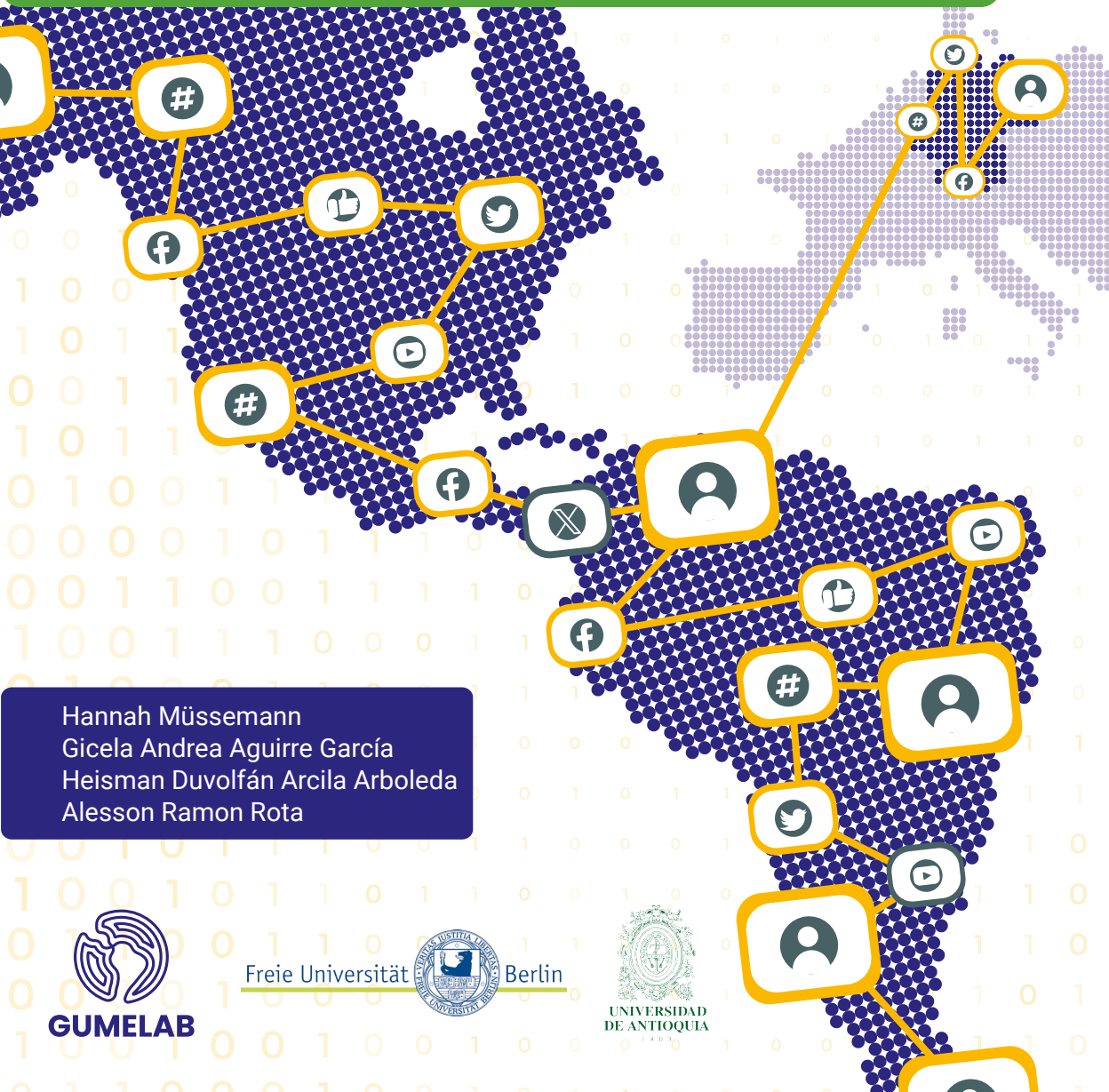
Manual para el uso de métodos digitales en proyectos de humanidades



Hannah Müssemann
Gicela Andrea Aguirre García
Heisman Duvolfán Arcila Arboleda
Alesson Ramon Rota

GUMELAB
Freie Universität Berlin
UNIVERSIDAD DE ANTIOQUIA

GUMELAB

Manual para el uso de métodos digitales en proyectos de humanidades

La experiencia del proyecto GUMELAB

Geschichtsvermittlung durch Unterhaltungsmedien in Lateinamerika. Labor für Erinnerungsforschung und digitale Methoden

(Transmisión de la historia a través de los medios de entretenimiento en América Latina. Laboratorio de Investigación de la Memoria y Métodos Digitales).

Hannah Müssemann
Gicela Andrea Aguirre García
Heisman Duvolfán Arcila Arboleda
Alesson Ramon Rota

Manual para el uso de métodos digitales en proyectos de humanidades. La experiencia del proyecto GUMELAB
ISBN: 978-3-00-078652-5
Primera edición, marzo de 2024

Stefan Rinke, *director del proyecto de investigación GUMELAB. Freie Universität Berlin*
Mónika Contreras Saiz, *directora y coordinadora del proyecto de investigación GUMELAB. Freie Universität Berlin*
Leonardo Augusto Pachón Contreras, *coordinador del proyecto de extensión entre la Freie Universität Berlin y la Universidad de Antioquia para cooperar en el desarrollo del componente de e-Research (métodos digitales) del proyecto GUMELAB*

El proyecto de investigación GUMELAB (Transmisión de la historia a través de los medios de entretenimiento en América Latina. Laboratorio de Investigación de la Memoria y Métodos Digitales), del Instituto de Estudios Latinoamericanos de la Universidad Libre de Berlín, ha sido financiado por el Ministerio Federal de Educación e Investigación (Bundesministerium für Bildung und Forschung, BMBF) en el marco del programa Entender la Sociedad-Formar el Futuro (Estudios Regionales).

www.gumelab.net/es

Equipo sistematizador
Gicela Andrea Aguirre García
Hannah Müssemann
Heisman Duvolfán Arcila Arboleda
Alesson Ramon Rota

Colaboradores en diferentes etapas del componente de e-Research (métodos digitales) del proyecto GUMELAB
Holle Meding (GUMELAB, Freie Universität Berlin)
Brayan Alexander Muñoz Barrera (Universidad de Antioquia)
Joseph F. Vergel-Becerra (Universidad de Antioquia)
Julián Andrés Montoya Carvajal (Universidad de Antioquia)
Marco A. Erazo (Universidad de Antioquia)

Juan David Villa
Corrección de estilo

Ana Milena Gómez C.
Diseño editorial

Contenido

Prólogo

El presente *Manual para el uso de métodos digitales en proyectos de humanidades. La experiencia del proyecto GUMELAB* ha sido redactado de manera colaborativa por un equipo interdisciplinario e internacional. Su génesis se encuentra en la extensa fase preparatoria de un proyecto de investigación en el campo de historia latinoamericana realizado en la Universidad Libre de Berlín (Freie Universität Berlin). Este proyecto, titulado *Transmisión de la historia a través de los medios de entretenimiento en América Latina. Laboratorio de investigación de la memoria y métodos digitales*, se identifica mediante el acrónimo alemán GUMELAB (*Geschichtsvermittlung durch Unterhaltungsmedien in Lateinamerika. Labor für Erinnerungsforschung und digitale Methoden*).

En el ámbito de la investigación histórica, la comprensión de cómo las personas se aproximan a su pasado y perciben su historia constituye una faceta esencial del objeto de estudio de la ciencia histórica, especialmente en el campo de la didáctica de la historia. En esta perspectiva, el proyecto GUMELAB profundiza en la problemática de la transmisión del pasado, focalizándose particularmente en telenovelas y series que abordan la historia reciente de algunos países de América Latina.

Estos formatos televisivos representan productos estratégicos dentro del ámbito del entretenimiento, al capturar la atención de amplias audiencias y generar diálogos y debates tanto en el día a día de sus espectadores como en la esfera pública. Asimismo, el visionado de estas producciones no solo puede moldear la percepción del pasado de las audiencias y sus interpretaciones, sino que también puede incidir en sus posturas políticas. Este impacto es especialmente significativo considerando que las principales temáticas abordadas en las telenovelas y series de interés para GUMELAB están vinculadas a pasados traumáticos inmersos en procesos de construcción y lucha por una memoria histórica.

Un escenario central para estos diálogos y debates es el entorno de las plataformas digitales, donde las audiencias participan activamente en las redes sociales. Cada interacción, ya sea un comentario, un tuit, un *like* o un *hashtag*, deja una huella que puede considerarse como una potencial fuente histórica para GUMELAB. Este enfoque implica un cambio significativo para las historiadoras y los historiadores, quienes se ven desafiados a abandonar los archivos y adentrarse en la búsqueda de fuentes de origen digital. Además, deben reflexionar sobre la metodología para la crítica documental que emplearán en este nuevo contexto.

El acceso de los datos disponibles en plataformas como X, Facebook, YouTube, entre muchas otras, representa un desafío metodológico para cada investigación. En el proyecto GUMELAB optamos por trazar nuestro propio camino, abordando no solo la cuestión de cómo acceder a los datos que consideraríamos como fuentes digitales, sino también involucrándonos en una reflexión heurística sobre lo que implica trabajar con estos datos dispersos en el ciberespacio.

Nuestra primera observación fue que, para explorar de manera más precisa el universo de nuestras fuentes en el ciberespacio, queríamos desarrollar métodos de orden digital para capturar los datos y crear nuestro propio archivo y heurística. Fue en este momento cuando recurrimos al equipo de físicos especializados en ciencias de datos de la Universidad de Antioquia (Medellín, Colombia), liderado por el profesor Leonardo Pachón. Juntos diseñamos el componente de *e-Research* de GUMELAB. Más adelante, durante el desarrollo de este componente, contamos con la colaboración de historiadores del Centro de Humanidades Digitales de la Universidade Estadual de Campinas (Brasil).

El trabajo colaborativo y toda la experiencia en el desarrollo del componente de *e-Research* fue materia de experimentación en el laboratorio de métodos digitales de GUMELAB. Todo lo que vivimos en el laboratorio fue tan enriquecedor en términos de aprendizaje que decidimos documentarlo en forma de un manual. Hannah Müssemann, Gicela García, Heisman Arcila y Alesson Rota asumieron esta tarea y en las siguientes líneas nos presentan

la ruta metodológica que elegimos, cómo se desarrolló, los inconvenientes que surgieron, los desvíos que tuvimos que tomar y cómo abordamos los retos presentados, tanto los que pudimos solucionar como los que quedaron sin resolver. El manual incluye un breve glosario que especifica los términos que fueron fundamentales para nosotros durante este proceso.

Una contribución significativa del manual es la presentación de ejemplos concretos que ilustran nuevas perspectivas analíticas que surgieron gracias al empleo de fuentes digitales. En esencia, el manual destaca lo que hemos logrado observar mediante el uso de métodos digitales, revelando aspectos que de otra manera hubieran pasado desapercibidos.

El manual cierra con una reflexión importante presentada en forma de recomendaciones dirigidas a colegas en el ámbito de las humanidades y las ciencias sociales que se encuentren ante la tarea de abordar fuentes digitales de origen primario. Estas fuentes, específicamente las que surgen en la web y no están resguardadas en ningún archivo oficial, plantean retos particulares. El documento ofrece reflexiones esenciales que deben ser consideradas al emprender un proyecto de esta naturaleza, abordando temas relevantes como los desafíos de la comunicación en el trabajo interdisciplinario y los obstáculos inherentes al manejo de estas fuentes digitales.

Dada la rápida evolución del mundo digital y la naturaleza efímera de muchos de sus contenidos, es probable que gran parte de la información recopilada en este manual se vuelva obsoleta en un futuro cercano. Sin embargo, este documento adquiere también la dimensión de un registro histórico que captura un momento crucial en la incorporación de métodos digitales en la investigación en humanidades. ¡Esperamos que les sea de gran utilidad!

Mónika Contreras Saiz
GUMELAB, Freie Universität Berlin

La era digital ha incentivado a historiadores y a otros académicos a reconocer cada vez más el valor del contenido y la dinámica de las redes sociales en las diversas plataformas digitales. Considerar la información que transita en estos espacios como fuentes digitales de información histórica (Brügger, 2018) ha impulsado la proliferación de nuevos tipos de datos históricos. De esta manera, cada interacción que deja huellas en textos e imágenes digitales puede servir como fuente legítima para la investigación histórica, abarcando áreas como la historia contemporánea, la construcción de la memoria pública, las historias personales, la opinión pública, las tendencias culturales, entre muchas más áreas de investigación. En la historia contemporánea, por ejemplo, estas fuentes se han utilizado para estudiar acontecimientos significativos del siglo XXI como la Primavera Árabe, el movimiento Occupy Wall Street, entre otros. Las plataformas digitales ofrecen información sobre la opinión y el discurso público que permite calibrar los sentimientos de amplios sectores de la población. Además, ayudan a trazar las tendencias culturales y sociales a lo largo del tiempo; los memes, los videos virales y otras formas de producción digital pueden arrojar luz sobre las preocupaciones, el humor y la estética de una época concreta. En todos los distintos tipos de plataformas digitales, no solo en los que interaccionan las redes sociales, sino también aquellos más personalizados, como los blogs y los sitios web personales, pueden ofrecer una visión profunda de las experiencias y perspectivas individuales. Asimismo, el gran volumen de datos que generan las interacciones en las redes sociales permite aplicar técnicas de análisis de datos a la investigación histórica que pueden proporcionar perspectivas y tendencias a gran escala que no son evidentes en fuentes tradicionales, cuyo volumen es notablemente reducido.

En el caso particular de América Latina, durante los últimos 20 años, producciones de televisión como telenovelas y series que tratan el pasado reciente se han convertido incluso,

a través del consumo de la cultura popular, en una forma de aprendizaje de la historia de sus países (Erlick, 2018, pp. 13-22).[1] Estas producciones, por su alcance masivo, influyen en la percepción del pasado de los espectadores y las espectadoras, y, por tanto, repercuten en su conciencia histórica y su formación política (Contreras Saiz, 2023). Las telenovelas movilizan los recuerdos y tienen una inmensa influencia en la representación del pasado y el presente no solo en América Latina, sino también en las comunidades latinoamericanas de exiliados y migrantes; así que dan forma a la memoria cultural más allá de la región.

Las plataformas digitales, por su parte, se consolidan cada vez más como espacios para el consumo de información y entretenimiento; entre otros, el acceso a telenovelas y series. Su alcance como espacios de interacción ha incidido en que funcionen como plataformas para la movilización del debate y la opinión pública con influencia en el ámbito de la memoria cultural (Birkner y Donk, 2018). De manera que la conexión entre la memoria y los medios de entretenimiento abre un fértil campo de investigación.

En el marco de estas transformaciones, el proyecto GUMELAB, financiado por el Ministerio Federal de Educación e Investigación (Bundesministerium für Bildung und Forschung, BMBF), aborda cuestiones vinculadas a la transmisión del conocimiento histórico, la construcción de memorias a

1 En el último decenio se han producido telenovelas y series que tratan aspectos de la historia contemporánea; entre ellos, las violaciones de los derechos humanos durante los regímenes militares de Argentina, Brasil y Chile; el conflicto armado de Colombia; la lucha contra el narcotráfico en Colombia y México; y la incidencia de Chávez en Venezuela. Algunas de estas producciones han atraído mucha atención y han desencadenado el desarrollo de nuevas series en plataformas por demanda como Netflix. Este tipo de producciones han sido denominadas como "telenovelas de la memoria" (Contreras Saiz, 2017). En el centro de estas telenovelas y series se encuentran los acontecimientos históricos del pasado traumático reciente, que la mayoría de la audiencia puede recordar directamente, o que han quedado en la memoria familiar a través de las historias de los parientes. El significado narrativo de estas telenovelas y series para el público está determinado en gran medida por la mediación de la memoria, en la que la narración histórica y la conciencia histórica se encuentran (Rüsen, 1997).

través de medios de entretenimiento como telenovelas y series, y su recepción nacional y transnacional. GUMELAB participa en el debate de los estudios históricos y culturales sobre la memoria y el recuerdo; especialmente sobre la transnacionalización de la memoria colectiva (Classen, 2009) y la memoria global (Assmann y Conrad, 2010).

Para abordar esta temática, GUMELAB se concentró en dos casos de estudio seleccionados mediante criterios nacionales y contextuales: Chile y Colombia. En el momento de conceptualizar este proyecto, en noviembre de 2019, ambos países estaban experimentando una agitación política, donde el tratamiento de su pasado dictatorial, en el caso de Chile, y del conflicto armado interno, en el caso de Colombia, ocupaban un lugar destacado en las protestas sociales. Ubicados en este contexto continuo, marcado por las tensiones propias de la negociación sobre la interpretación y reinterpretación del pasado violento y conflictivo, seleccionamos para el análisis cinco producciones de televisión; dos series chilenas: Los 80, más que una moda, (Wood Producciones, 2008-2014) y Dignity (Mega, 2019); y tres producciones relacionadas con la historia reciente colombiana: Pablo Escobar. El Patrón del Mal (Caracol, 2012), Tres Caínes (RCN, 2013) y la serie estadounidense-colombiana Narcos (Netflix, 2015-2017).

Adicionalmente, se incluyó un tercer caso de estudio que trasciende las fronteras nacionales. Con el objetivo de investigar cómo se reciben estas telenovelas y series más allá de las fronteras y entender la manera en que se forman recuerdos transnacionales del pasado de Chile y Colombia, decidimos enfocar el estudio en comunidades de migrantes latinos en Estados Unidos.

La elección de estos tres casos de estudios (Chile, Colombia, comunidades migrantes en Estados Unidos) y de las cinco producciones televisivas determinó los parámetros que delimitan el objeto de estudio, y con esto el espacio de observación en el inmenso mundo de las interacciones dentro de las redes sociales en las plataformas digitales. El análisis incluye todo tipo de comentarios, opiniones y discusiones originadas en Chile, en Colombia y en la comunidad latina que vive en Estados Unidos, relacionadas con estas cinco producciones audiovisuales.

Teniendo como centro las categorías de análisis *imágenes de la memoria*, *conciencia histórica* y *formación política*, la investigación indaga sobre cómo se afecta (i) la recepción y procesamiento del contenido de las telenovelas y series por el público; (ii) las dimensiones, extensión y tendencias de la transnacionalización de los sentidos de la memoria generadas por estas telenovelas y series; (iii) la conciencia histórica de la audiencia; y (iv) la formación política del público.

Coherente con la naturaleza del problema, el componente *e-Research* de GUMELAB constituye una orientación metodológica que traza un camino para el empleo y desarrollo de métodos digitales, por cuanto se orientan a grandes cantidades de datos, y el uso de tecnologías digitales y métodos computacionales para construir el acervo de fuentes provenientes de las redes sociales que interaccionan en las plataformas digitales X (antiguo Twitter) y YouTube, así como periódicos en línea y blogs encontrados a través de Google Search.[2] Estos datos se complementan con entrevistas cualitativas a profundidad.[3]

A la vez, como orientación metodológica, esperamos que los métodos digitales desarrollados en el marco del componente de *e-Research* proporcionen nuevas perspectivas de análisis y formas de entender los problemas a los que se aboca la investigación. Estos métodos pueden ser útiles para la minería de datos, organización de información, codificación y automatización de procesos, creación de bases de datos, visualización y análisis de datos con diferentes modelos matemáticos y de la estadística descriptiva y prescriptiva.

Aunque la interacción de las redes sociales en las plataformas digitales y las huellas que dejan en textos e imágenes desempeñan un papel cada vez

2 De ahora en adelante, cuando hacemos referencia a Google o Google Search como una red social, nos referimos a las páginas encontradas a través de esa herramienta.

3 Las entrevistas fueron llevadas a cabo con distintos segmentos de las audiencias de los cinco casos de estudio en los que se concentra GUMELAB, así como con personas vinculadas en la producción de estas obras, como guionistas, productores, actores y actrices. También incluimos entrevistas con investigadores de los temas que interesan a GUMELAB. Las hicimos en Chile, Colombia y Estados Unidos.

más importante como fuentes históricas legítimas, es fundamental abordarlas con el mismo rigor y análisis crítico que se aplicaría con las fuentes históricas tradicionales. En particular, el ejercicio histórico con fuentes digitales enfrenta retos, entre otros, (i) preservación, (ii) privacidad y cuestiones éticas, e (iii) información errónea y prejuicios. Los contenidos digitales pueden desaparecer rápidamente, por lo que su conservación es un reto importante. No todos los contenidos digitales se archivan, e incluso cuando se archivan, puede ser difícil encontrar piezas concretas de información más adelante. Además, el uso de información procedente de las redes sociales puede plantear problemas de privacidad y consentimiento. Se trata de un tema de debate permanente en el campo de la historia digital. Finalmente, como cualquier fuente, el contenido digital puede mostrar sesgos, mentiras o información errónea. Es esencial tener en cuenta estos factores a la hora de utilizar dichas fuentes.

El presente manual tiene como objetivo sistematizar la experiencia de los dos equipos de trabajo[4] que cooperaron en el marco del proyecto GUMELAB para la investigación histórica con métodos digitales sobre la recepción de telenovelas y series que transmiten la historia reciente y afectan los procesos de construcción de memoria histórica en el ámbito de los cambios paradigmáticos que introducen la cuarta y quinta Revolución Industrial, cuyas herramientas cada vez toman más fuerza en la investigación de problemas sociales (Botero *et al*., 2019). El proceso de sistematización se centró en discutir tanto los errores como los éxitos, con la intención de fomentar la

4 El primer equipo de trabajo está compuesto por historiadores y latinoamericanistas asociados a la Universidad Libre de Berlín (Alemania). Este grupo, denominado en adelante como Equipo de Humanidades, estableció las bases del proyecto GUMELAB en cooperación con otros socios estratégicos. A este grupo se sumaron activamente historiadores de la Universidade Estadual de Campinas. El segundo equipo está conformado por físicos especializados en ciencia de datos y vinculados a la Universidad de Antioquia (Medellín, Colombia). Este grupo, denominado en adelante como Equipo de Ciencias de Datos, trabajó conjuntamente con el Equipo de Humanidades en el diseño del componente de e-Research de GUMELAB. En lo sucesivo, al describir actividades específicas de uno de los equipos, lo haremos explícitamente, mientras que, cuando digamos el equipo GUMELAB, nos referiremos al esfuerzo conjunto de ambos equipos.

incursión de las humanidades en esta área, proporcionando recomendaciones para integrar estas experiencias en futuras investigaciones relacionadas.

El primer capítulo presenta la ruta metodológica de la investigación histórica con métodos digitales, que se inició con el diseño del sistema categorial adaptado al lenguaje de las redes sociales para la minería de datos, el etiquetado o categorización, el entrenamiento/reentrenamiento hacia la automatización de las búsquedas, con alcance en el diseño de modelos de clasificación y puesta en operación de la plataforma del banco de datos para el análisis de la información.

El segundo capítulo detalla el proceso técnico con las plataformas digitales en las que interaccionan las redes sociales, sus generalidades y la crítica como fuente, el proceso de organización de datos no estructurados y el proceso de *natural language processing* (NLP). El tercer capítulo presenta el modelo de entrenamiento de datos y clasificación mediante modelos de agentes cognitivos de acuerdo con las categorías centrales de análisis de la investigación. El cuarto capítulo presenta algunos resultados descriptivos relativos a la recepción del contenido de las telenovelas y series por el público. Finalmente, siguiendo el proceso de investigación cualitativa, el quinto capítulo busca centrar los aprendizajes, alertas y desafíos para el uso de métodos digitales aplicados a problemas de investigación de las humanidades y la historia como disciplina desde la configuración de los equipos de trabajo.

Este manual presenta a la comunidad académica de manera transparente el camino que el proyecto GUMELAB ha recorrido desarrollando métodos digitales, y algunos de los resultados preliminares en el desarrollo de un método de investigación en el ámbito de las humanidades digitales. En tal sentido, se centra más en el problema metodológico del abordaje de los métodos digitales, y no tanto así en el abordaje teórico del problema de investigación en específico.

En el marco de este proyecto, desmitificamos la inteligencia artificial (IA) en su trasfondo matemático revelando sus limitaciones, que implican grupos humanos resolviendo problemas de aplicación a las humanidades e

implementando desarrollos útiles para el propósito. Las técnicas cambian vertiginosamente sin siquiera iterar sus usos de aplicación a problemas específicos para avanzar sobre sus propios límites y posibilidades. De manera que la IA avanza sobre las principales vertientes del funcionamiento de las sociedades, mientras las reflexiones en el ámbito epistemológico sobre las implicaciones sociales en la educación y el trabajo siguen su propio curso marginal.

La propia noción de IA es paradójica. Se trata de un término acuñado en los Estados Unidos durante la Conferencia de Dartmouth, en 1956, y surgió inicialmente como un campo de estudio académico, pero rápidamente atrajo interés militar, apuntando a su aplicación en contextos de guerra y defensa. Esta evolución evidencia un aspecto retórico en el término *IA*: aunque sugiere capacidades cognitivas comparables con las humanas, muchas de sus aplicaciones prácticas aún son herramientas avanzadas de procesamiento de datos y automatización, que no alcanzan la plena autonomía o conciencia que la palabra *inteligencia* podría implicar. Este dualismo entre la promesa teórica de la IA y sus aplicaciones prácticas actuales revela la complejidad y los desafíos éticos inherentes a su desarrollo y uso.

El equipo de investigación del componente de *e-Research* de GUMELAB comparte los aprendizajes adquiridos a lo largo del proceso con la comunidad académica. El propósito es facilitar la selección directa de las opciones metodológicas más adecuadas y coherentes con los problemas de investigación y las fuentes disponibles.

E-Research

Es un término que generalmente se refiere a la investigación electrónica, en la que se incluyen el uso de medios, métodos y herramientas digitales en la investigación académica. Tres características son importantes en este tipo de investigación: (i) el manejo de una gran cantidad de datos; (ii) la creación de aplicaciones web y soluciones de publicación para la presentación y visualización de datos y resultados de investigación; y (iii) la multidisciplinariedad de los equipos de investigación.

Humanidades digitales

Son una categoría conceptual que pretende organizar las diversas aplicaciones, enfoques y visiones que se producen en las relaciones entre las humanidades y la investigación informática. Ellas implican el uso de métodos, tecnologías digitales y técnicas computacionales para analizar, visualizar y compartir datos (es decir, *e-Research*) en campos de estudio humanísticos, como literatura, historia, filosofía, ciencias sociales y artes. Estos métodos permiten a investigadores en humanidades trabajar con cantidades de datos que serían imposibles de manejar mediante técnicas tradicionales, y pueden proporcionar perspectivas complementarias y para entender las disciplinas humanísticas.

Inteligencia artificial (IA)

Para la investigación aplicada con métodos digitales, se refiere al uso de algoritmos y técnicas de IA en el análisis y el procesamiento de datos digitales con el objetivo de resolver problemas prácticos en diversas disciplinas. Esta práctica se puede aplicar en una variedad de campos, incluyendo ciencias sociales, humanidades, ciencias de la salud, negocios, ingeniería, entre otros más.

Algunos ejemplos de cómo la IA puede ser utilizada en la investigación aplicada con métodos digitales son estos:

1. **Procesamiento de lenguaje natural (PLN; *natural language processing*, NLP).** Las técnicas de PLN les permiten a las máquinas leer, comprender e interpretar el lenguaje humano. Esto puede ser útil en la investigación aplicada para analizar grandes volúmenes de texto, como las publicaciones en redes sociales, los artículos de noticias o la literatura académica.
2. **Análisis de imágenes y visión por computadora.** Los algoritmos de IA pueden ser utilizados para reconocer patrones y extraer información de imágenes y videos digitales. Esto puede ser aplicado en campos como la medicina (por ejemplo, para el análisis de imágenes médicas), la ecología (para el análisis de imágenes de satélite) o la seguridad (para el análisis de videovigilancia).
3. **Aprendizaje automático (*machine learning*).** El aprendizaje automático es un campo de estudio dentro de la IA que se centra en el desarrollo de algoritmos y modelos computacionales capaces de aprender y mejorar automáticamente a través de la experiencia y los datos. En lugar de ser programados de manera explícita, estos sistemas son entrenados con conjuntos de datos para reconocer patrones y realizar tareas específicas sin necesidad de instrucciones precisas. El aprendizaje automático abarca diferentes enfoques, como el aprendizaje supervisado, el no supervisado y por refuerzo, entre otros un poco menos usados, y se aplica en una amplia gama de campos, desde el reconocimiento de voz y la visión por computadora hasta la predicción de comportamientos y la toma de decisiones basadas en datos. Su objetivo es permitir que las máquinas aprendan y mejoren por sí mismas, para proporcionar soluciones inteligentes y adaptativas a problemas complejos.

Metadatos

Son información sobre información. Son datos que describen o proporcionan contexto a otros datos, como una etiqueta en un libro que indica su título, autor y fecha de publicación. Estos detalles ayudan a las personas a entender y organizar mejor los datos, como fotos, canciones o documentos, facilitando así su búsqueda y comprensión.

Entidad

Es un objeto o concepto distinto que puede ser representado en una base de datos como una persona, producto o evento. Son cruciales para las relaciones de datos, ya que ayudan a organizar y estructurar la información de manera lógica, a fin de facilitar la recuperación y el análisis de datos relacionados. Las entidades forman la base para crear tablas y establecer relaciones en las bases de datos, lo que garantiza la gestión eficiente de la información interconectada.

Endpoint de API

Un *endpoint* de una interfaz de programación de aplicaciones (API); puede definirse como un punto de interacción específico que facilita la comunicación y el intercambio de datos entre distintos sistemas de *software*. Este punto de interacción actúa como una dirección o puerto a través del cual los servicios de la API son accesibles. Los *endpoints* permiten que diferentes aplicaciones, tales como programas de *software*, sistemas operativos o dispositivos, soliciten y reciban información entre sí siguiendo un conjunto de protocolos y formatos predefinidos. Este proceso es fundamental para la integración y la funcionalidad efectiva de los sistemas tecnológicos en la era digital, lo cual hace más sencillos la interoperabilidad y el acceso eficiente a funciones y datos necesarios para diversas aplicaciones y usuarios.

Minería de datos

Es una disciplina que integra aspectos de la estadística, la inteligencia artificial (IA) y la gestión de bases de datos. Su finalidad es identificar patrones y relaciones relevantes en extensos conjuntos de datos. Mediante el uso de estas técnicas, la investigación aplicada puede revelar nuevas correlaciones o tendencias que no son inmediatamente perceptibles. Esta es una muestra de las múltiples aplicaciones de la IA en la investigación. En cada contexto, la IA capacita a los investigadores para examinar de forma más eficiente y precisa grandes cantidades de datos digitales, para el descubrimiento de nuevas perspectivas en la información que pueden ser fundamentales para abordar desafíos del mundo real.

Algoritmo

Un algoritmo aplicado a métodos digitales para investigación en humanidades es un conjunto de instrucciones o reglas lógicas definidas que las computadoras siguen para analizar, interpretar y presentar datos relacionados con temas humanísticos. Estos algoritmos pueden ser herramientas esenciales para los investigadores y las investigadoras en las humanidades, ya que sirven para detectar normas en diversos formatos de medios, como textos, imágenes, audio y otros medios digitales que quizá los métodos tradicionales no puedan abordar.

Agente cognitivo

Se refiere a una entidad capaz de percibir su entorno, procesar información, tomar decisiones y llevar a cabo acciones basadas en ese procesamiento. Un ejemplo sería un sistema de inteligencia artificial que pueda analizar patrones de tráfico y tomar decisiones con el fin de optimizar rutas de entrega para un servicio de paquetería.

Hemos decidido abandonar el uso de agentes cognitivos a favor de la adopción exclusiva de sistemas de inteligencia artificial (IA) más tradicionales. Este cambio refleja las limitaciones prácticas y los desafíos éticos enfrentados con agentes cognitivos, como la dificultad para simular fielmente la cognición humana y las complejidades en la toma de decisiones autónomas en ambientes variables. La IA convencional, aunque menos sofisticada en términos de autonomía y adaptación, ofrece un mayor control, aspecto crucial para aplicaciones seguras y confiables.

Modelo de aprendizaje

Un modelo o paradigma de aprendizaje en inteligencia artificial (IA) es una estrategia o marco conceptual que los algoritmos de aprendizaje automático utilizan para adaptarse a los datos y mejorar su rendimiento con el tiempo. En otras palabras, es la metodología que un algoritmo de IA sigue para aprender de los datos y detectar patrones. En general, estas formas de aprendizaje se llaman *aprendizaje automático* (*machine learning*). En los últimos años se ha popularizado el término *aprendizaje profundo* (*deep learning*) como un subcampo del aprendizaje automático, donde este último implica redes neu-

ronales con muchas capas. La elección del modelo de aprendizaje depende en gran medida del tipo de problema que se esté tratando de resolver, de la cantidad y calidad de los datos disponibles, y de las necesidades específicas del proyecto.

Existen varios **modelos o paradigmas de aprendizaje en** IA. Estos se pueden dividir en tres categorías principales:

1. **Aprendizaje supervisado.** Este es el modelo más común. En el aprendizaje supervisado, el algoritmo aprende de un conjunto de datos de entrada y salida ya etiquetados, y se le enseña a predecir las salidas de datos de entrada no vistos previamente. Los modelos de regresión y clasificación, como la regresión lineal, la regresión logística, las máquinas de soporte vectorial (SVM) y los árboles de decisión, son ejemplos de aprendizaje supervisado.
2. **Aprendizaje no supervisado.** En el aprendizaje no supervisado, el algoritmo se encarga de descubrir la estructura y los patrones subyacentes en los datos por sí mismo. No hay una salida específica que se esté buscando, sino que el objetivo es entender la estructura de los datos. Los métodos de *clustering*, como *K-means*, y los algoritmos de reducción de dimensionalidad, como el análisis de componentes principales (PCA), son ejemplos de aprendizaje no supervisado.
3. **Aprendizaje por refuerzo.** Este tipo de aprendizaje es un poco diferente. Un agente cognitivo aprende a tomar decisiones mediante la interacción con su entorno. El agente realiza acciones, recibe retroalimentación en forma de recompensas o castigos, y ajusta sus decisiones futuras en función de esta retroalimentación. El objetivo es maximizar la recompensa total a lo largo del tiempo.

También existen otras formas de aprendizaje, como el aprendizaje semi-supervisado (una combinación de aprendizaje supervisado y no supervisado), el aprendizaje por transferencia (donde un modelo preentrenado se adapta a una nueva tarea), el aprendizaje activo (donde el modelo selecciona los datos más útiles para aprender), y otros. Sin embargo, los tres mencionados anteriormente son los fundamentales y los ampliamente utilizados en la IA.

Interfaz de programación de aplicaciones (API, *application programming* interface)

Es un conjunto de reglas y protocolos que establece cómo deben interactuar diferentes piezas de *software*. Proporciona un medio para que los desarrolladores de *software* puedan utilizar funciones específicas de una aplicación, servicio o plataforma sin tener que conocer los detalles del código fuente subyacente. Las API actúan como una especie de puente o intermediario permitiendo que diferentes programas de *software* se comuniquen entre sí y compartan datos. Las API se pueden utilizar para una variedad de tareas, desde extraer datos de una base de datos hasta crear, leer, actualizar y eliminar recursos en un servicio web.

Por ejemplo, muchas redes sociales y plataformas en línea, como X o Google Maps, ofrecen API públicas que los desarrolladores pueden utilizar para interactuar con la plataforma y acceder a sus datos y funciones. Un desarrollador podría usar la API de X para recopilar tuits que contengan un *hashtag* específico, o usar la de Google Maps para obtener información sobre una ubicación.

En resumen, una API les permite a los diferentes componentes de *software* interactuar y comunicarse entre sí, facilitando la creación de aplicaciones más complejas y potentes.

Plataforma digital

Facebook, X (antiguo Twitter), Instagram, YouTube, TikTok, etc., no son en realidad redes sociales, sino plataformas digitales en las que interactúan las redes sociales.

Red social

Una red social, en el ámbito del *e-Reseach*, hace referencia a la comunicación que tiene lugar entre las personas que participan en una determinada plataforma digital. En esta comunicación se generan conexiones e intercambia información, creando un tejido virtual en el ciberespacio. Las redes sociales están compuestas por un conjunto de actores sociales, que pueden ser individuos o diferentes tipos de organizaciones adscritos a un determinado contexto social.

Second screening (segunda proyección)

Es un fenómeno que ocurre cuando se consume un contenido por una pantalla (televisor, tablet) mientras se usa al mismo tiempo otra pantalla (celular, computador). Un ejemplo clave para la investigación de GUMELAB: cuando la audiencia comenta en las redes sociales mientras ve un capítulo de una serie o telenovela.

Procesamiento de lenguaje natural (PLN; *natural language processing,* NLP)

Es una rama de la inteligencia artificial que se centra en el estudio de la interacción entre los seres humanos y las máquinas mediante el uso del lenguaje humano. Consiste en el desarrollo y la aplicación de algoritmos y técnicas que permiten a las computadoras comprender, interpretar y generar lenguaje natural de manera efectiva. Algunos ejemplos de los posibles usos del NLP en ámbitos investigativos pueden ser:

1. **Análisis de sentimientos.** Puede utilizarse para identificar y clasificar las emociones y opiniones expresadas en textos como reseñas de productos, comentarios en redes sociales o encuestas. Esto es útil para estudiar actitudes, percepciones y tendencias en la sociedad.
2. **Extracción de información.** Mediante técnicas de NLP es posible extraer información relevante y estructurada de textos no estructurados, como artículos científicos, informes o libros o documentos históricos. Esto facilita la recopilación de datos para investigaciones en ciencias humanas.
3. **Análisis de texto y discurso.** El NLP permite analizar y comprender la estructura y el significado de textos y discursos, incluyendo el estudio de la sintaxis, la semántica y la pragmática. Esto es útil para investigaciones en lingüística, psicología del lenguaje y estudios de comunicación.
4. **Traducción automática.** Los sistemas de NLP pueden realizar traducciones automáticas entre diferentes idiomas, lo que facilita la comunicación y el intercambio de conocimientos en contextos internacionales y multilingües.
5. **Generación de resúmenes.** Mediante el NLP es posible generar resúmenes automáticos de textos extensos, lo que facilita la revisión y el análisis de información en investigaciones científicas y literatura académica.

Estos son solo algunos ejemplos de cómo el procesamiento de lenguaje natural es aplicado en diferentes ámbitos. Su capacidad para comprender y procesar el lenguaje humano tiene un amplio potencial en la investigación y el avance del conocimiento en diversas disciplinas relacionadas con las ciencias humanas.

Complejidad

En el contexto de las redes, la complejidad se refiere a la estructura y el comportamiento de estas bajo diferentes supuestos; por ejemplo, en las redes neuronales artificiales o las redes de comunicación o transporte público. La complejidad en la inteligencia artificial y las redes neuronales presenta desafíos significativos en términos de diseño, implementación, optimización y comprensión de estos sistemas. Sin embargo, también proporciona oportunidades para desarrollar soluciones más sofisticadas y avanzadas que ayuden a abordar problemas complejos y lograr resultados más precisos y eficientes en una amplia gama de aplicaciones.

En este mismo contexto, la complejidad de la red neuronal se relaciona con la capacidad de los sistemas para manejar y procesar grandes volúmenes de datos, así como para realizar tareas cognitivas complejas. Un sistema de inteligencia artificial complejo puede requerir algoritmos sofisticados, modelos de aprendizaje profundo (*deep learning*) y una potencia computacional significativa para lograr resultados precisos y eficientes.

Redes neuronales computacionales

A menudo llamadas simplemente redes neuronales, son un tipo de tecnología utilizada en inteligencia artificial que está inspirada en el funcionamiento del cerebro humano. Nuestro cerebro está compuesto por células llamadas neuronas, las cuales se comunican entre sí para procesar información. Cuando aprendemos algo nuevo, las neuronas cambian la forma en que se conectan y comunican, lo cual es parte esencial de cómo aprendemos y recordamos. Las redes neuronales computacionales intentan imitar esta forma de trabajo del cerebro humano. Ellas tienen *neuronas artificiales*, que son programas o códigos en una computadora. Sin embargo, la idea de las redes neuronales

es una metáfora del cerebro humano porque, aunque intentan imitar la forma en que este trabaja, son mucho más simples y menos poderosas que el cerebro real. Son una forma de intento de réplica de nuestra capacidad de aprender y reconocer patrones, pero lo hacen de manera limitada y enfocada.

Métricas de desempeño

Son medidas numéricas que se usan para evaluar el rendimiento de un modelo o algoritmo. Estas métricas proporcionan una forma objetiva de cuantificar aspectos como la precisión, el error, la eficiencia y otros resultados importantes en el desarrollo y la evaluación de modelos de IA y ML (*machine learning*, aprendizaje automático).

Algunos ejemplos de métricas comunes utilizadas son:

1. **Precisión.** Es una métrica que indica la proporción de predicciones correctas realizadas por un modelo (verdaderos positivos). Se calcula dividiendo el número de predicciones correctas entre el número total de predicciones realizadas.
2. **Sensibilidad (*recall*).** Es la proporción de ejemplos positivos que son correctamente identificados por un modelo. Se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos.
3. **Especificidad.** Es la proporción de ejemplos negativos que son correctamente identificados por un modelo. Se calcula dividiendo el número de verdaderos negativos entre la suma de verdaderos negativos y falsos positivos.
4. **Error cuadrático medio (MSE, *mean squared error*).** Es una métrica comúnmente utilizada en problemas de regresión. Mide el promedio de los errores al cuadrado entre los valores predichos y los valores reales. Cuanto menor sea el MSE, mejor será el rendimiento del modelo.
5. **Curva ROC (*receiver operating characteristic*).** Es una métrica utilizada para evaluar la capacidad de discriminación de un modelo en problemas de clasificación binaria. La curva ROC muestra la tasa de verdaderos positivos frente a la tasa de falsos positivos a medida que varía el umbral de clasificación.

6. **Área bajo la curva ROC (AUC-ROC, *area under the ROC curve*).** Es una métrica que resume la curva ROC en un solo número. Cuanto mayor sea el valor del AUC-ROC, mejor será el rendimiento del modelo en términos de clasificación.

En los problemas específicos de NLP se utilizan algunas métricas específicas para evaluar el rendimiento de los modelos y algoritmos:

1. **F1-score.** Esta es una métrica que combina la precisión y el *recall* en una sola medida. Representa la media armónica de estas dos métricas y proporciona una forma equilibrada de evaluar el rendimiento de un modelo. Es especialmente útil cuando se desea tener en cuenta tanto los falsos positivos como los falsos negativos. Sin embargo, al combinar la precisión y el *recall* se pierde información sobre cada métrica por separado.
2. **BLEU (*bilingual evaluation understudy*).** Es una métrica utilizada para evaluar la calidad de las traducciones automáticas en NLP. Calcula la coincidencia de n-gramas (secuencias de palabras) entre la traducción generada y una referencia humana. El BLEU puede ser útil para comparar diferentes enfoques de traducción automática, pero puede tener limitaciones al no considerar la semántica y la coherencia del texto.
3. ***Perplexity* (perplejidad).** Es una métrica utilizada en modelos de lenguaje para evaluar la capacidad de un modelo para predecir secuencias de palabras. Mide cuán *sorprendente* es un conjunto de datos dado el modelo. Una menor perplejidad indica que el modelo es más capaz de predecir las secuencias de palabras. Sin embargo, la perplejidad puede no reflejar directamente la calidad semántica de las predicciones del modelo.

Estos son solo algunos ejemplos de métricas utilizadas en el campo de la IA y el *machine learning*. La elección de las métricas adecuadas depende del tipo de problema, del tipo de datos y de los objetivos específicos del proyecto.

Wordcloud (nube de palabras)

Es una técnica de visualización de datos utilizada principalmente para representar datos textuales de manera gráfica. En esta visualización, las palabras se muestran en diferentes tamaños, y el tamaño de cada palabra indica su frecuencia o importancia en el conjunto de datos. Las palabras más frecuentes o significativas aparecen en un tamaño más grande, mientras que las menos usadas se muestran más pequeñas. Este método no solo facilita la identificación visual rápida de los términos clave en un texto, sino que también ofrece una perspectiva intuitiva sobre la estructura temática o el énfasis en los datos textuales analizados. Es importante señalar que, a pesar de su utilidad en la representación visual, las nubes de palabras no proporcionan información detallada sobre las relaciones o el contexto de las palabras dentro del texto

Hashtag (#)

En el contexto de las redes sociales, un *hashtag*, representado por el símbolo #, se define como una etiqueta de metadatos utilizada para categorizar o agrupar contenidos. Al preceder una palabra o frase (sin espacios) con el símbolo #, el *hashtag* se convierte en un enlace cliqueable dentro de la plataforma de la red social. Esto les permite a los usuarios encontrar fácilmente mensajes y publicaciones relacionadas con un tema específico.

Los *hashtags* son herramientas clave para organizar la información y facilitar la búsqueda de temas específicos en las redes sociales. También son utilizados para promover campañas, eventos, tendencias, o para expresar sentimientos o comentarios relacionados con temas de actualidad. Además, gracias a ellos los usuarios pueden unirse a conversaciones globales o seguir debates en tiempo real. En resumen, los *hashtags* son fundamentales en las dinámicas de interacción y descubrimiento de contenido en las redes sociales.

Tag (@)

El término *tag (@)*, en el contexto de las redes sociales como X, Facebook e Instagram, se refiere al acto de mencionar o referenciar a otros usuarios mediante sus nombres de cuenta precedidos por el símbolo @ (por ejemplo, @nombreusuario). Cuando se utiliza un tag, el usuario mencionado recibe

una notificación y puede ver directamente dicho contenido, lo que facilita su capacidad de responder o interactuar con el mensaje.

Es importante diferenciar los *tags* y los *hashtags*. Mientras que los *hashtags* (#) se utilizan para categorizar o agrupar publicaciones bajo un tema específico, los *tags* (@) se emplean específicamente para involucrar a otros usuarios en una conversación o para dirigir una publicación hacia ellos. Esta herramienta ofrece una interacción más directa y personalizada en las redes sociales, lo que posibilita, a su vez, conversaciones y conexiones entre los usuarios.

SQL (*structured query language*) y noSQL (no *structured query language*)

SQL (lenguaje de consulta estructurada) es un lenguaje estándar utilizado para gestionar y manipular bases de datos relacionales, donde los datos se almacenan en tablas estructuradas e interrelacionadas. Es ideal para consultas complejas y operaciones en datos estructurados.

No-SQL (no solo SQL) se refiere a una variedad de tecnologías de bases de datos diseñadas para almacenar, recuperar y manipular datos de maneras que no se ajustan al modelo relacional tradicional. Típicamente son más flexibles y escalables, adecuadas para manejar grandes volúmenes de datos no estructurados o semiestructurados.

Query

En programación, especialmente en contextos de bases de datos, es una solicitud para acceder o modificar datos. Está escrita en un lenguaje específico, como SQL o Python, y se utiliza para realizar operaciones como buscar, insertar, actualizar o eliminar datos. Las *queries* les permiten a los usuarios interactuar con la base de datos de manera recortada, extrayendo solo la información considerada necesaria.

Lista de siglas

API	*Application programming interface*, interfaz de programación de aplicaciones.
BLEU	*Bilingual evaluation understudy*, subestudio de evaluación bilingüe.
CH	Conciencia histórica (categoría elaborada por la investigación GUMELAB).
CNN	*Convolutional neural network*, red neuronal convolucional.
CSV	*Comma-separated values,* Valores separados por comas (Estructura de un archivo de texto para almacenar o intercambiar datos de estructura simple).
DL	*Deep learning*, aprendizaje profundo.
EPM	Pablo Escobar. El Patrón del Mal (caso de estudio de GUMELAB).
FP	Formación política (categoría elaborada por la investigación GUMELAB).
GCP	Google Cloud AutoML.
IA	*Artificial intelligence*, inteligencia artificial.
IM	Imágenes de la memoria (categoría elaborada por la investigación GUMELAB).
LSTM	*Long short-term memory*
ML	*Machine learning,* aprendizaje de máquinas (o aprendizaje automático).
MONGO	Abreviatura para MongoDB, una base de datos NoSQL de código abierto, orientada a documentos
MSE	*Mean squared error*, error cuadrático medio.

NLP *Natural language processing*, procesamiento de lenguaje natural.

noSQL *No structured query language*, lenguaje de consulta no solamente estructurada.

NR No relevante (categoría elaborada por la investigación GUMELAB).

PCA *Principal component analysis*, análisis de componentes principales.

ROC *Receiver operating characteristic*, característica operativa del receptor.

SEO *Search engine optimization*, optimización de motores de búsqueda.

SQL *Structured query language*, lenguaje de consulta estructurada.

SVM *Support vector machine*, máquinas de soporte vectorial.

TF-IDF *Term frequency-inverse document frequency*, término de frecuencia- documento de frecuencia inversal.

1. Ruta metodológica para una investigación histórica con métodos digitales

1.1. Opciones metodológicas de la investigación hacia la automatización de procesos de búsqueda y clasificación de información

Según el *Digital 2022. Global overview report*, publicado por la agentura We Are Social y la herramienta de gestión de redes sociales Hootsuite en el año 2022, de toda la población del mundo, el 62,5% usaba internet entonces, 4% más que el año anterior (We Are Social y Hootsuite, 2022, p. 20).

Para el caso de Sudamérica, el 75% de la población usaba internet; mientras que en Centroamérica el 70% y en el Caribe el 66%. En el continente norteamericano, el 92% de la población era usuaria de internet (We Are Social y Hootsuite, 2022, p. 22). Países como Brasil, Colombia y Argentina estaban entre los cinco que usaban internet durante más tiempo al día: casi 10 horas diarias; mientras que Estados Unidos estaba en un poco más de siete horas. A modo de comparación, el tiempo medio en internet en todo el mundo era de seis horas y cincuenta y ocho minutos al día (We Are Social y Hootsuite, 2022, p. 27).

Entre las razones comunes de uso estaban búsqueda de información (61,00%), entrar en contacto con amigos o familiares (55,2%) y ver televisión o videos (51,5%), lo cual en parte explica el crecimiento anual en plataformas de *streaming* (We Are Social y Hootsuite, 2022, p. 29). El uso de plataformas digitales donde interactúan las distintas redes sociales fue muy importante para entrar en contacto con otros usuarios y ocupó el 95,2% de todas las páginas y aplicaciones utilizadas en 2022 (We Are Social y Hootsuite, 2022, p. 43).

Las plataformas digitales con mayor interacción mundial de redes sociales en ellas durante el año 2022 fueron, en su orden, Google, YouTube, Facebook y Twitter[1] (We Are Social y Hootsuite, 2022, p. 45). Cabe destacar que es necesario tener en cuenta la proporción de usuarios de las plataformas en comparación con el número total de usuarios de internet. Para calcular el

1 Ya que en 2021 la plataforma X todavía se llamaba Twitter.

alcance de las diferentes plataformas se analiza en el reporte qué porcentaje de los usuarios de internet en su totalidad se puede alcanzar con anuncios. Respecto a eso, anuncios en Facebook lograban un 58,8% (We Are Social y Hootsuite, 2022, p. 119), mientras que en YouTube un 51,8% (We Are Social y Hootsuite, 2022, p. 132) y en Twitter solamente un 8,8% (We Are Social y Hootsuite, 2022, p. 192) de todos los usuarios de internet en el mundo, lo que muestra el alcance de la plataforma.

Mirando los casos de estudio de la investigación de GUMELAB, se puede destacar que en 2022 en Chile y en Estados Unidos el 92% de la población tenía acceso a internet (Kemp 2022a; Kemp 2022c), mientras que en Colombia un 69,1% (Kemp 2022b). En Chile, un 81,1% de la población total usó YouTube a inicios de 2022 (Kemp 2022a). En Estados Unidos, YouTube fue usado por un 74% de la población total a inicios del mismo año, un 80,4% de todos los usuarios de dicho país (Kemp 2022c). En Colombia, YouTube alcanzó un 59,2% de la población, que corresponde a un 85,6% de los usuarios de internet (Kemp 2022b). Estos números muestran la importancia de la plataforma en la región para los usuarios y las usuarias de la red.

Consideramos usar Twitter por dos razones importantes. En primer lugar, en Colombia, Chile y también en Estados Unidos las opiniones e informaciones que se intercambian en esta plataforma digital tienen un gran eco en las noticias que circulan allí. En segundo lugar, a través de los tuits se tiene acceso a opiniones y discusiones simultáneas sobre el visionado que tienen las audiencias activas en Twitter respecto de las telenovelas y series. En Chile, Twitter era usado por el 15,1% de la población (Kemp 2022a), y en Colombia solo por el 8,4% (Kemp 2022b), mientras que en Estados Unidos por el 23% (Kemp 2022c). A pesar de que Twitter no suele ser la plataforma más usada en Latinoamérica ni mundialmente, constituye una fuente que contiene numerosas publicaciones sobre telenovelas y series, y valiosas interacciones entre espectadores, actores y actrices. También se crearon cuentas ficticias de los personajes (por ejemplo, @JuanHerrera de *Los 80*), donde comentan alrededor de las telenovelas y series que investigamos.

La interacción y la circulación de opiniones entre las audiencias activas en las plataformas de las redes sociales sobre las telenovelas y series que investi-

gamos constituyen una fuente importante para los objetivos de investigación de GUMELAB. Este tipo de fuentes se diferencian de las entrevistas, ya que las redes sociales funcionan también como un espacio transnacional regional para discutir contenidos de telenovelas y series. En este contexto, tienen lugar dos fenómenos descritos por la investigación en medios de comunicación: (i) el llamado *second screening*, que describe la forma de ver un contenido por una pantalla (televisor o tablet), mientras se comenta por otra (celular, computador); y (ii) el llamado *social television*, que resalta el componente social de estas interacciones, ya que la idea es validar y compartir emociones, ideas y reacciones con otros usuarios entre personas que no siempre se conocen más allá de la recepción de contenidos.[2] Ambos fenómenos cuentan con un componente importante para nuestra investigación, y es el efecto de la reacción simultánea. Por los datos sabemos en qué fecha y hora tuvo lugar la reacción, muchas veces hasta en qué lugar, informaciones que podemos cruzar con los visionados. Es decir, con la información de emisión de los capítulos de las telenovelas y series (sí fueron emitidas en televisión abierta) a las que las audiencias activas en redes sociales están reaccionando. La evaluación de los datos masivos encontrados en las redes sociales tiene la ventaja de que estas opiniones y declaraciones no se ven afectadas por la interacción con el investigador o la investigadora, que siempre ejerce una influencia inconsciente en los participantes y las participantes, como ocurre, por ejemplo, en las entrevistas al momento de formular las preguntas. Cuando se analizan datos de plataformas digitales, no hay interacción directa entre ellos, ya que los datos son analizados *a posteriori*. Pero, al mismo tiempo, trabajar con redes sociales pone en duda la representatividad de los datos.

Tal y como lo mencionamos en la introducción, la investigación parte de las categorías de análisis *imágenes de la memoria*, *conciencia histórica* y *formación política*, a partir de las cuales se abre el árbol temático de descriptores.[3]

2 Más información sobre el concepto de *second screening* en Midha y Nagy (2014, pp. 448-453); Wilson (2016, pp. 174-191); Selva (2016, pp. 159-173).

Para ver una demostración de cómo detectamos los fenómenos de *second screening,* consulte el subcapítulo sobre detección de patrones temporales y PNL.

3 Véase más en Contreras Saiz (2019, pp. 51-86).

A su vez, creamos una lista de palabras clave y una lista de combinaciones posibles para orientar organizadamente la recuperación de datos tendientes a la saturación de posibilidades de combinación temática/espacio/temporal para cada una de las series y telenovelas.

Esta fase de diseño es determinante para la calidad de los datos y el menor lastre de información que no está relacionada con la recepción de las telenovelas y series, sino más bien con temas tratados en ellas.[4] Esta estrategia facilita las etapas subsiguientes en la estructuración de los datos hacia la automatización de las búsquedas propiamente codificadas de acuerdo con el sistema categorial. En ese momento, de estructuración de los datos o de la organización de la información, lo hicimos bajo los parámetros del sistema categorial previamente diseñado, tendiente a la mayor y significativa codificación de información para la emisión de resultados susceptibles de análisis a la luz de las preguntas de interés y bajo los modelos y métodos de análisis pertinentes. Es muy importante que el proceso descrito hasta este punto se encuentre debida y delicadamente documentado, siempre teniendo en mente las preguntas centrales de la investigación, a fin de no perdernos en la compleja fase de extracción de información y estructuración de los datos, tal como lo describimos más adelante. Es decir, tal como en la investigación cualitativa tradicional, es el diseño de la investigación lo que orientará al final del proceso de estructuración de los datos para avanzar propiamente en el análisis del problema de investigación. En todo caso, se debe tener como punto de partida que, de fondo, estamos frente a un diálogo en construcción entre dos sistemas epistemológicos. Por otro lado, el proceso de automatización de las búsquedas y su clasificación son bastante similares a los métodos tradicionales de la investigación cualitativa; sin embargo, no se debe perder de vista que estamos frente a altos volúmenes de información, lo cual hace costo-eficiente la aplicación de métodos digitales, entre otras posibilidades analíticas.

4 Por ejemplo, en la serie chilena titulada *Los 80, más que una moda*. Las dos palabras, *los 80,* aparecen mucho en las redes, pero no necesariamente conectadas con algún comentario relacionado con la serie de televisión.

Ahora bien, entre el diseño de la investigación y el análisis de la información, la aplicación de métodos digitales para la investigación con grandes volúmenes de datos implica un arduo trabajo de campo para investigaciones con datos no estructurados como GUMELAB. Esta etapa de *trabajo de campo* fue organizada, genéricamente, a través de las fases de extracción de información, etiquetado o codificación de información útil y no útil para la estructuración de los datos, limpieza de datos hacia la creación de modelos de entrenamiento y reentrenamiento para la automatización de las búsquedas y su clasificación de acuerdo con las principales categorías de análisis de la investigación. Hasta este punto apenas empezaremos a darle orden a la información por medio de la creación de archivos en sentido estricto y a organizar en una nube para el llamado de la información, bajo los parámetros de búsquedas que la investigación particularmente requiera y proyecte de acuerdo con las posibilidades de usos futuros, como detallamos en los siguientes apartados y en el capítulo segundo.

La delimitación del periodo de extracción de información se basó en la fecha de la primera mención pública de la producción. *Pablo Escobar. El Patrón del Mal* se lanzó en la televisión abierta colombiana en 2012; a su vez, *Narcos* en 2015, *Tres Caínes* en 2013, *Dignity* en 2019 y *Los 80* en 2008. Recabamos información hasta el 31 de julio de 2021, establecido como límite temporal para el proyecto.

Teniendo muy bajas posibilidades de recabar información parametrizada para todas las plataformas digitales en las que interaccionan las redes sociales, dos criterios fundamentales orientaron la selección de fuentes para GUMELAB. Por un lado, la posibilidad de interacción del público en torno a los visionados de las telenovelas y series de análisis. Por otro lado, la capacidad de acceso amplio a los datos. Entendida la potencia de las plataformas digitales como espacio de interacción de redes sociales y, además, como fuente de información, seleccionamos las redes sociales que interaccionan en las plataformas Twitter, YouTube, Facebook y Google Search; sin embargo, Facebook debió ser descartada por tener acceso restringido, como lo explicaremos más adelante.

Dadas las limitaciones de punto de partida para la definición previa de las características poblacionales de las personas espectadoras y comentadoras de las telenovelas y series, usamos como criterio principal que fueran de la comunidad hispanohablante, y la ubicación geográfica de espectadores y comentadores de cada una de las telenovelas y serias; (i) público nacional de la respectiva telenovela o serie: Chile, Colombia y Alemania para la serie *Dignity*; (ii) internacional, es decir, toda la recepción proveniente de otros países; y (iii) público transnacional, como migrantes latinoamericanos en Estados Unidos, colombianos y colombianas, chilenos y chilenas en otras partes del mundo, y latinos y latinas en todo el planeta.

Posteriormente, de acuerdo con la información emergente, pudimos proveer una mejor caracterización de la población receptora y comentadora según las posibilidades de los datos generales provistos por la plataforma y facilitados por los usuarios: ¿quiénes hablan en las redes sociales?, ¿quiénes tienen acceso a internet?, ¿quiénes son los usuarios activos?, ¿qué tan representativas son las redes sociales en la creación de la esfera pública?, ¿será que hablan siempre *los mismos* por la dinámica de las redes sociales (por ejemplo, más hombres que mujeres)?

Con mayor profundidad, el equipo de investigación identificó cuentas específicas pertenecientes a personas con roles centrales o significativos en torno a las diferentes telenovelas y series, tales como cuentas oficiales de estas, cuentas de actores y actrices, cuentas de canales de televisión, etc.

Para el componente de *e-Research* de GUMELAB una fuente es el texto que escribe el usuario o la usuaria en cada una de las plataformas digitales que investigamos, así como la información relacionada con el grupo de usuarios, como nombre de la persona, nombre de usuario(a), ubicación, de acuerdo con las posibilidades que ofrezca la red, el contenido de la publicación, entre otros, que son los metadatos, como *la información sobre la información*, si se quiere.[5] Además, hay, o existe, otra información que con frecuencia viene

5 Para saber todos los metadatos posibles, vea la documentación en cada red social. Es importante destacar que la información proporcionada por las redes sociales a

aparejada a los comentarios o sobre la que se abre el debate (pueden ser imágenes, videos, enlaces). Todas estas categorías serán llamadas *entidades* y se corresponden con los encabezados de cada una de las columnas de los archivos .csv en los que finalmente reposan, y se les da una estructura a los datos extraídos. [Revisa aquí las estructuras de las bases de datos, a partir de dos scripts python para cada red social].

Las opciones metodológicas tomadas hasta este punto por el equipo de investigación de humanidades dieron paso al proceso de extracción de información de las plataformas digitales a cargo del equipo de ciencias de datos. Hasta este momento, los parámetros que guiaron las primeras búsquedas fueron *palabras clave*, sus combinaciones posibles, fuentes digitales, poblaciones, espacio, periodo de tiempo para cada una de las telenovelas y series. Los resultados de las extracciones se guardaron en archivos .csv y se fueron organizando bajo criterios comunes de las ciencias de datos en una nube en Google Drive, con la administración del equipo de ciencias de datos. Los datos obtenidos de cada búsqueda variaron de acuerdo con las características de la información provista por las redes sociales, y las propias de la configuración de usuario al momento de unirse a un tipo determinado de red. El equipo GUMELAB debió orientar un posprocesamiento de esta información mediante la creación de parámetros uniformes para todos los encabezados (entidades) de cada una de las columnas de los archivos .csv para todas las redes sociales.

La información específica a cada plataforma de redes sociales, recopilada para describir a los usuarios y sus interacciones, como comentarios y contenidos discutidos, se denomina *datos no estructurados*. Estos datos son obtenidos de la fuente y representados en archivos .csv a través de títulos en las columnas, conocidos como encabezados (o entidades), que varían según la red social. A pesar de que cada plataforma proporciona diferentes tipos de información, su organización requiere la estandarización de estos

través de las API cambia con el tiempo, de acuerdo con el desarrollo de las propias plataformas y sus públicos. Por lo tanto, nuevos cambios en el tipo de datos pueden requerir una nueva organización, es decir, una nueva estructuración.

encabezados para el análisis. Según las metodologías de GUMELAB, esto incluye: (i) datos generales de las redes sociales para la caracterización básica de usuarios, como nombre de usuario, fecha de creación de la cuenta, ubicación, etc.; (ii) características básicas de los comentarios, esenciales para el análisis de contenido; (iii) los comentarios o contenidos, clasificados bajo categorías centrales de análisis y sus respectivos descriptores.

El inicio del proceso de evaluación, tras obtener los primeros resultados, tuvo como fin determinar la calidad de la información recabada, con el propósito de validar o rechazar los datos extraídos. Este paso es crucial para el preentrenamiento algorítmico destinado a mejorar las búsquedas (véase el capítulo 2 para más detalles). Posteriormente, basados en las categorías centrales de análisis y los descriptores predefinidos, los cuales proporcionaron orden y estructura a la vasta cantidad de información, procedimos con la tarea de codificación o etiquetado de los datos. Este proceso consistió en asignar significados y categorías a los comentarios e información recogidos en las plataformas digitales. Es importante resaltar que este procedimiento está dirigido a entrenar algoritmos automatizados para la búsqueda y clasificación de información, así como para el desarrollo de modelos analíticos.

A continuación detallamos el proceso técnico del equipo de ciencias de datos para la automatización y creación de modelos.

Bibliografía:

Busse, Laura; Enderle, Wilfried; Hohls, Rüdiger; Meyer, Thomas; Prellwitz, Jens y Schuhmann, Annette (Hg.) (2018). *Clio-Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften.* 2.a edición. Clio-online und Humboldt-Universität zu Berlin (Historisches Forum, 23).

Cohen, Daniel J. y Rosenzweig, Roy (2006). *Digital history: a guide to gathering, preserving and presenting the past on the web.* University of Pennsylvania Press.Dougherty, Jack y Nawrotzki, Kristen (Hg.) (2013). *Writing History in the Digital Age.* University of Michigan Press.

Graham, Shawn; Milligan, Ian y Weingart, Scott (2016). *Exploring big historical data. The historian's macroscope.* Imperial College Press.

Lässig, Simone (2021). Digital history. *Geschichte und Gesellschaft,* 47(1), 5-34. 10.13109/gege.2021.47.1.5.

Theocharis, Yannis y Jungherr, Andreas (2021). Computational social science and the study of political communication. *Political Communication,* 38(1-2), 1-22. 10.1080/10584609.2020.1833121.

Turkle, William; Adam Crymble y Alan MacEachern (2009). *The programing historian.* NiCHE, Network in Canadian History & Environment. http://niche-canada.org/programming-historian/

Windsor, Leah Cathryn (2021). Avanzando en el trabajo interdisciplinar en Ciencias de la Comunicación Computacional. *Political Communication,* 38 (1-2), 182-191. 10.1080/10584609.2020.1765915.

1.2 Automatización de búsquedas

Al comienzo del proyecto carecimos de conocimiento sobre la cantidad aproximada de datos disponibles en las redes y plataformas seleccionadas. Para recopilar de manera exhaustiva todos los datos existentes, que presumimos serían considerablemente abundantes, fue necesario emplear herramientas computacionales. Esto incluyó el uso de un lenguaje de programación adecuado y de API o de bibliotecas de *web scraping* para realizar la recopilación de datos automatizadamente.

GUMELAB ha usado Python 3 como lenguaje de programación, ya que este es de alto nivel,[6] bastante fácil de usar y con una disponibilidad de bibliotecas que permite no solo rapidez en la implementación del código para la extracción de datos, sino también su limpieza, procesamiento y entrenamiento de modelos para la automatización de búsquedas en las redes sociales definidas.

6 Los lenguajes de programación de alto nivel están más cerca del lenguaje humano y abstraen la mayoría de los detalles del *hardware* de la máquina. Son más fáciles de aprender y usar, ya que tienen una sintaxis más simple y permiten que el programador o la programadora se concentre en la lógica del programa en lugar de concentrarse en los detalles del sistema. Por otro lado, los lenguajes de programación de bajo nivel están más cerca del código de máquina y proporcionan un control más directo sobre los componentes físicos del ordenador.

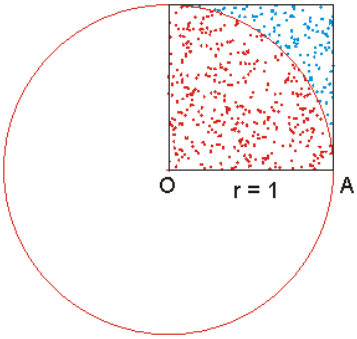
Además, Twitter (hoy X), YouTube y Google Search ofrecen API propias para Python, y no es necesario recurrir al *scraping* de sus webs. Si bien esto facilita el acceso a los datos, también impone ciertos límites a la hora de realizar las búsquedas. Dichas limitaciones dependen de cada caso específico, e incluso cambiarán con el paso del tiempo y las decisiones tomadas por cada corporación. Ejemplos de estas limitaciones son los costos de acceso a las API de Google Search, que imponen no solo una carga monetaria sobre el proyecto, sino además límites sobre la cantidad de peticiones diarias que pueden realizarse, o los límites a la cantidad de tuits accesibles o ventana de tiempo disponible en la API de esa red social.[7]

Con el fin de mitigar estas limitaciones se pueden crear estrategias diferentes; un ejemplo del uso de estos métodos en el caso de GUMELAB son las estrategias tipo Monte Carlo[8] en la selección de criterios de búsqueda para Google Search; de esta manera es posible barrer el espacio de parámetros homogéneamente, controlando el gasto y la cantidad de llamados a la API; así, si en algún momento se hace necesario suspender las búsquedas, no se habrá quedado por fuera del muestreo ninguno de los casos de estudio. En la **figura 1** representamos cómo se configuran las parametrizaciones de las búsquedas. Los puntos rojos y azules representan el diagnóstico realizado en diversas selecciones del alcance total del documento, mientras que un razonamiento matemático para prever los posibles resultados de lo que sería el archivo antes de crearlo. Con esto podemos mitigar los resultados no deseados.

7 Normalmente, los cambios en las API vienen acompañados de cambios en la documentación, pero en la actualidad la documentación para la API de Twitter no suele reflejar dichos cambios con la rapidez esperada.

8 Los Monte Carlo son métodos estadísticos numéricos usados para evaluar ecuaciones matemáticas muy complejas; estos están centrados en el hecho de que en un espacio ergódico un punto en movimiento aleatorio recorrerá todos los puntos del espacio dado el tiempo suficiente. Su nombre proviene de la ciudad de Monte Carlo (Mónaco), conocida como la capital de los juegos de azar. Consulte algunos ejemplos del uso de los métodos Monte Carlo aquí.

Figura 1



Nota: Los puntos rojos representan resultados dentro de las expectativas. Los puntos azules representan resultados fuera.

Fuente: Elaboración propia.

1.3 Refinamiento de búsquedas

¿Cómo eliminamos de las búsquedas realizadas la información que no se relacionaba con la recepción de las telenovelas y series seleccionadas? Uno de los grandes retos al hacer búsquedas de datos en cantidades altas y en plataformas tan heterogéneas como las elegidas por el proyecto GUMELAB son los datos indeseados que también se capturan; por ello es necesario plantear diferentes estrategias de limpieza y filtrado de datos para el refinamiento de aquellas.

En GUMELAB planteamos dos estrategias como primera medida de refinamiento de búsquedas: limpieza por similaridades y limpieza por modelado de tópicos. Cada caso lo aplicamos de acuerdo con las características de la fuente.

Limpieza por similaridad.[9]

Las técnicas de similaridad se utilizan para medir la similitud o distancia entre dos o más textos. Estas técnicas permiten cuantificar la proximidad entre las estructuras y el contenido de dichos textos. Existen muchas medidas de similaridad. En GUMELAB usamos:

- **Similaridad de Levenshtein.** La similaridad de Levenshtein se basa en la distancia de edición entre dos cadenas de texto. Esto se refiere a la cantidad mínima de operaciones necesarias, ya sea inserciones, eliminaciones o sustituciones de caracteres, para convertir una cadena de texto en otra. La similaridad de Levenshtein se calcula dividiendo la distancia de edición entre la longitud total de las cadenas. Cuanto menor sea la distancia de edición y, por lo tanto, mayor sea la similaridad de Levenshtein, mayor será la similitud entre los textos.
- **Similaridad del coseno.** La similaridad del coseno se utiliza para medir la similitud entre dos vectores de características numéricas; por lo tanto, los textos se deben representar como vectores de características antes de realizar la comparación. La similaridad del coseno se calcula como el coseno del ángulo entre los dos vectores de características. Cuanto más cercano sea el ángulo a cero, mayor será la similaridad del coseno, y, por lo tanto, mayor será la similitud entre los textos.

Ambas técnicas son útiles en diferentes escenarios y dependen del contexto de aplicación. La similaridad de Levenshtein es especialmente útil cuando se trata de comparar cadenas de texto de longitud similar, como palabras o frases cortas. Por otro lado, la similaridad del coseno es más adecuada para medir la similitud entre documentos más largos o textos que han sido representados como vectores de características.

9 Unas lecturas relacionadas con las métricas de distancia: https://towardsdatascience.com/3-basic-distance-measurement-in-text-mining-5852becff1d7 y https://blog.aibits.dev/cosine-similarity-a-guide-to-understanding-and-implementing-text-similarity-measurement-75a172e7806f

Para usar estas similaridades en la limpieza de GUMELAB tomamos la distancia entre los títulos de los videos recuperados y sus descripciones con el contenido de los textos recuperados, y definimos una distancia mínima necesaria para aceptar un texto como útil. Dicha distancia mínima debe ser definida empíricamente con la exploración de los datos. [Ver ejemplo con aplicación de código aquí y aquí].

Limpieza por modelado de tópicos.[10]

Las técnicas por modelado de tópicos se utilizan para identificar y extraer temas o tópicos latentes presentes en un conjunto de datos de texto. Estas técnicas permiten descubrir patrones subyacentes en grandes volúmenes de texto y agrupar documentos relacionados en categorías temáticas descubiertas de manera automática por los algoritmos.

Una de las técnicas más comunes para el modelado de tópicos es el modelo de tópicos *latent* Dirichlet allocation (LDA).[11] LDA es un modelo generativo que asigna probabilidades a palabras y documentos, asumiendo que los documentos son mezclas de varios tópicos y que los tópicos son distribuciones de palabras. Así, asumiendo que los textos relevantes quedarán asignados a tópicos similares y, por el contrario, los textos irrelevantes serán tópicos diferentes, es posible separar los textos y eliminar, por tanto, aquellos sin interés para la investigación.

10 Otras formas de modelado de tópicos pueden encontrarse en https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

11 Para ver un ejemplo del uso de LDA, vaya a https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

1.4 Etiquetado de categorías de análisis, ejemplos y problemas

Ya que las búsquedas se hacen de manera automatizada, siempre pueden traer datos que no sirven para la investigación. A fin de mejorar el algoritmo de búsqueda, se necesita de un etiquetado que diferencie entre "sí, útil" o "no útil".[12] Esta primera etiquetación requiere mucho tiempo, ya que es esencial el trabajo manual del equipo de investigación de humanidades. Para llegar a una relevancia estadística usamos una calculadora estadística para calcular cuántos datos era necesario validar en un nivel de confianza del 95%, con un margen de error del 5%.[13] Esto es recomendable, porque un nivel de confianza del 99%, con un margen de error, significa validar casi todos los datos. Sin embargo, dado que los datos no están distribuidos aleatoriamente, esto es solo una estimación.

En un segundo paso clasificamos ejemplos en las tres categorías: *imágenes de la memoria* (IM), *formación política* (FP) y *conciencia histórica* (CH).

Tabla 1. Ejemplos de clasificación entre diferentes categorías.

Tuit	Categoría
Los bandidos son hombres de palabra. Pero los políticos y los bandidos de corbata son más jodidos #escobar.	FP
Uta estoy viendo los 80 y nuevamente me emociona. Y me emociona y entristece ver que una serie que te mueve por 30 años de este país nos muestra qué poco ha cambiado nuestro sistema. #los80 #canal13	FP
Caí en la tentación y me mató la curiosidad: viendo la primera temporada de #Narcos. Emociones encontradas, más de tristeza al evocar una de las etapas más duras y crudamente salvajes de este país de Dios. @NetflixLAT	CH

12 En el caso de GUMELAB, útil significa que el dato sí habla de una de las cinco producciones de televisión seleccionadas, y *no útil* que los datos no tienen una correlación con ninguna producción investigada.

13 Un ejemplo es https://www.questionpro.com/es/calculadora-de-muestra.html

Tuit	Categoría
Doy con #MaryYMike, serie que recrea crímenes de la DINA durante la dictadura de Pinochet. Hace años me enganché con #Los80, serie maravillosa que replicó en Chile lo que España hizo con #Cuéntame. Admirable cómo la TV y el cine chileno están luchando por su Memoria Histórica.	CH
Yo era muy niño cuando el m19 asociado a #escobar quemó el palacio @navarrowolff puede instruirnos + sobre ese pasado asqueroso d la patria.	IM

El etiquetado de las diferentes categorías pudo tener dificultades porque no siempre fue posible realizar una asignación clara a las distintas categorías. Esto tenía efectos directos en el entrenamiento del modelo de clasificación, sobre todo cuando un texto correspondía a dos categorías distintas, lo que dificulta el entrenamiento del modelo. Existían también diferencias en el etiquetado dentro de los integrantes del equipo de humanidades y durante varias fases del proyecto.

Este procedimiento también lo utilizamos para las otras fuentes, como YouTube, Google y las entrevistas.

1.5 Procesamiento de NLP para la clasificación de información

Antes de realizar la clasificación de un texto, es necesario llevar a cabo una serie de pasos de limpieza y preprocesamiento para preparar los datos adecuadamente.[14] Estos pasos de limpieza y preprocesamiento son fundamentales para preparar los datos de texto antes de su paso por un modelo de clasificación. Sin embargo, es importante señalar que los pasos específicos pueden

14 Puede consultar en https://towardsdatascience.com/all-you-need-to-know-about-text-preprocessing-for-nlp-and-machine-learning-bc1c5765ff67 o https://medium.com/sciforce/text-preprocessing-for-nlp-and-machine-learning-tasks-3e077aa4946e para una descripción más profunda de estos pasos.

variar dependiendo del contexto y del problema que se esté abordando, así como del modelo usado. Además, es posible que se requieran otros pasos adicionales según las características particulares del texto y los requisitos del proyecto. Por ejemplo, en algunos casos puede ser necesario eliminar URL o emojis, o hacer corrección ortográfica.

Es recomendable adaptar y ajustar estos pasos según las necesidades y particularidades del conjunto de datos y la tarea de clasificación específica que se esté abordando.

- **Tokenización.** Es el proceso de dividir un texto en unidades más pequeñas llamadas *tokens*. Los *tokens* suelen ser palabras, pero también pueden ser caracteres o frases cortas. La tokenización sirve para analizar y procesar el texto a nivel de unidad básica, lo que facilita tareas posteriores como la eliminación de palabras irrelevantes o la generación de representaciones vectoriales. Por ejemplo, la frase "El gato está durmiendo" se tokenizaría en los *tokens el*, *gato*, *está* y *durmiendo*. De nuevo, la tokenización elegida dependerá en gran medida del modelo, o del objetivo de clasificación.
- **Procesamiento de *stopwords*.** Se les llama *stopwords* a las palabras que son muy comunes y que, por tanto, no aportan un significado distintivo al texto, y se consideran irrelevantes para tareas específicas, como la clasificación de texto. Ejemplos de *stopwords* son *a*, *al*, *el*, *y*, *o*. En esta etapa son eliminadas del texto para reducir la dimensionalidad y mejorar la eficiencia del procesamiento. Esta eliminación se realiza utilizando una lista predefinida de *stopwords*, que depende del idioma o del contexto (por ejemplo, existen diccionarios de *stopwords* específicos para X), o basándose en algoritmos que identifican palabras poco informativas.
- **Lematización.** Es el proceso de reducir las palabras a su forma base o raíz, conocida como lema. El objetivo de la lematización es normalizar el texto y reducir la variabilidad en las palabras para, a su vez, reducir así la dimensión del lenguaje usado, lo que facilita el análisis y la identificación de patrones. Por ejemplo, las palabras *correr*, *corrió* y *corriendo* se lematizarían todas como *correr*. La lematización se lleva a cabo aplicando reglas lingüísticas y conocimientos sobre la morfología de las palabras.

Esto puede implicar cambios en la terminación, el uso de diccionarios léxicos o el uso de algoritmos basados en aprendizaje automático.

- **Eliminación de caracteres no alfabéticos y normalización.** En algunos casos es necesario hacer una limpieza adicional del texto. Esta implica eliminar caracteres no alfabéticos, como signos de puntuación, números o caracteres especiales que no aportan información relevante para la clasificación del texto. Además, la normalización puede ser útil para asegurar que las palabras sean representadas coherentemente a lo largo del conjunto de datos. Esto puede implicar el hecho de convertir todas las letras a minúsculas para evitar distinciones entre mayúsculas y minúsculas, y tratar formas comunes de abreviaturas o variantes ortográficas.
- **Vectorizado (*embedding*).** En esta etapa las palabras se representan mediante vectores numéricos llamados *embeddings*. Los *embeddings* capturan el significado semántico de las palabras y permiten que el modelo de clasificación comprenda la relación entre ellas de forma matemática reemplazando las palabras textuales por un conjunto de vectores en un nuevo espacio (espacio de embebimiento). Hay varias técnicas para generar *embeddings*, como Word2Vec, GloVe o BERT. Estos métodos asignan valores numéricos a las palabras en función de su contexto y significado en el texto. Al utilizar *embeddings*, las palabras se convierten en vectores densos en un espacio de características, lo que facilita el procesamiento posterior. [Ver ejemplo con aplicación de código aquí].

1.6 Modelos de clasificación

Una de las tareas más usadas en la IA, de la cual pueden desprenderse muchas otras tareas, es la clasificación. En este contexto, el objetivo es que el algoritmo, al recibir un conjunto de clases previamente definido por el usuario, sea capaz de determinar a cuál de dichas clases pertenece un ejemplo nuevo que le sea entregado.Este tipo de tarea se entrena mediante técnicas supervisadas, por lo cual se hace necesario tener un conjunto de datos previamente clasificados por el humano que servirán como ejemplos para la optimización de los hiperparámetros de la arquitectura o el modelo elegidos.

Es posible elegir entre modelos de ML o de DL; dicha elección dependerá de varios factores, entre ellos: la cantidad y complejidad de los datos disponibles; la precisión (nivel de error) aceptable de acuerdo con la necesidad; la capacidad computacional disponible.

Dentro de las múltiples opciones ML, hay ejemplos como los modelos Naïve Bayes y las máquinas de soporte vectorial:

- **Naïve Bayes.**[15] Los clasificadores de este tipo se basan en el teorema de Bayes, haciendo la asunción *ingenua* (*naïve*) de que las características son independientes entre ellas dada la clase. Con esto a mano es posible usar los ejemplos previamente clasificados para aproximar la distribución *a posteriori* y, por tanto, tener una distribución de probabilidad para las clases dado un conjunto de características de entrada.
  Los modelos basados en Naïve Bayes son bastante rápidos comparados con otras técnicas más sofisticadas; además, al haber asumido que las características condicionales para cada clase son independientes, es posible estimar la distribución de cada clase de manera independiente, lo cual ayuda a reducir los problemas relacionados con la alta dimensionalidad. Sin embargo, si bien estos modelos son buenos clasificadores y son rápidos, no son tan buenos estimadores, y, por tanto, las probabilidades que asignan a las diferentes clases no suelen ser muy serias.
- **Máquinas de soporte vectorial (SVM, *support vector machine*).**[16] Son una técnica que puede ser usada tanto para clasificación como para regresión. La idea básica detrás de ellas es el uso de un conjunto de vectores desde los cuales se les medirá la distancia a las clases etiquetadas por el usuario, y el entrenamiento consiste en encontrar el conjunto de

15 Para una descripción más detallada del método, https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf; o para un ejemplo del uso de este método, https://medium.com/machine-learning-101/chapter-1-supervised-learning-and-naive-bayes-classification-part-1-theory-8b9e361897d5

16 Para una descripción más detallada del método, https://medium.com/@zachary.bedell/support-vector-machines-explained-73f4ec363f13

hiperparámetros para estos vectores que maximicen la distancia (gap) entre los elementos de diferentes clases.

Los modelos basados en SVM suelen ser efectivos incluso cuando la dimensionalidad del problema es alta; además, dado que usan los vectores de soporte entrenados en la función de decisión, la cantidad de hiperparámetros necesarios no es alta y, por tanto, no usan gran cantidad de espacio en memoria. Por otro lado, al poder usar funciones kernel,[17] las SVM son muy flexibles y se adaptan muy bien a diferentes contextos. Sin embargo, si el número de características es muy grande o el número de ejemplos es muy pequeño, la implementación de estrategias de regularización y la elección del kernel se hacen cruciales para su correcto funcionamiento. Otra desventaja es que las SVM no calculan directamente las probabilidades de pertenencia a la clase, y para este cálculo se hace necesaria la implementación de validación cruzada, que puede aumentar drásticamente la carga computacional.

Para las estrategias basadas en redes neuronales existen diferentes arquitecturas; las más usadas para el análisis de texto son:

- **Redes neuronales convolucionales (CNN, *convolutional neural network*).**[18] Son un tipo de arquitectura que consta de pequeños filtros, también llamados *kernels*, que se encargan de analizar diferentes características del ejemplo de entrada y, de esta manera, extraer características de forma jerárquica, para así ir creando un ensamble de características más complejas usando patrones más pequeños y simples incrustados en sus filtros. Las CNN suelen usarse en datos complejos compuestos por elementos que se relacionan entre sí con sus elementos vecinos; por ejemplo, en las imágenes, los píxeles aledaños suelen estar relacionados,

17 Cuando los puntos que se quieren clasificar no están separados de forma lineal, se les puede aplicar una función de transformación para hacerlos lo más lineales posible previo al uso de la SVM; a esta función se le conoce como *kernel*. Vease por ejemplo: https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe

18 Un ejemplo del uso de las CNN en clasificación de texto se encuentra en https://medium.com/paper-club/cnns-for-text-classification-b45bde0bb254

pues pertenecen a un elemento más general dentro de la imagen (un rostro, una parte de un paisaje, etc.), o en el texto, en el cual la forma de una palabra suele guardar relación con sus palabras circundantes; o, más aún, el inicio o final de una palabra puede tener mucha relevancia en su significado (por ejemplo, los prefijos o sufijos). Así, es posible, por ejemplo, usar las CNN para crear una clasificación de los datos recolectados por GUMELAB según las categorías definidas en el estudio.

- **Redes tipo *long short-term memory* (LSTM).**[19] Estas redes son una arquitectura de red neuronal que introduce conceptos nuevos a las redes recurrentes; en estos casos, a diferencia de las CNN, no se trabaja con el contenido *espacial*, sino con el contenido secuencial de los datos; es decir, en el caso de las CNN se ven las frases como un todo correlacionado (tal como si estuviera escrito), mientras que en las LSTM como una secuencia (tal como si alguien las fuera hablando palabra a palabra).

  La principal ventaja de las LSTM es el hecho de poder guardar en sus estados una porción más grande del vector inicial y, por tanto, *recordar* un contenido mayor de las frases que se le introducen, gracias a lo cual puede tomar un mayor contexto a la hora de hacer una clasificación.
- ***Transformers*.**[20] Los *transformers* son un tipo de red neuronal que implementa una técnica diferente a las de las CNN y las LSTM, y otras, para hacer frente al problema de la memoria a largo plazo. En el caso de los *transformers*, se usa una capa llamada *de atención*, con la cual el modelo se entrena para saber a qué partes del vector de entrada dar más importancia durante el procesado. De esta manera no se hace necesario el uso de convoluciones o alineación de secuencias.

  Esto les da varias ventajas a los *transformers* respecto a las anteriores; entre ellas está que son no-secuenciales, es decir, procesan la frase como un todo y no por partes. La posibilidad de tener autoatención permite que se calculen similaridades o relaciones entre las palabras de la misma frase, de tal suerte que se encuentren relaciones entre diferentes partes

19 Para una descripción más profunda de las LSTM, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

20 Para una discusión sobre los *transformers*, https://towardsdatascience.com/transformers-89034557de14

de la entrada. Finalmente, otra gran innovación de los *transformers* es el *embebimiento*[21] posicional, con el cual la representación de una palabra dependerá no solo de ella, sino de su posición en el texto; esto hace que la extracción de la información contextual sea mucho más eficiente cuando se combinan las diferentes partes del *transformer*.

Como se ha visto, cada tipo de arquitectura tiene sus ventajas y desventajas; todo dependerá del problema por resolver, de la estructura de los datos, de las necesidades específicas de la investigación y de la capacidad computacional disponible. Dentro de GUMELAB hicimos experimentos usando todas estas técnicas de clasificación con una muestra reducida de los datos, de tal manera que pudiéramos explorar las ventajas y desventajas de cada una. Finalmente, dada la variedad de textos debido a la heterogeneidad de las muestras recogidas, decidimos tener un modelo de clasificación tipo *transformer*[22] para las fuentes, con el fin de extraer más eficientemente las características de la estructura interna de los textos producidos en cada fuente; por ejemplo, la estructura de un tuit con su forma de hacer mención a otros usuarios o el uso de *hashtags* vs. un comentario de YouTube.

1.7 Creación de un banco de datos para las fuentes

Al tomar decisiones sobre la gestión de grandes volúmenes de datos en el ámbito de la investigación con métodos digitales, existen varios aspectos por considerar. Uno de ellos se refiere a la selección y utilización de bases de

21 *Los embebimientos* son técnicas que permiten transformar las palabras textuales en estructuras matemáticas que luego son procesadas por las redes neuronales: https://www.turing.com/kb/guide-on-word-embeddings-in-nlp y https://towardsdatascience.com/a-deeper-look-into-embeddings-a-linguistic-approach-89cc428a29e7

22 Hemos usado una arquitectura BERT, ya que presenta todas las ventajas de un *transformer,* siendo lo suficientemente pequeño para ser entrenado, guardado y ejecutado en una máquina de bajos recursos, a diferencia de los grandes modelos del estado del arte, que requirieron servicios de terceros para ser implementados. Para una descripción de BERT, https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

datos para almacenar los datos recolectados. La elección de la base de datos adecuada dependerá en gran medida del tipo de datos (estructurados o no estructurados) y del propósito para el cual serán utilizados (consumo mediante aplicaciones, frecuencia de acceso, etc.).En el caso de los datos estructurados, es común utilizar bases de datos relacionales tipo SQL (*structured query language*). Estas bases de datos permiten almacenar la información en formato tabular, facilitando así su consulta y manipulación, especialmente cuando se requiere realizar operaciones que implican la combinación de diferentes columnas. Además, las bases de datos relacionales reducen la redundancia de datos, lo que mejora el rendimiento en las consultas de este tipo de información.

Por otro lado, para los datos no estructurados, como documentos, archivos o combinaciones de diferentes formatos, se suelen emplear bases de datos no relacionales tipo noSQL. Estas bases de datos ofrecen mayor flexibilidad en el diseño de esquemas, lo cual resulta beneficioso cuando los datos pueden variar en su naturaleza o contenido durante el desarrollo del proyecto. Además, tienen una capacidad escalable que les permite manejar eficientemente un creciente volumen de datos almacenados.

Otro aspecto relevante es determinar la frecuencia con la que los datos serán accedidos, ya que esto influye en la decisión de almacenarlos como *datos calientes* (altamente disponibles y accesibles) o como *datos fríos* (disponibilidad reducida). Esta elección tiene implicaciones importantes en términos de los costos asociados con el mantenimiento de los datos en una u otra disponibilidad.Asimismo, existen detalles aparentemente insignificantes para el proyecto, pero que adquieren vital importancia a medida que este crece, se involucra un mayor número de participantes o aumentan la complejidad y cantidad de los datos. Uno de estos detalles es establecer estándares internos para la nomenclatura de los diferentes elementos, como las claves de las tablas, su denominación, el control de versiones y su ubicación. Por ejemplo, se podría establecer una convención de nombramiento para las tablas utilizando un esquema consistente y comprensible para todos los miembros del equipo de investigación: esto facilitaría la identificación y manipulación de los datos.

En el caso de GUMELAB hemos usado inicialmente estructuras tipo tabla en las cuales hemos desagregado las consultas de cada API, creando así un sistema estructurado guardado en diferentes archivos tipo csv, que permiten almacenar toda la información extraída de cada fuente en una sola tabla. Además de ello, definimos una base de datos no estructurada en MongoDB[23] para guardar los datos sin importar de qué fuente provenían, y así poder mostrarlos de una forma homogénea en una plataforma visual servida al equipo de investigación. Con esta base de datos noSQL podemos hacer cambios rápidos en los esquemas y mostrar diferentes datos o agregar campos requeridos de una manera fluida.

Bibliografía:

Akita, F. (s. f.). *Akitando* [archivo de video]. YouTube. https://www.youtube.com/@AkitandoLucchesi, A. (2014). Por um debate sobre história e historiografia digital. *Boletim Historiar,* (2).Nicodemo, T. L. y Cardoso, O. P. (2019). Meta-história para robôs (bots): o conhecimento histórico na era da inteligência artificial. *História da Historiografia: International Journal of Theory and History of Historiography, 12*(29). 10.15848/hh.v12i29.1443. https://www.historiadahistoriografia.com.br/revista/article/view/1443

Lucchesi, Anita. Por um debate sobre História e Historiografia Digital. *Boletim Historiar,* n. 2, 2014.

Nicodemo, T. L.; Cardoso, O. P. Meta-história para robôs (bots): o conhecimento histórico na era da inteligência artificial. *História da Historiografia: International Journal of Theory and History of Historiography,* Ouro Preto, v. 12, n. 29, 2019. DOI: 10.15848/hh.v12i29.1443. Disponível em: https://www.historiadahistoriografia.com.br/revista/article/view/1443

Olhar Digital. (s. f.). Olhar Digital [pódcast]. Spotify. https://open.spotify.com/show/6BEYHG5W4PJAZYSvZ9GuaM

Rota, R. Alesson; Bauer, Carolina. Implicações do uso da inteligência artificial para a prática historiográfica: agências, comunicação e sensibilidades. In SciELO Preprints. https://doi.org/10.1560/SciELOPreprints.8965

23 MongoDB es un sistema de base de datos noSQL orientado a documentos. Es nuestro gestor de archivos de orígenes y formatos diferentes.

2. Trabajar con redes sociales como fuentes históricas

En las últimas décadas, el concepto tradicional de *archivo* ha sido cuestionado debido a cambios en los enfoques historiográficos, como los estudios lingüísticos, la generación pos-1970 y los estudios poscoloniales. Estas discusiones, que se desarrollaron durante el siglo XX, llevaron a una revisión de la estabilidad de los documentos, destacando la importancia del tratamiento técnico, del lugar de almacenamiento y de las acciones realizadas sobre ellos. Los archivos personales pueden contener información singular y ofrecer perspectivas únicas respecto de eventos históricos, relaciones sociales y estructuras de poder. La protección de la privacidad de los datos personales y la anonimización son aspectos relevantes al tratar con archivos personales, especialmente cuando se discuten la difusión y el uso público de esta información.

La información que circula en la interacción de las redes sociales en plataformas como X o YouTube ha sido considerada una forma de archivo y documentación personal. Aunque hay similitudes entre las redes sociales y los antiguos álbumes de recortes en términos de compartir experiencias personales, existen diferencias significativas, como las configuraciones de privacidad y la distribución de contenido basada en algoritmos. Las redes sociales dejan rastros digitales de las interacciones sociales y eventos que moldean la vida de los usuarios. Sin embargo, la mayor parte de este contenido permanece personal y solo se vuelve visible para un público más amplio cuando es seleccionado por los algoritmos o cuando individuos específicos se convierten en temas de debates públicos. Los cambios en los enfoques historiográficos han traído cuestionamientos sobre la estabilidad de los archivos, destacando la importancia de los tratamientos técnicos y del contexto en el que se producen los documentos.

El archivo personal en la web, incluyendo las redes sociales, es una práctica relevante para la preservación de la memoria personal y para la comprensión de la historia digital. Los historiadores deben establecer estándares de archivo digital durante sus investigaciones para documentar su propia práctica histórica. En resumen, el concepto de *archivo* ha sido reevaluado frente a los cambios historiográficos y las transformaciones en

el entorno digital, y es necesario considerar aspectos como la privacidad, la anonimización y la preservación al tratar con archivos personales, incluyendo aquellos encontrados en las redes sociales.

Bibliografía:

Rota, Alesson R.; Nicodemo, Thiago L. (2023). Arquivos pessoais e redes sociais: o x construído como documento histórico. Estud. Hist. 36(79), 44-67. 10.1590/S2178-149420230204.

Rota, Alesson y Bonaldo, Rodrigo (2023). On the measure of synchronization: human experience and artificial agents on Twitter. *SciELO Preprints.* https://doi.org/10.1590/SciELOPreprints.7339

Rogers, Richard (2019). *Doing digital methods.* SAGE.

2.1 X (antiguo Twitter)

Generalidades

Hay plataformas como X que, dependiendo del caso académico, permiten acceso a la API. Para la investigación de GUMELAB hay que destacar que durante el proyecto ya se han cambiado los requisitos para tener acceso a la API. Cuando en marzo de 2020 presentamos el proyecto GUMELAB para ser financiado, pensamos que era necesario comprarle los datos, en aquel entonces, a Twitter por alrededor de 20.000 euros. En septiembre de 2021, X anunció que se podría usar una API académica[24] siempre y cuando el proyecto declarara para qué se usarían los datos. Esto hizo las búsquedas más fáciles y más económicas. Con la adquisición de la empresa por Elon Musk en octubre de 2022, el trabajo con X se ha hecho más complicado y más inestable, ya que no está claro cómo afectará los procesos de extracción de datos para proyectos de investigación en el futuro. El cambio institucional de

24 https://developer.twitter.com/en/use-cases/do-research/academic-research

X afectará sobre todo el acceso de datos, y tendrá una consecuencia directa en los parámetros de su API.

Para X establecimos distintas palabras clave con respecto al nombre de la telenovela o serie, y algunas variaciones para las cinco producciones de televisión, como, por ejemplo, "Escobar telenovela", "Escobar novela", "Patrón del mal", "Patrón del mal, telenovela", "Patrón del mal telenovela", que suelen ser diferentes formas de referirse a la telenovela colombiana de 2012. En el caso de la producción chileno-alemana *Dignity*, fue necesario buscarla bajo su título original en inglés ("Dignity serie"), y bajo la traducción del título en español ("Dignidad serie").

En el proceso descubrimos que era un riesgo buscar exclusivamente por el nombre de la producción. En muchos casos, el nombre atañe a otro tipo de información que no se relaciona con la investigación. Así que, para evitar extraer datos innecesarios, cruzamos el nombre de la telenovela con diferentes palabras clave. Estas palabras clave estaban relacionadas con los contenidos históricos que representan las telenovelas y series, asumiendo que podrían ser una temática discutida en la red (por ejemplo, plebiscito, Pinochet, víctimas, paramilitarismo, Luis Carlos Galán Sarmiento, Paul Schäfer, etc.). Después pensamos palabras clave que mostraran alguna forma de valoración en general. En este sentido, una búsqueda previa realizada manualmente por el equipo de humanidades fue clave para identificar qué palabras eran empleadas en combinación con la recepción, y que denotaran algún tipo de valoración. Recopilamos palabras como *historia*, *corrupción*, *memoria*, *derechos humanos*, *comunismo*, *izquierda*, *derecha*..., así como expresiones del lenguaje coloquial del país, tales como *bombazos* o *berraca*, en referencia al caso colombiano.

Para obtener datos que representaran una recepción transnacional, recopilamos todos los países donde se emitió la producción o desde donde podrían surgir usuarios de Twitter que postearan comentarios. Expresiones como "saludos desde + país" o "soy + venezolano" también las consideramos, ya que podrían indicar una recepción transnacional.

Buscando tags, hashtags *y palabras clave*

Para llegar a la información más exacta de las telenovelas y series cruzamos los nombres de las producciones con cuentas oficiales, como, por ejemplo, @PatronDelMalTV o @Los80_serie, igual que con las cuentas oficiales de los canales, como @CaracolTV, @DynamoCine, y de personajes que estaban en el desarrollo de las producciones (guionistas, productores y camarógrafos), como @cbrancato86 (creador y productor de *Narcos*) o @mwoodmontt (productora de *Los 80*). También hicimos una lista de las cuentas de los actores y actrices, ya que ellos promueven las producciones en la red: @SoyAndresParra (actor de Pablo Escobar en EPM), @paugaitan (actriz en *Narcos*) o @LORETOARAVENA (actriz en *Los 80*). Como los fans o los propios canales crean cuentas para los personajes ficticios, esas cuentas también las sistematizamos para llegar a datos útiles: @JuanHerera80s (personaje principal en *Los 80*) o @Chili y @yomellamopablo (personajes de *El Patrón del Mal*). En algunos casos también seleccionamos cuentas creadas para mostrar aversión en contra de una telenovela, como pasó con *Tres Caínes*, que tuvo una campaña en contra de su emisión (@Noen3caines). Asimismo coleccionamos cuentas de fans, porque estimamos que ahí habría mucha discusión sobre la producción (@narcosgallery).

Cada telenovela y serie generó un conjunto de *hashtags* (#) o etiquetas que conectaban la comunidad de personas discutiendo y opinando sobre ellas. De este modo, aunque no todos los usuarios se referían a las cuentas oficiales de los personajes, ya por el solo hecho de crear o emplear los mismos *hashtags* (#) quedaban conectados. Es por esto que los *hashtags* son tan importantes para buscar informaciones relevantes en Twitter (X). En ese sentido, capturamos todos los *hashtags* que se referían al nombre de la telenovela o serie (como #Narcos, #NarcosNetflix, #NarcosColombia), o que se referían a las compañías o personas que trabajaban en el set (#WoodProducciones, #BorisQuercia, productores de *Los 80*, o #joyn, canal de emisión de Dignity).

También podía ser útil la combinación de *hashtags* con personajes ficticios, como #DonGenaro (*Los 80*). En algunos casos también pensamos en *hashtags* que podrían estar relacionados con la crítica a estas producciones

(#NoTresCaines, #Esomeduele), con el apoyo a ellas (#YoApoyoTresCaines) o relacionados con movimientos políticos (#ChileDesperto). Este ejercicio fue especialmente productivo y necesario, sobre todo cuando, por ejemplo, se emitió por segunda vez en televisión abierta la serie chilena *Los 80* en el año 2019, coincidiendo con el estallido social[25] que tuvo lugar en aquel país por esa época.

El problema de encontrar datos (útiles): Pablo Escobar. El Patrón del Mal *en la red*

Cuatro de las cinco telenovelas/series seleccionadas tienen un problema muy concreto, el cual no habíamos contemplado: su título no es muy específico para obtener datos, lo que requiere una limpieza muy detallada, que, a su vez, consume mucho tiempo. Buscar "Pablo Escobar" en la red no es una muy buena idea, porque muchos usuarios se refieren a Pablo Escobar en la vida real, en otras películas, o para hablar de valores morales, etc. Después de ver y analizar los datos para evaluar manualmente la búsqueda hecha, nos dimos cuenta de que Escobar también es el nombre de una pequeña ciudad situada al nordeste de la provincia de Buenos Aires, en Argentina. Constatamos que los habitantes de "Belén de Escobar" (ese es el nombre completo de la ciudad), son muy activos en X. Algo parecido nos pasó con "Dignidad/Dignity", una palabra que se emplea mucho en las protestas sociales que se organizan en las redes sociales; "los 80", que se refiere muchas veces a la música o a la moda de esa década, pero no necesariamente a la serie chilena. "Narcos" ha tenido problemas parecidos a los de "Escobar", ya que la palabra "narcos" también se usa para designar a capos de la droga. La

25 Las protestas en Chile de 2019, conocidas como el estallido social, comenzaron como una respuesta al aumento del precio del metro en Santiago, anunciado en octubre de 2019, y rápidamente se convirtieron en un movimiento más amplio contra la desigualdad y los problemas estructurales en el país. Las masivas manifestaciones, marcadas por su intensidad y alcance nacional, dieron lugar a demandas de transformaciones políticas y sociales profundas, incluyendo el intento de redactar una nueva Constitución para reemplazar la heredada de la dictadura de Pinochet.

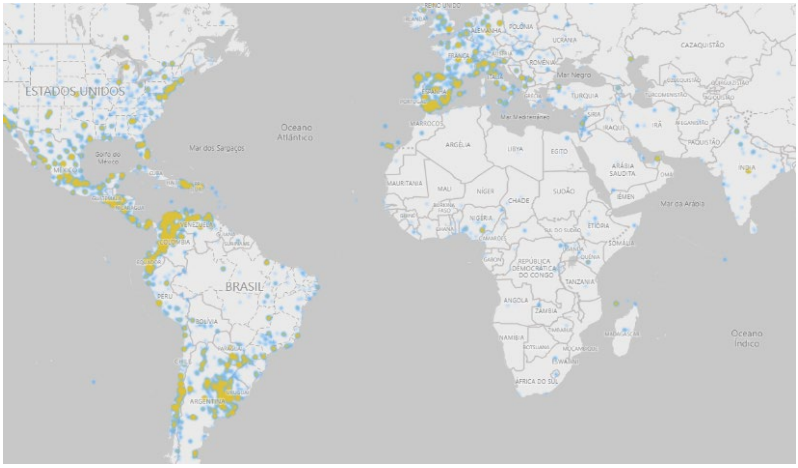
única telenovela que ha tenido un nombre muy específico fue *Tres Caínes*, donde la limpieza necesaria fue mínima.

Aplicación de los conceptos de minería de datos y procesamiento del lenguaje natural

La minería de datos es el proceso de explorar y analizar grandes conjuntos de datos en busca de la identificación de patrones y tendencias. Algunos ejemplos son el análisis de sentimientos, que consiste en identificar y clasificar las emociones expresadas en textos; la detección de tendencias, para saber cuáles fueron las palabras más utilizadas y sus articulaciones; la identificación de influenciadores, quienes son vectores de transmisión en la red. Veamos un ejemplo a través de Twitter (X). Después de descargar todos los datos (del apartado 1.1 al 1.7), buscamos identificar si existía alguna comunidad internacional que pudiera debatir sobre las telenovelas y series, como una forma de conocer mejor nuestro documento histórico. El total de 67.370 cuentas, distribuidas por varios países del mundo, parecía sugerir que sí, dado el gran volumen de datos. Para construir este gráfico, bastó con representar en un mapa de calor la entidad *place*, presente en la base de datos extraída de esta red social. [Ver ejemplo con aplicación de código aquí].

Obsérvese que en la **figura 2,** para confirmar nuestra interrogante sobre la comunidad internacional, aplicamos un método de detección de tendencias. En la figura a continuación verificamos todas las cuentas que comentan más de una serie, de modo que estamos buscando usuarios que tienen la costumbre de ver telenovelas/series; hay perfiles de varios tipos, que comentan dos, tres y hasta cuatro de las cinco producciones seleccionadas como objeto de estudio. [Ver ejemplo con aplicación de código aquí].

Figura 2



Nota: Colores cálidos (amarillo), grandes concentraciones; colores fríos (azul), pequeñas concentraciones.

Fuente: Elaboración propia.[26]

Desde un punto de vista analítico, este tipo de gráfico puede describirse como un gráfico no dirigido, donde los nodos representan entidades (en este caso, autores y las cinco producciones) y las aristas representan las relaciones entre estas entidades (aquí cuantificadas por el número de comentarios de los autores sobre las telenovelas y series). La proximidad de ciertas telenovelas y series entre sí puede ser un artefacto de la técnica de diseño del gráfico (como el algoritmo de Fruchterman-Reingold o la fuerza dirigida por Barnes-Hut),[27] que intenta optimizar la posición de los nodos

26 Varias informaciones recogidas en X, como *place, descripción del perfil* y *nombre de usuario* son espontáneas, es decir, no representan necesariamente datos verdaderos.

27 El algoritmo Fruchterman-Reingold se utiliza para la visualización de grafos en dos o tres dimensiones. Fue desarrollado por Thomas Fruchterman y Edward Reingold en 1991. Este algoritmo trabaja con la idea de que los nodos del grafo son como

para minimizar la superposición de las aristas y distribuir los nodos uniformemente en el espacio. Si las informaciones aparecen cerca una de otra, esto indica que comparten muchos autores en común o que un conjunto de autores interactúan frecuentemente con ambas producciones. Pero, claro, todo depende de la configuración utilizada. [Ver ejemplo con aplicación de código aquí].

En resumen:

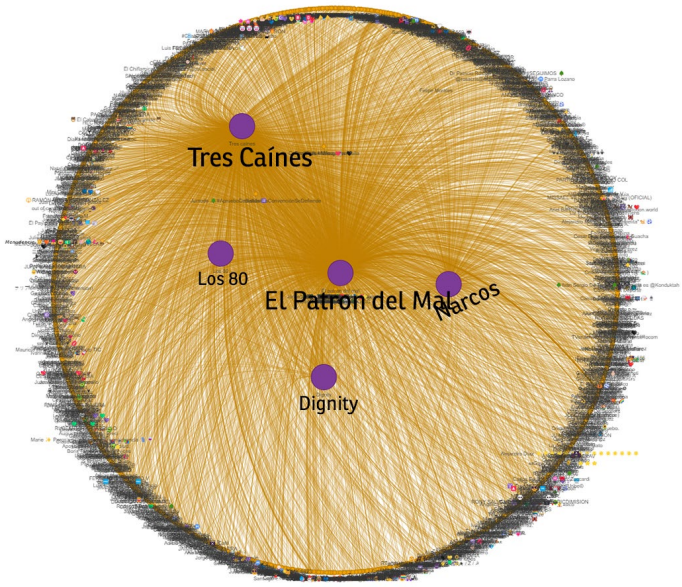
- **Nodos grandes y coloridos** representan telenovelas y series. El tamaño de estos nodos es uniforme, destacando la importancia equivalente de cada producción dentro del conjunto de datos analizados. El color puede utilizarse para facilitar la identificación visual de estos nodos importantes dentro de la red.
- **Nodos más pequeños** representan autores individuales. Estos son menos prominentes visualmente, lo que puede indicar que el enfoque del análisis está en las telenovelas/series, y no en la identidad de los autores.
- **Aristas,** las líneas que conectan los nodos, simbolizan las interacciones entre autores y producciones. El grosor de cada arista es proporcional al número de comentarios que un autor hizo en una telenovela o serie

partículas cargadas eléctricamente que se repelen entre sí y están conectadas por resortes que las atraen. El objetivo es encontrar una disposición espacial de los nodos en la que las fuerzas de repulsión y atracción estén en equilibrio, lo que resulta en una representación gráfica que minimice el cruce de aristas y haga que el grafo sea más legible. El algoritmo Barnes-Hut se utiliza para acelerar el cálculo de fuerzas en algoritmos de visualización de grafos, como el Fruchterman-Reingold, cuando hay un gran número de nodos y aristas en el grafo. Fue desarrollado por Josh Barnes y Piet Hut en 1986 para simulaciones de n-cuerpos en astronomía, pero también se puede aplicar a grafos. El algoritmo divide el espacio en celdas jerárquicas, creando un árbol de *quadtree* (para 2D) u *octree* (para 3D) para representar la distribución espacial de los nodos. Esto permite que las fuerzas entre grupos de nodos distantes se aproximen de manera eficiente, lo que ahorra tiempo de cálculo en comparación con el enfoque de calcular todas las interacciones individualmente. Utilizamos estos dos ejemplos conocidos para que el lector o la lectora tenga en cuenta que, cuando creamos gráficos a partir de datos, la construcción de la visualidad depende de características propias del entorno representacional, incluyendo la presencia de ausencia, distorsiones de las líneas y elecciones estáticas.

específica, con aristas más gruesas indicando un mayor número de interacciones.

Desde un punto de vista narrativo, la **figura 3** ilustra la robusta interacción de más de 67.000 cuentas únicas con las telenovelas y series investigadas. Ofrece una visión sobre la extensión y profundidad de la audiencia y sobre el compromiso en torno a estas producciones, permitiéndonos medir su impacto cultural en el público. La escala de compromiso vista aquí enfatiza la relevancia de las telenovelas y series como vehículos pertinentes para la narración de historias, generando discusiones y reflexiones en una base de fans comprometida.

Figura 3



Nota: Relación entre los usuarios y las telenovelas y series comentadas. *Pablo Escobar. El Patrón del Mal* y *Tres Caínes* fueron las series más comentadas, seguidas por *Narcos, Los 80* y *Dignity.*

Fuente: Elaboración propia.

Otro tipo de *análisis de tendencias* es visualizar los principales usuarios de Twitter (X) que publican algún contenido relacionado con las telenovelas y series. A través de este sencillo ejercicio heurístico, como lo muestra la figura, podemos identificar los siguientes patrones de cuentas: canales de difusión, cuentas de producción, personas comunes que exhiben las telenovelas y series, cuentas de usuarios comunes y cuentas falsas con los nombres de personajes famosos.

- **Cuentas de difusión.** Estas cuentas engloban tanto las creadas específicamente para las telenovelas o series como las pertenecientes a noticieros, programas de entretenimiento e individuos que crean y comparten contenido en dicha red abordando una variedad de temas para alcanzar a un amplio público e interactuar con los espectadores.
- **Cuentas de las empresas de producción.** Estos perfiles son creados por empresas de producción como Netflix y canales de televisión con el propósito de difundir información, tráileres e interactuar con los fans; sirven de representación en línea de sus series, películas u otros proyectos mediáticos.
- **Cuentas de personas comunes.** Cuentas de los individuos corrientes sin presencia oficial en los medios, quienes consumen contenido e interactúan como espectadores, seguidores y participantes en las discusiones relacionadas con las producciones seleccionadas.

En el siguiente gráfico buscamos cuantificar el análisis heurístico sobre los tipos de cuentas identificadas anteriormente. Para ello seleccionamos las primeras 500 cuentas que más publicaron sobre las telenovelas y series, y analizamos caso por caso, perfil a perfil, qué cuentas encajan en las categorías identificadas. En este universo ordenado según el criterio de mayor cantidad de publicación y limitado a un número de 500 (elegido aleatoriamente solo para demostración), vemos el importante papel de las cuentas de difusión en la promoción del debate en las redes sociales. Cabe destacar que esto significa solamente que ciertas cuentas de empresas o cuentas de difusión publican más que los usuarios comunes que más publican. Sin embargo, cuando miramos la totalidad de los datos, los usuarios comunes son la mayoría.

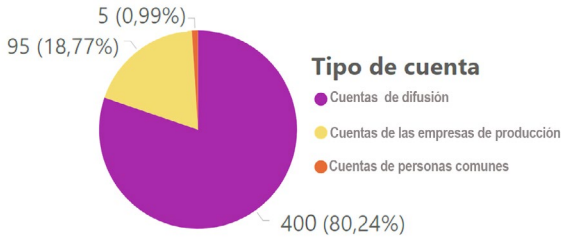
Figura 4

Autor	Publicaciones	Likes	Retweet	Reply	Telenovela
Escobar, el Patrón	935	1435	7403	2850	Escobar patron del mal
Nico Lorenzo ⭐⭐⭐	756	72	320	47	Escobar patron del mal
Caracol Televisión	677	522	3635	1697	Escobar patron del mal
Zona Norte Hoy	511	738	637	491	Escobar patron del mal
Hamilton	467	121	306	32	Escobar patron del mal
InfoBAN Noticias	437	310	317	92	Escobar patron del mal
Canal Provincial	425	33	35	6	Escobar patron del mal
Zucaritas TV	392	17	443	31	Escobar patron del mal
Rosana Piñeiro	379	3	12	0	Escobar patron del mal
#PasiónDeportiva1079	362	11	8	1	Escobar patron del mal
Canal RCN	319	293	693	865	Tres Caines
Andrés Parra	313	9235	8636	4921	Escobar patron del mal
Se cortó la luz!	307	124	185	56	Escobar patron del mal
Semanario Regional	305	35	131	11	Escobar patron del mal
PLUS RATING	267	46	87	15	Escobar patron del mal

Nota: Proporción de clasificación en las categorías de cuentas de personas comunes, cuentas de las empresas de producción y cuentas de difusión; muestra de las primeras 500 cuentas que más publicaron al respecto.

Fuente: Elaboración propia.

Figura 5



Nota: Proporción de clasificación en las categorías de cuentas de personas comunes, cuentas de las empresas de producción y cuentas de difusión; muestra de las primeras 500 cuentas que más publicaron al respecto.

Fuente: Elaboración propia.

Bibliografía:

Burgess, Jean y Baym, Nancy K. (2020). *Twitter. A biography.* New York University Press.

Pfaffenberger, Fabian (2016). *X als Basis wissenschaftlicher Studien. Eine Bewertung gängiger Erhebungs- und Analysemethoden der X-Forschung.* Springer VS.

Weller, Katrin; Bruns, Axel; Burgess, Jean; Mahrt, Merja y Puschmann, Cornelius (2014). *X and society.* Peter Lang (digital formations, vol. 89).

2.2 Facebook

Dentro de las posibles fuentes de datos que consideramos para GUMELAB, planteamos la opción de obtener acceso a los datos de los usuarios de Facebook[28] como parte de la investigación, ya que esta plataforma alberga numerosos grupos de discusión relacionados con diversas telenovelas y series, donde la interacción y las discusiones son constantes y significativas. Sin embargo, identificamos dificultades que obstaculizaron la obtención eficiente de dicho acceso. Para llevar a cabo esta tarea, se requiere lo siguiente:

- **Creación de una aplicación para conectarse con la API de Facebook, registro de esta y obtención de autorización.** Para acceder a los datos de Facebook (de los usuarios), se requiere crear una aplicación en la plataforma, obtener aprobación por parte de la plataforma mediante un registro y la autorización individual de cada usuario (para acceder a sus datos). Este proceso implica no solo el tiempo necesario para su desarrollo y aprobación, sino además una interacción directa con cada usuario, lo que puede resultar complicado y requerir una considerable inversión de tiempo.

28 La documentación asociada a las API de Facebook está en https://developers.facebook.com/docs/graph-api y https://developers.facebook.com/docs/facebook-login/guides/access-tokens

- **Gestión de *tokens* de autorización.** Una vez obtenida la autorización de cada usuario, es necesario generar y almacenar los *tokens* de autorización correspondientes. Estos *tokens* son utilizados para acceder a los datos del usuario, y deben ser guardados en una base de datos para su posterior uso. Mantener estos *tokens* actualizados implica un proceso de gestión constante, ya que tienen una duración limitada de tiempo, generalmente hasta 60 días.
- **Mantenimiento de los *tokens* actualizados.** Dado que los *tokens* de autorización tienen una vigencia limitada, es necesario implementar un servicio que se encargue de mantenerlos actualizados. Esto implica un esfuerzo adicional en términos de desarrollo y mantenimiento técnico para garantizar que sean renovados oportunamente y no se pierda el acceso a los datos.
- **Costo y tiempo asociados a la publicidad y registro de usuarios.** Para obtener un número significativo de usuarios que autoricen el acceso a sus datos de Facebook, se requeriría llevar a cabo campañas publicitarias en la misma plataforma para promover el registro y la autorización. Esto implica un costo adicional en términos de inversión publicitaria, así como la necesidad de dedicar tiempo y recursos para gestionar el proceso de registro de los usuarios.

Considerando todas estas dificultades, concluimos que el acceso a los datos de Facebook para el estudio no resultaba viable ni eficiente en términos de costo y tiempo. Era recomendable buscar fuentes alternativas de datos que estuvieran más disponibles y más accesibles. Al explorar otras fuentes de datos, pudimos encontrar información valiosa para el estudio sin los desafíos y limitaciones identificados.

A fines de 2023, tras las crecientes presiones ejercidas por la Ley de Servicios Digitales (Digital Services Act, DSA) de la Unión Europea, Meta Platforms Inc. tomó la decisión de abrir el acceso a su biblioteca de contenidos de Facebook e Instagram para investigadores universitarios y ONG. Esta legislación, que enfatiza la necesidad de transparencia en las plataformas digitales y la protección de datos personales, especialmente en el ámbito de la investigación, ha sido un catalizador para cambios significativos en la forma

en que operan los gigantes tecnológicos. La disponibilidad de una interfaz gráfica y una API por parte de Meta es un paso importante. Sin embargo, es crucial mantener una perspectiva crítica sobre esta medida. Aunque parece una concesión generosa para el avance de la investigación científica y social, no se debe ignorar que tal acción puede estar más motivada por una necesidad de cumplimiento regulatorio que por un compromiso genuino con la transparencia y el avance del conocimiento. Además, la necesidad de solicitar acceso a través del Consorcio Interuniversitario de Investigación Política y Social (ICPSR), de la Universidad de Michigan, implica cierto grado de exclusividad, potencialmente limitante con criterios desconocidos. Esta iniciativa, aunque prometedora, plantea preguntas sobre la autonomía e independencia de la investigación. En un entorno donde los datos son controlados por entidades corporativas, como la principal materia prima del siglo XXI, los investigadores pueden encontrarse en una posición delicada. Pero no solo ellos: en última instancia, toda la sociedad es la que pierde.

Bibliografía:

Burkhardt, Hannes (2021). *Geschichte in den Social Media: Nationalsozialismus und Holocaust in Erinnerungskulturen auf Facebook, X, Pinterest und Instagram. Beihefte zur Zeitschrift für Geschichtsdidaktik.* V&R unipress; Vandenhoeck & Ruprecht GmbH & Co.KG. 10.14220/9783737012515.

Franz, Daschel; Marsh, Heather Elizabeth; Chen, Jason I. y Teo, Alan R. (2019). Using Facebook for qualitative research: a brief primer. *Journal of Medical Internet Research,* 21(8), e13544. 10.2196/13544.

Hammerschmidt, Peter; Sagebiel, Juliane; Hill, Burkhard y Beranek, Angelika (2018). *Big Data, Facebook, X & Co. und Soziale Arbeit. 1. Auflage. Aktuelle Themen und Grundsatzfragen der Sozialen Arbeit.* Beltz Juventa.

2.3 YouTube

Generalidades

YouTube se lanzó en 2005 como plataforma para compartir videos. Al principio no se hacía mucho hincapié en la recopilación de datos, ya que el objetivo principal era permitir que la gente compartiera y viera videos. En 2007 YouTube lanzó su primera API, que permitía a los desarrolladores interactuar mediante programación con la plataforma. Esto allanó el camino para la recopilación automatizada de datos, por cuanto ahora los desarrolladores podían utilizar la API para extraer información sobre videos, canales y usuarios. A partir de entonces se crearon una serie de funciones dentro de la plataforma para permitirles a los creadores de contenidos ver estadísticas detalladas sobre sus videos con el fin de comprender mejor a la audiencia receptora. Se trata de un proceso de transformación en el que la plataforma ya no se limita a ser únicamente un depósito de videos, sino que adquiere la dimensión de un videoblog. En este entorno se desarrolla una red social donde se llevan a cabo diversas formas de interacción, como comentarios y respuestas. Además, se proporcionan espacios designados para que el creador del canal interactúe con el público a través de encuestas, publicación de imágenes o texto. Gracias a esta doble función, en la cual la plataforma permite tanto la visualización de contenidos (en el caso específico de GUMELAB, telenovelas y series) como la interacción activa, y considerando la amplia utilización global de YouTube, esta plataforma se presenta como una de las fuentes principales con las que GUMELAB trabaja.[29]

La historia de las herramientas de recopilación de datos de YouTube está intrínsecamente ligada al desarrollo de la propia plataforma YouTube y su conjunto de API. La API de YouTube ha sido el principal medio por el que

29 Ver contenidos en la plataforma de YouTube e interactuar con otros usuarios mediante comentarios no es considerado *second screening,* ya que normalmente se trata de una misma pantalla donde al tiempo se ve el contenido y se comenta.

los desarrolladores recopilan datos de YouTube.[30] Actualmente, con la API de YouTube (v3) es posible recopilar datos como título del video (*title*), descripción del video (*description*), etiquetas asociadas (*tags*), fecha y hora de publicación del video (*publishedAt*), ID del canal que publicó el video (*channelId*), título del canal que publicó el video (*channelTitle)*, ID de categoría del video (*categoryId*), si el video es una emisión en directo (liveBroadcastContent), idioma predeterminado del video (*defaultLanguage*), número de visionados del video (*viewCount*), número de *me gusta* del video (*likeCount*), número de *no me gusta* del video (*dislikeCount*), número de veces que el video se ha añadido como favorito (favoriteCount), número de comentarios sobre el video (*commentCount*) y lista de comentarios (*commentThreads. list*).[31] [Ver ejemplo con aplicación de código aquí].

Bibliografía:

Burgess, Jean y Joshua Green (2018). *YouTube: online video and participatory culture.* Digital Media and Society Series. Polity Press.

Fontoura, Odir (2020). Narrativas históricas em disputa: um estudo de caso no YouTube. *Estudos Históricos (Rio de Janeiro),* 33, 45-63.

30 El acceso a la API es público y gratuito, pero dependiendo de la cantidad de datos se puede demorar su recolección. Durante el proyecto GUMELAB YouTube anunció que proveería un acceso a una API académica (https://research.youtube/). Después de una evaluación interna decidimos que supuestamente la diferencia de la API académica y la API normal era mínima, entonces no valía la pena cambiar los procesos de la minería de datos.

31 Entre paréntesis están las entidades extraídas de YouTube, las cuales constituyen la manera mediante la cual obtenemos y estructuramos la información en nuestras bases de datos.

Prefiltrado y limpieza

Para llevar a cabo la investigación en YouTube, el primer paso consistió en identificar los videos relevantes de los cuales extraeríamos la información. En el caso de GUMELAB, buscábamos videos que mostraran capítulos completos de las producciones seleccionadas para nuestra investigación, o aquellos videos que seleccionaran las mejores escenas, con la expectativa de encontrar comentarios útiles para la investigación. Con este propósito empleamos tres funciones distintas de la API de YouTube.

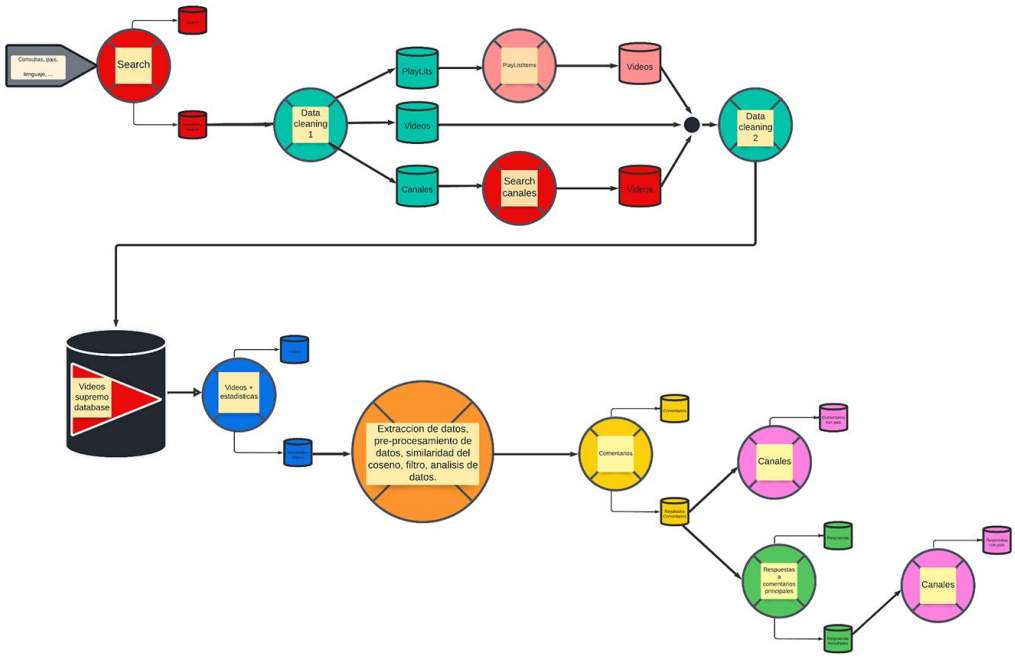
En primer lugar, una función para buscar listas de reproducción cuyos nombres o descripciones coincidieran con los criterios de búsqueda. Esto permitía extraer los identificadores (IDs) de los videos contenidos en dichas listas. Posteriormente empleamos otra función para encontrar los IDs de los videos que también cumplían con los criterios de búsqueda, pero que no formaban parte de una lista de reproducción.

Con todos los IDs recolectados procedimos a descargar los comentarios de cada video mediante la tercera función de la API de YouTube. Sin embargo, surgió un problema después de la descarga de los datos: algunas listas de reproducción no siempre contenían los videos de la telenovela/serie que prometían (por ejemplo, nombre de telenovela + capítulos completos). Esto se debía a que los videos podían ser retirados por la plataforma por problemas de derechos de autor; o bien las listas podían incluir videos que no coincidían con su título o descripción (por ejemplo, en una lista sobre la serie *Dignity,* los usuarios agregaron videos de música). Una característica distintiva de la lista de reproducción, en comparación con Twitter (X), es que permite organizar el caos generado por algoritmos a través de palabras clave. Estas palabras clave se asemejan a las utilizadas en otras búsquedas realizadas en Twitter (X) y Google.

Otro problema que surgió tuvo que ver con el acceso a los videos y, por ende, a sus comentarios desde distintas partes del mundo. Las búsquedas hechas por el equipo de ciencias de datos, ubicado en Colombia, solo estaban disponibles en Colombia y no en Alemania, donde se ubica el equipo de humanidades.

A partir de los títulos y descripciones de los videos encontrados mediante el motor de búsqueda, hicimos la extracción de emoticones, URL, menciones y *hashtags*. Para llegar a los videos útiles realizamos un preprocesamiento

Figura 6.



Nota: Flujo de trabajo para YouTube.
Fuente: Elaboración propia.

Figura 7.

Nota: *Wordcloud* de comentarios bajo *videos relevantes* (ejemplo, *Pablo Escobar. El Patrón del Mal*).

Fuente: Elaboración propia.

de los datos. Gracias al modelado de temas[32] pudimos establecer filtros para los diferentes casos de estudio, facilitando la agrupación de videos según el tema más relevante para cada uno, con técnicas como TD-IDF (término de frecuencias-documento de frecuencia inversal), bolsas de palabras (*wordclouds*) y bigramas[33] y trigramas.[34] Posteriormente seleccionamos los videos relacionados con tópicos relevantes, reduciendo así la presencia de comentarios ajenos a nuestra investigación. Dado que YouTube cuenta con tres *endpoints* diferentes para videos, comentarios y respuestas a comentarios, se hace necesaria la descarga en tres etapas. Después establecimos la relación entre ellos mediante los IDs de video y los IDs de comentario. Y finalmente

32 Para más detalles sobre la limpieza de datos, el numeral 1.3.

33 Combinación de dos palabras que vienen juntas, como "Andrés_Parra".

34 Combinación de tres palabras juntas, como "Luis_Carlos_Galán".

procedimos con el proceso de etiquetado, clasificando los comentarios como útil o *no útil*. [Ver ejemplo con aplicación de código aquí]

Gestión de datos

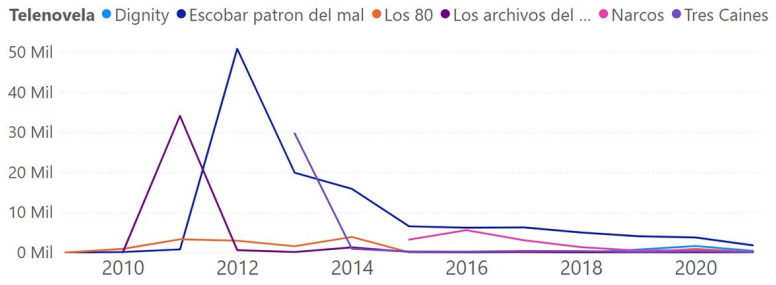
El total de datos de YouTube asciende a 289.000 líneas, con información sobre comentarios, respuestas, visualizaciones y todas las categorías anunciadas en la generalidad. En primer lugar, el proceso de recogida y tratamiento, donde todas las bases de datos de YouTube se unificaron de forma estructurada en una única tabla publicada en Data Repositories (enlace).

Detección de patrones: línea de tiempo y NLP

Como ha sido mencionado en este manual, el procesamiento de lenguaje natural (NLP) es un campo de la inteligencia artificial que se centra en la interacción entre computadoras y lenguaje humano, permitiendo que las máquinas procesen textos. Este tipo de procesamiento puede ser utilizado de diversas formas, incluyendo la limpieza de datos, que es un proceso esencial en la preparación de conjuntos de información para análisis. Al tratar con datos textuales, es común encontrar *ruido*, como errores de tipeo, caracteres especiales no deseados o incluso palabras irrelevantes. En este contexto se pueden aplicar técnicas de limpieza de datos para eliminar estas inconsistencias y preparar el texto de manera adecuada para un análisis posterior.

A continuación, un ejemplo de NLP aplicado al análisis de comentarios en YouTube. La siguiente figura ilustra la cantidad de comentarios allí sobre las cinco producciones a lo largo de los años. La mayoría de los videos acumulan más comentarios a medida que se acercan al presente, especialmente después de 2021. Este fenómeno podría tener diversas explicaciones, como la eliminación de videos con derechos de autor o el crecimiento de una cultura digital de videos en los últimos años. Este aumento, a su vez, podría vincularse al surgimiento de fenómenos como *youtubers* e *influencers*, además de la pandemia de COVID-19, que intensificó los procesos anteriores.

Figura 8.

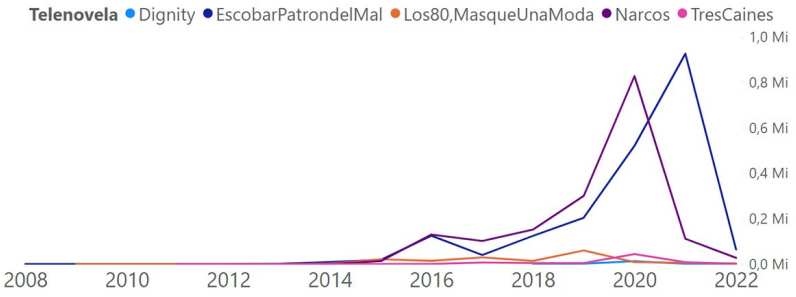


Nota: Publicaciones sobre las cinco telenovelas y series en YouTube.

Fuente: Elaboración propia.

Se puede comparar la repercusión de las mismas telenovelas y series con Twitter (X). La siguiente figura destaca que las publicaciones sobre estas producciones en Twitter (X) se configuran precisamente por el fenómeno denominado *second screening*, explicado en el glosario. En este contexto, el público comenta la telenovela/serie simultáneamente mientras la está viendo, en contraste con YouTube, donde la interacción se produce posteriormente.

Figura 9.



Nota: Publicaciones sobre las cinco telenovelas y series en X.

Fuente: Elaboración propia.

Para interpretar cualitativamente el gráfico de diversas formas, una de las alternativas es aislar las fechas donde se acentúan las curvas para cada producción.

Puede ver ejemplos en el 4.1, que analiza la recepción de la serie *Pablo Escobar. El Patrón del Mal* (EPM) a través de datos recopilados de YouTube. Realizamos 38 distintas, para un total de 154.161 datos, incluyendo *likes* y comentarios en los videos. Analizamos el *engagement* de la serie evaluando la relación entre *likes* y vistas, los capítulos que más gustaron y fueron comentados, y la distribución temporal de los comentarios. Además, examinamos la distribución geográfica de los comentarios, destacando el interés internacional en la serie; en el apartado 4.2 hay un ejercicio similar con datos de X. Asimismo, usamos procesos de minería de datos para localizar ejemplos cualitativos de publicaciones. En ese apartado se observa que la recepción y los temas discutidos variaban según el año y el contexto de emisión de la telenovela/serie. [Ver ejemplo con aplicación de código aquí]

2.4 Google Search

Como se ha mencionado anteriormente, planteamos el uso de artículos de prensa y blogs personales como parte de los datos para nuestro estudio; una de las herramientas más fáciles y rápidas para acceder a dichos documentos es la herramienta de búsqueda personalizada de Google: Google Search. Con esta es posible definir una serie de parámetros de búsqueda como idioma, geolocalización, palabras clave, páginas y palabras excluidas, entre otros.[35]

Generalidades

La herramienta Google Search, al ser utilizada para la recolección de datos en investigaciones académicas y de mercado, ofrece una ventana única al

35 Para ver todos los posibles parámetros de la API, la documentación se encuentra en https://developers.google.com/custom-search/v1/overview?hl=es-419

vasto mundo de la información en línea. No obstante, es importante recordar que los resultados obtenidos están influenciados por algoritmos específicos de Google, que priorizan ciertas páginas según diversos factores, como la relevancia, la popularidad del sitio y las prácticas de SEO (*search engine optimization*). Esto puede introducir un sesgo en los datos recolectados, lo que requiere una consideración crítica de su uso y significado.

Prefiltrado y limpieza

Antes de analizar los datos, es crucial realizar una etapa de prefiltrado y limpieza. Esta fase implica la eliminación de información irrelevante o redundante, asegurando que los datos finales sean precisos y útiles para el propósito de estudio. Este proceso también debe considerar la eliminación de *ruido* en los datos, como pueden ser páginas web irrelevantes o duplicadas, para no distorsionar los resultados de la investigación.

Gestión de datos

La gestión de los datos recopilados implica su almacenamiento, organización y análisis. Herramientas como Google Sheets o bases de datos personalizadas pueden ser utilizadas para organizar la información. El análisis posterior de estos datos, utilizando métodos estadísticos o de aprendizaje automático, puede revelar patrones y tendencias que serían imperceptibles a simple vista.

Por ejemplo, una investigación sobre la percepción pública de un tema específico puede utilizar Google Search para identificar las fuentes de información más comunes y cómo estas influyen en la opinión pública. Otro ejemplo podría ser el análisis de la competencia en el mercado digital, para identificar las estrategias de SEO más efectivas empleadas por los competidores.

Es fundamental considerar la naturaleza y las limitaciones de los datos recopilados a través de Google Search. Los resultados de búsqueda están influenciados por el comportamiento del usuario y los algoritmos de Google,

lo que puede llevar a una representación sesgada de la información. Además, la dinámica de la web está en constante cambio, lo cual significa que los datos son inherentemente efímeros y deben ser contextualizados dentro de un marco temporal específico.

Este método de recolección de datos ofrece tanto oportunidades como desafíos. Permite el acceso a una gran cantidad de información, pero también exige un análisis crítico y reflexivo para evitar conclusiones erróneas o superficiales. Los investigadores y las investigadoras deben ser conscientes de estos desafíos y aplicar metodologías rigurosas para validar e interpretar sus hallazgos.

En conclusión, el uso de Google Search como herramienta de recolección de datos para el análisis ofrece un potencial significativo, pero debe ser abordado con un entendimiento crítico de sus limitaciones y sesgos inherentes. La combinación de una metodología sólida y un enfoque crítico es esencial para garantizar que los resultados sean válidos y significativos.

[Ver ejemplo con aplicación de código aquí]

3. Entrenamiento de datos, modelos agentes cognitivos

Para comprender mejor cómo las narrativas de las telenovelas y series de este estudio resuenan en la sociedad, recopilamos comentarios de espectadores en las plataformas de Twitter (X) y YouTube. Estos comentarios, al ser recolectados de forma masiva, presentan un reto para su estudio de manera manual; por ello propusimos estrategias de clasificación mediante modelos de aprendizaje automático, con los cuales es posible proporcionar una visión única y complementaria de las percepciones del público. Sin embargo, la tarea de clasificación se presenta desafiante debido a la complejidad de las narrativas y a la sutil diferencia entre categorías de análisis: FP (formación política), IM (imágenes de la memoria), CH (conciencia histórica) y NR (no relevante), que en adelante serán nombradas como *clases asignadas*. *No relevante* se refiere a todos los datos que, pese a estar relacionados con las telenovelas y series que investigamos, no aportan información a ninguna de las tres categorías anteriormente mencionadas.

La tarea de clasificación, *codificación categorial* en el lenguaje de la investigación cualitativa, es una de las tareas de aprendizaje automático supervisado; por tanto, era necesario proveer algunas muestras como ejemplos de las tres clases, para lo cual el equipo de humanidades etiquetó ~500 ejemplos de cada telenovela/serie para cada una de las fuentes (Twitter/X y YouTube); con esto fue posible entrenar dos modelos para cada una de las producciones.

Además de la separación de los datos por telenovela/serie y por fuente, también decidimos utilizar dos enfoques diferentes para los modelos; el primer enfoque fue usar Google Cloud AutoML (GCP) y el segundo fue la biblioteca PyTorch; esta decisión refleja la diversidad de desafíos presentes en este estudio y nos dio la posibilidad de tener dos alternativas para la tarea de clasificación.

GCP representa un enfoque amigable y automatizado; destaca la gestión de grandes volúmenes de datos, así como la facilidad para la integración en otros flujos y su escalabilidad; sin embargo, esta plataforma no ofrece mucha libertad a la hora de personalizar los modelos y hacer un ajuste fino. Otra característica importante de esta plataforma es que los modelos están alojados en la nube de Google. Por otro lado, PyTorch proporciona

una plataforma flexible para adaptarse a la complejidad de los comentarios, lo que facilita una mayor personalización y un ajuste fino, y esto, a su vez, ofrece mayor control sobre el modelo y puede ser crucial cuando las clases tienen muchas similitudes y es necesario diferenciar sutilmente entre ellas; no obstante, requiere habilidades técnicas más avanzadas, así como también más recursos de desarrollo. Los modelos de PyTorch deben ser manejados por cada usuario y, por tanto, pueden ser usados localmente o ser desplegados en una plataforma en la nube, siendo siempre esto responsabilidad del usuario, a diferencia de los modelos GCP.

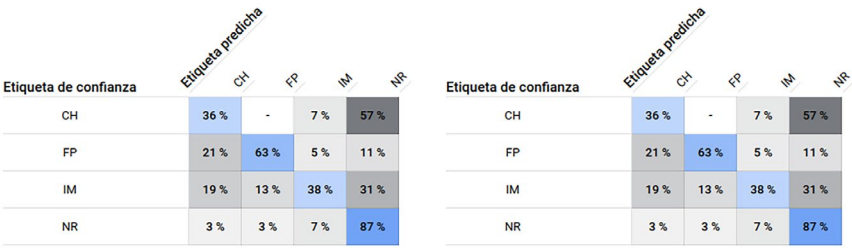
Para crear un modelo en GCP, una vez tuvimos los datos etiquetados por el equipo de humanidades, creamos una instancia para guardarlos en la nube, y los pasos posteriores son ejecutados de manera automatizada por la plataforma. Estos son:

- Preprocesar los datos, que incluye *tokenización*, normalización de datos y extracción de características.
- Seleccionar el modelo. GCP evalúa diferentes modelos de aprendizaje automático con hiperparámetros por defecto, para seleccionar el que mejor se adapta a los datos y al problema específico.
- Realizar un ajuste de los hiperparámetros para optimizar el rendimiento del modelo seleccionado en el paso anterior.
- Validar y evaluar el rendimiento en cada paso de ajuste con validación cruzada para garantizar la generalización.
- Finalmente, desplegar el mejor modelo en la plataforma de GCP y poner a disposición del usuario para realizar predicciones en la nube.

Por el contrario, para crear el modelo en PyTorch definimos la arquitectura con todos los hiperparámetros y dimensiones de la red, además de que hicimos su optimización nosotros mismos. Para este caso llevamos a cabo pruebas con arquitecturas tipo RNN, LSTM y CNN, y diferentes configuraciones de hiperparámetros. Optamos por usar un modelo tipo CNN con la siguiente arquitectura y dimensiones:

- Una capa de embebimiento con un tamaño de vocabulario de 10.000 y una dimensión de 50.
- Cuatro capas convolucionales 2D, cada una con 100 filtros y el tamaño de cada filtro en 2, 3, 4 y 5; con esto tuvimos en cuenta 2-gramas, 3-gramas, 4-gramas y 5-gramas, respectivamente.
- Una capa de linearización con activación ReLU y dimensión de salida 4 (el número de clases más una clase *indeterminada*).

Figura 10.



Nota: Matrices de confusión para modelos GCP (izquierda) y modelo CNN (derecha) para los comentarios de YouTube.

Fuente: Elaboración propia.

La **figura 10** tiene una muestra de las matrices de confusión creadas para los modelos que nos permiten estudiar la eficacia para distinguir entre dos conceptos. Por ejemplo, en ambos casos la etiqueta de CH suele ser fácilmente desestimada como un comentario común sin importancia (38% y 57% de las veces, respectivamente). Estos modelos pueden mejorarse un poco más aumentando la cantidad de datos de muestra etiquetados; y no obstante que arquitecturas tan robustas tengan tanta dificultad en diferenciar entre una clase y otra, también es una muestra de la complejidad del problema, debido a las diferencias sutiles entre las clases y la diversidad narrativa y estructural de los comentarios.

Al mismo tiempo cabe destacar que las categorías de análisis, a la luz de referentes teóricos concretos, pueden ser interpretadas de diferentes maneras, por lo cual hubo desequilibrios a la hora de clasificar ejemplos concretos de una categoría específica, incluso dentro del mismo equipo de humanidades. También existen afirmaciones de las redes sociales en las plataformas que pueden ser clasificadas en más de una categoría de análisis, lo cual dificulta el entrenamiento de los modelos, con la consecuencia de que a largo plazo confundirá la clasificación exacta. Es necesario abordar ese sesgo metodológico en la investigación a futuro.

A medida que exploramos la clasificación de comentarios en relación con estas telenovelas y series no solo estamos analizando la recepción pública de las obras, sino también comprobando cómo las técnicas de automatización pueden ayudar a estos estudios cuando se combinan con el análisis de grandes cantidades de datos, sin dejar de lado que siempre será necesario el ojo experto de las ciencias humanas y sociales para darles sentido a los descubrimientos y aplicación de métodos de análisis propios de la ciencia de datos. El uso de modelos no solo responde a la tarea de clasificación de categorías de análisis, sino que pone a disposición la información para la aplicación de modelos matemáticos en torno a preguntas sociales específicas; para el caso, la intersección entre la historia y la cultura de las sociedades latinoamericanas. [Ver ejemplo con aplicación de código aquí].

4. Perspectivas analíticas expandidas mediante la aplicación de métodos digitales

En el proyecto GUMELAB el trabajo con YouTube y Twitter (X) ha sido enriquecedor para conocer la audiencia activa en redes sociales de las telenovelas y series que investigamos en un marco global, así como para entender las tendencias y poder reconstruir la historicidad de la recepción de las audiencias desde una mirada más internacional. Especialmente con las producciones que han tenido éxito en una escala global, este tipo de fuentes ayuda a identificar las tendencias en la audiencia de diferentes países y a conocer los temas que más la conmovieron en varias localizaciones, entre otros. La clasificación automatizada de temas recurrentes en comentarios y respuestas a los comentarios retroalimentó y mostró perspectivas analíticas para ser profundizadas con métodos cualitativos. A continuación, algunas de estas perspectivas identificadas gracias a los datos obtenidos en YouTube y Twitter (X), no sin antes aclarar que se trata de resultados preliminares. Al momento de cerrar la edición de este manual nos encontramos en la fase de posprocesamiento de datos y análisis.

4.1 Perspectivas analíticas expandidas por los datos obtenidos en YouTube

Para abordar este aparte se emplean como caso de estudio solo los datos obtenidos para la telenovela[36] *Pablo Escobar. El Patrón del Mal* (EPM). Esta producción retrata la vida del narcotraficante colombiano Pablo Escobar Gaviria desde su infancia hasta su fallecimiento. La telenovela también aborda los enfrentamientos protagonizados por Escobar con la Policía colombiana y otros carteles del narcotráfico, como el Cartel de Cali. Además, presenta los ataques, asesinatos, magnicidios y el sufrimiento de las víctimas que padecieron bajo el narcoterrorismo en su vida diaria.

36 Considerando la duración en número de capítulos y la frecuencia de su primera transmisión a través del Canal Caracol en la televisión abierta de Colombia, presentamos la producción *Pablo Escobar. El Patrón del Mal* bajo el formato de telenovela en lugar de serie. La emisión constó de 113 capítulos, transmitidos de lunes a viernes, desde el 28 de mayo hasta el 19 de noviembre de 2012.

Del conjunto de información extraída de la plataforma YouTube, desde el 11 de julio de 2009 hasta el 30 de julio de 2021 sobre la telenovela EPM, realizamos 38 consultas distintas, de las cuales solo 19 fueron seleccionadas como consultas base. Estas búsquedas recabaron información para dos países, Chile y Colombia, y generaron un total de 154.161 datos recolectados vinculados a un conjunto de 1974 videos, 18.902 comentarios y 6.443.831 *likes*.

Dentro de este conjunto de videos identificamos dos tipos: 456 videos que contienen algún capítulo de la telenovela EPM, que denominamos *videos con capítulo reportado*, y 1518 videos que contienen distintos tipos de información relacionada con las telenovelas, que denominamos *videos generales.* Estos videos son muy diversos: pueden ser recopilación de las escenas seleccionadas por los usuarios, resúmenes y comentarios sobre algún aspecto de la telenovela, etc. Los videos generales acumularon un total de 6.296.484 y 147.347 comentarios, mientras que los videos con capítulo reportado recibieron un total de 998.157 *likes* y 33.555 comentarios.

A continuación algunos resultados sobre la interacción generada por la telenovela, basados en la correlación entre la razón *likes* y número de visualizaciones, los capítulos con más apreciados, la distribución temporal de los comentarios relacionados con la telenovela, los capítulos más comentados y los países de emisión con mayor participación a través de comentarios. El análisis de estos aspectos proporciona en general pistas sobre la recepción de una producción de televisión y sugiere cuáles podrían haber sido los capítulos y las escenas clave en la trama que han iniciado discusiones en la plataforma. Esta perspectiva de análisis permite delimitar y a la vez estructurar la investigación cualitativa cuando examinemos cómo una telenovela o serie afecta las imágenes y recuerdos del pasado, la conciencia histórica y la formación política de las audiencias.

En el análisis de la interacción y *enganche* de la audiencia activa en YouTube con la telenovela EPM resultó especialmente relevante examinar la correlación entre *likes* y número de visualizaciones, como lo muestra la

figura 11.[37] Dicha figura revela que, en términos generales, según la proporción de visionados y *likes*, la telenovela EPM mantiene sus audiencias activas en YouTube *enganchadas* a lo largo de sus 113 capítulos. En otras palabras, quienes ven los capítulos en YouTube manifiestan su aprecio mediante la interacción positiva con *likes*. El promedio de *likes* por visualización se situó en aproximadamente el 60%. Destacan los capítulos 1, 35, 52, 68, 84, y 85, con una proporción de *likes* por visualización cercana al 80%. El capítulo 1 se centra en la infancia de Pablo Escobar, mientras el capítulo 35 relata el asesinato de un corresponsal del periódico colombiano *El Espectador* que estaba investigando un caso de corrupción. Por su parte, el capítulo 68 presenta la planificación de los asesinatos de Luis Carlos Galán y el coronel Quintana.[38] Los capítulos siguientes abordan los enfrentamientos entre el Cartel de Medellín, liderado por Pablo Escobar, y los capos del Cartel de Cali (capítulo 52), así como la persecución y muerte del Mariachi a manos de las autoridades (capítulos 84 y 85).

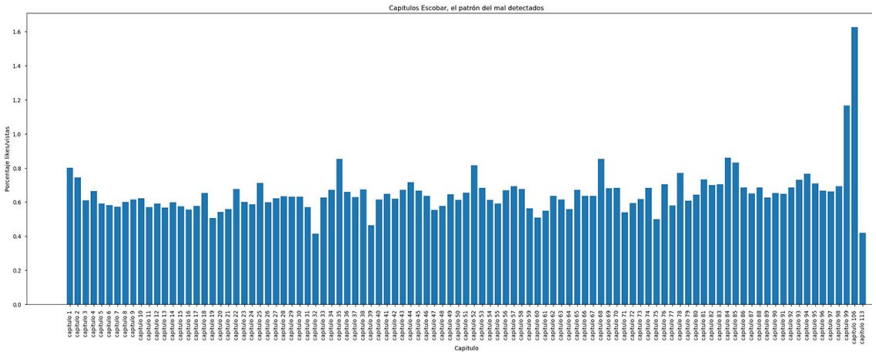
No obstante, los capítulos 99 y 106 se destacaron al generar el mayor nivel de interacción del público con un 50% superior al promedio. El capítulo 99 retrata uno de los eventos más significativos en contra de la prensa colombiana y su élite: el secuestro de la renombrada periodista Diana Turbay por parte de Los Extraditables, hija del expresidente Julio César Turbay y

37 Los análisis de YouTube se basan en la versión nacional de *Pablo Escobar. El Patrón del Mal,* que consta de 113 capítulos. Existe también una versión internacional que se encuentra disponible en la plataforma Netflix, la cual se compone de 74 capítulos, aparentemente 39 menos. Sin embargo, al comparar ambas versiones, constatamos que la diferencia radica en 29 minutos de contenido. En otras palabras, la versión nacional presentó 29 minutos más de contenido que la internacional. La disparidad de 39 capítulos está estrictamente relacionada con la duración de los capítulos emitidos: en la nacional, el promedio es de 27 minutos, mientras que en la internacional es de 42 minutos.

38 Luis Carlos Galán Sarmiento fue un político colombiano que se postuló como candidato a la presidencia de Colombia en 1982, 1986 y 1989. Fue asesinado por orden de Pablo Escobar el 18 de agosto de 1989. En la telenovela existe un personaje que lleva su mismo nombre y representa parte de su historia. El personaje coronel Oswaldo Quintana Quintero representa al general Valdemar Franklin Quintero, de la Policía Nacional, quien fue asesinado por orden de Pablo Escobar el mismo día, 18 de agosto de 1989.

cuya muerte se produjo en medio de una operación militar (relatada en los siguientes capítulos). Por otro lado, el capítulo 106 relata la entrega de Pablo Escobar a la justicia colombiana y las condiciones de su sometimiento, que incluyeron la construcción de su propia cárcel, conocida como La Catedral.

Figura 11.

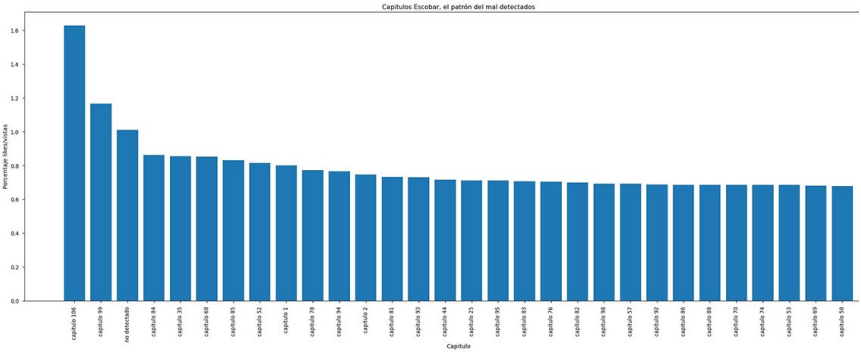


Nota: correlación entre *likes* y número de visualizaciones de *Escobar. El Patrón del Mal.*

Fuente: Elaboración propia

El análisis anterior es convalidado cuando limitamos el procesamiento de información a solo el 30% de los datos más relevantes, visto a través de la **figura 12**, y muestra que los capítulos 106 y 99 de EPM tienen el mayor porcentaje de *likes* sobre vistas.

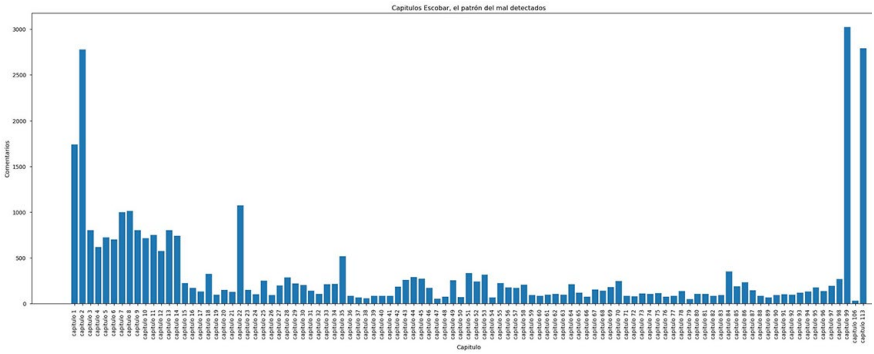
Figura 12.



Nota: Porcentaje de *likes* sobre vistas de *Escobar. El Patrón del Mal.*
Fuente: Elaboración propia

La **figura 13** da cuenta de la distribución temporal de comentarios, leídos en millones, detectados y asociados al inicio, clímax y finalización de la telenovela EPM. Ella muestra el interés por comentar los capítulos de acuerdo con el orden temporal de emisión, y evidencia estabilidad en los comentarios recibidos en los primeros capítulos, del 1 al 14, más allá del 1 y del 2, que despertaron un mayor interés por comentar. Es importante tener en cuenta que el número de comentarios no detectados fue de 5.700.000 aproximadamente, con un margen de error que tiene influencia en el análisis.

Figura 13.



Nota: Distribución temporal de comentarios de *Escobar. El Patrón del Mal,* leídos en millones

Fuente: Elaboración propia

Los capítulos más comentados, pero sobre todo reproducidos y en muchos casos reeditados para su publicación en YouTube, sugieren para la investigación cualitativa priorizar el análisis de estos capítulos a fin de entender por qué resultaron tan importantes para las audiencias activas de esa plataforma. También nos dejan abrir una comparación con las interacciones de las audiencias en otras redes sociales y con las entrevistas que hemos realizado con distintos segmentos de audiencias en Chile, Colombia y Estados Unidos.

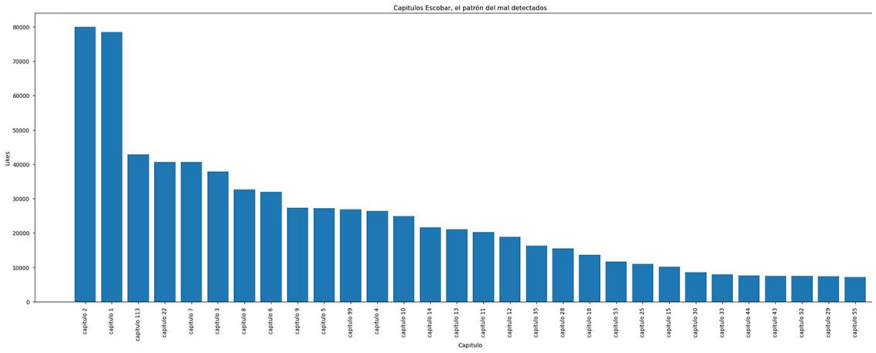
Con frecuencia, las personas que parecen tener referencias y haber vivido circunstancias en medio de los diversos hitos que recrea la telenovela EPM comparten sus reflexiones y sentimientos asociados a los recuerdos que les suscita. Las temáticas que más resaltan nos ayudan a marcar una ruta de análisis. Por ejemplo, en un video vinculado al capítulo 84, uno de los de mayor proporción de visualizaciones/*like*, quienes comentaron relevan aspectos tales como la corrupción e ineficacia institucional expresando su

desprecio. Pero, al mismo tiempo, es prevalente el sentimiento de duelo por la muerte del personaje el Chili, quien representó al sicario y miembro del Cartel de Medellín John Jairo Arias Tascón, alias Pinina. Entre las expresiones de condolencia los comentarios aluden a los valores de fidelidad, hombría asociada a la valentía o falta de miedo ante las situaciones de peligro, y la sagacidad del delincuente, que posiciona culturalmente el Cartel de Medellín en el mundo de la delincuencia y la criminalidad: "Es triste la muerte del chili", "gente q no conoce al chili: awebo entremos en este funeral aver [sic] q cosa pasa xd", "Es el momento en que los hombres guardamos un minuto de silencio", "Chili, leal hasta la muerte", "El papá de los bandidos", "El más inteligente y fiel a su patrón, pero esa muerte me dejó el corazón partido". En los comentarios relacionados con la muerte del Chili, la audiencia activa en YouTube destaca la lealtad y confianza que existe entre bandidos, especialmente por una escena en la que el Topo, personaje que representa otro sicario del Cartel de Medellín, manifiesta su tristeza por la muerte del Chili.

También registramos, aunque con menor frecuencia, comentarios que expresaban satisfacción por la muerte del personaje: "Satisfacción [de] ver la muerte de uno de esos malnacidos asesinos", "Buena participación del Chili, era un sicario y tenía que morir", "Una lacra menos, faltan más", "¿Pobre Chili? ¡No se dan cuenta [sic] que era un sicario y mataba a gente inocente!". Sin embargo, no observamos opiniones o reflexiones elaboradas respecto de las condiciones estructurales, subjetivas u otras dimensiones de análisis sobre la problemática histórica en Medellín o en el ámbito global. De este modo, la información ausente en estas interacciones digitales enriquece el análisis cualitativo y genera nuevas preguntas.

La **figura 14** muestra los capítulos más comentados por encima del percentil 70; es decir, sobre el 30% más relevante de los capítulos comentados. Estos fueron, en su orden, los capítulos 99, 113, 2, 1 y 22.

Figura 14.



Nota: Capítulos *Escobar. El Patrón del Mal* mas comentados por encima del percentil 70.

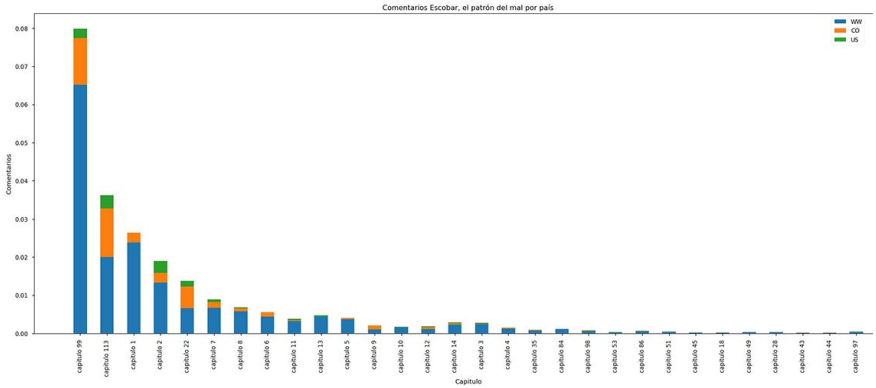
Fuente: Elaboración propia

Los capítulos sobre Diana Turbay (99) y sobre la entrega de Pablo Escobar a las autoridades (113) siguen siendo importantes para que la gente decida comentar sobre lo visto. Ambos capítulos concentran la mayor oportunidad de analizar la opinión pública en torno a las categorías centrales de análisis del proyecto GUMELAB: *imágenes de la memoria*, *conciencia histórica* y *formación política*, y *no relevante*.

Con respecto a la ubicación geográfica desde donde se emiten los comentarios, la siguiente figura muestra que ocurre en lugares no identificados en todo el mundo, denominados como WW (*world wide*), donde se producen más comentarios, seguidos de Colombia y Estados Unidos para los capítulos 99, 113, 1 y 2, lo cual coincide con los capítulos más comentados por encima del percentil 70. Esto muestra no solo el éxito internacional de la telenovela y que cuenta con una audiencia global que comenta y discute sobre el contenido, sino, además, los temas que más interesan a estas audiencias. Esta significativa información expande una nueva perspectiva de análisis que nos

invita a realizar la correlación entre las temáticas y los lugares geográficos de emisión, lo cual sin duda aportará importantes hallazgos con relación a la transnacionalización de la memoria.

Figura 15.



Nota: Comentarios *Escobar. El Patrón del Mal*, por país.
Leído en millones, por encima del percentil 70.

Fuente: Elaboración propia

Otro método adicional que exploramos con los comentarios extraídos de YouTube consistió en agrupar todos los comentarios sobre los videos asociados con un mismo capítulo y representarlos en una nube de palabras (*wordcloud*). Esta técnica de delimitación por capítulos proporciona una visión de las opiniones del público de manera más específica, brindando así una comprensión más precisa de los aspectos relevantes de cada capítulo. Este enfoque podría contribuir significativamente a la aplicación de técnicas de clasificación. En el caso del capítulo 113, a continuación aparecen las palabras "mataron", "grande", "mejor" y "corrupto", las más usadas, que podrían dar pistas para la siguiente investigación.

Figura 16.

Nota: Nube de palabras con los comentarios capturados en YouTube para los videos relacionados con el capítulo final de la telenovela EPM, capítulo 113.

Fuente: Elaboración propia.

4.2 Perspectivas analíticas expandidas por los datos obtenidos en X (antiguo Twitter)

Analizando los datos obtenidos en X (Twitter), ya fueron explicados algunos resultados y ejes de análisis en el apartado 2.1. Twitter ha sido muy útil para reconstruir la historicidad de la recepción y la temporalidad de los tuits. En el marco temporal de la emisión de los datos (2009 hasta 2021), pudimos observar cómo cambian la recepción y los temas discutidos dependiendo del año de la interacción en la red y la transmisión de la telenovela o serie. Esto es especialmente interesante cuando alguna se emite en momentos diferentes, ya que puede generar siempre nuevos tuits con otras temáticas y otras conexiones a la realidad de la audiencia con su actualidad. Por ejemplo, en el contexto del estallido social y las protestas de 2019 acaecidas en Chile, se emitió de nuevo la serie *Los 80* en Canal 13, que retrata la vida de una

familia durante la dictadura militar. La audiencia conectó lo que vio con las protestas actuales y tuiteó contenido más político. Fue notable una tendencia a emociones como rabia y decepción de ver que no había cambiado mucho la historia, además de la conexión de los acontecimientos de la dictadura reflejados en la trama de la serie con las luchas recientes para tratar de cambiar la Constitución, como se puede ver en el siguiente tuit:

> #los80 #canal13
> Que triste ver este capítulo de los 80, todavía ni existía, mis papás eran chicos.
> Lo veo y veo que Chile no ha cambiado en nada, lo veo y es ver un día cualquiera año 2019/2020. Los mismos problemas, la misma injusticia y desigualdad.

Al mismo tiempo, reconstruir la recepción en esta plataforma permite entender la temporalidad de dicha recepción. Las opiniones sobre las telenovelas y series no empiezan a partir de la transmisión del primer capítulo, sino que ya antes las producciones son parte de discusiones y debates gracias a los tráileres, noticias y, en general, campañas publicitarias de expectativa que lanzan los canales o plataformas *streaming* que las transmitirán; esto ha sido un hallazgo que expandió nuestra perspectiva analítica. Dada la simultaneidad entre el tiempo de visionado y la interacción en redes sociales, con estos datos es posible reconstruir mejor el contexto en que se tuiteó, algo inviable en las entrevistas, ya que las personas no siempre recuerdan exactamente las emociones y conexiones que hicieron en el momento justo en que vieron un capítulo. Más bien, siempre lo reflejan en el momento de la entrevista desde el presente, con interpretaciones que pueden ser influenciadas por otros acontecimientos, e incluso por el entrevistador o entrevistadora.

Gracias a los *hashtags* y *tags* más usados pudimos entender los temas más discutidos al respecto. La discusión en Twitter (X) es de otra naturaleza en comparación con YouTube, y por los *hashtags* y *tags* es posible comprender mejor el hilo de discusión, por la interconexión entre varias cuentas y usuarios. Por ejemplo, un hallazgo al que llegamos gracias a Twitter (X) es el rol que juegan los políticos actuales en la recepción de la audiencia. Muchas veces la audiencia conectó la opinión de la telenovela/serie con políticos actuales, a

pesar de que no eran específicamente parte de la trama ficcional. La audiencia conecta la historia nacional con los acontecimientos y políticos actuales, lo cual muestra una conexión entre el pasado y el presente. Un ejemplo es el siguiente tuit, donde el usuario hace referencia al político Álvaro Uribe Vélez y a la telenovela *Pablo Escobar. El Patrón del Mal*. Podemos estimar que se refiere a los delitos de las dos personas.

> Viendo @AlvaroUribeVel el patrón del mal... Upss perdón. #Escobar, el patrón del mal. @CaracolTV

Gracias a la recopilación de datos en el ámbito global, logramos observar la conexión y la transferencia del contenido de la telenovela/serie, aplicándolo a las realidades históricas de otros contextos nacionales. Esto se evidencia especialmente cuando abordamos temas denunciados de alguna manera en la producción, como la violación de derechos humanos o la corrupción, los cuales la audiencia solicitaba para su propio país:

> ¿Para cuándo la serie de #Narcos Honduras, @NetflixLAT?

Otra perspectiva de análisis que se abrió con los datos obtenidos en Twitter (X) fue la consideración del peso que tienen los personajes ficticios en la memoria colectiva de la audiencia. Estos personajes poseen una complejidad que le permite a ella empatizar, encontrar gracia, fascinarse y, hasta cierto punto, identificarse más fácilmente con ellos que con los propios personajes históricos, también representados en las telenovelas y series investigadas por GUMELAB. Ejemplos de personajes ficticios relevantes son el Chili, de *Pablo Escobar. El Patrón del Mal*; o el papá Juan Herrera y su hijo Félix Herrera, de la serie *Los 80*. La cantidad de tuits que mencionan a estos personajes ficticios proporciona retroalimentación a la investigación, con nuevas pistas para las preguntas que se podrían plantear en entrevistas de recepción o de producción. Esto facilitó el proceso de reconstrucción de la producción y ayudó a analizar si las intenciones del equipo de producción se ven reflejadas en la audiencia o si han surgido nuevas interpretaciones a partir de personajes

cuyas características no fueron concebidas intencionalmente por el equipo creativo. Sin la contribución de los datos obtenidos en Twitter (X) quizás no habríamos visualizado la importancia que tienen los personajes ficticios para la audiencia en un nivel global. También hay opiniones sobre los diferentes actores para expresar elogios elaborados desde la audiencia.

Muchas frases elocuentes, tanto de los personajes como de los mismos cabezotes, generaron un gran eco en Twitter (X). En el caso de EPM, las frases que más se reprodujeron en los tuits provienen del personaje de Pablo Escobar y del personaje de la madre de Escobar. Asimismo, la famosa frase "quien no conoce su historia está condenado a repetirla", del cabezote de la telenovela, resonó profundamente, al igual que la letra de la canción del cabezote de *Narcos*: "Soy el fuego que arde tu piel, soy el agua que mata tu sed". Además, pudimos rastrear escenas clave que generaron intensas discusiones, donde las personas intercambiaron recuerdos personales de los hechos históricos retratados o se validaron las opiniones y emociones. Estas escenas podrían ser hechos históricos retratados en la trama o escenas ficticias que se transforman en imágenes de la memoria, y que analizamos con más detalle en el proyecto. Todos estos hallazgos abrieron nuevas líneas de análisis para la investigación cualitativa.

5. Recomendaciones para el uso de métodos digitales en investigación de ciencias humanas y sociales

En proyectos de humanidades que trabajan con métodos digitales es siempre importante hacerse la pregunta sobre el valor agregado de usar dichos métodos; es decir, ¿qué podemos ver gracias al empleo de los métodos digitales que no habríamos visto de otra manera? Por ejemplo, en términos cualitativos, ¿qué nuevos descubrimientos han aportado los nuevos métodos?

El proyecto GUMELAB ha sido innovador por dos temas centrales. Primero, por investigar las telenovelas y series dentro del marco académico en Alemania; y, segundo, por incorporar métodos digitales en un proyecto de la disciplina de la historia, algo todavía poco común. Una de las principales cuestiones abordadas por GUMELAB giró en torno a la relación historia y fuentes históricas digitales en la construcción social del hipertexto; es decir, ¿cómo el uso del hipertexto influencia nuestras interacciones sociales y cómo es la relación entre redes sociales y construcción de fuentes para el análisis histórico?

La historia digital es una forma de aproximación a la historia global de conexiones, interacciones y comparaciones en diferentes partes del mundo sobre fenómenos sociales presentes en el vasto espacio de la web 2.0 y versiones subsiguientes, que desplaza el laboratorio de la investigación en historia a la web. La disciplina de la historia se enfrenta a una diversidad de lenguajes en los que está representada la historia, pero también a la debilidad de la permanencia y la digitalización de la realidad material. Así como la construcción y conservación de los materiales históricos que son producidos vertiginosamente en estas redes, lo que William J. Turkle (2005) ha denominado el *archivo infinito*, y el papel de la minería de datos textuales como una manera de enfrentar el problema. Se sugiere visitar el apartado de lecturas recomendadas a fin de destacar en este espacio los aprendizajes y retos de las humanidades digitales desde el componente *e-Research* GUMELAB suscritos a debates en curso aún lejos de saldarse.

Avanzar globalmente en condiciones positivas de acceso a conocimiento y su producción por parte de las sociedades y la academia en diferentes latitudes del mundo es condición para la superación de las brechas de marginalidad y pobreza, que, sin duda, se profundizan con el dominio de

los métodos y herramientas provistas por la Cuarta y la Quinta Revolución Industrial. En general, el idioma principal para la ciencia de datos, así como para los principales repositorios de información, es el inglés. La implementación de métodos digitales en la investigación científica en ciencias humanas y sociales necesita de recursos y requisitos técnicos que no se distribuyen homogéneamente en la sociedad y en los procesos formativos de las disciplinas de las humanidades; si bien aún con limitaciones, las brechas de alfabetización digital en el mundo forman parte de las agendas de gobierno en Latinoamérica.

Por otro lado, enfoques académicos decoloniales pueden aportar a la gestión de conocimiento con carácter circular, con acceso abierto que contribuya a diseminar resultados de investigación entre la comunidad académica y la sociedad global. Para el caso, el proyecto GUMELAB no es solo un proyecto interdisciplinar, sino que también está conectado internacionalmente y con una apuesta por el trabajo con datos de redes sociales en español.

Campos de aplicación de investigación en humanidades digitales

Es frecuente asimilar el análisis de *big data* y la investigación con métodos digitales en humanidades, toda vez que ambos implican el uso de datos digitales y tecnologías para obtener información. Sin embargo, las diferencias son de carácter epistemológico: el primero está orientado por preguntas y problemas de las ciencias exactas, en tanto que la segunda por problemas de las diferentes disciplinas que conforman las humanidades, y destaca la interpretación contextual crítica.

Una precisión necesaria respecto del análisis de *big data* es que se orienta a la extracción automática de patrones determinados por teorías generales y modelos de las ciencias exactas, naturales y aplicadas. A menudo se centra en la identificación de patrones, correlaciones y tendencias en grandes conjuntos de datos para obtener información útil para la toma de decisiones basada en datos, la predicción de tendencias y la optimización de

procesos con volúmenes de información medidos en *terabytes*, *petabytes* o incluso *exabytes* de datos, y retos tecnológicos altos con alcances cuánticos. En proyectos de humanidades, y también en el proyecto GUMELAB, no se trata de tal cantidad de datos, y, por ende, no se puede hablar de *big data*. Cabe destacar que no todas las universidades públicas tendrán la capacidad de almacenar datos a nivel *big data*, por lo cual no contarán con las mismas precondiciones de las empresas privadas.

Por su parte, las humanidades digitales aluden a la aplicación de métodos digitales sin prescindir de la interpretación contextual y crítica de fuentes orientada por teorías interpretativas de la realidad con distintas bases epistemológicas. En tal sentido, busca comprender fenómenos sociales adaptados al seno de las disciplinas humanas, sus traslapes e intersecciones. Entonces, por su naturaleza obedece a tipos de investigaciones mixtas e implica una variedad de enfoques cualitativos y cuantitativos adaptados a las preguntas específicas de la investigación en humanidades para el análisis de contenido, análisis textual, análisis de redes, visualización de datos y enfoques hermenéuticos.

Las perspectivas de las humanidades digitales emergen y continúan en expansión de la mano de los avances de la Cuarta Revolución; es decir, de la integración de tecnologías digitales, la automatización, la inteligencia artificial y su aplicación a las comunicaciones en relación con la masividad, la variedad de formatos en que se produce información, la agilidad en la transmisión y, por supuesto, la diversidad y pluralidad humana en interacción.

Los campos de aplicación de las humanidades digitales están orientados por el interés de (i) estudiar problemas propios de la sociedad cibernética como un nuevo problema y paradigma de conocimiento; (ii) dotar de instrumentos el procesamiento de datos masivos, voluminosos, diversos en formato y materialidad, que de otra manera no sería posible abordar; y (iii) aplicar métodos de procesamiento del campo de las tecnologías digitales y métodos de análisis matemáticos y estadísticos a problemas de las disciplinas que integran las humanidades.

La cuestión fundamental para el diseño de una investigación en el campo de conocimiento de las humanidades digitales es definir si el problema de investigación es aplicado a los problemas o métodos que abordan estas humanidades. En este caso, el componente de *e-Research* de GUMELAB partió de inscribir su problema de investigación al campo analítico de las interacciones en la WWW, de donde se deriva una cascada de opciones metodológicas circunscritas a este ámbito (capítulo 1). Pero no siempre es así en todos los casos: pueden existir investigaciones que solo usen métodos digitales sin que su objeto de investigación pertenezca al ámbito de la web.

La interdisciplinaridad

Entre los aspectos fundamentales de las investigaciones en humanidades digitales se encuentran el reconocimiento y la exploración previa de diversas tecnologías y las posibilidades que ofrecen de acuerdo con los mencionados intereses y la pertinencia para problemas específicos de los campos disciplinares de aplicación que correspondan. En tal sentido, los profesionales científicos sociales se encuentran en los límites de la transdisciplinariedad, con capacidad de reconocimiento de las bases epistemológicas, métodos y herramientas de investigación que ofrecen la programación, la inteligencia artificial, y sus alcances aplicados al procesamiento, análisis y automatización de procesos.

Lo más frecuente es que los equipos de trabajo tengan niveles diferentes de apropiación disciplinar, a menos que constituyan un grupo de investigación consolidado. Siempre se requieren directores de investigación con experiencia en ambos campos del conocimiento científico de las humanidades, con experiencia en metodologías de investigación social y científica de datos e informática con investigación aplicada. Uno de sus principales retos será la capacidad para integrar enfoques disciplinares y gestionar las diferencias epistemológicas. La comunicación entre equipos inter- y transdisciplinarios es fundamental para el ambiente productivo de trabajo. Proyectos que trabajan con métodos digitales necesitan de participantes con distintos conocimientos, que tengan en mente la pregunta de investigación general

y los referentes teóricos para el análisis. Los proyectos de métodos digitales se enriquecen con el intercambio entre diferentes disciplinas, que ayudan a ver y analizar con perspectivas más integrales. Esta ganancia de incorporar métodos digitales en proyectos de humanidades no viene sin secuelas, ya que precisa una comunicación intensa entre epistemes. Uno de los principales riesgos de no lograr el diálogo transdisciplinar es terminar *trabajando para la máquina*; es decir, un equipo perdido en procesos y procedimientos sin tener en mente el horizonte de trabajo colectivo.

Para esto es fundamental establecer una relación de confianza a la hora de trabajar en proyectos interdisciplinarios. La comunicación es sumamente importante para que los integrantes del equipo vinculados a las humanidades y las ciencias sociales, así como aquellos vinculados a las ciencias de datos, adquieran un cierto entendimiento sobre el desarrollo del trabajo de sus colegas. Esto resulta especialmente importante para apreciar los numerosos pasos *invisibles* implicados en la extracción, limpieza y modelado de los datos por parte del equipo de ciencias de datos. Al mismo tiempo, es esencial reconocer el tiempo dedicado por el equipo de humanidades a la ubicación, lectura, discusión y procesamiento de la literatura necesaria para construir la ruta teórica, así como el tiempo empleado en la recopilación de información empírica no digital (entrevistas y consultas en archivos). La interdisciplinaridad en el proyecto GUMELAB ha sido tanto una ganancia como un desafío. En algunos casos, incluso se ha presentado como un obstáculo. El entendimiento entre las disciplinas es un proceso continuo que no se logra con solo unas pocas reuniones. En el proyecto GUMELAB hemos subestimado este aspecto, especialmente porque resulta crucial contar en ambos equipos con al menos un integrante que posea conocimientos previos en los campos de la programación y de investigación en humanidades o ciencias sociales. También hemos subestimado la distancia geográfica. Aunque el equipo de ciencias de datos se encuentra en Colombia y el equipo de humanidades en Alemania, hemos logrado mucho mediante la comunicación virtual. Sin embargo, nos hemos dado cuenta de que nuestras interacciones son considerablemente mejores y más fluidas cuando logramos trabajar cara a cara de manera constante, como ocurrió durante las estancias de algunos integrantes del equipo de ciencias de datos en Alemania.

Al mismo tiempo, es importante considerar que la contratación de programadores para proyectos académicos puede ser costosa, especialmente debido a la competencia por la retención de talento. Actualmente hay una alta demanda laboral de las empresas de la industria tecnológica e informática, que además complica el reclutamiento y permanencia de programadores en proyectos de investigación académica. Adicionalmente, los costos asociados con la infraestructura, incluyendo equipos y espacios de almacenamiento de información, no pueden ser subestimados.

El diseño metodológico de la investigación

El trabajo con métodos digitales tiene muchas ventajas, sobre todo, como se ha dicho, cuando se requiere trabajar con gran cantidad de datos. En cualquiera de los alcances del uso de métodos digitales, bien como instrumentalización para el procesamiento de datos, bien con alcance analítico soportados en modelos matemáticos y estadísticos, es preciso tener un pulcro diseño metodológico. Pese a las ventajas de la utilización de métodos digitales, el proceso de investigación en sí mismo puede ser objeto de nuevo conocimiento, por lo cual es recomendable que el proceso metodológico considere el uso responsable y continuo de una bitácora, entre otras, por la posible reiteración de métodos para el procesamiento de datos camino a la generación de información analizable.[39] Llevar una bitácora implica no solo unificar el lenguaje en la comunicación, sino que, además, mantenerla al día requiere tiempo adicional. En el marco del proyecto GUMELAB, nuestra bitácora estuvo constituida por el uso de la herramienta visual Trello y los protocolos de las reuniones entre los equipos de humanidades y ciencias de datos.

39 Los datos pueden ser números, palabras, imágenes, sonidos, entre otros, y por sí mismos no tienen un significado intrínseco. Los datos adquieren relevancia cuando son procesados y contextualizados, lo que les permite convertirse en *información.* Por su parte, la información es el resultado del procesamiento de datos susceptibles de interpretar. Es un conjunto de datos organizados y estructurados de manera significativa para que tenga sentido y sea útil. La información tiene la capacidad de transmitir conocimiento y generar comprensión (Turban *et al.,* 2014).

El grado de complejidad de la investigación puede aumentar de acuerdo con el tipo de datos no estructurados y su captura, el mayor alcance en el uso de métodos de programación y la creación de agentes cognitivos, que requieren entrenamiento, y el uso de IA.

El diseño metodológico debe reflejar la trazabilidad del flujo paso a paso desde las preguntas de investigación, los objetivos, las fuentes de información, el arsenal de datos y de herramientas de procesamiento, almacenamiento y métodos de análisis. Se debe tener en cuenta lo propio para cada fase de investigación: levantamiento de información, procesamiento de información y sus etapas de acuerdo con los métodos previstos, análisis de información y diseño de cruces analíticos bajo modelos de análisis coherentes con las preguntas y objetivos de investigación, generación de formas de visualización de información compleja comunicable y escritura o comunicación de resultados. Este tipo de investigaciones suelen ser valiosas particularmente en la generación de repositorios y fuentes de información de futuras investigaciones, entre otros usos. De manera que vale la pena, finalmente, proyectar las opciones futuras de usabilidad de los repositorios de información que se hayan generado, la sostenibilidad y costos bajo principios de democratización del conocimiento.

Para elegir las mejores opciones metodológicas es esencial evaluar la relación entre los objetivos de la investigación, el tiempo disponible y los costos asociados de iterar el proceso. Esto incluye la categorización manual de altos volúmenes de datos para el entrenamiento de algoritmos, así como la consideración de aplicación de IA o agentes cognitivos especializados para automatizar procesos mecánicos en la investigación, como la captura y limpieza de datos y la codificación bajo el sistema teórico categorial, entre otros posibles. En resumen, se requiere de IA para desarrollar IA, lo cual implica una inversión de tiempo y recursos adicionales. De lo contrario, las tareas manuales, como limpieza de datos y validación de datos, requerirán un considerable esfuerzo de los investigadores, que afectará negativamente el progreso en otras áreas de trabajo de la investigación.

Los datos

El mundo aún avanza en la captura de datos estructurados y representa un importante negocio de la industria informática que usa las comunicaciones y el *marketing* para forzar a los cibernautas a ceder datos a cambio del uso de herramientas informáticas y de telecomunicaciones cada vez más imprescindibles. En tal sentido, el acceso a datos generalmente tiene costo, así que esto constituye un reto central en la investigación para las humanidades digitales. En medio de esto, algunas redes sociales, los medios digitales de comunicación y los repositorios digitales de diverso tipo son las fuentes de datos más comunes.

El hecho de que el laboratorio de fuentes en la web sea más vasto tiende a convertirse en lugar común para la definición de investigaciones en humanidades digitales, pero esto no necesariamente significa que sea más rico en información. En todo caso, su riqueza dependerá de un cierto tipo de problemas y conjunto de métodos de la historia que avancen en la fundamentación de la historia digital como campo de conocimiento. Así, usar información digital como tuits no es propiamente hacer investigación digital, sino usar fuentes digitales.

Gracias a la digitalización de fuentes como archivos históricos y documentos, una gran cantidad de datos se hizo accesible para las investigaciones de cualquier disciplina. Este tipo de fuentes tienen la ventaja de ofrecer información más estructurada para el análisis, por cuanto se presenta a modo de fragmentos de texto o grafos sobre las categorías específicas de análisis. Y no constituyen un mayor reto para la ciencia de datos en la fase de extracción y organización de la información. Otro tipo de fuentes son las que ofrece en general la web o *digital born*, tales como datos de las redes sociales, prensa *online* y otras plataformas donde se genera información no necesariamente parametrizada, pero que puede ser extraída en bases de datos externas.

Las investigaciones con fuentes digitales pueden partir de (i) la digitalización del mundo material, (ii) el uso de datos digitales previamente estructurados en bases de datos o (iii) de datos digitales desestructurados para

responder a problemas específicos. GUMELAB debió organizar su archivo digital porque su problema no partió de la digitalización del mundo material ni de datos digitales predeterminados, sino de textos nativos digitales (*digital born*) en la web como fuente. Este enfoque le implicó al equipo de investigación la ardua tarea de estructurar los datos para crear el archivo con una fracción de información del mundo de la web extraída de las plataformas digitales donde interactúan redes sociales, de cuya representación no podrá tener una estimación con referencia al universo en el que se circunscribe.

En ese sentido, los retos que detectamos para construir fuentes digitales según la experiencia de GUMELAB son (i) la transparencia en el acceso a la fuente, (ii) los cambios técnicos en requisitos y permisos de las API para la extracción de datos, (iii) la dificultad de definir el universo de análisis, (iv) la representatividad de los datos, (v) la falta de garantía sobre la permanencia de los datos en las plataformas digitales, (vi) la limpieza de datos o la eliminación de datos *basura*, es decir, que no se relacionan para nada con la investigación.

(i) Transparencia

Un problema central relacionado con la transparencia en la obtención de datos surge en el ámbito de la industria generadora de datos, donde resulta difícil anticipar los sesgos introducidos por empresas privadas al momento de realizar las búsquedas o generar los llamados de la información a través de las API. Plataformas como X, YouTube, Facebook, Google Search son empresas privadas que operan bajo una lógica de mercado donde lo secreto puede ser clave para ganar ventaja competitiva con otras compañías. Esto puede referirse a la creación de algoritmos, datos de usuarios, etc., y al ocultamiento de los protocolos de manejo de información a los cuales los equipos de investigación no pueden acceder. Como consecuencia, aunque se pague por los datos, la calidad de los entregados o vendidos, según el caso, es desconocida, lo que implica que la investigación siempre tendrá este punto ciego que no podrá superar.

Los algoritmos que recopilan y clasifican los datos pueden introducir sesgos. Por ejemplo, los algoritmos de recomendación de información en las redes sociales pueden estar introduciendo sesgos desconocidos de segmentación. En suma, la investigación con métodos digitales conlleva limitaciones en términos de transparencia, las cuales pueden entrar en conflicto con las exigencias de disciplinas como la ciencia histórica.

(ii) Cambios en las API

Es necesario monitorear los cambios técnicos en las API durante la etapa de extracción de información. Plataformas como X ofrecen acceso a la API luego de una evaluación del proyecto académico. En el transcurso del proyecto GUMELAB aquella cambió varias veces los requisitos de acceso a la API, como ya se explicó.

De fondo, trabajar con métodos digitales significa estar al tanto de todos los cambios técnicos que se llevan a cabo, e introducir principios de flexibilidad y pragmatismo en la investigación, sobre todo con el avance del Chat GPT, que salió al público a finales de 2022 e hizo importante preguntarse por el uso de la nueva tecnología dentro del proyecto.

Acceder a redes sociales que permanecen en la empresa Meta, como Facebook e Instagram, no fue posible debido a los escándalos como el de Cambridge Analytica. También ha habido amenazas jurídicas de parte de Facebook contra investigaciones que iban a usar los datos de la empresa para sus proyectos, como fue el caso de Algorithm Watch; el resultado fue que la investigación tuvo que ser detenida para evitar una demanda judicial (Algorithm Watch, s. f.).

Un aspecto positivo de trabajar con los datos de X y YouTube es que los comentarios tienen formato de texto. En X los datos se reducen a 280

caracteres, antes solamente 140.[40] Instagram, Facebook y TikTok se caracterizan por el uso de imágenes, música y movimiento, útiles para un tipo de análisis más complejo.

(iii) Limitación en la identificación del universo completo de datos

La cuantificación del universo de datos que dan cuenta de la interacción de las redes sociales en las plataformas será incierta. Esto plantea un desafío para establecer criterios de validación que aseguren la representatividad de la muestra requerida para responder a las preguntas de investigación. Esta consideración representa una advertencia y limitación metodológicas que deben ser comunicadas en los resultados de la investigación.

(iv) Representatividad de los datos

La descripción del universo de fuentes se limita a los encabezados de los datos de los usuarios, es decir, a las entidades, que permitan la extracción realizada a través de la API. Con frecuencia esto es bastante limitado en la caracterización de la población sujeto de investigación. Así, los usuarios de estas plataformas digitales no son representativos de la población general, lo que debe limitar la generalización de los hallazgos. Por otro lado, en medio del volumen abrumador de datos, surge el reto de determinar cuáles son verdaderamente representativos. ¿Cómo evitar el peligro de focalizarse únicamente en las voces más *escuchadas* o en el contenido más viral, corriendo el riesgo de perder perspectivas más matizadas y menos populares? ¿Cuáles son los métodos informáticos disponibles para abordar estos problemas?

40 Elon Musk creó X Premium. A sus usuarios se les permite el uso de 25.000 caracteres.

(v) Incertidumbre sobre la permanencia de los datos

También es posible que los datos ya extraídos se pierdan con el transcurso del tiempo, y que no puedan volver a ser encontrados en las plataformas. Esto puede ocurrir porque los enlaces dejan de funcionar o porque los usuarios han bloqueado o eliminado sus cuentas, entre otras razones. Por ejemplo, algunos tuits que fueron buscados y recopilados manualmente en 2015, y cuyos resultados de análisis se encuentran en el seguiente artículo de Mónika Contreras Saiz[41], ya no están disponibles.

(vi) Obtención de datos útiles

La especificidad técnica y la calidad de los datos, fundamentales para aumentar su utilidad, no está bajo control del equipo investigador. Aunque la interacción en las redes sociales genera grandes volúmenes de datos, su calidad puede resultar cuestionable. Por un lado, los datos pueden ser *ruidosos*, es decir, consistir en caracteres sin sentido. Además, la polisemia de las palabras, como se ha explicado ya, puede dar lugar a la captura de datos no relacionados directamente con la pregunta de investigación (véase el ejemplo de la ciudad argentina llamada Escobar en el contexto de la extracción de datos sobre *Escobar. El Patrón del Mal*). Se requieren, entonces, diversos métodos de limpieza para mantener la integridad y la utilidad de los datos.

A este conjunto de retos se le suma un desafío más, relacionado con la contextualización de los datos. Tuits, comentarios, publicaciones o imágenes, una vez extraídos, se aíslan del contexto de emisión y del hilo del debate. Para una debida crítica de una fuente histórica es necesario conocer el contexto en que surge, pero muchos de los datos capturados pueden llegar a carecer de contexto, o ser sumamente dispendioso reconstruirlo para cada uno de

41 Contreras Saiz, M. (2019). Conciencia histórica, pensamiento crítico y telenovelas en Latinoamérica. En E. Varela Sarmiento, *Escenarios para el desarrollo del pensamiento crítico.* Clacso, Universidad de La Salle, 51-86.

los datos. De este modo, constituye un gran desafío extraer conocimientos históricos significativos a partir de datos fragmentados y sin contexto.

Al igual que la historia oral, la historia digital parte de una fuente que no ha sido acumulada en archivos históricos tradicionales (las fuentes digitales) y requiere proyectar, por lo menos, un repositorio para la información (archivo digital o laboratorio de fuentes). En tal sentido, para el caso del componente de *e-Research* del proyecto GUMELAB, constituye un esfuerzo inútil buscar explicar la representación del archivo digital en el universo de la web. En cambio, se debe avanzar en la estructuración del archivo digital/laboratorio de fuentes acotado en el universo de los parámetros utilizados para la extracción de datos y el alcance proyectado para que como archivo ofrezca posibilidades a investigaciones futuras. Sobre todo considerando que los datos recopilados muy seguramente no volverán a ser hallados en su totalidad en las plataformas digitales. El futuro de ellas es incierto y puede cambiar de un momento a otro, a tal punto de desaparecer por completo.

Para ello es fundamental partir de las discusiones sustanciales que tienen lugar entre las investigadoras y los investigadores, y que son pertinentes para otorgar valor histórico al archivo digital y responden las preguntas de investigación iniciales, al seno de las cuales nació y se estructuró el archivo digital. Este enfoque debe extenderse para abordar la usabilidad de la información y la permanencia del archivo, sirviendo como un laboratorio de fuentes para futuras investigaciones. En tal sentido, se requiere un diseño cuidadoso tanto en el componente teórico como en el metodológico para estructurar el archivo de forma efectiva.

La aplicación de métodos digitales a la investigación en humanidades con datos no estructurados obliga a tener un mayor rigor metodológico para no perderse en el proceso de extracción, limpieza y estructuración de los datos, por lo cual es recomendable la creación de un *documento de orientaciones metodológicas* que registre las opciones metodológicas frente a los problemas técnicos y de ciencia de datos que se presentan en el camino hacia la extracción, estructuración de los datos y automatización de procesos de investigación y archivística para la creación de repositorios finales. Entre otras,

esto es central para caracterizar el archivo y el alcance de sus posibilidades de consulta. Esto, además, debe atarse a un *documento maestro* de definiciones clave en el ámbito del archivo histórico que orienten los procesos de clasificación, rotulación, archivo y conservación de documentos de consulta. Este manual será un aporte en la creación de futuros documentos maestros.

6. Bibliografía

Introducción

Assmann, A. y Conrad, S. (2010). *Memory in a Global Age. Discourses, practices and trajectories.* Palgrave Macmillan. https://doi.org/10.1057/9780230283367

Birkner, T. y Donk, A. (2018). Collective memory and social media. Fostering a new historical consciousness in the digital age? *Memory Studies, 42*(2), 1-17. https://doi.org/10.1177/1750698017750012

Botero, J. D., Guo, W, Mosquera, G., Wilson, A., Johnson, S., Aguirre García, G. y Pachón, L. (2019). Gang confrontation: The case of Medellin (Colombia). *PloS ONE, 14*(12), e0225689. https://doi.org/10.1371/journal.pone.0225689a

Brügger, N. (2018). Web history and the web as a historical source. *Zeithistorische Forschungen/Studies in Contemporary History, 15*(1), 171-193. https://zeithistorische-forschungen.de/2-2012/4426

Classen, C. (2009). Balanced Truth: Steven Spielberg's "Schindler's List" among history, memory and popular culture. *History and Theory, 47,* 77-102. https://www.jstor.org/stable/25478838

Contreras Saiz, M. (2017). Narcotráfico y telenovelas en Colombia: entre narconovelas y "telenovelas de la memoria". *Hispanorama. Zeitschrift des Deutschen Spanischlehrerverbandes August,* (157), 26-31. https://www.lai.fu-berlin.de/homepages/contreras/Contreras_2017_Telenovelas-de-la-memoria-1_compressed.pdf

Contreras Saiz, M. (2023). Telenovelas, series y formación política en Latinoamérica. En M. Pardo y S. Peters, *Educación política. Debates de una historia por construir.* Instituto Colombo-Alemán para la Paz (CAPAZ). https://www.gumelab.net/Publikationen-_-Presse/Publikationen/099_Libro_Educacion_politica_Pardo_Peters.pdf

Erlick, J. C. (2018). *Telenovelas in pan-Latino context.* Routledge. https://doi.org/10.4324/9781315545608

Rüsen, J. (1997). Historisches Erzählen. En K. Bergmann *et al., Handbuch der Geschichtsdidaktik.* Kallmeyer.

1. Ruta metodológica para una investigación histórica con métodos digitales

Kemp, S. (2022a). Digital 2022: Chile. https://datareportal.com/reports/digital-2022-chile

Kemp, S. (2022b). Digital 2022: Colombia. https://datareportal.com/reports/digital-2022-colombia

Kemp, S. (2022c). Digital 2022: The United States of America. https://datareportal.com/reports/digital-2022-united-states-of-america

Midha, A. y Nagy, J. (2014). The value of earned audiences: how social interactions amplify TV impact. *JAR* Bd, *54*(4), 448-453. https://doi.org/10.2501/JAR-54-4-448-453

Selva, D. (2016). Social television. *Television & New Media* Bd, ***17***(2), 159-173. https://doi.org/10.1177/1527476415616192

We Are Social y Hootsuite. (2022). Digital 2022 Global Overview Report. https://datareportal.com/reports/digital-2022-global-overview-report?utm_source=DataReportal&utm_medium=Country_Article_Hyperlink&utm_campaign=Digital_2022&utm_term=Chile&utm_content=Global_Promo_Block

Wilson, S. (2016). In the Living Room. *Television & New Media* Bd, *17*(2), 174-191. https://doi.org/10.1177/1527476415593348

5. Recomendaciones para el uso de métodos digitales en investigación de ciencias humanas y sociales

Algorithm Watch. (s. f.). Facebook macht dicht: Forschung braucht Zugang zu Plattformen. https://algorithmwatch.org/de/offener-brief-forschung-zu-plattformen/.

Contreras Saiz, M. (2019). Conciencia histórica, pensamiento crítico y telenovelas en Latinoamérica. En E. Varela Sarmiento, *Escenarios para el desarrollo del pensamiento crítico.* Clacso, Universidad de La Salle, 51–86. https://biblioteca.clacso.edu.ar/clacso/gt/20200306054249/Escenarios-para-el-desarrollo-del-pensamiento-critico.pdf

Turban, E., Sharda, R. y Delen, D. (2014). *Decision support and business intelligence systems.* Pearson.

Turkle, W. (2005). Teaching young historians to search, spider and scrape. *Digital History Hacks (2005-08). Methodology for the Infinite Archive.* http://digitalhistoryhacks.blogspot.com/2005/12/

7. Los autores y las autoras

Hannah Müssemann
(Freie Universität Berlin, Alemania)

Es candidata al doctorado en Historia en el Instituto de Estudios Latinoamericanos de la Universidad Libre de Berlín. Estudió Filología Alemana y Portuguesa en Mainz y Lisboa, así como Estudios Latinoamericanos en la Universidad Libre de Berlín. Actualmente trabaja como investigadora en el proyecto GUMELAB, donde coordina el componente *e-Research* junto con la Universidad de Antioquia. Investiga la recepción de telenovelas y series de televisión en la diáspora latinoamericana en Estados Unidos. Su investigación se centra en métodos digitales y estudios de la memoria en un contexto global.

Gicela Andrea Aguirre García
(Freie Universität Berlin, Alemania)

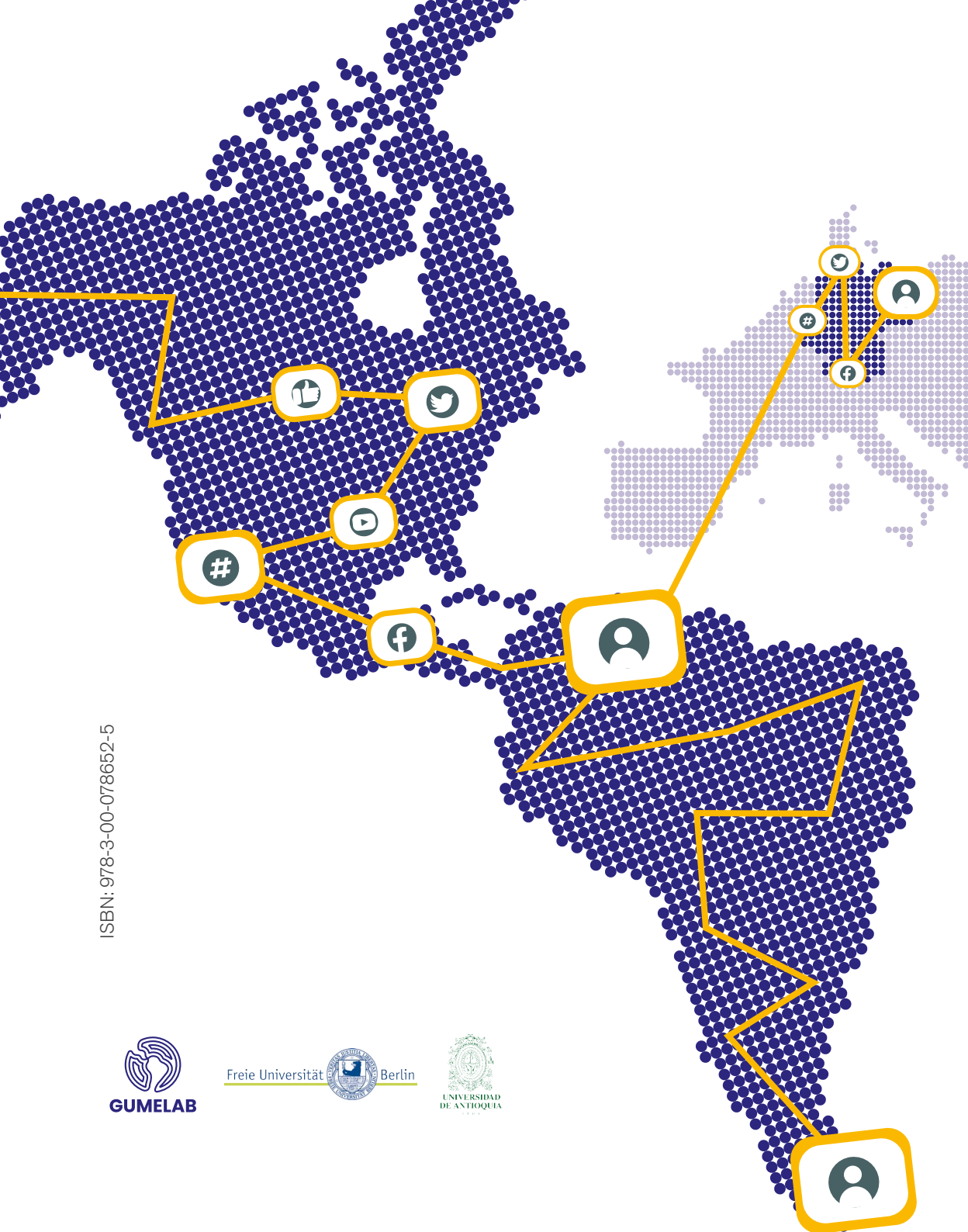
Es candidata al doctorado en Historia del Instituto de Estudios Latinoamericanos de la Universidad Libre de Berlín, financiada por el programa Pasaporte a la Ciencia, de Colciencias. Estudió Trabajo Social y es magíster en Ciencia Política en la Universidad de Antioquia. Es la coordinadora de Gestión del Conocimiento de la Corporación Concuidadanía, docente universitaria y científica investigadora de la Universidad de Antioquia en colaboración con el componente *e-Research*-GUMELAB. Investiga sobre cultura de paz, dinámicas del conflicto armado, delincuencia urbana y criminalidad. Su investigación doctoral se centra en las dinámicas de transformación de las bandas delincuenciales con métodos digitales.

Heisman Duvolfán Arcila Arboleda
(Guane Enterprises/Universidad de Antioquia)

Estudió Física en la Universidad de Antioquia y actualmente es gerente de Productos en el Área de Análisis de Lenguaje en Guane Enterprises y miembro del componente *e-Research* del proyecto GUMELAB desde la Universidad de Antioquia. Su investigación se centra en el uso de los modelos de lenguaje, las técnicas de procesamiento de lenguaje natural y la inteligencia artificial en problemas sociales, logísticos y energéticos.

Alesson Ramon Rota
(Universidade Estadual de Campinas, Brasil)

Es candidato al doctorado en Historia por la Universidade Estadual de Campinas (Unicamp), financiado por la Fundación Paulista de Investigación (FAPESP). Máster en Historia por la Unicamp y licenciado en Historia por la Universidad Federal de Rio Grande (FURG), con un periodo de intercambio en la Universidad de Coimbra. Investigador visitante en la Freie Universität Berlin. Colaborador en los laboratorios de investigación CHD-Unicamp y GUMELAB-FU. Ganador del Premio de Monografía 2016 de la Sociedad Brasileña de Teoría e Historia de la Historiografía (SBTHH). Miembro de los grupos de investigación Historia y Lenguajes Políticos: Razón, Sentimientos y Sensibilidades, e Historiografías Periféricas en Perspectiva Global.



ISBN: 978-3-00-078652-5

GUMELAB

Freie Universität Berlin

UNIVERSIDAD DE ANTIOQUIA