DOI: https://doi.org/10.29105/gmjmx23.44-556 Artículos

CONSTRUCCIÓN DE CORPUS EN ENTORNOS DIGITALES: ESTRATEGIAS MIXTAS PARA ESTUDIAR LA COMUNICACIÓN POLÍTICA EN X

CONSTRUCTION OF CORPUS IN DIGITAL ENVIRONMENTS: MIXED STRATEGIES FOR STUDYING POLITICAL COMMUNICATION ON X

Enrique Pérez Reséndiz
Universidad Autónoma de la Ciudad de México, México
https://orcid.org/0000-0001-8330-3856

Autor para correspondencia: Enrique Pérez Reséndiz, email: enrique.perez.resendiz@uacm.edu.mx

Resumen

El análisis de entornos sociodigitales en el campo de la comunicación política presenta desafíos particulares, derivados tanto de la velocidad con que circula la información como de la magnitud de datos producidos en plataformas como X (antes Twitter). Este artículo se centra en una reflexión de corte metodológico en relación a la producción y configuración de subjetividades políticas en torno al caso Ayotzinapa y su relación con la plataforma X para mostrar cómo se construyó un corpus de análisis. A partir de una estrategia metodológica mixta que combina la etnografía digital con la extracción de datos de la plataforma, se discuten los criterios de delimitación seguidos para configurar el corpus: pertinencia temática, temporalidad de los acontecimientos y selección de hashtags vinculados con la protesta y la denuncia pública del caso. Asimismo, se revisan tres enfoques metodológicos habituales en la conformación de corpus en plataformas sociodigitales —estudios de caso, la reutilización de muestras previas y las muestras de conveniencia— para contrastarlos con la estrategia adoptada en esta investigación. Con ello, el texto pretende contribuir a la reflexión sobre los retos metodológicos de la investigación en comunicación política digital y aportar una ruta de trabajo para el análisis de interacciones y discursos en escenarios digitales.

Palabras clave: corpus de análisis, representatividad, sesgos algorítmicos, etnografía digital, hashtags.

Abstract

The analysis of sociodigital environments in the field of political communication presents particular challenges, derived both from the speed at which information circulates and from the vast amount of data produced on platforms such as X (formerly Twitter). This article develops a methodological reflection on the production and configuration of political subjectivities around the Ayotzinapa case and its relationship with the X platform in order to illustrate how an analytical corpus was constructed. Based on a mixed methodological strategy that

combines digital ethnography with data extraction from the platform, the article discusses the delimitation criteria followed to configure the corpus: thematic relevance, event temporality, and the selection of hashtags linked to protest and public denunciation. It also reviews three common methodological approaches to corpus construction in sociodigital platforms —case studies, the reuse of samples from previous research, and convenience sampling— in order to contrast them with the strategy adopted in this study. In doing so, the text contributes to ongoing reflections on the methodological challenges of research in digital political communication and provides a roadmap for analyzing interactions and discourses in digital environments.

Keywords: corpus construction, data representativeness, algorithmic bias, digital ethnography, hashtags.

Recibido: 30/09/2025
Aceptado: 15/12/2025

Introducción

En el panorama contemporáneo de la investigación social, la relación entre tecnologías digitales y cultura introduce desafíos significativos para el estudio de los entornos sociodigitales. Estos retos abarcan tanto la comprensión de las prácticas que se configuran en dichos espacios como las implicaciones metodológicas y técnicas vinculadas al trabajo de campo y a la construcción de corpus de análisis. Abordarlos resulta fundamental para avanzar en el entendimiento de las dinámicas sociales que se despliegan en el ámbito digital. Las dificultades del análisis en plataformas sociodigitales son parte de la misma estructura y dinámica de estos entornos. Una de las características principales es la rápida producción y circulación de información; los grandes volúmenes de datos representan un reto tanto en términos técnicos como metodológicos para el análisis social (Rodríguez Cano, 2022). Específicamente la construcción y delimitación de muestras y de corpus de análisis se enfrentan a los siguientes cuestionamientos: ¿es posible aplicar el criterio de representatividad a los grandes volúmenes de datos surgidos de estas plataformas?, ¿es posible la representatividad al trabajar con datos que constantemente se están renovando y generando en tiempo real?, ¿qué criterios se deben establecer en los cortes temporales/espaciales para el análisis de los datos, si una de las características de estas plataformas es que trascienden, o por lo menos cambian, los flujos espacio-temporales?

A las preguntas anteriores también habría que añadir la calidad y fiabilidad de los datos, y la problemática con la falta de estructuración de los mismos, es decir, que los datos que se van generando en estos entornos no siguen un formato estructurado, lo que dificulta no sólo su extracción y sistematización sino también su análisis.

La intención de este texto es presentar una reflexión acerca de cómo se ha abordado esta

decisión metodológica en algunos trabajos de investigación. Posteriormente se explican los criterios bajo los cuales se construye el corpus de análisis en el marco de la investigación titulada "Experiencia y tecnopolítica: análisis sobre la construcción de subjetividades políticas en el caso #TodosSomosAyotzinapa". El documento presenta, en primer lugar, una serie de criterios, que a nivel teórico-epistemológico se considera que deben ser tomados en cuenta en la construcción de un corpus de análisis con datos de plataformas sociodigitales. En segundo lugar, se hace una breve revisión de otros trabajos de investigación, poniendo énfasis en los criterios establecidos en la construcción de sus corpus de análisis. Finalmente, se reflexiona en torno al propio trabajo de investigación y la inclusión de distintos criterios y elementos metodológicos para la construcción de un corpus de análisis propio.

Elementos estructurales básicos para la construcción de muestras *ad hoc* en la investigación en entornos sociodigitales

Las muestras *ad hoc* son un tipo de muestreo no probabilístico diseñado para atender un objetivo puntual de investigación. Su nombre, proveniente del latín *"para esto"*, lo cual quiere decir que no siguen un procedimiento estandarizado, sino que se construyen a partir de criterios definidos por el investigador. En la literatura metodológica este tipo de diseño muestral se ha relacionado con la muestra teórica relacionada con la teoría fundamentada (de la Garza Toledo, 2012), y con la muestra intencional u orientada utilizada en estudios cualitativos (Otzen & Manterola, 2017). Ambas estrategias comparten la búsqueda deliberada de casos que aporten información relevante para el desarrollo analítico. Las muestras *ad hoc* suelen emplearse en estudios exploratorios, pruebas piloto o investigaciones que requieren obtener datos de manera ágil y con recursos limitados. Entre sus características destacan la flexibilidad y la rapidez (Hernández Sampieri et al., 2014), ya que permiten seleccionar a los participantes de acuerdo con su pertinencia para el estudio (por ejemplo, estudiantes de un curso particular, consumidores de una marca o usuarios de una aplicación). Sin embargo, al no basarse en principios de aleatoriedad, carecen de representatividad estadística, por lo que los resultados no pueden generalizarse a toda la población. Aún con esta limitación, son útiles para obtener primeras aproximaciones, validar instrumentos o indagar en grupos muy concretos.

En el campo de las investigaciones sobre entornos y plataformas sociodigitales, las muestras *ad hoc* resultan especialmente útiles porque permiten seleccionar, de manera intencional, a los usuarios o colectivos que cumplen con características muy específicas vinculadas al objeto de estudio, como quienes participan en un *hashtag* determinado, administran grupos en redes sociales o producen contenidos en comunidades digitales. Dada la naturaleza dinámica y cambiante de estos espacios, el muestreo *ad hoc* ofrece la flexibilidad necesaria para adaptarse a fenómenos emergentes y acotar poblaciones difíciles de delimitar con métodos

probabilísticos. Aunque sus resultados no son generalizables al conjunto de usuarios de la plataforma, sí posibilitan obtener información relevante y focalizada que contribuye a comprender prácticas, interacciones y significados en contextos digitales concretos.

La utilización de una muestra *ad hoc* en investigaciones desarrolladas en entornos digitales suele generar diversas objeciones. Antes de optar por este tipo de selección, resulta pertinente considerar ciertos elementos estructurales que orientan su diseño. A partir de una reflexión propia, surgida de la revisión de la literatura en muestreos no probabilísticos se proponen tres ejes que permiten valorar su pertinencia y alcance en contextos sociodigitales:

- El primero tiene que ver con el volumen y dinamismo de los datos en el contexto actual, donde se genera una gran cantidad de datos en un corto período de tiempo sobre temas específicos. Este problema se intensifica debido a la naturaleza dinámica y abierta de las plataformas sociodigitales, en las que la información fluye constantemente y se genera en una amplia variedad de formatos, como texto, imágenes, videos y enlaces. El proceso de delimitación y construcción de un corpus de análisis es esencial para garantizar la relevancia y la coherencia de los resultados obtenidos. Lo anterior se traduce en un planteamiento concreto: es imposible dar cuenta de todos los datos que se generan. Transparentar este criterio es fundamental, es decir, señalar cuáles son los criterios de inclusión y exclusión utilizados, así como los métodos para recolectar, clasificar y organizar los datos.
- El segundo elemento por considerar es el del criterio de representatividad. ¿Cuántos posts, fotografías, enlaces, notas, comentarios, likes o shares son suficientes para ser considerados como una muestra representativa?, ¿debe considerarse el total de usuarios de la plataforma?, ¿el total de usuarios de una plataforma en una región en específico o dónde aparezca el fenómeno de interés para la investigación?, ¿debe considerarse el total de los usuarios que utilizan un *hashtag*, por ejemplo? Las preguntas anteriores obligan a repensar el criterio de representatividad, o por lo menos adaptarlo a las diferentes investigaciones, pues su aplicación de manera tajante puede derivar en que no puede construirse una muestra realmente representativa en el diseño del corpus. Una posible solución a este desafío podría ser reemplazar el concepto de representatividad estadística por el de muestra teórica. La muestra teórica, a menudo asociada con la teoría fundamentada, implica delimitar la muestra según el criterio de saturación, donde agregar más elementos no produce datos distintos (Ardila Suáez & Rueda Arenas, 2013). Esto permite flexibilidad para ampliar la muestra y recolectar datos en inmersiones posteriores al campo, descubriendo así nuevos datos de manera constante.

- El sesgo algorítmico es un fenómeno que se refiere a la tendencia de los algoritmos de las plataformas en línea a influir en las decisiones y experiencias de los usuarios, lo que puede llevar a la amplificación o perpetuación de sesgos existentes en la sociedad. Este fenómeno es especialmente relevante en el contexto de las redes sociales y otras plataformas digitales, donde los algoritmos determinan qué contenido se muestra a los usuarios, en qué orden y con qué frecuencia. Lo anterior resta aleatoriedad en la muestra y en la construcción del corpus, volviendo al punto anterior y la discusión sobre si es posible hablar o no de representatividad.

Los criterios anteriores representan apenas el punto de partida de consideración metodológica para la definición de criterios de delimitación, es decir, antes de diseñar los parámetros en la recopilación de datos se deben contemplar estas tres dimensiones y hacerlas explícitas en el proceso de investigación.

Al proporcionar detalles claros y exhaustivos sobre los procedimientos utilizados en la recopilación, el análisis y la interpretación de los datos, se tiene como resultado que otros trabajos de investigación puedan entender, evaluar y replicar el estudio de manera adecuada. La transparencia metodológica también contribuye a la confianza en la investigación.

¿Qué dice la literatura?

La literatura especializada ofrece distintas pistas en torno a la construcción de corpus de análisis a partir de datos extraídos de las distintas plataformas sociodigitales. Estas alternativas metodológicas son de diversa índole y se construyeron en función y necesidades de cada investigación; considerando esto, el primer punto que se puede señalar es el de la heterogeneidad de criterios (*Tabla 1*).

De acuerdo con Pattier (2024), aunque cada investigación ha desarrollado un planteamiento propio, existen tres grandes tendencias para la delimitación del corpus de información construido a partir de plataformas sociodigitales. El primero se relaciona con la lógica del estudio de caso, ampliamente utilizada en la literatura, en buena medida por la facilidad de acceder a uno o varios perfiles o canales digitales, así como por la posibilidad de seleccionar el caso con base en criterios de conveniencia. Aunque aporta datos de valor, es importante señalar la dificultad de generalización de resultados inherente a este planteamiento metodológico.

Sobre esta primera "salida" se puede rescatar que el estudio de caso supone un acceso relativamente fácil a los datos específicos, pues se trabaja con perfiles, canales o temas que suponen cierta cercanía. De la misma manera, otra ventaja de los estudios de caso radica en la contextualización a profundidad que puede hacerse, pues al centrarse en uno o en un par de casos específicos se pueden explorar con detalle las condiciones que producen el fenómeno de interés. En el extremo opuesto la

oposición más clara sería, sin duda, la imposibilidad de la generalización de resultados, dado que el estudio de caso supone un conjunto limitado de datos (lo cual se puede solventar si la investigación no busca representatividad). Por otro lado, el sesgo de selección supone otra objeción, puesto que la elección del corpus de análisis puede estar influenciada por criterios como la conveniencia o la disponibilidad de datos. Si bien se ha discutido bastante sobre la imposibilidad de la objetividad en la ciencia y particularmente en las ciencias sociales, un corpus construido *ad hoc* corre el riesgo de una interpretación subjetiva al extremo.

La segunda salida señalada por Pattier es la utilización de la muestra de otras investigaciones, lo que conlleva una facilidad del establecimiento de la misma debido a que ya está disponible en las publicaciones de referencia. Sin embargo, esta opción hace depender el rigor metodológico del proceso de obtención de la muestra y de los criterios de delimitación de otras investigaciones. A lo anterior también hay que añadir que la utilización de criterios metodológicos ajenos a la propia investigación puede ir en detrimento de ésta o conducir la construcción de los corpus de análisis en distintas direcciones. Además, la utilización de criterios metodológicos ajenos y, por consiguiente, la heterogeneidad de las muestras dificulta, sin duda, la comparación con otros trabajos de corte similar.

Finalmente, la tercera salida consiste en la delimitación de una muestra de tamaño reducido, entendida no en términos cuantitativos estrictos, sino como un conjunto acotado de casos que permita realizar un análisis manejable y coherente con los objetivos del estudio. La imposibilidad de capturar toda la información en los entornos sociodigitales junto con la inexistencia de información poco fiable y de las complicaciones técnicas[i], han impulsado esta posibilidad metodológica de obtener una pequeña muestra de una manera sencilla y por conveniencia. Esto presenta la ventaja de facilitar el acceso a participantes y datos de manera rápida, lo que puede ser especialmente útil en investigaciones con recursos limitados o plazos ajustados. Además, este enfoque puede ser adecuado para explorar fenómenos emergentes o poco estudiados, permitiendo una mayor flexibilidad en la selección de participantes. Sin embargo, este método conlleva el riesgo, como ya se ha señalado, de introducir sesgos de selección, ya que los participantes pueden no ser representativos de la población objetivo, lo que limita la generalización de los resultados o la posibilidad de justificar de manera clara los criterios de selección.

Big Data y extracción de datos

Desde la perspectiva del *Big Data*[ii], Hernández-Leal et al. (2017) señalan la necesidad de incluir los elementos del protocolo "HACE" (*Heterogeneous, Autonomous, Complex* y *Evolving*) independientemente del número de datos que se pretenda analizar. La heterogeneidad implica una variedad de representaciones para los mismos individuos, mientras que la diversidad de características se refiere a la variabilidad en la representación de cada observación específica. Las aplicaciones del *Big Data* se caracterizan

principalmente por fuentes de datos autónomas con control distribuido y descentralizado. Esto significa que cada fuente de datos tiene la capacidad de recopilar información de forma independiente y sin la necesidad de un ente de control centralizado.

En el contexto de la construcción de un corpus de análisis de redes sociales, particularmente en el caso de X, la heterogeneidad se refiere a la diversidad de representaciones de los usuarios y sus interacciones en la plataforma. Esto incluye una variedad de perfiles de usuarios, tipos de contenido compartido (*posts*, *repost*, respuestas y *likes*), así como la diversidad de temas y opiniones expresadas.

Por otro lado, la diversidad de características se relaciona con la variabilidad en la forma en que cada observación (por ejemplo, un *post* o una interacción entre usuarios) se representa en el corpus. Esto puede incluir diferencias en el formato del texto, la presencia de imágenes o enlaces adjuntos, así como metadatos asociados como la ubicación geográfica o el tiempo de publicación.

Siguiendo con las posibles aplicaciones del *Big Data*, la construcción de un corpus de análisis de redes sociales como X se ve influenciada por la descentralización y el control distribuido de las fuentes de datos. En X, cada usuario es esencialmente una fuente de datos autónoma que genera contenido de forma independiente. Esto significa que cada usuario contribuye con su propio conjunto único de datos, lo que resulta en un corpus heterogéneo y diverso que refleja la amplia gama de perspectivas y experiencias presentes en la plataforma, y que se traduce a su vez en una diversidad de opiniones y posturas respecto a determinado tema.

El hecho de que X opere con un control distribuido y descentralizado implica que no hay un actor central que controle toda la información generada en la plataforma. Esto facilita la recopilación de grandes volúmenes de datos para la construcción de un corpus de análisis de redes sociales, permitiendo la realización de investigaciones sobre temas como la opinión pública, la difusión de información y la interacción social en línea.

Momentos mediáticos y comunidades

Taraborrelli & Sokil (2021) analizan los discursos en el marco de la apertura de sesiones legislativas del Congreso Argentino en 2020 en Twitter[iii], para ello diseñaron diversas etapas en la construcción del corpus.

Inicialmente se estableció un marco temporal, desde el 1 de marzo de 2020 a las 9:00 a.m. hasta el 2 de marzo de 2020 a las 23:59 p.m. Además, se llevó a cabo una búsqueda exhaustiva de palabras clave relevantes relacionadas con este evento en los principales medios de comunicación nacionales y redes sociales. Estas palabras clave fueron utilizadas como filtro para acceder a los servicios de Twitter "Stream API" y "API Search", permitiendo la recopilación de una muestra aleatoria de enunciados que cumplían con los criterios de búsqueda establecidos.

Posteriormente, se construyó un corpus de trabajo a partir de la recopilación de enunciados seleccionados durante el periodo de tiempo definido y mediante el uso de las palabras clave como filtro.

Este corpus sirvió como base de datos para el análisis de redes sociales y la identificación de comunidades en Twitter. El proceso involucró una cuidadosa selección de datos, un filtrado preciso y la construcción de un corpus robusto que permitió el análisis detallado de las interacciones y discursos relacionados con el evento de interés político.

Las ventajas de este enfoque incluyen la precisión temporal y contextual en la recopilación de datos, lo que permitió una comprensión más profunda de las conversaciones en torno al evento específico y su impacto en las redes sociales. Además, el uso de palabras clave relevantes fue útil para garantizar que la muestra recopilada se considerara relevante y representativa del tema de interés, lo que facilita un análisis más enfocado y significativo de las interacciones en Twitter. Por otro lado, entre las desventajas, se puede mencionar la limitación inherente al uso de palabras clave, que puede perder ciertos matices o contextos relevantes en las conversaciones, y la dependencia de los servicios de Twitter API[iv], que pueden estar sujetos a cambios en sus políticas o restricciones de acceso que afecten la recopilación de datos.

Adicionalmente, el trabajo de Taraborrelli & Sokil (2021) evidencia que las comunidades identificadas en el estudio presentan una fuerte concentración de la capacidad de enunciación. Se destaca que, en las tres comunidades analizadas, pocos usuarios producen el contenido mientras que muchos lo ponen en circulación. Además, se menciona que las comunidades difieren en tamaño y en la clase de usuarios que las componen, pero comparten la característica de tener una distribución concentrada del poder enunciativo.

Las comunidades juegan un papel fundamental en la construcción de un corpus de análisis en redes sociales como X por varias razones. En primer lugar, permiten la identificación de grupos de usuarios que comparten intereses o temas de conversación específicos, lo que facilita la selección precisa de contenido para el corpus. Además, al enfocarse en las comunidades con una alta interacción entre sus miembros, se prioriza el análisis de interacciones relevantes, proporcionando una visión detallada de las dinámicas de la red. Analizar estas comunidades también proporciona una comprensión más profunda de la estructura de la red social, lo que orienta la construcción de un corpus más representativo. Finalmente, al considerar las comunidades en la selección de contenido, se asegura la inclusión de una variedad de perspectivas y temas en el corpus, lo que contribuye a un análisis más completo y diverso de la plataforma.

Métodos mixtos

Bonilla (2022) define el corpus de análisis como un conjunto de datos recolectados y organizados de manera sistemática para su posterior análisis cualitativo. En el contexto del análisis de discursos digitales en X, el corpus se refiere a la recopilación de "*posts* relevantes" que han sido filtrados, categorizados y etiquetados para facilitar su estudio y comprensión en profundidad. Para el desarrollo de un análisis centrado en algunos discursos de esta plataforma sociodigital, la autora propone la

construcción de un modelo mixto de lo que denomina como un "corpus manejable" (Bonilla, 2022, p. 4), a partir de la recolección de datos, la visualización y filtrado de la información, así como la categorización y etiquetado de los enunciados verbales e icónicos presentes en los posteos.

Para construir un corpus manejable para un análisis cualitativo de los discursos digitales en X, se pueden seguir los siguientes pasos:

- Registro de la etnografía virtualv: Inicialmente, se debe realizar un registro detallado de las observaciones y registros obtenidos a través de la etnografía virtual, identificando las cuentas más activas sobre los temas de interés utilizando hashtags relevantes.
- Recolección de datos: Utilizar la API de X junto con herramientas como Python para recolectar los datos de interés.
- Visualización y filtrado de datos: Emplear herramientas para visualizar y filtrar los datos recolectados, facilitando la identificación de patrones y la organización de la información de manera más clara.
- Construcción del corpus: Una vez filtrados los datos, se procede a la construcción del corpus, organizando los posts de manera coherente y estructurada para facilitar su posterior análisis cualitativo.

La importancia de combinar recursos cualitativos y cuantitativos en el análisis de datos en X radica en la posibilidad de obtener una comprensión más completa y profunda de los discursos digitales presentes en esta plataforma. Al integrar ambos enfoques, se pueden aprovechar las fortalezas de cada uno para enriquecer el análisis y obtener resultados más significativos.

Los recursos cualitativos permiten explorar en detalle el contenido de los *posts*, identificar patrones, interpretar significados y contextos, y comprender las motivaciones que hay detrás de los discursos. Por otro lado, los recursos cuantitativos brindan la oportunidad de analizar grandes volúmenes de datos, identificar tendencias, realizar análisis estadísticos y cuantificar la presencia de ciertos temas o palabras clave en la plataforma.

Al combinar los distintos enfoques se puede realizar un análisis completo y holístico de los discursos y prácticas en X permitiendo una comprensión más profunda de las interacciones, opiniones y dinámicas presentes en esta plataforma. Esta integración de recursos cualitativos y cuantitativos enriquece la investigación y contribuye a una interpretación más sólida de los datos recolectados en X.

Tabla 1.

Comparativa de los criterios y elementos utilizados en la construcción de un corpus de análisis a partir de datos de X

Investigación	Ventajas	Desventajas
Pattier (2024) Plantea tres "salidas": 1) un estudio de caso, es decir centrarse en un perfil, en una etiqueta, en un vídeo, *post*, etc. con criterios establecidos previamente, 2) utilización de criterios de otras investigaciones; y 3) delimitación de una muestra pequeña.	- Cercanía y conocimiento de los datos; - Contextualización y profundidad en los datos; - Facilidad en el establecimiento de los criterios del corpus de análisis; - Facilidad en el acceso a datos; y - Criterios útiles en el análisis de fenómenos emergentes o poco analizados.	- Imposibilidad de generalizar resultados; - Sesgos de selección; - Dependencia de los criterios de otras investigaciones; y - Dificultad en la adaptabilidad de los criterios de otras investigaciones.
Hernández-Leal et al. (2017) Aplicación del protocolo HACE en la construcción de corpus desde el Big Data.	- Gran volumen de datos que se convierte en un gran conjunto de datos; y - Heterogeneidad en la recolección de datos.	- Uso de softwares especializados para la construcción de corpus.
Taraborrelli & Sokil (2021) Identifican momentos claves en la conformación de discursos en twitter para agruparlos posteriormente en comunidades discursivas.	- Análisis en profundidad, particularmente de la dinámica de la red en la agrupación en comunidades; y - Precisión contextual y temporal.	- Dependencia de los servicios de Twitter API; y - Requerimientos técnicos, pueden ser necesarios conocimientos previos de extracción de datos y programación.
Bonilla (2022) Combina elementos cuantitativos con la extracción de datos mediante APPI de Twitter, y cualitativos como la etnografía digital.	- Mayor profundidad, contextualización y comprensión de los datos al combinar enfoques.	- La combinación de datos puede dificultar su integración; y - Requerimientos técnicos, pueden ser necesarios conocimientos previos de extracción de datos y programación.

La investigación en cuestión: Criterios descartados y transparencia metodológica

El estudio de la comunicación política en entornos digitales requiere abordar no sólo la circulación de información, sino también la manera en que se configuran subjetividades políticas colectivas a través de la interacción en plataformas sociodigitales. Las reflexiones previas sobre los desafíos de análisis de datos masivos, la delimitación de corpus y la comprensión de los procesos de participación digital proporcionaron un marco conceptual y metodológico que orientaron la investigación. Estas perspectivas permitieron situar el análisis en un contexto más amplio de transformación de la comunicación política en la era digital, donde los discursos y prácticas en línea son tanto objeto de estudio como herramientas para comprender la movilización social y la construcción de sentido.

En el marco de la investigación referida, el caso Ayotzinapa se abordó como un ejemplo paradigmático de interacción entre movilización presencial y comunicación digital. La plataforma X se conviertió en un espacio privilegiado para observar la producción de narrativas de protesta, denuncia y solidaridad, así como para identificar las tensiones y polarizaciones que emergen en torno a los eventos de relevancia política y social. Este enfoque permitió explorar cómo se configuran y negocian subjetividades políticas en entornos digitales, atendiendo tanto a los contenidos compartidos como a las formas de interacción entre diversos actores.

Con base en estos planteamientos, este apartado describe cómo se construyó y delimitó el corpus de análisis. Se explican los criterios de selección de *hashtags*, temporalidad de los acontecimientos, representatividad de los datos y estrategias para mitigar posibles sesgos algorítmicos. La exposición metodológica no sólo clarifica el procedimiento seguido, sino que también establece el vínculo con el resto del artículo, mostrando cómo estas decisiones permiten abordar de manera sistemática el análisis de subjetividades políticas digitales en el caso Ayotzinapa.

Volumen y velocidad de datos, y la imposibilidad de un corpus representativo

La obtención de todos los *posts* de un *hashtag* para la construcción de un corpus de análisis se enfrenta a desafíos técnicos y metodológicos significativos. Las limitaciones técnicas de las API de las plataformas de redes sociales, en especial X, que imponen restricciones en la cantidad de datos que se pueden extraer en un período de tiempo determinado, hacen que sea difícil recopilar una muestra completa y actualizada. Además, el flujo constante de nuevos contenidos en tiempo real y la posibilidad de que algunos *posts* estén disponibles sólo para usuarios autorizados o sean eliminados por los propios usuarios plantean obstáculos adicionales para obtener una muestra representativa y confiable.

Esta imposibilidad de obtener todos los *posts* asociados a un *hashtag* resalta la naturaleza dinámica

y efímera de las redes sociales, donde la disponibilidad y la calidad de los datos pueden variar considerablemente. Reconocer estas limitaciones es crucial al diseñar estudios de análisis de redes sociales y en la construcción de corpus de análisis. Asimismo, esta imposibilidad también implica la dificultad para construir un corpus representativo estadísticamente. Dado que la muestra de *posts* recopilada podría no ser exhaustiva y estar sujeta a sesgos como la disponibilidad de datos y la actividad de los usuarios, existe el riesgo de que los resultados del análisis no sean generalizables o representativos de la población de *posts* en su totalidad. En el marco de esta investigación, lo anterior resulta de suma importancia por la complejidad en la obtención de datos de una etiqueta que tiene en circulación más una década, como lo es #TodosSomosAyotzinapa. Esta limitación puede comprometer la validez y la fiabilidad de los hallazgos obtenidos a partir del corpus, ya que la falta de representatividad podría llevar a conclusiones erróneas o incompletas sobre las tendencias observadas en el conjunto completo de *posts* relacionados con el *hashtag* en cuestión. Por lo tanto, la dimensión de la representatividad estadística no es una opción viable para dicha investigación. Además, cada año, conforme se acerca el día 26 de septiembre se generan nuevos datos en relación al *hashtag,* lo que implicaría una constante actualización del diseño muestral en términos estadísticos.

Sesgos de selección

Los sesgos de selección en un corpus de análisis pueden justificarse mediante una comprensión integral de las limitaciones técnicas y operativas involucradas en la recopilación de datos digitales. Estas limitaciones pueden incluir restricciones impuestas por las API de las plataformas de redes sociales, como límites en la cantidad de datos que se pueden extraer en un período de tiempo determinado, así como decisiones prácticas relacionadas con los recursos y el tiempo disponibles para llevar a cabo la investigación. Es crucial reconocer que estas limitaciones resultan en sesgos inevitables en la construcción del corpus de análisis.

Además, los sesgos de selección pueden surgir también debido a las características intrínsecas de los datos a analizar. Por ejemplo, en esta investigación se asumió una especie de subrepresentación de algunos grupos sociales debido a diferencias en la participación en X. En función de lo anterior y con fines de garantizar la transparencia y la precisión en la presentación de los resultados se asumió dicho sesgo.

Por lo anterior, proporcionar una descripción detallada de los métodos utilizados para recopilar y seleccionar los datos, incluidos los criterios de inclusión y exclusión, las fuentes de datos y cualquier sesgo en potencia garantizará la transparencia metodológica. Por lo anterior, a continuación se presentan los criterios para la construcción del corpus de análisis de la investigación en cuestión.

Criterios de selección en la construcción del corpus de análisis

Siguiendo la propuesta metodológica de Bonilla (2022), se adoptó una estrategia mixta que combinó la etnografía digital con la extracción sistemática de datos de la plataforma X para la construcción del corpus. Este enfoque partió de la premisa de que el análisis de subjetividades tecnopolíticas requiere tanto la observación cercana de las dinámicas de interacción digital como la sistematización de grandes volúmenes de información que permitan identificar patrones discursivos y de comportamiento. La etnografía digital, concebida como una práctica de permanencia y participación en los entornos digitales (Bárcenas Barajas & Preza Carreño, 2019), otorgó la posibilidad de comprender desde adentro la forma en que los usuarios se involucran con el caso Ayotzinapa, los repertorios de acción que despliegan y los sentidos que dotan a su interacción en torno al *hashtag* #TodosSomosAyotzinapa. Al mismo tiempo, el uso de herramientas de extracción y análisis de datos contribuyó a ampliar la mirada, generando un corpus masivo que hace posible observar no sólo las expresiones particulares, sino también la estructura relacional que emerge de la interacción entre múltiples nodos.

La combinación de estas dos aproximaciones permitió alcanzar una doble perspectiva metodológica: por un lado, una mirada cualitativa y situada, centrada en la experiencia de los usuarios y en las narrativas que producen y, por otro, una aproximación cuantitativa y estructural, capaz de mapear la intensidad de la actividad, los momentos de mayor concentración de interacciones y la relevancia de determinados actores en la circulación del discurso. El corpus híbrido, de este modo, se convirtió en una herramienta que no sólo recoge lo que se dice, sino también cómo se dice, quiénes lo dicen y en qué contextos de interacción se producen esas enunciaciones. Este diseño metodológico respondió a la necesidad de articular las escalas micro y macro del fenómeno digital, evitando quedarse únicamente en el nivel anecdótico o, por el contrario, en la abstracción estadística.

Un criterio adicional y central en la delimitación del corpus fue la selección de "episodios relevantes", idea retomada a partir de los planteamientos de Taraborrelli & Sokil (2021). Estos autores destacan la importancia de identificar momentos que condensan la intensidad de la interacción y que funcionan como puntos de inflexión en la conversación digital. En el caso del estudio del Ayotzinapa, los episodios relevantes se entendieron como coyunturas críticas que movilizan la memoria colectiva, intensifican la participación y generan nuevos sentidos en torno al caso. La inclusión de estos episodios en la construcción del corpus permitió analizar cómo los discursos no sólo se mantienen, sino que también se transforman en momentos de crisis, conmemoración o confrontación política.

Paralelamente, y con el propósito de atender la dimensión temporal y de continuidad de las prácticas de memoria digital, se retomó el planteamiento de Rovira-Sancho & Morales-i-Grass (2023) sobre el ecosistema de *hashtags*. En este

sentido, se llevó a cabo un proceso de recolección y sistematización de publicaciones realizadas cada 26 de mes, abarcando el periodo comprendido entre el 26 de septiembre de 2023 y el 26 de septiembre de 2024. Esta estrategia longitudinal permitió registrar cómo la memoria del caso Ayotzinapa se mantiene activa a través de prácticas de conmemoración mensual que se articulan en torno a *hashtags* y consignas compartidas. Dichas prácticas funcionan como recordatorios colectivos y como mecanismos de visibilización constante, otorgando continuidad al reclamo de justicia más allá de los picos coyunturales.

Los episodios relevantes considerados hasta el momento incluyen:

- Desaparición de los estudiantes (26 y 27 de septiembre de 2014);
- Primeras investigaciones e informes sobre el caso (septiembre de 2014);
- Trabajos del Equipo Argentino de Antropología Forense (septiembre-octubre de 2014);
- Construcción de la "Verdad histórica" (noviembre de 2014);
- Trabajos del Grupos Interdisciplinario de Expertos Independientes (GIEI) (enero-marzo de 2015);
- Sentencia de la comisión de la verdad por el caso Ayotzinapa (mayo 2018);
- Toma del poder legislativo por parte de Andrés Manuel López Obrador (julio-diciembre de 2018);
- Creación de la Comisión para la Verdad y Acceso a la Justicia (diciembre de 2018)
- Designación de Omar Gómez Trejo como fiscal para el caso Ayotzinapa (junio de 2019);
- Informe del GIEI y primeros enfrentamientos entre AMLO y padres de los normalistas desaparecidos (junio de 2023);
- Asesinato del normalista Yanqui Kothan (marzo de 2024); y
- Diez años del caso Ayotzinapa (septiembre de 2024).

La integración de estos momentos al corpus, en conjunto con la sistematización mensual de publicaciones, permitió construir un mapa de intensidades que mostró cómo se configuran las subjetividades políticas en el entramado digital. Así, el corpus no se limitó a ser un repositorio de datos, sino que se constituyó como un espacio analítico donde convergen las dinámicas de lo coyuntural y lo rutinario, lo efímero y lo permanente, lo individual y lo colectivo. De esta manera, el diseño metodológico se orientó a capturar la complejidad de las interacciones digitales en torno a Ayotzinapa, atendiendo a las dimensiones discursivas, relacionales y temporales que atraviesan la construcción de subjetividades tecnopolíticas.

Finalmente, también se incorporó como criterio en la construcción del corpus la heterogeneidad de los datos (Hernández-Leal et al., 2017), tratando de contrastar las diferentes opiniones y posturas en X respecto a cada episodio del caso Ayotzinapa.

Considerando lo anterior, se recuperó el planteamiento de Pattier (2024) en torno a las "salidas" para la construcción del corpus. En primer

lugar, se trató del seguimiento de un *hashtag* específico (#TodosSomosAyotzinapa) bajo criterios particulares, como la heterogeneidad y su ubicación en ciertos contextos (episodios relevantes), consiguiendo con esto una muestra *ad hoc* que en su análisis permitiría identificar los nodos para dar paso al siguiente momento metodológico (entrevistas y reconstrucción de historias de vida), y, así, dar cuenta de los elementos constitutivos de las subjetividades políticas de los participantes del *hashtag* # TodosSomosAyotzinapa.

La intención de utilizar *hashtags* como puerta de entrada surgió de los planteamientos de Zires (2014) en torno a considerarlos como "espacios virtuales de articulación" entre distintos usuarios y participantes de las movilizaciones en red. En este sentido, estos dispositivos digitales constituyen un escenario inédito de interacción pública cuya función inicial es la de visibilizar causas, sujetos y luchas. En el mismo sentido, Rovira-Sancho & Morales-i-Grass (2023) también entienden a los *hashtags* como "identificadores meta-discursivos" performativos que difunden y conectan marcos de protesta, potenciando con ello la voz de actores marginados o, que de otra manera, no entrarían en el debate de lo público.

Los *hashtags* resultan de vital importancia para la movilización en red puesto que facilitan la organización y la difusión de información, crean visibilidad, fomentan la articulación y la participación, amplifican el impacto de la protesta y construyen una narrativa colectiva.

Una vez identificados los nodos centrales en la red, se dió paso al análisis del hábitat de las densidades, entendido como la posibilidad de reconocer múltiples capas y niveles de interpretación (Rodríguez Cano, 2022) en torno a las prácticas y relaciones que se desarrollan en los entornos sociodigitales. Este planteamiento permitió no sólo observar la estructura relacional que emerge en el ecosistema de *hashtags*, sino también atender a la pluralidad de sentidos que circulan en torno al caso Ayotzinapa. En este punto, la etnografía digital se ha revelado como una herramienta fundamental para captar el amplio marco de discursos y prácticas que atraviesan la interacción en Internet, permitiendo dar cuenta tanto de la dimensión simbólica como de las dinámicas de acción colectiva que se articulan en la esfera digital.

El segundo momento metodológico de la investigación correspondió precisamente a la consolidación de este enfoque etnográfico, ahora ampliado hacia una estrategia que combinó lo digital con lo presencial. Este tránsito permitió conectar los repertorios de acción observados en línea con las experiencias situadas de quienes participaron en ellos. Así, junto con la etnografía digital, se incorporaron entrevistas en profundidad y reconstrucciones de historias de vida, con el propósito de explorar cómo se configuran las subjetividades políticas en la intersección entre los entornos digitales y los contextos presenciales. Este cruce de escenarios metodológicos dialoga con la propuesta de la activación digital (Bustamante-Farías, 2014), en tanto que subraya la importancia de entender cómo las prácticas en línea se articulan con dinámicas de movilización y producción de sentidos en el espacio social más amplio.

En particular, las entrevistas constituyeron una herramienta importante en este segundo momento, pues perimitieron acceder a narrativas personales que enriquecen y complejizan el análisis. A través de ellas fue posible comprender cómo los participantes interpretan su involucramiento, cuáles son sus motivaciones y cómo las experiencias de interacción digital se entrelazan con trayectorias biográficas y con procesos de subjetivación política más amplios. Este enfoque no sólo aportó densidad contextual, sino que abrió un espacio de reflexión en el que las voces de los actores adquieren centralidad en la explicación del fenómeno, otorgando al estudio una perspectiva más humana y situada sobre las formas en que la memoria, la protesta y la política se configuran en los entornos sociodigitales.

Al centrarse en los individuos como agentes activos en la producción y reinterpretación de discursos políticos, este enfoque buscó desentrañar los procesos que subyacen en la formación de subjetividades políticas. Las entrevistas proporcionaron un espacio para explorar las conexiones entre la participación en dinámicas de movilización en red y la construcción de subjetividades políticas, abriendo la puerta a una comprensión más holística y detallada de este complejo entramado.

Esta estrategia metodológica no sólo perimitó capturar la interacción dinámica entre los individuos y el entorno comunicativo en constante evolución, sino que también resaltó la mutua influencia entre las dinámicas políticas y las subjetividades individuales. A través de las voces y experiencias directas de los participantes, se buscó arrojar luz sobre cómo las prácticas políticas contemporáneas, especialmente aquellas vinculadas a la movilización en red, contribuyen a la construcción y transformación de las subjetividades políticas de los individuos. En este sentido, el enfoque adoptado no se limitó a describir interacciones en plataformas digitales, sino que buscó comprender los procesos de significación, apropiación y resignificación que los sujetos realizan en el cruce entre lo digital y lo presencial.

Reflexiones finales sobre el alcance metodológico y analítico

El análisis etnográfico y narrativo de las experiencias recogidas permitió reconocer cómo las subjetividades políticas se constituyen en diálogo con repertorios de acción colectivos, con marcos de memoria compartidos y con las posibilidades técnicas y comunicativas que ofrecen los entornos digitales. Al mismo tiempo, la integración de entrevistas e historias de vida aportó una densidad interpretativa que permitió comprender cómo lo político se inserta en trayectorias biográficas, en emociones y en proyectos personales, evidenciando que la acción en red no es un fenómeno aislado, sino parte de un entramado más amplio de relaciones sociales y culturales.

De este modo, el estudio metodológicamente híbrido que combina observación etnográfica digital, extracción de datos, entrevistas y reconstrucción biográfica aportó no sólo a la comprensión del caso

Ayotzinapa en el ámbito sociodigital, sino también a un marco más general para pensar cómo las subjetividades políticas se configuran en el contexto de sociedades atravesadas por la tecnología y la comunicación en red. De esta manera, la investigación no sólo pretende aportar al entendimiento de la constitución de subjetividades políticas en la era digital, sino también ofrecer criterios metodológicos para la construcción de corpus de análisis en plataformas sociodigitales, donde la velocidad, el volumen y la diversidad de los datos puede resultar abrumadora. Se busca contribuir a la definición de estrategias que permitan seleccionar, organizar y sistematizar información relevante, integrando criterios de representatividad, episodios significativos y patrones de interacción, de manera que los estudios sobre fenómenos políticos en entornos digitales puedan abordar la complejidad de los mismos entornos sin perder consistencia analítica ni rigor interpretativo.

Referencias bibliográficas

Ardila Suárez, E., & Rueda Arenas, J. (2013). La saturación teórica en la teoría fundamentada: su delimitación en el análisis de trayectorias de vida de víctimas del desplazamiento forzado en Colombia. *Revista Colombiana de Sociología, 36*(2), 93-114. https://www.redalyc.org/pdf/5515/551556228007.pdf

Bárcenas Barajas, K. y Preza Carreño, N. (2019). Desafíos de la etnografía digital en el trabajo de campo onlife. *Virtualis. Revista de cultura digital, 10*(18), 134-15. https://doi.org/10.2123/virtualis.v10i18.287

Bonilla, L. (2022). Claves para analizar datos en Twitter. Recolección y procesamiento de corpus. *Cuadernos de Lingüística Hispánica*, (39), 1-21. https://doi.org/10.19053/0121053X.n39.2022.14283

Bustamante-Farías, Ó. D. (2014). *Mediatización de la protesta: la activación digital como modalidad de comunicación política. Viaje al centro del movimiento estudiantil 2011 en Chile* [Tesis doctoral]. Instituto Tecnológico y de Estudios Superiores de Occidente. http://hdl.handle.net/11117/1270

de la Garza Toledo, E. (2012). Grounded theory. Cantidad, Calidad y comprensión de significados. In E. de la Garza Toledo, & G. Leyva (Eds.), *Tratado de metodología de las ciencias sociales: perspectivas actuales* (pp. 397-419). Fondo de Cultura Económica, Universidad Autónoma Metropolitana.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación*. McGraw-Hill.

Hernández-Leal, N., Duque-Méndez, N., & Moreno-Cadavid, J. (2017). Big Data: una exploración de investigaciones, tecnologías y casos de aplicación. *TecnoLógicas, 20*(39). https://doi.org/10.22430/22565337.685

Otzen, T., & Manterola, C. (2017). Técnicas de muestreo sobre una población a estudio. *International Journal of Morphology, 35*(1), 227-232. https://intjmorphol.com/es/resumen/?art_id=4049

Pattier, D. (2024). ¿Callejón sin salida?: El problema muestral en el estudio de las redes sociales digitales. *Comunicación y Sociedad, 21*, e8580, 1-23. https://doi.org/10.32870/cys.v2024.8580

Rodríguez Cano, C. A. (2022). *Hipermétodos. Repertorios de la investigación social en entornos digitales*. Universidad Autónoma Metropolitana, Unidad Cuajimalpa. http://dccd.cua.uam.mx/libros/investigacion/HIPERMETODOS%202022%20Digital.pdf

Rovira-Sancho, G., & Morales-i-Gras, J. (2023). Femitags in the networks and in the streets: 50 hashtags for feminist activism in Latin America. *Profesional de la Información, 32*(3), e320319. https://doi.org/10.3145/epi.2023.may.19

Taraborrelli, D., & Sokil, J. P. (2021). *Twitter: la construcción de un corpus de trabajo* [Ponencia]. XIV Jornadas de Sociología, Facultad de Ciencias Sociales, Universidad de Buenos Aires. https://cdsa.aacademica.org/000-074/59

Zires, M. (2014). Violencia, redes sociales y procesos de subjetivación política. El caso de #verfollow en Veracruz, México. Argumentos. Estudios Críticos de la Sociedad, (75), 119-146. https://argumentos.xoc.uam.mx/index.php/argumentos/article/view/165

Notas

[i] Por ejemplo la API de Twitter se cerró en abril de 2023, impidiendo con ello que softwares especializados, como MAXQDA, puedan rastrear la información de la plataforma y limitando en gran medida el acceso a datos y su posterior análisis.

[ii] El Big Data se asocia comúnmente con los grandes volúmenes de datos, sin embargo su uso puede aplicarse también a datos provenientes de diferentes fuentes o datos que se generan con rapidez independientemente de las fuentes o de su volumen (Hernández-Leal et al., 2017).

[iii] Se conserva el nombre Twitter porque en el marco del desarrollo de la investigación la plataforma aún tenía ese nombre.

[iv] Que ya no existe de manera gratuita

[v] La autora utiliza específicamente el término etnografía virtual; para el caso de esta investigación se usa el término etnografía digital.