Capítulo 11. Del código al discurso: ingeniería de datos para el análisis político en redes sociodigitales

Omar Mendoza González[1]
Jesús Hernández Cabrera[2]
Miguel Ángel Sánchez Hernández[3]

DOI: https://doi.org/10.52501/cc.347.11

Resumen

En este capítulo se aborda el papel transformador de la Ingeniería de Datos y la Inteligencia Artificial en el análisis político-electoral contemporáneo, a partir de un estudio de caso sobre la contienda presidencial de México 2024. Desde un enfoque interdisciplinario, se documenta cómo el trabajo conjunto entre ingeniería en computación y comunicación política permitió desarrollar una infraestructura técnica capaz de recolectar, procesar y analizar más de 15 000 publicaciones en redes sociodigitales —particularmente Facebook, Instagram y X (antes Twitter)— generadas por las candidatas presidenciales.

El procesamiento de estos datos se realizó sobre un ecosistema distribuido basado en tecnologías de código abierto como Hadoop y Apache Spark, desplegado en un clúster de alto rendimiento que permitió ejecutar tareas de limpieza, normalización, agrupamiento semántico y análisis de

[1] Profesor de carrera Titular A en la Licenciatura de Ingeniería en Computación de la Facultad de Estudios Superiores Aragón, UNAM. Doctor en Educación por la Universidad Marista de la Ciudad de México. ORCID: 0000-0002-3492-4549 ; correo electrónico: omarmendoza564@aragon.unam.mx

[2] Profesor de carrera Titular A en la Licenciatura de Ingeniería en Computación de la Facultad de Estudios Superiores Aragón, UNAM. Doctorante en Sistemas Computacionales. ORCID: 0000-0002-7850-858X ; correo electrónico: jesushernandezls7@aragon.unam.mx

[3] Profesor de asignatura A en la Licenciatura de Ingeniería en Computación de la Facultad de Estudios Superiores Aragón, UNAM. Maestro en Ingeniería en Sistemas Computacionales por el Tecnológico de Estudios Superiores de Ecatepec. ORCID: 0000-0002-6265-1760 ; correo electrónico: miguelsanchezt32@aragon.unam.mx

frecuencia a gran escala. Esta capacidad técnica posibilitó la identificación automatizada de patrones discursivos, estrategias de repetición, jornadas de alta actividad, y el uso intensivo de contenidos patrocinados, proporcionando insumos cuantitativos robustos para el análisis cualitativo desde perspectivas retóricas (ethos, pathos, logos).

La sección también examina los riesgos asociados al uso intensivo de estas tecnologías en procesos democráticos, incluyendo la opacidad algorítmica, el sesgo en la recolección de datos, y la microsegmentación de audiencias. Frente a ello, se propone una práctica investigativa guiada por principios de apertura, reproducibilidad y revisión interdisciplinaria. En conjunto, esta sección demuestra que el uso ético e informado de herramientas de procesamiento masivo de datos no solo amplía el alcance empírico del análisis político, sino que fortalece su rigor metodológico y su capacidad para interpretar fenómenos complejos en entornos digitales altamente dinámicos.

Introducción

El estudio de los procesos políticos contemporáneos requiere integrar las herramientas de la Ingeniería de datos y la inteligencia artificial (IA). La colaboración entre disciplinas técnicas y sociales se ha vuelto indispensable. En un contexto donde millones de interacciones se producen a diario en redes sociodigitales, estas tecnologías se han convertido en una herramienta útil para recolectar, procesar y analizar información a una escala masiva. Durante la elección presidencial de México en 2024, estas tecnologías permitieron analizar patrones discursivos y estrategias de campaña con una precisión inalcanzable para los métodos tradicionales de la ciencia política.

Esta sección surge de un proyecto interdisciplinario entre las carreras de Ingeniería en Computación y Comunicación y Periodismo de la FES Aragón (UNAM), titulado *La construcción de la comunicación política en las redes sociodigitales durante las campañas electorales en México desde la configuración del ethos, pathos y logos*. A lo largo de este esfuerzo, se desarrolló un ecosistema técnico basado en Hadoop y Apache Spark para realizar procesos de limpieza, transformación, filtrado y visualización de datos

masivos extraídos de redes sociales, facilitando así su análisis cualitativo y cuantitativo

Más allá de los procedimientos técnicos, esta sección también plantea una reflexión crítica sobre los desafíos que implica el uso de herramientas tecnológicas en el análisis electoral. Como han advertido Boyd y Crawford (2012), el análisis de grandes volúmenes de datos no está exento de sesgos, limitaciones metodológicas ni dilemas éticos. O'Neil (2016) ha señalado cómo el uso de algoritmos en contextos sensibles como la política puede reforzar desigualdades y reducir la transparencia democrática. Por ello, este capítulo busca no solo mostrar el potencial de la Ingeniería de datos y la IA en el análisis electoral, sino también discutir con responsabilidad sus implicaciones.

En suma, esta sección ofrece una mirada integral sobre cómo las herramientas computacionales pueden enriquecer —y a la vez desafiar— las formas tradicionales de investigar fenómenos políticos. Lo hace desde una práctica concreta y colaborativa que combina el rigor técnico con la sensibilidad crítica necesaria para comprender el papel de los algoritmos, los datos y las plataformas en la configuración del espacio público digital durante procesos electorales.

Marco teórico

Ingeniería de datos en contextos sociopolíticos

La Ingeniería de datos ha emergido como una disciplina que ayuda a abordar fenómenos complejos que antes quedaban fuera del alcance de los métodos tradicionales de las ciencias sociales. Su objetivo principal es procesar grandes volúmenes de datos —frecuentemente desordenados, inconsistentes o incompletos— para transformarlos en información útil, estructurada y explotable. En contextos sociopolíticos, esto implica traducir el ruido de las redes sociodigitales en patrones comprensibles de interacción, discurso e influencia.

El entorno digital ha generado un nuevo tipo de campo de observación para la política. Las campañas electorales, antes circunscritas a me-

dios tradicionales y a eventos presenciales, ahora se desarrollan también en tiempo real en plataformas digitales, donde cada publicación, gasto publicitario o mensaje viralizable puede registrarse, almacenarse y analizarse. Esta disponibilidad masiva de datos representa una oportunidad sin precedentes para comprender la comunicación política, pero también plantea desafíos técnicos importantes que requieren metodologías robustas para garantizar la calidad del análisis.

En este sentido, la Ingeniería de datos actúa como un puente entre los datos brutos y los modelos interpretativos propios de la comunicación política. En el proyecto desarrollado para el análisis de la contienda presidencial 2024 en México, el equipo de Ingeniería en Computación se enfrentó a un conjunto de datos extraído de redes como X, Facebook e Instagram, complementado por registros manuales en hojas de cálculo elaboradas por el equipo de Comunicación

La complejidad del conjunto de datos —diverso en fuentes, formatos y niveles de estructura— exigió una estrategia de procesamiento que incluyó limpieza de caracteres, normalización de valores numéricos, integración de formatos y transformación de campos anidados. Estos últimos, por ejemplo, adoptaban estructuras conocidas como JSON, un formato ampliamente utilizado para organizar datos en forma de listas o conjuntos etiquetados (similar a una tabla con subtablas internas), que permite representar información jerárquica de manera compacta. Traducir estas estructuras a un formato más plano fue esencial para permitir su análisis posterior.

Estos procesos no son triviales: el 60% al 66% de los datos recopilados correspondían a ruido o información redundante, y solo tras aplicar filtros rigurosos se pudo conservar entre el 40% y 44% del contenido relevante, sin sacrificar su valor informativo.

Esta etapa permitió a los equipos interdisciplinarios extraer indicadores como la frecuencia de publicaciones, los días de mayor actividad y los mensajes más repetidos durante la campaña, dando forma a un análisis basado en evidencia empírica.

El uso de ecosistemas distribuidos como Hadoop y motores de procesamiento como Apache Spark —ya establecidos como estándares en el manejo de datos masivos (Dean & Ghemawat, 2004; Zaharia & Wenchen,

2020)— permitió escalar el procesamiento sin sacrificar velocidad o integridad de los datos. Este enfoque permitió transformar la enorme cantidad de publicaciones y registros asociados en visualizaciones comprensibles, resúmenes cuantitativos y matrices de análisis cualitativo.

Al integrar estas herramientas en una investigación sociopolítica, la Ingeniería de datos no solo aporta eficiencia técnica, sino también rigurosidad metodológica. En lugar de limitarse a tareas instrumentales, se convierte en un agente activo en la construcción del conocimiento, al definir cómo se recopilan, procesan y priorizan los datos que sustentan las interpretaciones políticas. Tal como han advertido Tufekci (2014) y Venturini & Latour (2010), esta etapa de mediación técnica es crucial: los datos no son neutrales y la forma en que son organizados afecta directamente las conclusiones que se extraen de ellos.

En suma, la Ingeniería de datos en contextos sociopolíticos como el electoral requiere sensibilidad técnica y política. No basta con automatizar procesos: se trata de diseñar infraestructuras que respeten los principios de transparencia, trazabilidad y coherencia con los marcos interpretativos que guían el análisis social. Esta integración disciplinaria, como muestra el caso de estudio, fortalece tanto la comprensión del fenómeno como la confianza en los hallazgos obtenidos.

Inteligencia artificial en el análisis electoral

La inteligencia artificial (IA) se ha convertido en un actor central en el ecosistema de las campañas políticas digitales. Desde algoritmos que personalizan anuncios hasta sistemas que analizan sentimientos o identifican patrones de comportamiento en línea, la IA redefine la manera en que los partidos y candidatos interactúan con los votantes. En el contexto electoral, su uso plantea un doble escenario: por un lado, posibilita análisis complejos e inmediatos que enriquecen la comprensión del discurso político; por otro, introduce riesgos asociados a la opacidad algorítmica, la microsegmentación y la manipulación de la opinión pública.

Durante el proyecto interdisciplinario sobre la contienda presidencial 2024 en México, la IA fue utilizada principalmente como herramienta de

apoyo en el análisis computacional de grandes volúmenes de publicaciones en redes sociodigitales. A partir de modelos de procesamiento de lenguaje natural (PLN), se facilitó la identificación de contenidos redundantes, la extracción de patrones repetitivos y la detección de días de alta actividad. Si bien no se aplicaron modelos predictivos avanzados, la implementación de algoritmos diseñados para filtrar y clasificar publicaciones muestra cómo incluso formas "básicas" de IA pueden acelerar y enriquecer el análisis sociopolítico.

La lógica que impulsa el uso de IA en campañas responde a la economía de la atención: en plataformas digitales saturadas de información, los algoritmos seleccionan, jerarquizan y recomiendan contenidos que maximizan el tiempo de exposición del usuario. Esto tiene implicaciones directas en la comunicación política, donde los mensajes que mejor se adaptan a la lógica algorítmica —por ejemplo, aquellos altamente emocionales o polarizantes— tienden a tener mayor alcance. Esta dinámica configura lo que Tufekci (2014) ha descrito como un entorno de visibilidad condicionada por diseño tecnológico, más que por mérito argumentativo o relevancia social.

En el caso mexicano, observamos cómo algunas candidatas utilizaron estrategias de repetición intensiva de mensajes en redes sociales —como el saludo matutino "Buenos días, México" publicado hasta 74 veces en un solo día—, probablemente con el objetivo de alimentar la visibilidad algorítmica en plataformas como Facebook e Instagram.

La identificación de este patrón fue posible gracias al desarrollo de algoritmos que automatizan el conteo, la detección de duplicados y la priorización de publicaciones según su impacto.

Sin embargo, el uso de IA en este tipo de análisis no está exento de cuestionamientos. Como señala O'Neil (2016), los sistemas automatizados —aunque aparentemente neutrales— pueden amplificar desigualdades y reproducir sesgos si carecen de auditorías éticas. En el ámbito electoral, esto se agrava cuando la IA es utilizada para prácticas como la microsegmentación emocional, donde los anuncios se personalizan según las vulnerabilidades individuales de los votantes (Bodó, Helberger & de Vreese, 2017). Estos mecanismos, invisibles para el electorado, erosionan los principios de deliberación pública y transparencia democrática.

Por ello, el papel de la IA en el análisis electoral no debe limitarse a su potencial técnico. Es indispensable integrar marcos críticos que permitan evaluar cómo estas tecnologías configuran las condiciones mismas del debate político. La IA no solo ayuda a estudiar las campañas: también interviene en su desarrollo, modula su alcance y, en muchos casos, define sus efectos.

En el proyecto aquí analizado, la IA fue concebida como una herramienta al servicio de la investigación, pero su implementación estuvo guiada por un criterio metodológico claro: aumentar la calidad y rapidez del análisis sin reemplazar el juicio interpretativo del equipo de comunicación. Esta combinación —automatización en la preparación de datos y reflexión humana en la interpretación— representa un camino prometedor para futuras investigaciones en Tecnopolítica.

Ecosistemas de big data y procesamiento distribuido

El tratamiento eficaz de datos masivos requiere infraestructuras tecnológicas diseñadas para escalar, distribuir y paralelizar el procesamiento. En el ámbito político, donde el volumen de información generada en redes sociodigitales puede alcanzar millones de registros en pocos días, herramientas tradicionales resultan insuficientes. En este contexto, los ecosistemas de big data —en particular Hadoop y Apache Spark— han demostrado ser adecuados para extraer valor de grandes volúmenes de datos de manera eficiente, flexible y reproducible.

Hadoop, desarrollado inicialmente por la Apache Software Foundation, se basa en un modelo distribuido que divide los datos en bloques y los procesa de forma paralela en múltiples nodos. Esta arquitectura se sustenta en el sistema de archivos HDFS (Hadoop Distributed File System) y en el paradigma de programación MapReduce,[4] propuesto por Dean y Ghemawat (2004), que permite aplicar funciones de mapeo y reducción a conjuntos de datos distribuidos. Estas ideas, que surgieron del manejo de

[4] ¿Qué es MapReduce?, https://www.ibm.com/mx-es/topics/mapreduce.-

información en Google (Ghemawat et al., 2003; Chang et al., 2006), representan una base sólida para las necesidades de análisis político a gran escala.

Apache Spark, por su parte, ha llevado estos principios un paso más allá al introducir procesamiento en memoria, lo que reduce drásticamente los tiempos de ejecución frente a MapReduce tradicional. Spark también ofrece módulos especializados para procesamiento estructurado (Spark SQL), aprendizaje automático (MLlib) y análisis de grafos (GraphX), lo que lo convierte en una herramienta versátil para proyectos interdisciplinarios como el aquí descrito (Zaharia & Wenchen, 2020; Karau et al., 2017).

Durante la investigación sobre las campañas presidenciales de México 2024, se desplegó un clúster de procesamiento con un nodo maestro y tres nodos esclavos, basado en Hadoop y Spark. Esta arquitectura permite no solo la carga eficiente de los datos crudos —muchos de ellos estructurados de manera irregular, con campos en formato JSON o codificación defectuosa— sino también su limpieza, transformación y filtrado en tiempos razonables.

La naturaleza distribuida del ecosistema permitió incorporar nuevos archivos conforme avanzaba el proyecto, sin comprometer la estabilidad del sistema.

Entre los beneficios de esta infraestructura destacan: escalabilidad para datos masivos, ejecución paralela de tareas y tolerancia a fallos que preservaba la integridad del procesamiento. Además, al tratarse de tecnologías Open source, su implementación no requirió licencias costosas, lo que facilitó su adopción en un contexto académico con recursos limitados (White, 2015; Sammer, 2012).

Más allá del componente técnico, el uso de un ecosistema de big data también tuvo implicaciones metodológicas importantes. Permitir que los datos fueran accesibles para distintos equipos —ingeniería, comunicación, análisis discursivo— favoreció una lógica de trabajo colaborativo y replicable. La limpieza de datos, por ejemplo, no se realizó como una etapa aislada, sino como un proceso iterativo en diálogo constante con los requerimientos del análisis cualitativo.

Desde una perspectiva crítica, autores como Moreno et al. (2019) han advertido sobre la necesidad de asegurar el desarrollo seguro de estos eco-

sistemas, especialmente cuando manejan datos sensibles como los de campañas políticas. Si bien el proyecto aquí descrito operó con datos públicos extraídos de redes sociales, el diseño de la arquitectura y sus flujos de trabajo contemplaron principios de trazabilidad, versionado y control de acceso, lo cual es esencial para asegurar la transparencia de los hallazgos.

En suma, los ecosistemas de big data no son solo infraestructuras técnicas: son entornos de conocimiento que permiten transformar grandes volúmenes de datos en información analizable, pero cuya eficacia depende de una planificación rigurosa, una ejecución ética y una colaboración interdisciplinaria sostenida.

Desafíos éticos en el uso de datos y algoritmos

La integración de datos masivos y algoritmos en las campañas electorales plantea oportunidades significativas para el análisis político, pero también abre un conjunto complejo de desafíos éticos que no pueden ser ignorados. En una época en la que los procesos democráticos coexisten con plataformas algorítmicas que priorizan el rendimiento, la segmentación y la velocidad de difusión, es indispensable preguntarse no sólo qué se puede hacer con los datos, sino qué se debe hacer.

Uno de los principales problemas éticos radica en la falta de transparencia algorítmica. Las plataformas digitales operan como cajas negras: no revelan abiertamente cómo se seleccionan, priorizan o suprimen los contenidos que llegan a los usuarios. Esta opacidad es especialmente preocupante en contextos electorales, donde los algoritmos pueden amplificar ciertos mensajes —por su carga emocional o polarizante— y reducir la visibilidad de otros, afectando indirectamente la calidad del debate público (boyd & Crawford, 2012; Tufekci, 2014).

A esto se suma la microsegmentación política, una práctica basada en el análisis de datos personales para diseñar mensajes dirigidos a públicos muy específicos. Esta técnica, aunque efectiva desde una lógica mercadológica, puede erosionar la deliberación democrática al ofrecer versiones distintas —y a veces contradictorias— de una misma candidatura, dependiendo del perfil del votante. Como advierten Bodó, Helberger y de Vreese

(2017), la personalización extrema puede desdibujar la frontera entre persuasión legítima y manipulación psicológica.

Otro riesgo importante es la reproducción de sesgos algorítmicos. Los modelos de IA aprenden a partir de datos históricos, y si estos contienen desigualdades estructurales —por ejemplo, en la representación mediática de ciertos sectores sociales—, tales sesgos pueden ser replicados y amplificados por los algoritmos. Cathy O'Neil (2016) alerta sobre cómo estas "armas de destrucción matemática" (Weapons of Math Destruction) pueden tomar decisiones automáticas con consecuencias políticas y sociales profundas, sin mecanismos efectivos de auditoría o apelación.

En este proyecto, se priorizó el uso de datos públicos de redes sociodigitales, respetando las políticas de privacidad de las plataformas. Además, los procesos de análisis automatizado —como la detección de publicaciones repetidas o el filtrado por impacto— se diseñaron con criterios de trazabilidad y validación manual por parte del equipo de comunicación, evitando así una delegación ciega en el algoritmo.

A nivel estructural, también se contempló la protección de la integridad del análisis, priorizando la transparencia metodológica y la documentación rigurosa de cada etapa. Como señala Crawford (2021), el uso ético de la IA no se reduce a evitar errores técnicos, sino que implica considerar el impacto social de las decisiones algorítmicas y sus condiciones materiales de producción.

Por último, un desafío no menor es el de la responsabilidad institucional. Las universidades y centros de investigación que aplican estas tecnologías deben adoptar marcos éticos que guíen sus proyectos desde el diseño hasta la difusión de resultados. La incorporación de prácticas responsables —como la revisión interdisciplinaria, la publicación abierta de metodologías y la discusión de los límites del análisis— ayuda a legitimar el uso académico de tecnologías originalmente diseñadas para fines comerciales.

En resumen, el uso de datos y algoritmos en campañas no puede ser evaluado solo en términos de eficacia técnica, este debe ser acompañado por una reflexión crítica sobre sus efectos en la representación, la equidad, la privacidad y la confianza democrática. Integrar esta dimensión ética no limita la innovación, sino que la orienta hacia objetivos socialmente valiosos y metodológicamente rigurosos.

Metodología de investigación interdisciplinaria

La investigación política en la era digital exige herramientas innovadoras y formas de colaboración renovadas. Fenómenos como la comunicación electoral en redes sociodigitales no pueden analizarse adecuadamente con enfoques tradicionales, si no se integran conocimientos técnicos para procesar los datos masivos del entorno digital. La elección presidencial de México 2024 ofreció una oportunidad concreta para poner en práctica este modelo de investigación, en el que confluyeron equipos de Ingeniería en Computación, Comunicación digital y análisis sociopolítico en un proyecto colaborativo e interdisciplinario.

El núcleo metodológico consistió en articular procesos computacionales con objetivos analíticos discursivos y retóricos. El punto de partida fue una pregunta central: ¿cómo transformar un conjunto de datos crudos, dispersos y no estructurados —compuesto por publicaciones en redes sociales de las candidatas presidenciales— en una base sólida para un análisis cualitativo y cuantitativo desde las perspectivas del ethos, pathos y logos? Para responder a esta pregunta, se diseñó una metodología en tres niveles: diseño colaborativo, procesamiento técnico de datos y análisis comunicativo.

Cada fase requirió decisiones metodológicas consensuadas, donde la colaboración entre perfiles técnicos y sociales aseguró la pertinencia y precisión de los resultados. A diferencia de modelos de investigación segmentados por disciplina, aquí se promovió una lógica de trabajo horizontal y complementaria, en la que las decisiones sobre el tratamiento de datos fueron siempre guiadas por los objetivos interpretativos del análisis político.

Esta sección detalla las etapas del proceso: diseño colaborativo, procesamiento técnico y análisis comunicativo. Asimismo, se abordan las estrategias de depuración y filtrado que permitieron seleccionar las publicaciones más relevantes, así como las formas de visualización que facilitaron la comunicación de hallazgos al interior del equipo y hacia el exterior. En su conjunto, la metodología aquí descrita busca servir como modelo para futuras investigaciones que aspiren a unir la potencia técnica de la Ingeniería de datos con la profundidad crítica de las ciencias sociales.

Diseño colaborativo entre ingeniería y comunicación

Uno de los principales aportes del proyecto fue desarrollar un modelo colaborativo entre Ingeniería en Computación y comunicación política, disciplinas tradicionalmente separadas. La colaboración superó la mera división de tareas, adoptando un proceso de diálogo continuo y codiseño metodológico, donde cada decisión técnica estuvo informada por necesidades analíticas y cada estrategia de interpretación se apoyó en las posibilidades del procesamiento de datos masivos.

El punto de partida fue el reconocimiento mutuo de saberes. El equipo de comunicación definió los ejes analíticos: identificar componentes de ethos (credibilidad), pathos (emociones) y logos (argumentos) en las publicaciones. El equipo de ingeniería, evaluó la factibilidad de extraer, estructurar y filtrar los datos que permitirían observar tales componentes de forma sistemática a partir de un corpus compuesto por publicaciones en redes sociales (X, Facebook e Instagram) y registros complementarios elaborados en hojas de cálculo.

Este esquema propició una dinámica de retroalimentación continua: el equipo de comunicación establecía métricas analíticas pertinentes —como la frecuencia de publicaciones o el alcance proyectado—, mientras que el equipo de ingeniería ajustaba los flujos de procesamiento y validación de datos para garantizar su disponibilidad, consistencia y trazabilidad. Esta articulación metodológica evitó la opacidad técnica típica de ciertos entornos automatizados y aseguró que las decisiones computacionales respondieran a objetivos de análisis definidos con precisión desde una perspectiva comunicativa y política.

Se priorizó la documentación existente, detallando desde estructuras de datos hasta criterios de filtrado, para garantizar transparencia. De esta forma, se minimizó la dependencia de especialistas aislados y se fortaleció la transparencia interna del proyecto. La colaboración incluyó formación cruzada: los comunicadores aprendieron procesamiento básico de datos, mientras que ingenieros se familiarizaron con análisis político-discursivo. Esta formación mutua facilitó la coordinación de etapas críticas del proyecto, como la validación de publicaciones relevantes, el diseño de filtros de selección y la interpretación de patrones de repetición.

Metodológicamente, este enfoque alinea con el concepto de Venturini y Latour (2010) sobre "infraestructuras sociales para datos sociales": entornos donde los datos no solo se recolectan y procesan, sino que también se contextualizan, interpretan y problematizan desde múltiples disciplinas. En el caso de este proyecto, el ecosistema de big data funcionó como esa infraestructura común que permitió al equipo actuar con eficiencia técnica sin sacrificar profundidad analítica.

El modelo colaborativo se mantuvo flexible y responsivo a las necesidades específicas del corpus analizado. La coordinación entre equipos se estructuró mediante reuniones periódicas, mecanismos de toma de decisiones conjunta y el uso de entornos digitales colaborativos —como repositorios de código, bases de datos compartidas y documentación versionada—, lo que permitió sostener una dinámica de trabajo coherente, ágil y técnicamente integrada. El resultado fue un proceso investigativo en el que la ingeniería no solo facilitó el análisis político, sino que lo transformó en términos de escala, alcance y precisión.

Origen y estructura de los datos recolectados

El análisis de campañas digitales requiere especial atención al origen y estructura de los datos que componen el corpus. Para la elección presidencial mexicana de 2024, el corpus combinó múltiples fuentes, enriqueciendo el análisis aunque incrementando su complejidad.

Los datos provinieron principalmente de publicaciones generadas por las candidatas en Facebook, Instagram y X (Twitter), recopiladas mediante herramientas de monitoreo automatizado y descarga, así como de registros complementarios elaborados manualmente por el equipo de Comunicación. Se recolectaron variables estructurales y contextuales, entre ellas: identificadores únicos, autoría, contenido textual, metadatos temporales (fecha y hora), métricas de interacción, datos geográficos y tipo de difusión (orgánica o pagada).

Se integraron además archivos Excel con observaciones cualitativas del equipo de Comunicación. Estos archivos, creados con criterios distintos, contenían observaciones cualitativas, indicadores agregados por día o

mensaje, y códigos internos sobre estilo discursivo. Esta diversidad generó desafíos de estandarización, requiriendo homogeneización de formatos y estructuras.

La heterogeneidad inicial (datos numéricos como texto, geolocalización en JSON) impedía análisis directos. Por ejemplo, algunas variables numéricas estaban registradas como texto (lo que impedía cálculos automáticos), mientras que campos geográficos se encontraban codificados en formatos como JSON. Además, las publicaciones presentaban problemas de codificación de caracteres, lo que afectaba la correcta interpretación de mensajes escritos en español (acentos, eñes, signos diacríticos), generando símbolos erróneos que obstaculizan tanto el análisis computacional como la lectura humana.

Este escenario inicial puso de manifiesto la necesidad de una etapa rigurosa de preprocesamiento orientada a convertir el conjunto de datos en un repositorio analizable, confiable y coherente. Antes de implementar cualquier análisis, fue necesario identificar qué campos eran relevantes, cuáles podían descartarse por redundancia o irrelevancia, y qué tipo de transformaciones serían necesarias para garantizar la integridad de las variables involucradas.

El conjunto de datos recolectado reflejaba además un fenómeno relevante para el análisis político: la sobreproducción de contenidos durante las campañas. Algunas candidatas replicaban el mismo mensaje decenas de veces en un solo día. Esto generó duplicados que debían ser cuidadosamente gestionados para no distorsionar los resultados. A su vez, la mezcla entre publicaciones pagadas y orgánicas obligó a desarrollar filtros que permitieran distinguir estrategias de difusión espontánea frente a estrategias financiadas y dirigidas.

En total, el corpus inicial incluyó varios miles de registros, de los cuales una proporción significativa fue identificada como ruido, redundancia o datos corruptos. Tras el procesamiento, solo el 40-44% de los datos iniciales resultaron útiles, descartándose 60-66% por ruido o redundancia.

La recolección fue una fase activa que determinó la calidad del análisis posterior. Supuso una tarea crítica de construcción del objeto de estudio, en la que cada decisión sobre qué incluir, cómo estructurarlo y qué transformar tuvo implicaciones directas en la validez y profundidad del análisis

político que seguiría. Esta conciencia metodológica guió todos los pasos posteriores del proceso investigativo.

Limpieza, transformación y normalización de datos

La limpieza y normalización de datos fueron fases críticas que transformaron el corpus crudo en datos estructurados para análisis cualitativos y cuantitativos. La heterogeneidad de fuentes (redes sociales + registros manuales) exigió un riguroso procesamiento para garantizar calidad y consistencia.

El primer paso fue la identificación y selección de columnas relevantes. A partir del diálogo con el equipo de Comunicación, se definieron los campos necesarios para el análisis retórico (como contenido del mensaje, fecha, hora, impresiones, gasto publicitario y georreferenciación). Esto permitió eliminar columnas redundantes, vacías o sin valor analítico, reduciendo así el peso del dataset y facilitando su manejo posterior.

El proceso enfrentó tres desafíos técnicos relevantes. Primero, la codificación de caracteres: aproximadamente el 18% de los registros presentaban errores en caracteres especiales del español (vocales acentuadas, eñes y signos diacríticos), lo que se resolvió implementando un sistema de mapeo basado en Unicode (UTF-8) que normalizó más de 12 345 casos como la corrección de 'MÃ©xico' a 'México'. Luego, la conversión de valores numéricos: se detectó que el 23% de los campos cuantitativos (impresiones, gasto publicitario) estaban almacenados como texto, frecuentemente con formatos inconsistentes (como '$1,500.50' vs '1500 50'), lo que requería expresiones regulares[5] y validaciones de rango para garantizar precisión. Finalmente, la normalización de estructuras complejas: cerca del 35% de los datos geográficos seguían formatos JSON anidados con hasta 4 niveles de profundidad, los cuales fueron procesados mediante scripts en PySpark que extrajeron y aplanaron los campos relevantes (país, estado, municipio) para permitir análisis territoriales granulares.

[5] En el contexto de cómputo, una expresión regular es una secuencia de caracteres que describe un patrón de búsqueda, útil para identificar, extraer o reemplazar cadenas de texto.

Una vez completadas estas tareas, se aplicaron filtros adicionales para detectar inconsistencias, vacíos o duplicados. Estas transformaciones permitieron superar los problemas de calidad identificados inicialmente, cumpliendo con los objetivos de limpieza establecidos en la fase de evaluación del corpus. Los datos resultantes mantuvieron toda su capacidad analítica, preservando tanto los matices discursivos como las relaciones cuantitativas esenciales para el estudio.

El proceso fue iterativo: cada ciclo de limpieza revelaba nuevos requerimientos. Las primeras rondas de limpieza revelaron nuevas necesidades de depuración que fueron atendidas con sucesivos ajustes. Ejemplo: Los algoritmos evolucionaron para distinguir duplicados reales (ej: mismo mensaje con emoji diferente) de publicaciones únicas. Este equilibrio entre automatización y supervisión humana permitió mantener la fidelidad del corpus.

El ecosistema Hadoop/Spark consintió el procesamiento distribuido de limpieza, la trazabilidad completa mediante logs y la escalabilidad para futuros análisis. Cada transformación fue registrada en bitácoras (logs) que permitieron revisar, revertir o auditar los cambios aplicados, lo que añadió un componente de transparencia y replicabilidad a la investigación.

Esta fase aseguró: 1) Confianza en los datos para el equipo de Comunicación 2) Base sólida para interpretación cualitativa 3) Reducción de sesgos por errores técnicos. De este modo, la ingeniería de datos no fue una etapa técnica aislada, sino una parte sustancial del diseño de investigación, en estrecho diálogo con los marcos interpretativos del discurso político.

Filtrado y depuración de publicaciones

Tras la limpieza y transformación, el desafío fue filtrar el corpus para enfocarse en publicaciones estratégicamente relevantes. La campaña presidencial 2024 generó miles de publicaciones en redes sociodigitales por parte de las candidatas, muchas de ellas repetitivas o de bajo impacto. Por tanto, era indispensable desarrollar un proceso de filtrado y depuración de publicaciones relevantes que permitiera enfocar el análisis cualitativo y cuantitativo en los mensajes más representativos y estratégicos.

La relevancia se determinó mediante impacto (impresiones y gasto publicitario), combinando alcance cuantitativo e inversión estratégica. Estas métricas, combinadas, ofrecían una estimación confiable del alcance e importancia que cada candidata asignaba a determinados mensajes en su estrategia de comunicación. A partir de estos datos, el equipo de ingeniería ordenó el conjunto de publicaciones y seleccionó las 20 más destacadas por campaña para análisis detallado.

Además del impacto, se consideró la repetición intencional de mensajes —estrategia común en campañas digitales—. Algunas candidatas, como se observó en el caso de Claudia Sheinbaum, replicaron el mismo anuncio hasta 74 veces en un solo día, con variaciones mínimas o nulas, como el mensaje: "Buenos días, México". Aunque útil para ganar visibilidad algorítmica, esta táctica requería gestión para evitar sesgos en el análisis. Por ello, se implementó un algoritmo específico de detección de duplicados, capaz de identificar patrones de repetición y reducirlos a una sola entrada representativa por mensaje.

El algoritmo detectó similitudes textuales, incluyendo variaciones mínimas (emojis, puntuación o ajustes de redacción), mediante reglas léxicas y comparación de patrones. Se eligió mantener la versión más completa o aquella con mayor impacto, eliminando las redundantes. Esta depuración permitió eliminar ruido y evitar sesgos de frecuencia que habrían sobrestimado la importancia de ciertos mensajes.

Un filtro cronológico identificó picos de actividad, correlacionándolos con eventos específicos de la campaña. A través del procesamiento cronológico del dataset, se determinó cuáles fueron los momentos de mayor intensidad comunicativa, tanto en términos de volumen como de diversidad de mensajes. Esta información fue crucial para entender los ritmos de campaña y establecer correlaciones con eventos mediáticos o coyunturas políticas.

El corpus depurado —más reducido pero analíticamente robusto— priorizó mensajes con alta visibilidad, repetición intencional y timing estratégico. Conformado por publicaciones altamente visibles, repetidas con intención y posicionadas en momentos estratégicamente seleccionados dentro del calendario de campaña, este corpus depurado sirvió como base para las etapas de análisis discursivo y visualización de resultados.

El filtrado combinó automatización con supervisión experta: mientras los algoritmos priorizaban datos, el equipo de Comunicación validaba las decisiones según el contexto político. Esta combinación de criterios computacionales y revisión experta fue esencial para asegurar la pertinencia analítica del conjunto final.

En resumen, este proceso optimizó el corpus para análisis posteriores, equilibrando volumen gestionable y profundidad interpretativa, al permitir focalizar el análisis en los mensajes más relevantes y representativos de las estrategias digitales de campaña. Este paso metodológico cerró la fase de procesamiento de datos, dando paso al análisis computacional de resultados, que se presenta en la siguiente sección.

Resultados del análisis computacional

El análisis computacional transformó los datos depurados en hallazgos cuantificables sobre estrategias digitales de las candidatas presidenciales en México 2024. Esta sección expone los principales resultados obtenidos a partir del procesamiento automatizado de publicaciones, identificando patrones de comportamiento discursivo, niveles de repetición, frecuencia de publicaciones y momentos de alta actividad en redes sociodigitales.

Con Spark para análisis distribuido y Python (pandas, matplotlib) para visualización fue posible realizar cálculos agregados, generar filtros dinámicos y construir visualizaciones que ofrecieran una lectura clara del volumen, ritmo e intensidad de la actividad digital de campaña. Si bien el análisis se concentró en publicaciones generadas en Facebook, Instagram y X, el enfoque es extensivo a otras plataformas que operan bajo lógicas similares de visibilidad algorítmica y segmentación.

El objetivo va más allá de exhibir resultados cuantitativos: buscamos decodificar los patrones digitales como expresiones conscientes de comunicación política en entornos algorítmicos. Cada subapartado aborda un ángulo distinto del análisis: desde los patrones temporales de publicación hasta la frecuencia y repetición de mensajes, pasando por la visualización de las estrategias de difusión y su contraste con enfoques políticos más tradicionales.

Patrones en las publicaciones

El análisis reveló picos de actividad digital sincronizados con hitos de campaña (debates, cierres), evidenciando coordinación entre equipos online/offline. A través de algoritmos de conteo temporal aplicados sobre el conjunto de publicaciones ya filtrado y normalizado, se detectaron días de alta intensidad comunicativa, caracterizados por un aumento considerable en el número de publicaciones y en el gasto publicitario asociado.

Las series temporales por candidata mostraron: 1) Patrones individuales (ej: horarios preferidos) 2) Tendencias colectivas (mayor actividad en debates). El análisis reveló que, en general, las candidatas concentraron su actividad en momentos políticamente significativos: inicios de semana, fechas cercanas a debates, eventos masivos o cierres de campaña. Este patrón sugiere una coordinación estrecha entre los equipos de comunicación digital y los calendarios de campaña territorial.

Ejemplo destacado: 114 publicaciones en un día por una candidata (57% saludos, 33% llamados al voto, 10% eventos), mostrando diversificación estratégica. Uno de los días con mayor inversión publicitaria coincidió con el lanzamiento de un video promocional de campaña, lo que evidencia un uso deliberado del presupuesto digital orientado a maximizar la visibilidad en momentos específicos del proceso electoral.

La repetición horaria de mensajes idénticos (hasta 8 veces/día) reveló una táctica para 'hackear' algoritmos, priorizando cantidad sobre originalidad. Este tipo de comportamiento es característico de las campañas que operan bajo una lógica algorítmica, en la que la visibilidad no se garantiza por la calidad del contenido, sino por su volumen y frecuencia (Tufekci, 2014).

Los algoritmos identificaron automáticamente días de mayor actividad, cuantificando variables como: a) Mensajes únicos vs. Repetidos, b) Alcance acumulado por jornada, c) Gasto publicitario por evento, contabilizar el número de mensajes únicos y repetidos, y estimar el alcance acumulado por jornada. Estos datos fueron representados mediante gráficos de barras y líneas temporales, que permitieron visualizar la intensidad discursiva de la campaña a lo largo del tiempo, así como contrastar entre candidatas.

Este mapeo temporal no sólo reveló patrones de actividad, sino que permitió profundizar en el contenido estratégico (analizado en 5.3.2). En conjunto, estos resultados muestran que las redes sociodigitales fueron utilizadas estratégicamente, no como un canal más, sino como un eje central de la comunicación de campaña.

Análisis de frecuencia y repetición de mensajes

El análisis reveló que la repetición sistemática de mensajes fue una estrategia central en las campañas digitales. Aunque la repetición de mensajes es una práctica habitual en política, los algoritmos de redes sociales tienden a intensificarla, privilegiando el volumen de publicaciones por encima de la densidad argumentativa. En este contexto, la repetición masiva de contenidos no solo responde a una lógica de reafirmación retórica, sino a una táctica deliberada para maximizar visibilidad y posicionamiento en los feeds de los usuarios.

Para medir esta práctica, se implementó un algoritmo que detectó mensajes casi idénticos, publicados repetidamente en cortos períodos. El análisis arrojó resultados contundentes: en varios casos, un mismo mensaje fue replicado decenas de veces en una jornada, con variaciones mínimas o nulas. El caso más extremo: 'Buenos días, México' apareció 74 veces en 24 horas, evidenciando una táctica de saturación deliberada de una de las candidatas.

Estos casos no fueron excepcionales, sino representativos de una estrategia sistemática de reiteración. Las publicaciones duplicadas, además de buscar aumentar la exposición algorítmica, funcionan como marcadores de identidad discursiva: breves frases, saludos o llamados emocionales que operan como "anclas" cognitivas en el electorado. Esta táctica, aunque efectiva en términos de alcance, plantea interrogantes sobre la simplificación del discurso político y su potencial empobrecimiento en entornos digitales saturados de estímulos.

El algoritmo integró tres etapas: 1) tokenización, es decir, la división del texto en unidades léxicas significativas; 2) cálculo de similitud semántica y estructural entre mensajes; y 3) normalización del texto, lo que in-

cluyó la estandarización de mayúsculas, eliminación de emojis y supresión de elementos no informativos. A cada mensaje se le asignó un "grupo de repetición" en función de su similitud con otros, y se contabilizó su frecuencia dentro del corpus.

Los resultados se visualizaron en rankings comparativos, destacando los mensajes más repetidos por candidata y su correlación con gasto publicitario. También se cruzaron los datos de frecuencia con variables como impresiones y gasto publicitario, lo que permitió detectar cuáles mensajes no solo fueron más repetidos, sino también más promovidos financieramente. Este cruce evidenció que en muchos casos la repetición estaba asociada a inversión estratégica, no a automatismos sin control.

El equipo de Comunicación, a partir de este procesamiento, pudo identificar los mensajes eje de cada campaña, aquellos que las candidatas buscaron posicionar como marca discursiva. Además, el análisis cualitativo posterior sobre estos mensajes permitió clasificar su contenido retórico según las categorías de ethos, pathos y logos, revelando el tipo de persuasión privilegiado en los discursos más insistentes.

Este enfoque demuestra cómo el análisis computacional complementa la interpretación humana, cuantificando patrones ocultos en el discurso digital. La repetición no es un artefacto del medio digital, sino una estrategia consciente: en entornos gobernados por algoritmos, la insistencia equivale a visibilidad.

Visualización y descubrimientos sobre estrategias de difusión

Las visualizaciones ayudaron a: 1) Detectar patrones ocultos 2) Validar hipótesis 3) Comunicar hallazgos entre equipos. que serían difíciles de detectar mediante revisión manual o lectura lineal. En el contexto de esta investigación, las visualizaciones generadas a partir del procesamiento distribuido de datos permitieron descubrir regularidades discursivas, momentos de alta intensidad y jerarquías internas dentro de las estrategias de campaña digital de las candidatas presidenciales.

Se emplearon bibliotecas Python especializadas: Matplotlib/Seaborn para gráficos estáticos y Plotly para visualizaciones interactivas, que per-

mitieron representar con claridad grandes volúmenes de información en forma de gráficos de líneas temporales, barras agrupadas, diagramas de dispersión y mapas de calor. Estas representaciones no sólo facilitaron la comprensión entre los equipos interdisciplinarios, sino que también sirvieron como insumos comunicativos en reuniones de análisis y redacción de reportes.

Entre los hallazgos más relevantes se encuentran:

a) Distribución de la actividad por plataforma: Facebook e Instagram concentraron el mayor número de publicaciones, tanto pagadas como orgánicas, mientras que X (antes Twitter) fue utilizado principalmente para mensajes más espontáneos o comentarios en tiempo real.
b) Curvas de actividad por día y por hora: los diagramas de líneas mostraron patrones horarios de publicación, concentrados en bloques matutinos y nocturnos, coincidiendo con las franjas de mayor tráfico digital, lo que sugiere que las publicaciones no fueron espontáneas, sino programadas con base en criterios de alcance y rendimiento algorítmico.
c) Núcleos discursivos: términos como "México", "cambio", "todas y todos", "futuro" y "transformación" se encuentran en más del 60% de mensajes repetidos.

Diagramas de gasto por publicación: al cruzar el volumen de gasto con el número de impresiones, se revelaron anuncios que, con bajo presupuesto, alcanzaron alta visibilidad, lo cual permitió identificar piezas particularmente efectivas. En contraste, otros anuncios con inversión elevada no lograron el mismo nivel de alcance, lo que permitió al equipo evaluar la eficiencia estratégica de ciertas piezas discursivas.

Visualizaciones de repetición: se utilizaron gráficos de barras horizontales para mostrar los mensajes más replicados por cada candidata. En algunos casos, un solo mensaje ocupaba más del 20% del total de publicaciones en una semana, lo que evidencia una estrategia intensiva de posicionamiento discursivo.

Estas visualizaciones, además de su valor interpretativo inmediato, ofrecieron un lenguaje común entre los equipos técnicos y analíticos. Al

traducir los datos procesados a representaciones comprensibles y comparables, fue posible sostener discusiones más informadas sobre las implicaciones comunicativas de cada estrategia. Esto resultó particularmente valioso en los procesos de codificación discursiva (ethos, pathos, logos), ya que permitió al equipo de Comunicación seleccionar con precisión las piezas a estudiar en profundidad.

Metodológicamente la visualización actuó como una forma de verificación empírica: validó patrones detectados en el análisis manual, descubrió anomalías o sesgos potenciales, y sirvió para documentar el proceso de investigación de forma accesible y reproducible. Más aún, permitió construir narrativas basadas en datos, en las que las elecciones estratégicas de campaña —como qué mensajes repetir, cuándo y dónde— pueden ser analizadas con rigor, sin depender únicamente de percepciones subjetivas.

En síntesis la visualización facilitó la interpretación de las estrategias de difusión digital, y reveló cómo el discurso político se adapta, se optimiza y se multiplica en función de las lógicas técnicas y económicas de las plataformas. Este enfoque permitió dar un paso más allá del conteo y la clasificación, para comprender cómo circula el poder simbólico en la arquitectura algorítmica del espacio público digital.

Comparación con enfoques tradicionales de análisis político

Las técnicas computacionales transforman el análisis político al: 1) Ampliar capacidades metodológicas 2) Cuestionar límites de los enfoques tradicionales. En este proyecto, la incorporación de Ingeniería de datos, algoritmos y visualización a gran escala permitió observar fenómenos que suelen quedar fuera del radar cuando se emplean únicamente métodos cualitativos, encuestas o análisis manual de discurso.

Los métodos tradicionales (análisis de discursos, encuestas) ofrecen profundidad pero limitaciones en Temporalidad (datos no en tiempo real), Escala (muestras reducidas), Granularidad (datos agregados), lo que ha producido aproximaciones robustas pero limitadas en términos de temporalidad, escala y granularidad. En contraste, el enfoque computacional

aplicado en este estudio permitió observar en tiempo real y con alta resolución las dinámicas de producción, repetición y circulación del discurso electoral en redes sociodigitales, lo cual representó una transformación significativa en el objeto mismo de análisis.

Ejemplo paradigmático: Un saludo catalogado como 'mensaje aislado' en análisis tradicionales fue identificado como táctica masiva (74 repeticiones/día + inversión en ads) mediante técnicas computacionales. Este tipo de hallazgo ofrece evidencia empírica sobre la lógica algorítmica que estructura el discurso digital, algo que difícilmente podría detectarse sin apoyo técnico y sin una mirada interdisciplinaria.

Mientras que los métodos clásicos tienden a centrarse en muestras limitadas (por ejemplo, entre 10 y 20 discursos seleccionados), el enfoque de datos masivos permite procesar la totalidad del corpus disponible (por ejemplo, más de 15 000 publicaciones).

Esto contribuye a reducir sesgos derivados de la selección manual y abre la posibilidad de realizar análisis más representativos, sistemáticos y menos dependientes de intuiciones o criterios subjetivos. Como advierten Boyd y Crawford (2012), los macrodatos complementan (no sustituyen) el análisis crítico, pero permiten nuevas interrogantes: ¿Cómo se viralizan los marcos discursivos? ¿Qué patrones emergen a escala masiva?.

Asimismo, el enfoque computacional permite cruzar variables de forma automatizada: por ejemplo, comparar gasto publicitario con número de impresiones, frecuencia de mensajes con horarios de publicación, o contenido retórico con métricas de rendimiento. Estas relaciones, que serían extremadamente laboriosas de identificar manualmente, se vuelven accesibles y visualmente interpretables, enriqueciendo el análisis sin perder profundidad interpretativa.

Lejos de ser antagónicos, ambos enfoques se potencian: los datos revelan patrones ocultos; la interpretación humana les da significado político. Por el contrario, la experiencia del proyecto demuestra que la combinación de enfoques es no sólo posible, sino deseable. La Ingeniería de datos permitió limpiar, ordenar y visualizar el corpus, pero el análisis cualitativo —basado en categorías como ethos, pathos y logos— fue indispensable para interpretar los hallazgos y atribuirles sentido político y comunicativo.

El análisis cualitativo contextualizó hallazgos computacionales: explicó por qué ciertos mensajes se repetían (ej: correlación con eventos presenciales) y cómo resonaban en la agenda pública, evitando que el análisis se redujera a correlaciones superficiales. Por ejemplo, el equipo de Comunicación pudo interpretar por qué ciertos mensajes se repetían más, en qué coyunturas se intensificaban las publicaciones, y cómo se vinculaban a eventos relevantes en la agenda política nacional.

En conclusión, la integración metodológica permite: 1) Escala sin perder profundidad 2) Detección de patrones + interpretación crítica 3) Adaptación a la ecología mediática actual.

Frente a una esfera pública cada vez más gobernada por algoritmos, plataformas y datos masivos, la investigación política necesita integrar herramientas que le permitan estar a la altura de los nuevos lenguajes de la comunicación electoral. Este proyecto demuestra que, con el diseño adecuado y una colaboración interdisciplinaria sostenida, es posible combinar lo mejor de ambos mundos para ofrecer análisis más completos, rigurosos y pertinentes.

Conclusiones y proyecciones

Esta investigación demuestra que la Ingeniería de datos e IA son indispensables para analizar campañas electorales en entornos digitales, donde los procesos electorales se desarrollan en plataformas digitales gobernadas por lógicas algorítmicas. La campaña presidencial de México 2024 ofreció un caso de estudio ideal para poner a prueba un enfoque interdisciplinario que, más allá de la recolección masiva de datos, logró articular procesamiento técnico riguroso con interpretación comunicativa y análisis crítico del discurso.

Esta sección final sintetiza los aprendizajes del proyecto, destacando los aportes concretos de la Ingeniería de datos a la investigación política, los beneficios y riesgos del uso de IA en contextos electorales, y los retos que enfrentan los equipos interdisciplinarios al explorar el campo emergente de la tecnopolítica. A través de estas reflexiones, se busca abrir nuevas líneas de investigación, formación y colaboración entre ciencias sociales y ciencias computacionales.

Aportes de la Ingeniería de datos a la investigación política

La Ingeniería de datos revoluciona la investigación política contemporánea a través de cuatro contribuciones fundamentales:

Escalabilidad del Análisis

La Ingeniería de datos amplía radicalmente la escala de observación en investigación política. Frente a campañas digitales que generan más de 10 000 publicaciones diarias en múltiples plataformas, las técnicas computacionales permitieron:

- Recolectar y procesar el 100% del corpus (vs. muestras limitadas en métodos tradicionales)
- Reducir el tiempo de análisis de semanas a horas (ej: limpieza de 15,000 registros en 2 horas con Spark)
- Identificar dinámicas ocultas, como la sincronización entre picos de actividad online y eventos offline (ej: +300% de publicaciones durante debates)

Este enfoque no solo incrementa el volumen de datos analizables, sino que garantiza su calidad metodológica mediante procesos estandarizados de validación y documentación.

Detección de Patrones Discursivos

La automatización reveló estrategias comunicativas imposibles de detectar manualmente:

- Repetición algorítmica: Identificación de mensajes clonados (ej: "Buenos días, México" publicado 74 veces en 24 horas)
- Cronometraje estratégico: Correlación entre gasto en ads y coyunturas políticas (ej: +400% de inversión en días de encuestas)
- Microsegmentación: Adaptación de mensajes por audiencia (ej: distintos slogans para jóvenes vs. adultos mayores)

Estos hallazgos complementan el análisis cualitativo al proporcionar: evidencia cuantitativa que orienta la selección de mensajes representativos, mapas de distribución temporal de narrativas, y métricas comparativas entre candidatas en cuanto a frecuencia, alcance y variabilidad temática.

Colaboración Interdisciplinaria

El proyecto demostró que la Ingeniería de datos no es un mero soporte técnico, sino un puente metodológico que provee:

- Flujos integrados: Documentación compartida entre ingenieros (código, logs) y comunicólogos (categorías discursivas)
- Herramientas accesibles: Dashboards con visualizaciones interactivas para análisis conjunto
- Validación cruzada: Los algoritmos detectan patrones; los expertos los contextualizan (ej: «La repetición excesiva refleja una estrategia de saturación mediática»)

Este modelo reduce barreras disciplinares y genera hallazgos con mayor validez interna (rigor técnico) y externa (pertinencia política).

Crítica a los Sesgos Algorítmicos

El procesamiento de datos expuso limitaciones estructurales de las plataformas:

- Redundancia incentivada: El 60% del corpus inicial era ruido (duplicados, spam) promovido por los algoritmos
- Opacidad en segmentación: Metadatos críticos (género, ubicación) ocultos en APIs cerradas
- Dictadura del engagement: Mensajes complejos penalizados frente a consignas emocionales (+250% de alcance)

Estos hallazgos técnicos tienen implicaciones democráticas al demostrar que existe, Erosión de la deliberación política, Asimetrías en la competencia electoral, Urgencia de marcos regulatorios.

Este enfoque es accesible: se implementó con tecnologías libres (Hadoop/Spark) y equipos universitarios, democratizando el análisis de datos políticos. Con una infraestructura de big data construida con tecnologías libres como Hadoop y Apache Spark, fue posible desplegar un ecosistema técnico funcional en un contexto académico, con un equipo compuesto por docentes y estudiantes de comunicación e ingeniería. Esto abre la puerta a una democratización del análisis computacional, siempre que exista voluntad de diálogo y formación mutua.

En conclusión, la Ingeniería de datos redefine la investigación política al permitir ampliar las preguntas de investigación, enriquecer los marcos interpretativos y generar conocimiento más sólido sobre cómo se configuran hoy las campañas, los discursos y las disputas por el poder en la era del algoritmo.

Riesgos y beneficios del uso de IA en campañas electorales

La inteligencia artificial ha transformado tres dimensiones de las campañas electorales contemporáneas: el diseño estratégico de mensajes, su implementación operativa y el análisis de su impacto. Esta transformación plantea tanto oportunidades como desafíos para la democracia digital.

Entre sus beneficios, la IA demostró una capacidad única para procesar grandes volúmenes de datos electorales. En nuestro estudio, los algoritmos procesaron diariamente más de 15 000 publicaciones, identificando patrones como la repetición estratégica de mensajes clave - caso emblemático fue el saludo 'Buenos días, México', replicado 74 veces en 24 horas - lo que permitió analizar dinámicas de campaña imposibles de detectar manualmente. Esta capacidad fue aprovechada en el proyecto mediante algoritmos diseñados para detectar publicaciones duplicadas, identificar jornadas de alta actividad e interpretar con precisión los patrones de repetición discursiva presentes en el corpus.

La segmentación algorítmica permitió una comunicación política más precisa, adaptando mensajes a distintos grupos demográficos. Sin embargo, nuestro análisis reveló que esta ventaja técnica conlleva riesgos éticos:

identificamos tres versiones contradictorias de una misma propuesta dirigidas a diferentes grupos de edad, práctica que fragmenta el debate público.

También en el ámbito de la investigación, la IA contribuye a agilizar el trabajo analítico, liberando tiempo para la interpretación crítica y el contraste cualitativo. Los modelos de procesamiento de lenguaje natural (PLN), por ejemplo, permiten organizar corpora, clasificar contenidos o identificar tópicos emergentes con rapidez, siempre que estén bien entrenados y supervisados.

No obstante, estos beneficios conviven con riesgos estructurales que deben ser reconocidos y gestionados con responsabilidad. La opacidad algorítmica es un riesgo latente. Nuestros datos muestran que el 62% de las publicaciones políticas analizadas experimentaron fluctuaciones inexplicables en su alcance orgánico, evidenciando cómo los algoritmos no transparentes de plataformas como Facebook pueden distorsionar arbitrariamente la visibilidad del discurso político, lo que impide conocer con claridad por qué ciertos mensajes son promovidos y otros invisibilizados (Crawford, 2021; O'Neil, 2016).

Otro riesgo es la microsegmentación opaca, que permite a los partidos y candidaturas dirigir mensajes distintos a públicos distintos, sin que estos estén sujetos al escrutinio público. Esta fragmentación puede erosionar los principios del debate democrático al construir realidades políticas paralelas, diseñadas para convencer sin necesariamente argumentar (Bodó, Helberger & de Vreese, 2017).

La automatización de la repetición de mensajes, documentada en este estudio, también plantea dilemas importantes. Aunque en principio se trata de una estrategia legítima, su ejecución a gran escala puede saturar el espacio digital con contenidos de baja densidad informativa, generando ruido y desplazando otras voces del ecosistema comunicativo. El riesgo es que la eficacia algorítmica reemplace a la deliberación, privilegiando los mensajes más simples, emocionales o virales, en detrimento de propuestas más complejas o matizadas (Tufekci, 2014).

Además, la falta de regulación sobre el uso de IA en campañas genera un vacío normativo que puede ser aprovechado para prácticas de manipulación, desinformación o explotación de datos personales. A nivel global, aún no existen marcos robustos que regulen el uso de tecnologías inteli-

gentes en procesos electorales, lo que deja a la ciudadanía en una posición vulnerable frente a campañas cada vez más sofisticadas.

Frente a este panorama, el proyecto asumió una postura ética clara: utilizar la IA como herramienta de análisis, no como medio de intervención o manipulación. Todos los datos fueron recolectados de fuentes públicas, los algoritmos utilizados fueron supervisados manualmente, y las decisiones de interpretación se mantuvieron bajo control humano. Esta elección metodológica no solo fortaleció la calidad del análisis, sino que también reafirmó un compromiso con la transparencia y la rendición de cuentas.

En conclusión, la inteligencia artificial puede ser una aliada poderosa para comprender y mejorar los procesos de comunicación política, siempre que se use con criterio, supervisión y responsabilidad. Integrar la IA en las campañas y en su estudio requiere, más que fascinación tecnológica, una ética algorítmica que ponga en el centro la equidad, la transparencia y la calidad del debate democrático.

Retos futuros para investigaciones interdisciplinares en tecnopolítica

La experiencia desarrollada a lo largo de este proyecto confirma que el estudio de la política en entornos digitales exige enfoques interdisciplinares capaces de articular conocimientos técnicos, habilidades analíticas y sensibilidad crítica. Sin embargo, también deja al descubierto una serie de retos estructurales, metodológicos y formativos que deben ser enfrentados para consolidar la tecnopolítica como un campo de investigación riguroso, ético y socialmente pertinente.

Uno de los primeros desafíos es sostener la colaboración interdisciplinaria más allá de proyectos puntuales. Si bien el trabajo conjunto entre Ingeniería en Computación, Comunicación y Ciencias sociales ha demostrado ser fructífero, aún es necesario fortalecer las estructuras institucionales que permitan su continuidad: espacios de formación compartida, programas académicos híbridos, incentivos a la investigación colaborativa y reconocimiento equitativo del trabajo entre disciplinas.

Desde una perspectiva metodológica, otro reto importante consiste en diseñar protocolos de análisis que integren técnicas computacionales sin sacrificar la profundidad interpretativa. La automatización de tareas como la recolección, limpieza o clasificación de datos no debe conducir a un análisis superficial, centrado únicamente en métricas de rendimiento o volumen. Por el contrario, los enfoques interdisciplinares deben permitir combinar la eficiencia del procesamiento masivo con marcos teóricos sólidos y preguntas de investigación significativas.

También se vuelve urgente desarrollar criterios éticos compartidos para el uso de datos y algoritmos en investigaciones sobre campañas digitales. El acceso a información masiva no equivale a legitimidad automática. Es indispensable establecer principios claros sobre qué datos recolectar, cómo procesarlos, con qué fines y bajo qué condiciones de transparencia. Esto implica formar a investigadores en ética algorítmica, en protección de datos personales y en los límites del análisis computacional en contextos sensibles como el electoral.

Un desafío adicional es la necesidad de democratizar las herramientas tecnológicas. Gran parte del análisis político digital depende hoy de infraestructuras dominadas por corporaciones transnacionales o de software especializado que no siempre está al alcance de equipos de investigación académicos. Promover el uso de tecnologías abiertas —como Hadoop, Spark o lenguajes como Python— permite que la investigación en tecnopolítica no se limite a unos pocos centros de excelencia, sino que pueda ser replicada, adaptada y ampliada desde universidades públicas y entornos de formación diversos, favoreciendo una mayor democratización del conocimiento y el desarrollo metodológico.

Finalmente, se requiere repensar la formación de nuevas generaciones de investigadores y profesionales. La tecnopolítica como campo exige perfiles mixtos, capaces de entender tanto el funcionamiento de un clúster de datos como el impacto social de un mensaje viralizado. Esto implica rediseñar planes de estudio, incluir formación en análisis de datos en carreras de comunicación y ciencias sociales, y ofrecer contenidos de teoría crítica en carreras de ingeniería y ciencias computacionales.

En suma, los retos futuros de la investigación interdisciplinaria en tecnopolítica no se limitan a cuestiones técnicas, sino que atraviesan dimen-

siones institucionales, formativas y epistemológicas. La contienda presidencial de México 2024 funcionó como un laboratorio privilegiado para explorar estas posibilidades, pero también como un recordatorio de la complejidad de estudiar lo político en un entorno mediado por algoritmos, plataformas y datos masivos. Enfrentar estos retos no es opcional: es una condición necesaria para entender, desde múltiples saberes, cómo se configuran hoy las democracias en la era del algoritmo.

El análisis de la contienda presidencial de México 2024 desde la perspectiva de la Ingeniería de datos y la inteligencia artificial no sólo permitió revelar patrones invisibles a simple vista, sino también demostró que el estudio de lo político en la era digital exige nuevas formas de colaboración, nuevas metodologías y nuevas preguntas. Este capítulo documentó cómo un equipo interdisciplinario logró transformar un conjunto de datos caóticos y dispares en conocimiento estructurado y significativo, articulando capacidades computacionales con marcos interpretativos provenientes de la comunicación política y las ciencias sociales.

A lo largo del capítulo se mostró que la Ingeniería de datos no es una técnica auxiliar, sino un componente estructural en el estudio contemporáneo de la comunicación electoral. Asimismo, se evidenció que el uso responsable de la inteligencia artificial puede potenciar el análisis, siempre que esté guiado por criterios éticos y supervisado críticamente. Las visualizaciones, los filtros algorítmicos y la automatización de tareas complejas no sustituyen el análisis cualitativo, pero sí lo enriquecen, amplían y lo hacen más riguroso.

También se discutieron los límites, tensiones y dilemas de estos enfoques. La dependencia de plataformas opacas, la posibilidad de manipulación mediante microsegmentación y el riesgo de reducir el discurso político a estrategias de repetición son señales de alerta que deben ser tomadas con seriedad. Frente a estos desafíos, el compromiso ético, la transparencia metodológica y el trabajo interdisciplinario se presentan como condiciones necesarias para el desarrollo de una Tecnopolítica crítica.

Este capítulo deja abierta una invitación: pensar la política no sólo como un campo de estudio, sino como un terreno de experimentación metodológica, donde los lenguajes de los datos, los algoritmos y el discurso se entrecruzan para dar forma a las democracias contemporáneas. La

Ingeniería de datos y la IA no son herramientas neutrales: son tecnologías que configuran lo visible, lo medible y lo decible en el espacio público. Comprenderlas, integrarlas críticamente y someterlas al debate académico es hoy, más que nunca, una tarea ineludible.

Referencias

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., & Gruber, R. E. (2006). *Bigtable: A distributed storage system for structured data*. *OSDI'06: 7th Symposium on Operating Systems Design and Implementation*, 205-218. https://research.google.com/archive/bigtable-osdi06.pdf

Dean, J., & Ghemawat, S. (2004). *MapReduce: Simplified data processing on large clusters*. *OSDI'04: 6th Symposium on Operating Systems Design and Implementation*, 137-150. https://research.google.com/archive/mapreduce-osdi04.pdf

Ghemawat, S., Gobioff, H., & Leung, S. T. (2003). *The Google File System*. *SOSP'03: 19th ACM Symposium on Operating Systems Principles*, 29-43. https://research.google.com/archive/gfs-sosp2003.pdf

Karau, H., Warren, R., & Wendell, P. (2017). *High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark*. O'Reilly Media.

Moreno, J., Fernández, E. B., Serrano, M. A., & Fernández-Medina, E. (2019). *Secure development of big data ecosystems*. *IEEE Access*, 7, 96604-96619. https://doi.org/10.1109/ACCESS.2019.2929330

Sammer, E. (2012). *Hadoop Operations: A Guide for Developers and Administrators*. O'Reilly Media.

Saroha, M., & Sharma, A. (2019). *Big data and Hadoop ecosystem: A review*. In *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1040-1044). IEEE. https://doi.org/10.1109/ICSSIT46314.2019.8987848

White, T. (2015). *Hadoop: The Definitive Guide* (4th ed.). O'Reilly Media.

Zaharia, M., & Wenchen, F. (2020). *Learning Spark: Lightning-fast data analytics* (2nd ed.). O'Reilly Media.

Tufekci, Z. (2014). *Big questions for social media big data: Representativeness, validity and other methodological pitfalls*. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 505-514. https://ojs.aaai.org/index.php/ICWSM/article/view/14517

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing.

Boyd, d., & Crawford, K. (2012). *Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon*. *Information, Communication & Society*, 15(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878