EISSN 185 12578

Ciencias Sociales Computacionales
Un estado de la cuestión y una agenda de investigación

Germán Rosati,[1] Adriana Chazarreta,[2] Laia Domenech Burin,[3] Florencia Piñeyrúa[4] y Tomás Maguire[5]

Resumen

El presente trabajo tiene como objetivo trazar un panorama general de las Ciencias Sociales Computacionales (CSC). Se argumenta que se trata menos de una disciplina *ad-hoc* que de un giro metodológico en las Ciencias Sociales. Se destacan tres características fundamentales de las CSC y se ilustran con algunas investigaciones llevadas adelante en la Escuela IDAES-UNSAM.

Palabras clave: ciencias sociales computacionales; aprendizaje automático; metodología de la investigación

Abstract

This paper aims to provide an overview of Computational Social Sciences (CSS). It argues that CSS is less of an *ad-hoc* discipline and more of a methodological shift in the Social Sciences. Three fundamental characteristics of CSS are identified and exemplified through research conducted at EIDAES-UNSAM.

Keywords: computational social sciences; machine learning; research methodology

1 Factor~data - Escuela Interdisciplinaria de Estudios Sociales - Universidad Nacional de San Martín, Consejo Nacional de Investigaciones Científicas y Técnicas, german.rosati@gmail.com, ORCID: 0000-0002-9775-0435.

2 Factor~data - Escuela Interdisciplinaria de Estudios Sociales - Universidad Nacional de San Martín, Consejo Nacional de Investigaciones Científicas y Técnicas, adchazarreta@gmail.com, ORCID: 0000-0002-4737-9578.

3 Factor~data - Escuela Interdisciplinaria de Estudios Sociales - Universidad Nacional de San Martín, laiadomenechburin@gmail.com, ORCID: 0000-0003-4576-3143.

4 Factor~data - Escuela Interdisciplinaria de Estudios Sociales - Universidad Nacional de San Martín, Consejo Nacional de Investigaciones Científicas y Técnicas, pinieyrua@gmail.com, ORCID: 0000-0002-2043-8240.

5 Factor~data - Escuela Interdisciplinaria de Estudios Sociales- Universidad Nacional de San Martín, tomasmaguire@gmail.com, ORCID: 0000-0001-6511-4728.

EISSN 185 12578

Introducción. ¿Qué son las CSC?

Siempre resulta difícil determinar el comienzo de un campo disciplinar. En efecto, el uso y procesamiento de información estadística secundaria han sido parte integral del repertorio metodológico de las ciencias sociales desde sus orígenes. Basta pensar en las aproximaciones de los clásicos de la sociología, como *El Suicidio* (Durkheim, 1897), *El Capital* (Marx, 2008) o la encuesta a obreros rurales de Weber (Carabaña Morales, 1990).

Sin embargo, las primeras aplicaciones intensivas de la computación en ciencias sociales surgieron con el desarrollo de una serie de herramientas técnicas a partir de la Segunda Guerra Mundial. Allí podemos fechar un primer hito en el desarrollo de las Ciencias Sociales Computacionales (CSC). Según Cioffi-Revilla (2017), la aparición de la computación digital transformó profundamente las ciencias sociales, proporcionando un recurso esencial para el procesamiento de información a una escala poco imaginable años antes.[6] Como ejemplo, se destaca el desarrollo del análisis factorial (una técnica pionera en tareas de reducción de dimensionalidad) y del *General Inquirer*, un método de análisis de contenido textual que sirvió como precursor de las modernas técnicas de procesamiento de lenguaje natural. Ese impulso, que interactuó con el desarrollo de métodos cuantitativos nuevos (sobre todo basados en el uso de encuestas demográficas y de opinión), también ha estado en el centro del desarrollo de las disciplinas sociales.

A su vez, Cioffi-Revilla (2011, 2017) identifica algunas áreas de trabajo principales de las CSC: extracción automática de contenido, sistemas de información geográfica-social, análisis de redes sociales, "complejidad social" y modelos de simulación social. Cada una de estas áreas se compone de "conglomerados de conceptos, principios, teorías y métodos de investigación". Sin embargo, la vinculación entre la teoría y las CSC no resulta tan clara.

Por ejemplo, el análisis/extracción de contenido se ha aplicado, entre otros, en estudios de medios de comunicación y en estudios de conflictividad social. Si bien se ha hablado mucho del Análisis de Redes Sociales (ARS) como una teoría social, lo cierto es que la teoría matemática de grafos subyacente se ha aplicado a problemas tan disímiles como la difusión de opiniones, la construcción de modelos de lenguaje y la conformación del capital social. Esta diversidad es aún más evidente en los modelos de simulación social, los cuales han abordado problemas que van desde decisiones migratorias hasta la toma de decisiones en agricultura, pasando por la formulación de hipótesis sobre segregación residencial.

En consecuencia, cada una de las áreas mencionadas parece estar más relacionada con un enfoque metodológico que con postulados teóricos fuertes. Su aplicación es relativamente independiente de las hipótesis de trabajo y, por ende, de las teorías subyacentes. Con la posible excepción, y hasta cierto punto, de la esfera de complejidad social, el resto

6 "Within the span of a single generation the volume of knowledge across the social sciences increased by many orders of magnitude thanks to the advent of the modern digital computer" (Cioffi-Revilla, 2017, p. 19).

de las áreas de las CSC se caracterizan por estar más cerca de las herramientas técnico-metodológicas que de las grandes teorías sociales.

Este es, desde nuestra perspectiva, un rasgo fundamental de las CSC: su ámbito de influencia, y por ende su potencia, parece residir, en primera instancia, en el momento operativo, técnico y metodológico del proceso de investigación. Se trataría más de un "giro metodológico" que de un área completamente nueva. Lógicamente, este aspecto "instrumental" de las CSC no les quita potencial teórico; por el contrario, el avance técnico y metodológico en una disciplina suele habilitar nuevas preguntas, además de permitir responder las ya existentes.

En la última década, las CSC han tomado un nuevo impulso gracias al notable incremento en el volumen de información disponible para la investigación social. Las redes sociales, las tecnologías *mobile*, la "internet de las cosas" y todas aquellas fuentes que pueden englobarse en el impreciso término de *big data* son ejemplos de estas nuevas fuentes. Gracias al desarrollo acelerado de la capacidad de cómputo disponible, se ha hecho posible recopilar, almacenar y analizar grandes cantidades de datos sociales de diversas fuentes, como las redes sociales, los registros gubernamentales y los archivos históricos. A su vez, han proliferado las formas de comunicación e interacción sustentadas en Internet, por lo que buena parte de las interacciones sociales, económicas, políticas y culturales de millones de usuarios pasaron a estar digitalizadas. De esta manera, se comenzaron a generar repositorios masivos de datos de interacciones entre personas en tiempo real y con un nivel de desagregación individual. Así, aparecieron nuevas clases de datos que se desviaban del típico formato estructurado diseñado específicamente para las ciencias sociales tradicionales, como las encuestas. Esta apertura de nuevas fuentes ha ido acompañada de la creación de herramientas y técnicas novedosas de análisis de datos. Quizás, el aprendizaje automático en sus diversas formas sea una de las herramientas más novedosas. El alcance de las CSC en esta etapa parece desbordar aquellas áreas iniciales.

Para ilustrar esta afirmación, tomaremos como ejemplo algunas investigaciones en curso y algunos resultados de investigación producidos en el marco de factor~data, un laboratorio de experimentación en CSC perteneciente a la Escuela Interdisciplinaria en Altos Estudios Sociales de la Universidad Nacional de San Martín. Parte fundamental de la tarea de factor~data se centra en la identificación de problemas de investigación (nuevos y clásicos) abordables mediante estas técnicas computacionales y fuentes de datos nuevas y clásicas. De esta forma, se llevan adelante investigaciones originales que permiten la producción de conocimiento sobre las estructuras sociales y sus dinámicas.

Una primera línea de investigación es el eje de territorio, espacio y desigualdad. A partir del uso de datos espaciales provenientes de diferentes fuentes (datos abiertos estatales, aplicaciones de ruteo, imágenes satelitales y sus derivados) y su análisis mediante herramientas vinculadas al aprendizaje automático, se propone indagar en la expresión de desigualdades en el territorio y en las marcas antrópicas sobre los sistemas de bosques nativos.

EISSN 185 12578

La segunda línea de investigación explora la viabilidad de diferentes técnicas de *natural language processing* (NLP) –modelado de tópicos, *word embeddings*, etc.– para su aplicación en problemas de las ciencias sociales. Se busca indagar en el enriquecimiento de los análisis de discursos tradicionales a partir de estas nuevas técnicas.

Tres fuentes y tres partes integrantes de las CSC

¿Qué características parecen adquirir las CSC en esta nueva etapa? En principio, mantienen la combinación de conocimientos de las Ciencias Sociales y las Ciencias Computacionales o la más moderna "Ciencia de Datos" (Engel, Quan-Haase, Xun Liu y Lyberg, 2021). Y si bien desde algunas posiciones se habla de las CSC no como una disciplina, sino como una "red fluida" o un "movimiento" dentro de las Ciencias Sociales (Geise y Waldherr, 2021), existe consenso en torno a tres componentes esenciales que las definen: 1) el planteo de preguntas e hipótesis que surgen de las ciencias sociales; 2) la integración entre fuentes de datos heterogéneas y con diferentes grados de estructuración y 3) el uso de metodologías basadas en recursos computacionales que permiten la automatización de gran parte del proceso de investigación.

En relación al primer elemento, es importante marcar que las CSC son, en primer lugar y ante todo, Ciencias Sociales. Es decir, los problemas que abordan son los mismos que abordaban los clásicos (como la estructura social, la conflictividad o la desigualdad) y otros más modernos (por ejemplo, las representaciones y discursos sociales o cuestiones referentes a las subjetividades). Así, es difícil plantear la existencia de una "ruptura" teórica entre CSC y el resto de las ciencias sociales.

Por poner un ejemplo posible entre muchos otros, De Francisci Morales, Monti y Starnini (2021) abordan la temática de la polarización política a partir de redes sociales, tomando como insumo interacciones entre grupos de partidarios de Trump y Clinton en Reddit durante las elecciones presidenciales de Estados Unidos de 2016. Los autores reconstruyeron la red de interacción entre estos usuarios en la principal comunidad de discusión política, r/politics. Encuentran que, a pesar de la polarización política, estos grupos tienden a interactuar más entre sí que al interior de ellos, es decir, la red exhibe heterofilia en lugar de homofilia. Este hallazgo surge de la comparación con un modelo nulo de interacciones sociales aleatorias, implementando tanto una red que preserva la actividad de los usuarios, como un modelo de regresión logística para la predicción de enlaces que tiene en cuenta posibles factores de confusión. En general, los hallazgos muestran que Reddit ha sido una herramienta para la discusión política entre puntos de vista opuestos durante las elecciones de 2016. Este comportamiento contrasta con las llamadas "cámaras de eco" observadas en otros debates polarizados sobre diferentes temas en varias plataformas de redes sociales. Aquí, la polarización se asocia con mayores interacciones entre grupos que tienen opiniones opuestas. Sin embargo, esta relación entre polarización y heterofilia podría no ir más allá del ámbito digital.

EISSN 185 12578

EISSN 185 12578

La segunda característica de las CSC se vincula con la utilización de fuentes diversas y heterogéneas en relación al grado de estructuración. Pensemos en dos casos extremos: la Encuesta Permanente de Hogares y un corpus textual extraído de comentarios en un foro de lectores de un diario de circulación nacional, obtenido mediante un procedimiento de *web scraping*.[7] El *scraping*, literalmente "raspado" o "rascado", consiste en la descarga y formateo de la información disponible en sitios web, información que generalmente no se encuentra en condiciones de ser trabajada de forma cuantitativa (Mitchell, 2015).

Vale aclarar que con la noción de "grado de estructuración", nos referimos a diferentes niveles y etapas en el proceso de producción y análisis de un dato.

El primero de los casos probablemente sea uno de los más habituales en las ciencias sociales. La primera manifestación visual del grado de estructuración se observa en la tabla misma: filas y columnas; unidades de análisis, variables y valores; en fin, la estructura tripartita (Galtung, 1966) o cuatripartita (Samaja, 2004) del dato. Pero, a su vez, todo el proceso de producción de esa tabla de datos se caracteriza por un alto grado de control: el diseño del instrumento de relevamiento (cuestionario) pasó por una serie de discusiones y decisiones (muchas a escala internacional) y el cuestionario está altamente organizado en preguntas cerradas y (unas pocas) abiertas. Al menos en teoría, los encuestadores deben aplicar de la misma forma el instrumento a todas las personas encuestadas (deben leer las preguntas tal cual están redactadas). Al mismo tiempo, la EPH tiene un proceso de selección de las unidades que forman parte de la encuesta sumamente controlado: el azar (la aleatoriedad de la muestra) garantiza la posibilidad de "expandir" los resultados de la muestra a la población (Sosa Escudero, 2019).

El segundo ejemplo presenta atributos prácticamente opuestos. La estructura tripartita del dato no se hace evidente. Son solo trozos de texto libre de los que solamente podemos identificar las unidades de análisis, pero no se encuentra estructurado en formato de variables. A su vez, el proceso de producción es "espontáneo": los usuarios comentan de forma libre (y no como respuesta a una pregunta; en todo caso, el único estímulo es la lectura de la noticia en cuestión o, quizás, como respuesta a otro usuario) en el foro. La producción de esos textos no sigue el formato de un cuestionario. La recolección de esa información también tiene un bajo grado de normalización. Es simplemente un programa que descarga esos textos y los formatea como un corpus. A su vez, esa espontaneidad también afecta la selección de las unidades o el muestreo. No hay un proceso sistemático y aleatorizado de selección. Solamente, se registran los comentarios de usuarios que participan del foro y no hay una definición clara del universo: ¿son los lectores del diario?, ¿los lectores digitales?, ¿existen diferencias y sesgos entre las personas que leen y las que comentan?

7 El corpus fue analizado en Rosati, Chazarreta, Domenech Burin y Maguire (2022).

EISSN 185 12578

Otro rasgo fundamental de las CSC es la combinación e integración de diferentes fuentes y diferentes tipos de datos. Es habitual (y necesaria) la combinación de datos cuyos orígenes y procesos de producción son divergentes. Domenech Burin (2023) busca entrenar un modelo que permita detectar zonas de deforestación. Para esto utiliza tres fuentes diferentes. Como variables independientes toma información derivada de imágenes satelitales (índice de vegetación, NDVI) y datos sobre trayectorias de uso del suelo basadas en información de la European Space Agency (Rosati, 2023), ambas fuentes, en formato ráster. Como variable dependiente, usa polígonos (en formato vectorial) que marcan las zonas deforestadas.[8]

Esta diferencia entre datos con diferente grado de estructuración es paralela a otra: la diferencia entre información de carácter primario (creada específicamente para abordar un problema de investigación específico) y secundario (producida por otros organismos, empresas, instituciones y/o personas y cuyo objetivo inicial es diferente al de la investigación en curso y que son reutilizados y transformados en función de la misma).[9] En este sentido, la tendencia parece ser hacia la integración entre ambos tipos de datos. Así, el trabajo de Rosati, Chazarreta, Domenech y Maguire (2022) analiza comentarios de lectores mencionados más arriba y es un ejemplo de este tipo de combinación de datos primarios (la conformación de un corpus de comentarios de lectores de diarios nacionales) y secundarios (el uso de una encuesta de consumo de medios para avanzar en una caracterización de dichos lectores).

La última característica mencionada de las CSC se vincula al uso intensivo de técnicas computacionales que buscan avanzar en la automatización de diferentes etapas del proceso de investigación. Una primera diferencia puede observarse en la etapa de recolección de información y construcción del dato. A diferencia de las técnicas de recolección de datos más tradicionales (encuestas, censos, entrevistas, relevamientos etnográficos, etc.) que tienden a consumir una gran cantidad de recursos y de trabajo humano,[10] las CSC abren la posibilidad de escalar y automatizar parte del trabajo de recolección del material mediante el uso de web scrapers, consulta a API, etc. Así, Piñeyrúa (2021) y Maguire (2021) exploran (en dos contextos electorales diferentes) las potencialidades y limitaciones del uso de técnicas de procesamiento de lenguaje natural (NLP) y web scraping para el análisis de medios

8 Desarrollado por el Laboratorio de Análisis Regional y Teledetección (LART) de la Facultad de Agronomía de la Universidad de Buenos Aires (FAUBA), el Instituto Nacional de Tecnología Agropecuaria (INTA) y la Red Agroforestal Chaco Argentina (Redaf) (http://www.monitoreodesmonte.com.ar/).

9 Salganik (2018, p. 7) hace un paralelismo entre datos primarios (*custommades*) y secundarios (*readymades*) y las obras de Miguel Ángel y Marcel Duchamp.

10 Una encuesta (ya sea automática o no), las entrevistas en profundidad, las campañas de campo etnográfico requieren una gran cantidad de tiempo para su relevamiento y, en muchos casos, requieren movilizar una logística bastante importante en términos de recursos, viajes, tareas de post-edición del material, etc.

EISSN 185 12578

desde la propuesta teórica de la Agenda Setting. Los autores intentan mostrar cómo este conjunto de técnicas computacionales permiten ampliar la escala del trabajo de forma eficiente y pueden ser útiles para morigerar algunas dificultades metodológicas presentes en el análisis de la agenda mediática digital, tales como el pequeño tamaño de las muestras o la replicabilidad presentes en la codificación y análisis de piezas periodísticas debido a la gran cantidad de recursos necesarios para desarrollar dichas tareas (Orozco Gómez y González, 2012). También, el trabajo de Rosati (2022), Rosati y Domenech Burin (2022) y Rosati, Chazarreta, Domenech y Maguire (2021) hacen uso de técnicas de web scraping para constituir los respectivos corpus de letras de tango, rock y comentarios de lectores.

También la etapa de análisis se ve afectada por el "giro computacional" en las CSC. La aplicación de diferentes técnicas de aprendizaje automático probablemente sea uno de los emergentes (desde ya, no el único) más evidentes. Muchas veces, el producto a analizar es una tabla de gran cantidad de columnas (alta dimensionalidad) o una gran cantidad de filas. Allí es sumamente necesario, para la etapa analítica, algún proceso que sintetice dicha información, la cual en muchos casos puede ser altamente descriptiva y de la que no se dispone de modelos conceptuales o mecanismos causales claros. En este punto, suelen entrar en juego las técnicas de aprendizaje no supervisado. Estas no disponen de una variable dependiente que se intenta modelar. Las tareas más comunes de este tipo de herramientas suelen ser dos: clusterización y reducción de dimensionalidad[11]. Chazarreta (2022a) construyó un índice que permite aproximarse a la medición de los diferentes niveles de separación entre capital y propiedad en las empresas industriales manufactureras. Para este índice, utilizó los datos de la Encuesta Nacional de Dinámica del Empleo y la Innovación (ENDEI) I (2010-2012) y, para seleccionar las variables relevantes, propuso un análisis de correspondencias múltiples (ACM), una técnica de reducción de dimensionalidad para variables categóricas (como es el caso de las disponibles en esta encuesta). Mediante el uso del ACM, buscó poder representar de forma "sintética" en una única variable diferentes características asociadas a la problemática. Los resultados parecen mostrar la existencia de dos dimensiones de la separación, aunque a partir del análisis de la varianza se decidió considerar solo la primera: esta hace referencia al tipo de estructuras de organización o gestión de la empresa al incluir mediciones sobre la toma de decisión descentralizada (delegación de autoridad y de responsabilidades) y sobre la formalización de roles, responsabilidades y métodos de evaluación de desempeño. A su vez, Rosati (2023) utiliza técnicas de clustering y distancias de edición para construir una tipología de trayectorias de uso del suelo en Argentina.

Por el contrario, las técnicas de aprendizaje supervisado tienen una variable dependiente a predecir. El trabajo de Domenech Burin (2023) mencionado más arriba utiliza

11 Una explicación al respecto puede encontrarse en James, Witten, Hastie y Tibshirani (2017).

EISSN 185 12578

y compara diferentes técnicas de aprendizaje supervisado (regresión logística, *xgboost* y *random forest*) para construir un clasificador tomando como input derivados de imágenes satelitales, que permita identificar y predecir zonas en las que se produce deforestación. Chazarreta (2022b) para analizar cuáles son los determinantes socio-productivos que afectan el grado de la separación de la propiedad y la dirección del capital en empresas industriales manufactureras de Argentina realiza una regresión logística, que permite estimar la probabilidad para cada empresa de que corresponda a grado alto de separación. A su vez, entrenó otro modelo, basado en *random forest*, resultando que tanto la performance predictiva como el orden de importancia de las variables predictoras son similares a los de la regresión logística.

Un párrafo aparte merecen las técnicas de procesamiento de lenguaje natural (NLP, por sus siglas en inglés). Estas habilitan la aplicación de métodos cuantitativos de análisis para una amplia diversidad de tareas (clasificación de textos, detección de temas y tópicos, detección de estructuras semánticas, etc.). Asimismo, permiten escalar el trabajo con datos textuales y pasar de una lectura cercana a una lectura distante para lidiar con la "enormidad de lo no leído" (Moretti, 2015).[12] Efectivamente, técnicas como el modelado de tópicos o los *word/sentence embeddings* permiten analizar de forma semiautomática corpus de gran escala, haciendo menos necesaria la lectura de cada uno de los textos (tarea que se hace imposible cuando el corpus adquiere cierto tamaño). Varios de los trabajos mencionados anteriormente hacen uso de este tipo de técnicas y herramientas.

Comentarios finales

En el presente trabajo, hemos intentado hacer un breve resumen del estado de situación de las CSC. Describimos cómo estas emergieron y se potenciaron a partir del incremento en la disponibilidad de fuentes de datos y del poder de cómputo. Luego, mencionamos tres rasgos de las CSC e ilustramos algunos rasgos de las mismas a partir de varias investigaciones llevadas adelante en el laboratorio factor~data de la EIDAES-UNSAM.

En ese sentido, argumentamos que, lejos de constituir una disciplina *ad-hoc*, pueden ser pensadas más bien como una especie de "giro metodológico", incorporando nuevas herramientas para resolver problemas y preguntas tradicionales y habilitando nuevos interrogantes. A su vez, la aparición de nuevas fuentes de información (vinculadas sobre todo a la llamada "revolución digital"), impulsaron el trabajo con datos de diferentes grados de estructuración y sumamente heterogéneos, y permitieron hacer avances en la automatización y estandarización de diferentes etapas del proceso de investigación.

12 "Mucha gente ha leído más y mejor que yo, por supuesto, pero eso tampoco basta: aquí hablamos de cientos de lenguas y literaturas. Todo indica que leer 'más' no es la solución. En especial, porque hemos comenzado a descubrir (...) la enormidad de lo no leído..." (Moretti, 2015, p. 59).

EISSN 185 12578

En el trabajo, hemos enfatizado considerablemente las ventajas y características de las CSC. Sin embargo, es necesario mencionar, para cerrar, algunas trampas de este nuevo abordaje metodológico.

Por un lado, los datos no "hablan" por sí solos a menos que se les hagan las preguntas adecuadas, preguntas que suponen un bagaje conceptual específico. A su vez, el trabajo con datos (ya sean "grandes" o "pequeños"; "nuevos" o "viejos") presenta los mismos riesgos y problemas metodológicos y está sujeto a las mismas necesidades de validación y consistencia. En este punto, las características que Salganik (2018) define como características del llamado *big data* son de utilidad para pensar estos problemas. De esos diez rasgos, algunos nos advierten los potenciales problemas que la investigación social en la "era digital" puede presentar: en general, trabajamos con datos incompletos que no tienen toda la información que querríamos (hay variables o atributos que no están relevados), algunos de los cuales son poco accesibles en tanto se trata de información producida por grandes empresas (Twitter, probablemente sea el caso más obvio) quienes limitan fuertemente el acceso a los mismos. A su vez, muchos de estos datos no son representativos: no han sido generados mediante procedimientos de muestreo aleatorios, lo cual los hace poco aptos para efectuar generalizaciones por fuera de los mismos pero muy útiles para hacer análisis descriptivos. En general, se trata de datos "sucios", que contienen información irrelevante para nuestros propósitos e inconsistencias entre los mismos datos. Todas estas características (que no son necesariamente exclusivas de los llamados *big data* y se encuentran también en otras fuentes como encuestas o censos) nos plantean la necesidad de considerar tres aspectos fundamentales en cualquier investigación científica: 1) relevancia de una coherencia conceptual y teórica; 2) importancia de un proceso consistente de operacionalización de tales conceptos y 3) ineludibilidad de una crítica rigurosa de las fuentes de datos a utilizar.

Quisiéramos dejar planteada una última cuestión. Un efecto del incremento en la cantidad de fuentes de datos y en la complejidad de las técnicas de análisis (vinculados especialmente al aprendizaje automático) ha sido una correlativa pérdida de "comunicabilidad" de los resultados. En efecto, a diferencia de un análisis de regresión lineal en el que los coeficientes beta son claramente interpretables o un árbol de decisión, un *random forest* o una red neuronal carecen de esta interpretabilidad y obligan a pensarlos (al menos inicialmente como una caja negra). Esta característica hace tanto a la posibilidad de entender de forma correcta los hallazgos y la información que arroja sobre el fenómeno analizado la técnica, como también la comprensión de aquellas situaciones en que los modelos pueden fallar. Aunque existen muchas herramientas y una investigación muy activa en el campo llamado *machine learning interpretable* (Molnar, 2023), el problema queda planteado como un área a profundizar.

Referencias

Carabaña Morales, Julio (1990). Un texto poco clásico de un autor clásico: la *Ausblick* de

Weber sobre la situación de los obreros agrícolas al Este del Elba. *Revista Española de Investigaciones Sociológicas*, 49, 223-231.

Chazarreta, Adriana (2022a). Aproximación empírica a la separación de la propiedad y el control del capital. Construcción de un índice de las estructuras organizativas de las empresas industriales. 2012, *Revista SaberEs*, *14*(2), 195-213.

Chazarreta, Adriana (2022b). Estimación de los determinantes en la separación de la propiedad y la dirección del capital de las empresas industriales manufactureras. Argentina, 2016. *Anuario CEEED*, *14*(17), 113-142.

Cioffi-Revilla, Claudio (2010). Computational Social Science. *WIREs Computational Statistics*, 3, 259-271.

Cioffi-Revilla, Claudio (2017) *Introduction to Computational Social Science. Principles and Applications.* Suiza: Springer Nature.

De Francisci Morales, Gianmarco; Corrado Monti, y Michele Starnini (2021). No echo in the chambers of political interactions on Reddit. *Scientific Reports*, 11(2818).

Domenech Burin, Laia (2023). Mapeo de desmontes en bosques nativos de Argentina. Propuesta de mejoras en el Sistema de Alerta Temprana de Deforestación. *FUNDAR*.

Durkheim, Émile (1897). *Le suicide. Étude de sociologie*. París: Félix Alcan.

Engel, Uwe; Anabel Quan-Haase; Sunny Xun Liu y Lars Lyberg (2021). Introduction to the Handbook of Computational Social Science. En U. Engel, A. Quan-Haase, S. Xun Liu, y L. Lyberg (comps.) *Handbook of Computational Social Science* (pp. 1-14). Nueva York: Routledge.

Galtung, Johann (1966). *Teoría y método de la investigación social*. Buenos Aires: EUDEBA.

Geise, Stephanie y Annie Waldherr (2021). Computational communication science: lessons from working group sessions with experts of an emerging research field. e En U. Engel, A. Quan-Haase, S. Xun Liu, y L. Lyberg (comps.) *Handbook of Computational Social Science* (pp. 66-82). Nueva York: Routledge.

James, Gareth; Daniela Witten, Trevor Hastie, y Robert Tibshirani (2017). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Maguire, Tomás (2021). *Aprendizaje automático y modelización de tópicos: un estudio de caso sobre la agenda mediática en contexto de las elecciones. Argentina, 2015*. Tesis de Licenciatura en Sociología, Escuela Interdisciplinaria de Estudios Sociales, UNSAM.

Mitchell, Ryan (2015). *Web scraping with python: Collecting data from the modern web*. California: O'Reilly.

Marx, Karl (2008). *El Capital*. México: Siglo XXI

Molnar, Christoph (2023). *Interpretable Machine Learning*. Munich: LeanPub.

Moretti, Franco (2015). *Lectura distante.* Buenos Aires: Fondo de Cultura Económica.

Piñeyrúa, Florencia Nathalia (2021). Aportes desde el procesamiento de lenguaje natural para incrementar la escalabilidad en los estudios sobre tópicos de noticias digitales securitarias. *Revista Comunicación, Política y Seguridad*, 3, 111-142.

Orozco Gómez, Guillermo y Rodrigo González (2012). *Una coartada metodológica. Abordajes*

cualitativos en la investigación en comunicación, medios y audiencias. Kindle Edition.

Rosati, Germán (2022). Procesamiento de Lenguaje Natural aplicado a las ciencias sociales: Detección de tópicos en letras de tango. *Revista Latinoamericana de Metodología de la Investigación Social*, *12*(23), 38-60.

Rosati, Germán (2023). Analizando trayectorias de uso del suelo. Una propuesta de clusterización. *Geograficando*, *19*(1). UNLP.

Rosati, Germán y Laia Domenech Burin (2022, 1-5 noviembre). Los temas del rock nacional. Una aproximación mediante técnicas de minería de texto. Ponencia presentada en *XIV Jornadas de Sociología*, Facultad de Ciencias Sociales, UBA, Buenos Aires.

Rosati, Germán; Adriana Chazarreta; Laia Domenech y Tomás Maguire (2021). Una aproximación a los temas acerca de la COVID-19. Aplicación de técnicas de procesamiento de lenguaje natural sobre comentarios de lectores de noticias digitales. *Papeles de Trabajo*, 15(28), 64-91.

Salganik, Matthew (2018). *Bit by bit. Social research in the digital age*. Oxford: Princeton University Press.

Samaja, Juan (2004). *Epistemología y metodología*. Buenos Aires: EUDEBA.

Sosa Escudero, Walter (2019). *Big Data. Breve manual para conocer la ciencia de datos que ya invadió nuestros días*. Buenos Aires: Siglo XXI.

EISSN 185 12578