias individuales que enriquecen la interpretación sociológica.

En suma, la convergencia entre minería de procesos y análisis secuencial constituye un campo en expansión que fortalece la capacidad de las ciencias sociales para comprender fenómenos complejos, dinámicos y estructurados en el tiempo. Al integrar enfoques provenientes de la informática, la estadística y la sociología, se construye un andamiaje metodológico más robusto, capaz de vincular el nivel micro de las trayectorias con los marcos macroestructurales donde estas se inscriben. Este horizonte no solo amplía las fronteras analíticas, sino que también refuerza el potencial de la investigación social para incidir en el diseño de políticas públicas más sensibles a los procesos reales que atraviesan los sujetos y colectivos.

Conclusiones

En este artículo hemos buscado enfocar la atención en las transformaciones sociales estructurales que están impactando tanto en lo que refiere a las fuerzas tecnológicas como en los dispositivos socio cognitivos a partir de los cuales los cientistas sociales caracterizamos lo social. Entendemos que la emergencia de las inteligencias generativas, y los sostenidos y acelerados desarrollos en el campo de la inteligencia artificial, nos convocan a abordar los desafíos inmediatos y mediatos que se nos presentan. Para ordenar estas ideas, organizamos las conclusiones en tres ejes: (a) Transformación tecnológica, (b) Reconfiguración epistémica y (c) Implicaciones institucionales y formativas.

a) Transformación tecnológica

La incorporación paulatina, pero creciente y sistemática, de las herramientas que hemos abordado permite no sólo la utilización de una variada gama de recursos técnicos de enorme potencial, sino la posibilidad de integrar tanto diversas fuentes de datos como estrategias metodológicas.

Además, la posibilidad de generar código eficiente a partir de los modelos fundacionales no sólo acorta la posibilidad de acceso y uso a modelos computacionales hasta hace muy poco muy difíciles de utilizar para quienes pertenecen al campo de las ciencias sociales, sino que nos permite supervisar el proceso de construcción técnica del objeto, así como formular soluciones funcionales ad hoc para los problemas y/o desafíos que estas complejas herramientas presentan a la hora de su instrumentación.

La disponibilidad de acceso a entornos de ejecución como los de Google Colab no sólo potencian las posibilidades de acceso a recursos abiertos, sino que permiten integrar el ciclo de los datos, desde el proceso de captura y carga de las fuentes de información, pasando por las etapas de limpieza y tratamiento hasta la construcción y aplicación de modelos computacionales complejos, eficientes y escalables. La posibilidad de integrar texto y celdas de código facilita también la documentación in situ de las estrategias metodológicas. Igualmente, estos formatos emergentes se integran muy bien en los procesos y entornos colaborativos cada vez más presentes en el trabajo colectivo.

b) Reconfiguración epistémica

En este sentido, resulta imprescindible reconocer que los cambios tecnológicos no se reducen a la incorporación de nuevas herramientas, sino que producen

transformaciones en los modos de producción de conocimiento, en las formas de interacción entre comunidades científicas y en la configuración de los propios objetos de estudio. La inteligencia artificial y los modelos generativos no solo amplían las fronteras metodológicas, sino que tensionan nuestras categorías de análisis, obligándonos a revisitar críticamente conceptos y enfoques tradicionales en las ciencias sociales.

Esto nos coloca ante un escenario en el que la frontera entre lo técnico y lo epistemológico se difumina, generando nuevas oportunidades para la investigación interdisciplinaria. La capacidad de intervenir en la programación, y no solamente en el análisis de los resultados, dota a los cientistas sociales de un papel activo en la definición de los modelos, evitando así la dependencia de soluciones cerradas y externas.

Consideramos que, un nuevo tipo de vigilancia epistemológica, o quizá una vigilancia epistemológica expandida a nuevas dimensiones, aparece como condición de posibilidad de un uso adecuado de estos recursos, destacando la importancia del conocimiento de los dominios científicos específicos como condición de posibilidad de la construcción del objeto, mediada por estas estrategias de construcción co-constitutivas. Es probable que las tareas inmediatas vayan en este sentido y requieran de nuestros mejores esfuerzos.

c) Implicaciones institucionales y formativas

Al mismo tiempo, esta integración plantea un reto de gobernanza científica y de formación, pues exige desarrollar competencias digitales en los investigadores y estudiantes, así como generar espacios institucionales que garanticen la circulación, la evaluación crítica y el uso responsable de estas herramientas. El potencial de estas tecnologías se materializa únicamente cuando existe una infraestructura académica y profesional capaz de sostener su apropiación reflexiva.

De este modo, se trata de consolidar un nuevo ecosistema investigativo donde la colaboración, la apertura y la reproducibilidad se convierten en valores centrales. La adopción de entornos compartidos, junto con repositorios abiertos de datos y código, no solo favorece la transparencia metodológica, sino que fortalece la capacidad de construir conocimiento colectivo y comparado en el marco de redes internacionales de investigación.

En conclusión, las ciencias sociales enfrentan una coyuntura histórica en la que la innovación tecnológica y la transformación social avanzan de manera simultánea y mutuamente constitutiva. Con esta expresión nos referimos a que no se trata de procesos paralelos o independientes, sino que cada uno influye activamente en el otro: los desarrollos tecnológicos generan nuevas formas de organización, interacción y producción social, mientras que las demandas, tensiones y dinámicas sociales condicionan a su vez las orientaciones, los usos y los límites de las innovaciones tecnológicas. Es, por tanto, una relación bidireccional en la que sociedad y tecnología se coproducen y transforman recíprocamente. El desafío no es solo adaptarse a nuevas herramientas, sino repensar nuestras prácticas de investigación, formación y producción de conocimiento en clave crítica, inclusiva y ética, para asegurar que estas transformaciones se traduzcan en avances sustantivos para la comprensión de la sociedad y su transformación.

Bibliografía

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21(1), 93–113. https://doi.org/10.1146/annurev.so.21.080195.000521

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the Third International Conference on Web and Social Media*, 3(1), 361–362.

Berti, A. (2022). *Nanofundios*. Universidad Nacional de Córdoba.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Bolaños, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: Opportunities and challenges. *Artificial Intelligence Review*, 57, 259. https://doi.org/10.1007/s10462-024-10902-3

Bourdieu, P. (1997). La ilusión biográfica. En *Razones prácticas*: *Sobre la teoría de la acción* (pp. 74–83). Anagrama.

Bourdieu, P. (1999). *Meditaciones pascalianas*. Anagrama.

Bourdieu, P. (2003). *El oficio de científico*. Anagrama.

Bourdieu, P. (2011). Porvenir de clase y causalidad de lo probable. *En Las estrategias de la reproducción social* (pp. 77–134). Siglo XXI.

Bourdieu, P., Passeron, J.-C., & Chamboredon, J.-C. (1973). *El oficio de sociólogo*. Siglo Veintiuno Editores.

Brooker, P. (2019). *Programming with Python for social scientists*. SAGE Publications.

Brown, M. A., Gruen, A., Maldoff, G., Messing, S., Sanderson, Z., & Zimmer, M. (2024). Web scraping for research: Legal, ethical, institutional, and scientific considerations. *arXiv*. https://doi.org/10.48550/arXiv.2410.23432

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V., Moat, S., Nadal, J. P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M., & Helbing, D. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, 214, 325–346.

Couldry, N., & Mejias, U. A. (2022). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

Elhemali, M., Gallagher, N., Gordon, N., Idziorek, J., Krog, R., Lazier, C., Mo, E., Mritunjai, A., Perianayagam, S., Rath, T., Sivasubramanian, W., Sorenson, J. C., Sosothikul, S., Terry, D., & Vig, A. (2022). *Amazon DynamoDB: A scalable, predictably performant, and fully managed NoSQL database service*. Amazon Science.

Oliveira, G., & Cardoso Sampaio, R. (2023). Ciencias sociales computacionales y análisis de contenido: Reflexiones a partir de la producción latinoamericana. *Revista de Estudios Brasileños*, 10(21), 151–167. https://doi.org/10.14201/reb20231021151167

Graff, M., Moctezuma, D., & Téllez, E. (2025). Bag-of-words approach is not dead: A performance analysis on a myriad of text classification challenges. *Natural Language Processing*, 11, 100154. https://doi.org/10.1016/j.nlp.2025.100154

Goyal, S. (2009). *Connections: An introduction to the economics of networks*. Princeton University Press.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24, 395–419.

Gutiérrez, A. (2008). Redes e intercambio de capitales en condiciones de pobreza: Dimensión relacional y dimensión vincular. Redes: *Revista Hispana para el Análisis de Redes Sociales*, 14, 1–26.

Gutiérrez, A., et al. (2021). *De la grieta a las brechas*. Eduvim.

Hagberg, A., Schult, D., & Swart, P. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy 2008)* (pp. 11–15).

https://api.semanticscholar.org/CorpusID:16050699

Hu, Z., Hou, W., & Liu, X. (2024). Deep learning for named entity recognition: A survey. *Neural Computing and Applications*, 36(5), 8995–9022. https://doi.org/10.1007/s00521-024-09646-6

Kahmann, C., Niekler, A., & Wiedemann, G. (2021). Application of the interactive Leipzig Corpus Miner as a generic research platform for the use in the social sciences. *arXiv*. https://arxiv.org/abs/2110.02708

Karwatowski, M., & Pietron, M. (2022). Context-based lemmatizer for Polish language. *arXiv*. https://doi.org/10.48550/arXiv.2207.11565

Keraghel, I., Morbieu, S., & Nadif, M. (2024). Recent advances in named entity recognition: A comprehensive survey and comparative study. *arXiv*. https://doi.org/10.48550/arXiv.2401.10825

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 1–12. https://doi.org/10.1177/2053951714528481

Latif, I., Newkirk, A. C., Carbone, M. R., Munir, A., Lin, Y., Koomey, J., Yu, X., & Dong, Z. (2024). Empirical measurements of AI training power demand on a GPU-accelerated node. arXiv. *arXiv*:2412.08602.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). *Computational social science. Science*, 323(5915), 721–723.

Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology*, 43, 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

López, C., Balmaceda, T., et al. (2024). *Ok, Pandora: Seis ensayos sobre inteligencia artificial*. El Gato y La Caja.

Lozares, C. (1996). La teoría de redes sociales. Papers: *Revista de Sociología*, 48, 103–126. https://doi.org/10.5565/rev/papers/v48n0.1814

Lozares Colina, C., Verd Pericás, J. M., Martí Olivé, J., & López Roldán, P. (2024). Relaciones, redes y discurso: Revisión y propuestas en torno al análisis reticular de datos textuales. *Revista Española de Investigaciones Sociológicas*, (101), 175–200. https://doi.org/10.5477/cis/reis.101.175

Luscombe, A., Dick, K., & Walby, K. (2022). Algorithmic thinking in the public interest: Navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity*, 56(3), 1023–1044. https://doi.org/10.1007/s11135-021-01164-0

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University – Computer and Information Sciences*, 36(4). https://doi.org/10.1016/j.jksuci.2024.102048

Marres, N., & Weltevrede, E. (2013). Scraping the social? Issues in live social research. *Journal of Cultural Economy*, 6(3), 313–335. https://doi.org/10.1080/17530350.2013.772070

Meta. (2024). *Meta Ad Library Dataset* [Data set]. Meta. https://www.meta.com/ad-library/

Mitchell, T. (1997). *Machine learning*. McGraw-Hill.

Molina, J. L. (2001). El análisis de redes sociales: Una introducción. *Revista Hispana para el Análisis de Redes Sociales*, 1(1), 1–42.

Morillas, A. (1995). Aplicación de la teoría de grafos al estudio de los cambios en las

relaciones intersectoriales de la economía andaluza en la década de los 80. En *Contabilidad regional y tablas input-output de Andalucía 1990* (pp. 1–20). IEA.

Müller, T., Cotterell, R., Fraser, A., & Schütze, H. (2024). Joint lemmatization and morphological tagging with LEMMING. *arXiv*. https://doi.org/10.48550/arXiv.2405.18308

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1), 3–42. https://doi.org/10.1177/0049124117729703

Pérez, J. M., Luque, F., Zayat, D., Kondratzky, M., Moro, A., Serrati, P., Zajac, J., Miguel, P., Debandi, N., Gravano, A., & Cotik, V. (2022). Assessing the impact of contextual information in hate speech detection. *arXiv*. https://doi.org/10.48550/arXiv.2210.00465

Roberti, E. (2017). Perspectivas sociológicas en el abordaje de las trayectorias: Un análisis sobre los usos, significados y potencialidades de una aproximación controversial. *Sociologías*, 19(45), 276–312.

Rosati, G., Chazarreta, A., et al. (2013). Ciencias sociales computacionales: Un estado de la cuestión y una agenda de investigación. *Papeles de Trabajo*, número especial, 59–69.

Sadin, E. (2018). *La siliconización del mundo*. Caja Negra.

Salim, M. N., & Mustafa, B. (2022). A survey on word representation in natural language processing. *AIP Conference Proceedings*, 2394(1), 050006. https://doi.org/10.1063/5.0121147

Schubert, E., Spitz, A., Weiler, M., Geiß, J., & Gertz, M. (2017). Semantic word clouds with background corpus normalization and t-distributed stochastic neighbor embedding. *arXiv*. https://doi.org/10.48550/arXiv.1708.0356

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society*: Series A (Statistics in Society), 179(2), 481–511. https://doi.org/10.1111/rssa.12125

Van der Aalst, W. M. P. (2016). *Process mining: Data science in action* (2nd ed.). Springer. https://doi.org/10.1007/978-3-662-49851-4

Vilela, R. B., Ribeiro, A., & Batista, N. A. (2020). Word cloud as a tool for content analysis: An application to the challenges of the professional master's degree courses in health education. Millennium – *Journal of Education, Technologies and Health*, 2(11), 29–36. https://doi.org/10.29352/mill0211.03.00230

Vissoci, J. R. N., Rodrigues, C. G., Andrade, L., Santana, J. E., Zaveri, A., & Pietrobon, R. (2013). A framework for reproducible, interactive research: Application to health and social sciences. *arXiv*. https://arxiv.org/abs/1304.5688

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Xie, Y., & Avila, S. (2025). The social impact of generative LLM-based AI. *Chinese Journal of Sociology*, 11(1), 31–57. https://doi.org/10.1177/2057150X251315997

Yepes-Cayuela, L. (2018). *La operativización del concepto de trayectoria con TraMineR: Una introducción al análisis de secuencias y al Optimal Matching* (INCASI Working Papers No. 4).