PUBLICACIONES DE LA
ASOCIACIÓN ARGENTINA
DE HUMANIDADES DIGITALES

Narrativas digitales de la COVID-19 en Twitter: de los datos a la interpretación

Digital Narratives of COVID-19 on Twitter: From Data to Interpretation

Susanna Alles Torrent
susanna_alles@miami.edu
University of Miami
https://orcid.org/0000-0002-3616-2285

Gimena del Rio Riande
gdelrio@conicet.gov.ar
CONICET
https://orcid.org/0000-0002-8997-5415

Romina De León
rdeleon@conicet.gov.ar
CONICET
https://orcid.org/0000-0003-2495-7213

Marisol Fila
mafila@umich.edu
University of Michigan
https://orcid.org/0000-0003-3445-2702

Nidia Hernández
nidiahernandez@conicet.gov.ar
CONICET
https://orcid.org/0000-0001-7557-6846

Jerry Bonnell
j.bonnell@miami.edu
University of Miami
https://orcid.org/0000-0002-7404-9160

Dieyun Song
dxs1138@miami.edu
University of Miami
https://orcid.org/0000-0003-2263-6475

PublicAAHD - 2020

Vol. 1

ISSN

2718-7470

ABSTRACT

The bilingual project Digital Narratives of COVID-19 brings together researchers, programmers, and students from the University of Miami and CONICET (Argentina). DHCovid aims to analyze and interpret Twitter data (in English and Spanish) on the SARS-CoV-2 global pandemic from the end of April 2020 to May 2021, through quantitative methods and tools used in the field of DH. This article explores the different methods and tools that the project has used, from basic search platforms and semi-automated text-mining to more complex and more or less supervised ones. The paper additionally discusses how this set of tweets was collected to study the narratives and emerging issues about the pandemic in South Florida and specific Spanish-speaking countries (Argentina, Mexico, Peru, Colombia, Ecuador, Spain). Furthermore, it presents the GitHub and Zenodo data repository as well as some of the tools developed by the project. Finally, work with data mining, frequency analysis of terms and concordances, and topic modeling will be exhibited.

KEYWORDS

Twitter, Data, Narratives, Data Mining, Analysis.

RESUMEN

El proyecto bilingüe Narrativas digitales de la COVID-19 (DHCovid) reúne a participantes de la Universidad de Miami (EE. UU.) y CONICET (Argentina). Tiene como objetivo analizar e interpretar datos (en inglés y español) sobre la pandemia global de SARS-CoV-2 procedentes de Twitter desde finales de abril de 2020 hasta mayo 2021, por medio de métodos y herramientas cuantitativos utilizados en el campo de las Humanidades Digitales. Se presentarán distintos métodos y herramientas empleados, desde plataformas simples de búsqueda y minería semi-automatizada a otras más complejas y humanamente semi-supervisadas. Asimismo, se describirá cómo se ha realizado la recopilación de tweets para estudiar narrativas y temas sobre la pandemia en diversas zonas de habla hispana (Argentina, México, Perú, Colombia, Ecuador, España); donde se haya el repositorio de datos de GitHub y Zenodo así como otras herramientas desarrolladas por el proyecto. Finalmente, se exhibirá lo trabajado con minería de datos, análisis de frecuencia de términos y concordancias, y topic modeling.

PALABRAS CLAVE

Twitter, datos, narrativas, minería de datos, análisis.

1. TWITTER COMO OBJETO DE ESTUDIO DE LAS HUMANIDADES DIGITALES

Aunque Twitter podría ser entendido como un gran corpus textual plausible de ser minado con herramientas digitales y analizado con ojos humanistas para comprender mejor ciertas narrativas a escala global, sin embargo, no ha sido objeto de estudio de demasiados proyectos de investigación en Humanidades Digitales (HD). Tal vez una de las dificultades a la hora de trabajar con esta red social ha sido justamente la imposibilidad de comprenderlo en su sentido estricto, en tanto cuerpo o, como su entrada en el Diccionario de la Lengua Española indica: "Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación"[1]. Podría agregarse aquí que, para los proyectos de HD, un corpus funciona como colección de textos o imágenes seleccionadas y curadas, donde una pregunta de investigación da origen a ese recorte y su posterior estudio con métodos y/o herramientas digitales o computacionales. Twitter es, en este sentido, y como toda red social, un espacio en constante mutación y expansión, donde la comunidad es la productora de textos, imágenes, videos y hasta vínculos hacia otros lugares de la web. Además, se trata de microtextos de menos de 280 caracteres, donde el sentido surge del agrupamiento de los mensajes individuales a través del uso de hashtags comunes, mención a usuarios, o respuestas.

Si bien en el campo de las Ciencias Sociales Computacionales la analítica cultural (Manovich, 2007) o la culturomía (Aiden y Michel, 2013) han combinado procesamiento computacional para manejar volúmenes de información cultural con marcos teóricos de las Ciencias Sociales, no hay en las HD ninguna línea específica dedicada a estudiar el contenido de redes sociales en términos de gran narrativa que permita un macroanálisis (Jockers, 2013) o lectura distante (Moretti, 2005). El uso más extendido de la red social en el campo ha sido el de la conformación de una comunidad académica, principalmente anglófona, que comparte allí trabajos, ideas y puntos de vista. Este movimiento, habitualmente entendido como parte del *Twitter for scholarly networking* o *#AcademicTwitter*[2], ha sido analizado con el fin de comprender mejor quiénes son los humanistas digitales y dónde y con quienes trabajan (Grandjean, 2016; Quan-Haase et al., 2015; Martin y Mc Kay Peet, 2016).

Sorprendentemente, la pandemia de COVID-19, que inundó todas las redes sociales con mensajes de todo tipo y tono, no atrajo aún grandes proyectos a las HD. Sin embargo, se destacan dos iniciativas relevantes que unieron estos temas el pasado año: un congreso llevado a cabo en Twitter oportunamente titulado *DH in the Times of Virus* que, el 2 de abril de 2020, reunió bajo el hashtag *#DHgoesviral* a una gran parte de la comunidad académica de humanistas digitales anglófonos en dicha red social. Y, por otro lado, los tradicionales *DH Awards*, que incluyeron una ca-

[1] Accesible desde: https://www.rae.es/drae2001/corpus.
[2] Algunos detalles sobre el uso de Twitter en la investigación se encuentran accesibles desde: https://digitalhumanities.berkeley.edu/twitter-scholarly-networking.

tegoría *Best DH Response to COVID-19*[3]. En su gran mayoría, se trataron de proyectos tipo archivo o series de visualizaciones puntuales.

2. NARRATIVAS DIGITALES DE LA COVID-19: UN PROYECTO COLABORATIVO SOBRE TWITTER DESDE LAS HUMANIDADES DIGITALES

2.1. Metodología y estrategia de minería de datos

Digital Narratives of COVID-19 / Narrativas digitales de la COVID-19 (DHCovid)[4] es un proyecto de HD que busca investigar la diversidad de discursos en la red que giran en torno a la pandemia del coronavirus. La iniciativa ha sido financiada por la Universidad de Miami (Estados Unidos) y se ha desarrollado junto con el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET, Argentina) entre abril de 2020 y mayo de 2021, donde participaron investigadores, programadores y estudiantes de sendas instituciones.

Una de las principales líneas de investigación del proyecto consiste en la recolección y el análisis de las conversaciones sobre la pandemia en Twitter tanto en inglés como en español. Por lo cual, el trabajo de minería de datos se inició a finales de abril de 2020 y ha continuado hasta mayo de 2021, ofreciendo así un amplio arco cronológico que comprende el pico más alto de la pandemia hasta la distribución de vacunas a un porcentaje significativo de la sociedad. La recolección de los tweets se ha llevado a cabo a través de la API de Twitter, que ofrece la posibilidad de minar su *feed* o flujo, aunque con algunas restricciones en su versión gratuita en cuanto a tiempo y cantidad[6].

A la hora de emprender el trabajo se consideraron dos factores: primero, el hecho que, a diferencia del inglés, el español no contaba con un corpus de Twitter específico para el coronavirus[7]; y segundo, la diversidad lingüística del español hablado en América Latina, Estados Unidos y España. En consecuencia, se decidió crear, por un lado, un dataset o conjuntos de datos recuperando un volumen significativo de los tweets en español relacionados con la crisis de la COVID-19 y, además, crear colecciones específicas de tweets de áreas concretas (Argentina, Colombia, Ecuador, México, Perú, España). Mientras que, para el caso del inglés, se descartó la posibilidad de minar todos los tweets en esa lengua porque, como se ha mencionado, existían, desde el inicio de la pandemia (enero 2020), varias iniciativas dedicadas a este propósito.

---

[3] Accesible desde: http://dhawards.org/dhawards2020/results/.
[4] La página del proyecto se encuentra accesible desde: https://covid.dh.miami.edu/ en español y en inglés.
[5] API por sus siglas en inglés *Application Programming Interfaces*, es decir, que se trata de un conjunto de definiciones y protocolos utilizados para desarrollar e integrar el software de las aplicaciones, permitiendo la comunicación entre dos aplicaciones de software a través de un conjunto de reglas.
[6] La documentación sobre el uso de la API de Twitter puede consultarse en: https://developer.twitter.com/en/docs. Mientras que algunos de los límites impuestos se explican en la sección *Rate limits* accesible desde: https://developer.twitter.com/en/docs/twitter-api/rate-limits.
[7] Existen diferentes iniciativas que han estado minando Twitter y recuperando tweets sobre la pandemia, algunos ejemplos son: Banda et al., 2021, Chen et al., 2020, Kerchner y Wrubel, 2020; o Lamsal, 2020.

Así, el proyecto ofrece datasets de Twitter estructurados en tres colecciones principales:

1. Tweets en español de todo el mundo (un total de 20.121.933 tweets en abril de 2021).
2. Tweets con ubicaciones geográficas en seis áreas seleccionadas de habla hispana que abarcan América del Norte y Central (México, Colombia, Ecuador), América del Sur (Argentina, Perú) y Europa (España).
3. Tweets geolocalizados en inglés y español del área metropolitana de Miami en el sur de Florida.

La idea subyacente a la planificación que se realizó para la minería de datos, tanto para el corpus general en español como para los de áreas específicas, ha sido la de poner en diálogo los puntos de vista globales y regionales, así como arrojar luz sobre las especificidades lingüísticas en algunas de las diferentes comunidades de habla hispana en todo el mundo. Estas tres secciones forman una red analítica interconectada que, se cree, pueden servir para delinear las narrativas transnacionales y transatlánticas sobre la pandemia de la COVID-19.

En primer lugar, para ensamblar el corpus de Twitter (recopilación de datos), un script PHP mina el flujo de datos de dicha red social a través de su API y recupera una serie de identificadores (ID) de tweets específicos. La estrategia para la recuperación de estos consta de cuatro variables principales: idioma, palabras clave, región y fecha[8]. Posteriormente, dichos ID se almacenan en una base de datos relacional MySQL[9] donde se hidratan, es decir, se recuperan todos los metadatos asociados con los tweets, incluido el cuerpo del texto. En tercer lugar, un script adicional organiza los ID en la base de datos por día, idioma y región, además, crea un archivo de texto plano para cada combinación con los identificadores correspondientes en forma de lista; es decir, genera archivos diarios y los organiza en carpetas, donde cada directorio representa un día. Estos se cargan directamente al repositorio público del proyecto en GitHub y pueden ser consultados y descargados en acceso abierto[10].

Una vez que se recuperan los ID de los tweets y se incluyen en la base de datos, se continua con la fase de preprocesamiento de datos. Esta etapa, explicada de una forma simplificada, consiste en estandarizar la estructura y el formato de los datos: se pasa todo el texto a minúsculas, se eliminan acentos, signos de puntuación, menciones de usuarios (@users) para proteger la privacidad y se sustituyen todos los enlaces por un genérico URL. Asimismo, este paso no está exento de

[8] En cuanto a idioma, decidimos minar tweets en español a nivel global y en las áreas ya mencionadas (Argentina, Perú, México, Colombia, Ecuador, España); añadimos también la zona del sur de la Florida, añadiendo en este caso, tweets en inglés. Las palabras clave que establecimos para la recuperación de los tweets fueron: en inglés, covid, coronavirus, pandemic, quarantine, #stayathome, outbreak, lockdown y #socialdistancing; mientras que, para el español, se buscaron tweets que contuvieran uno de estos términos: covid, coronavirus, pandemia, cuarentena, confinamiento, #quedateencasa, desescalada, and #distanciamientosocial.

[9] En una base de datos relacional MySQL, los datos son fragmentados en áreas de almacenamiento separadas, denominadas tablas.

[10] El repositorio del proyecto DHCovid19 se encuentra accesible desde: https://github.com/dh-miami/narratives_covid19/tree/master/twitter-corpus.

problemas, como pueden ser el uso de tildes y ciertos grafemas específicos del español (como la ñ) que no siempre aparecen transcriptos correctamente. Los emojis, por su parte, también han presentado un desafío, por lo cual se decidió transliterarlos en su correspondiente juego de caracteres en UTF-8. Igualmente, se decidió unificar todas las diferentes grafías de COVID-19 bajo una forma única[11]. Este procesamiento, en definitiva, permite obtener una colección limpia y ordenada de tweets organizados por idioma, por día y por área, facilitando su uso para análisis estadísticos o de modelado de temas.

Conjuntamente, se creó un sitio web en WordPress donde se ofrece acceso a los diferentes repositorios, scripts para ejecutar análisis y una serie de entradas de blog donde se presenta el trabajo, reflexiones sobre la pandemia con el uso de herramientas digitales propias o desarrolladas por otras iniciativas, e información acerca de cómo usar los scripts[12]. En resumen, los interesados pueden interactuar con el proyecto a través de tres canales: primero, a través del ya mencionado repositorio de GitHub, donde se encuentran los datasets que contienen los ID organizados por día y área. Cada uno de estos recopila nueve archivos diferentes, que se actualizan diariamente: a. tweets de seis áreas de habla hispana (México, Argentina, Colombia, Ecuador, Perú, España); b. tweets en inglés y español del área de Miami; c. tweets en español de todo el mundo. Segundo, una copia de este conjunto de datos se publica con fines de citación y preservación a largo plazo en el repositorio de acceso abierto Zenodo[13]. Por último, se puede observar una interfaz pública beta que permite la personalización y recuperación de datos, desde allí pueden descargarse tweets por fechas e idiomas enumerados anteriormente[14].

3. DHCOVID EN ACCIÓN

3.1. *Coveet.py: un script para el análisis de frecuencias y concordancias*

Inicialmente, para comprender las narrativas digitales que rodean a la COVID-19, se ha trabajado con métodos de análisis de frecuencia, con el objetivo de examinar qué palabras se utilizan, con qué frecuencia aparecen ciertas frases y cómo los discursos sobre la pandemia varían de un país a otro en Twitter (Gelfgren, 2016). Para ello, se desarrolló una herramienta en Python llamada *coveet.py*[15], que permitió explorar estas preguntas y llevar a cabo tres tareas:

1. Recuperar los datos textuales de Twitter desde la API del proyecto según país,

[11] Por ejemplo, se hallaban formas como COVID-19, COVID19, covid19, Covid19 con o sin guion, etc
[12] Accesible desde: https://covid.dh.miami.edu/blog/.
[13] Una primera versión estable del conjunto de datos (en formato zip), publicada el 13 de mayo de 2020, está disponible en Zenodo (Allés Torrent et al., 2020) y contiene solo los ID de los tweets publicados entre el 24 de abril de 2020 y el 12 de mayo de 2020. Una vez finalizado el proyecto, se subirá la versión completa del conjunto de datos .
[14] En este enlace https://covid.dh.miami.edu/get/, los usuarios pueden descargarse una colección de tweets personalizada en función de sus intereses de investigación.
[15] Este script se encuentra disponible en el repositorio de GitHub, https://github.com/dh-miami/narratives_covid19/tree/master/scripts/freq_analysis.

idioma y fecha;

2. Limpiar dichos datos a través de lematización y eliminación de palabras vacías (*stopwords*), para así obtener un formato ágil para las tareas posteriores de análisis textual.

3. Analizar los datos mediante la aplicación de técnicas de Procesamiento del Lenguaje Natural (PLN), como, por ejemplo, calcular palabras frecuentes y hashtags según la fecha.

Una de las particularidades de *coveet.py* es su modularidad, es decir que cada componente se construye de forma independiente entre sí, de manera que permite la incorporación de nuevas secciones de técnicas de PLN con un esfuerzo mínimo. Esta característica es importante para poder satisfacer nuevas demandas interpretativas y para que la herramienta pueda ser reutilizable en estudios futuros.

Si bien los datos cuantitativos producidos por *coveet.py* son significativos para explorar el corpus de Twitter, su cuantía para fines interpretativos se ha puesto en valor a través de visualizaciones adecuadas (Sinclair y Rockwell, 2016). Con este fin, se han utilizado las siguientes herramientas de visualización a través del paquete *matplotlib*[16], una matriz de gráficos de barras para visualizar los *n-grams*[17] principales y palabras únicas, y gráficos de concordancia para estudiar cada aparición de una palabra determinada en conjunto y en su contexto particular. Estas visualizaciones están contenidas en una serie de Jupyter Notebooks[18] y son accesibles en línea, a la vez pueden ser ejecutadas a través de Binder[19]. Esta metodología proporciona una especie de laboratorio de interpretación de datos en vivo y bajo demanda para acceder e interactuar con *coveet.py* y las visualizaciones obtenidas como resultado, como puede observarse en la siguiente figura:

[16] Esta librería de visualización con Python, *matplotlib*, es accesible en línea: https://matplotlib.org/.
[17] Los n-grams son una subsecuencia de n elementos de una secuencia dada, pueden contener cualquier tipo de unidad lingüística (fonemas, sílabas, etc.) o unidades textuales, sin embargo, los más frecuentes son los de palabras. .
[18] Esta librería de visualización con Python, *matplotlib*, es accesible en línea: https://matplotlib.org/.
[19] Binder transforma cualquier repositorio de GitHub que contenga una Jupyter Notebook y crea un entorno ejecutable que puede compartirse en tiempo real con otros, se encuentra accesible desde: https://mybinder.org/.

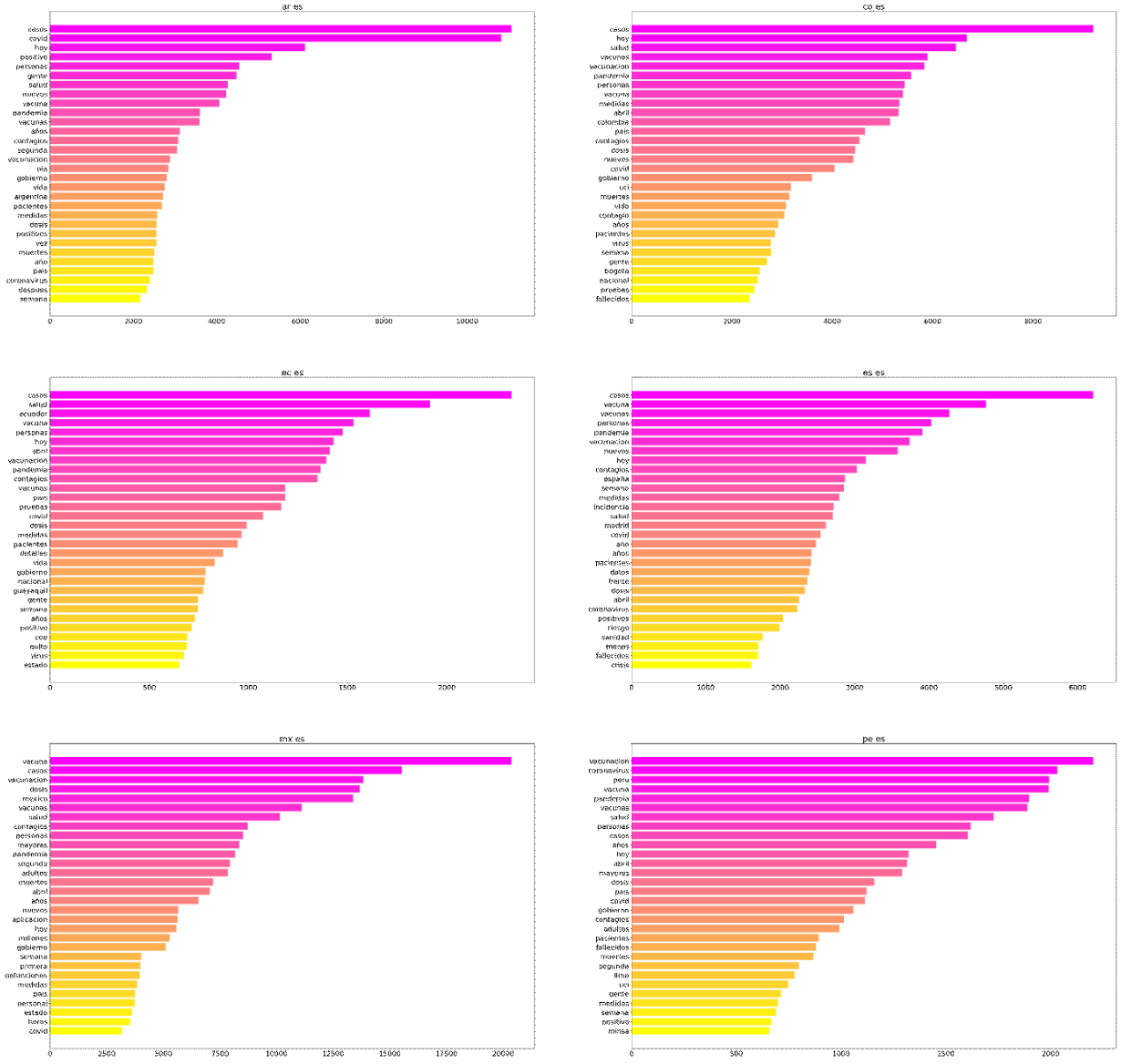

Figura 1. Palabras más frecuentes a lo largo del mes de abril de 2021 en Argentina, Colombia, Ecuador, España, México y Perú. Elaboración propia.

3.2. Topic modeling sobre la COVID-19

Dentro de la línea de investigación del proyecto dedicada al trabajo de detección mediante técnicas de PLN se incorporaron herramientas de topic modeling, teniendo en cuenta dos aspectos centrales para las Humanidades, por un lado, la importancia de complementar los métodos de lectura distante con un estudio detenido de los datos de entrada que permita asignar significados pertinentes a los resultados del procesamiento automático; y por el otro, la potencialidad de esta metodología concebida, más bien como, catalizador de nuevos interrogantes acerca de un corpus, y no sólo como una herramienta concluyente sobre su estructura semántica.

El topic modeling es un procedimiento estadístico de organización de grandes corpus en el que un algoritmo recibe datos no etiquetados para agruparlos según patrones comunes subyacentes. Esta técnica de aprendizaje automático apareció como una respuesta frente a la necesidad creciente surgida a partir de la explosión del Big Data de organizar temáticamente colecciones textuales diversas, no estructuradas y de gran escala de manera rápida y continua. En el ámbito

académico, por ejemplo, posibilita abordar grandes corpus accediendo primero a los principales temas y luego a los documentos, colaborando así con la construcción de vías de lectura para la investigación.

En el caso particular del corpus de DHCovid, el crecimiento acelerado de la gran masa de tweets y de sus metadatos, hizo indispensable el uso de herramientas de lectura distante. El interés del proyecto en el aspecto narrativo planteó la necesidad de complementar estos metadatos contextuales con información sobre el contenido del tweet y más específicamente con información temática. Por otra parte, la variabilidad de temas de interés entre los diferentes países y el vertiginoso ritmo de cambio de tópicos en las redes sociales en general, como podrá distinguirse en los posteriores ejemplos, requerían tecnologías que detectaran temas que emergieran de los datos sin necesidad de clasificaciones fijadas de antemano.

Para el procesamiento, se creó una representación vectorial de tipo *bag-of-words* combinada con bigramas, donde cada vector está compuesto de la cantidad de ocurrencias de cada token/bigrama en el tweet[20] y se aplicó el modelo de Asignación Latente de Dirichlet (ALD o por sus siglas en inglés LDA, Latent Dirichlet Allocation)[21] a cada colección. Para ello, se entrenaron modelos para tópicos de 3 a 50 con detección automática de densidad de tópicos por documento y densidad de palabras por tópico. La calidad de los modelos fue evaluada usando un *scoring* automático e inspección manual, para cada colección se obtuvo automáticamente el modelo con el mejor número de tópicos y luego se analizaron manualmente tópicos y tweets correspondientes[22].

Con el objetivo de nivelar los desbalances en la cantidad de tweets de distintos países, se intentó crear muestras comparables con la misma cantidad de tweets para cada región, sin embargo, el techo fijado por las regiones de menor cantidad de datos muchas veces redujo la calidad de los tópicos. Otro desafío para el trabajo fue el cambio diacrónico en el corpus, que obligó a realizar una revisión en la adecuación de los tópicos propuestos inicialmente[23].

Cabe señalar, que el análisis de las experiencias comparativas entre todas las regiones del corpus excede el propósito de este artículo; asimismo, a modo de muestra de las dificultades de interpretación de los resultados del topic modeling, se presenta una experimentación reducida de características similares. En la figura 2 se pueden apreciar los tópicos para el día 25 de abril de 2020 en Colombia, donde cada color indica un tópico y cada círculo una palabra, cuando estas tienen un mayor peso dentro de un tópico son representadas por círculos mayores. Asimismo, se

---

[20] Sólo se incluyen los bigramas más relevantes (dos palabras consecutivas con una frecuencia mayor de 20 ocurrencias en la colección).

[21] El script mencionado se encuentra accesible en: https://github.com/dh-miami/narratives_covid19/blob/master/scripts/topic_modelling.

[22] La medida de *scoring* usada es cv. La evaluación manual de tópicos es una práctica frecuente (Stefanidis et al., 2017; Abdo et al., 2020) que consiste en observar las principales palabras clasificadas por el algoritmo en ese tópico y proponer un título representativo.

[23] El estudio de la evolución temporal de los temas también implicó decisiones sobre la ventana temporal seleccionada para la generación de modelos, así como el criterio de identidad de temas entre períodos. Tomar períodos muy extensos en un medio tan dinámico como Twitter puede generar superposición de temáticas.

debe destacar que en el corpus general la información sobre número de casos positivos, muertes, pacientes recuperados, hospitalizaciones, camas disponibles, etc., fue denominada *Estadísticas*, y se trató de un tópico consistente en todas las regiones y períodos, pues emergió con claridad en todas las experiencias realizadas; además se puedo observar en el modelo una preponderancia en recuentos (número de casos, infectados, muertos), medidas sanitarias y cuestiones políticas (como medidas de prevención). De manera similar, se pudo estimar que los tweets de este tópico provenían normalmente de cuentas institucionales o de medios de comunicación y se focalizaron en diferentes escalas (global, nacional, local, institucional, etc.) a lo largo de la pandemia.

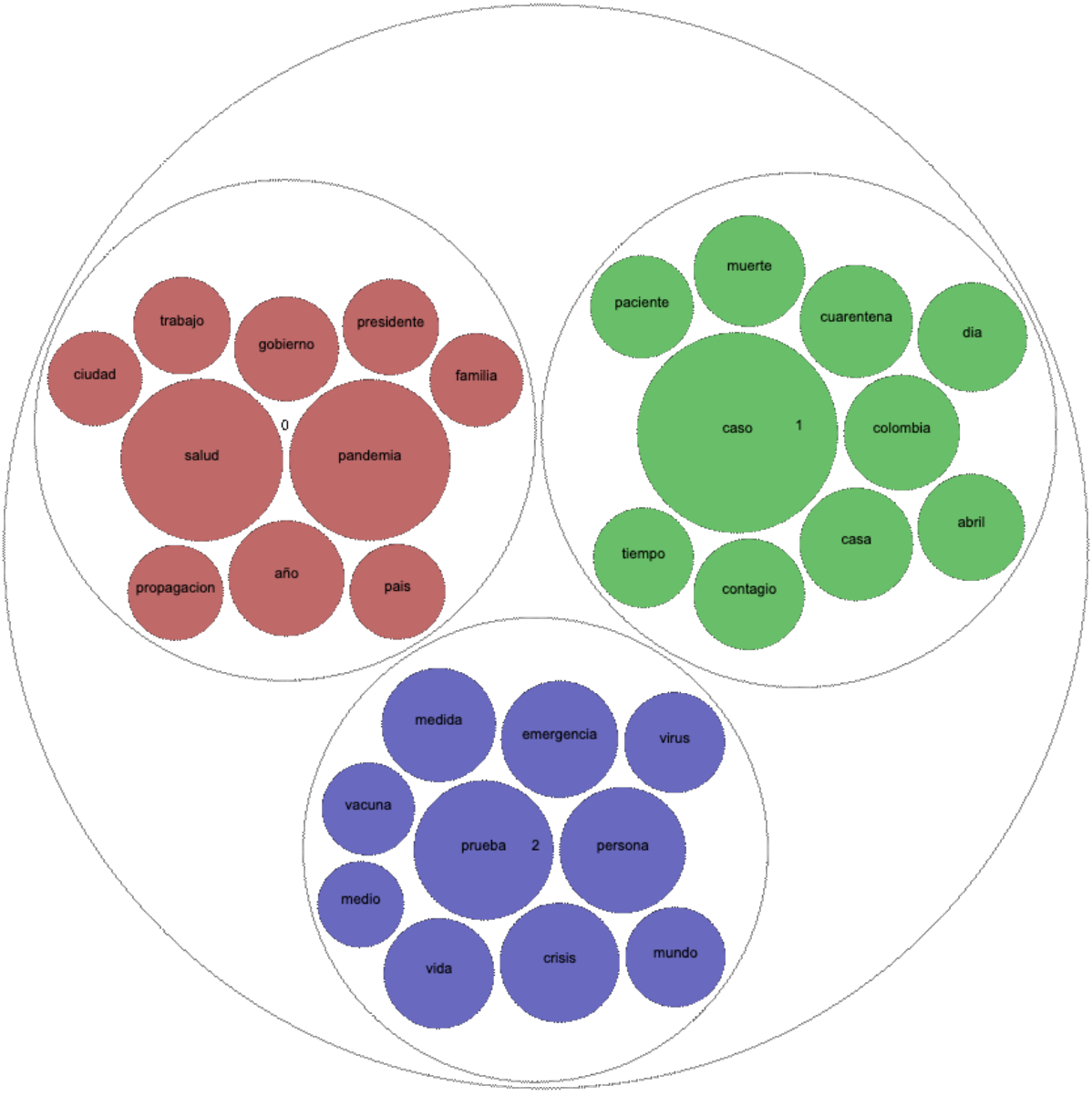

Figura 2. Tópicos de Colombia, 25 de abril de 2020. Elaboración propia.

Ahora bien, si se toma la misma visualización para el 25 de abril de 2020 en Argentina (figura 3), los resultados pueden parecer más opacos para un lector ajeno a la realidad de dicho país, pues para poder interpretar estos tópicos es imprescindible conocer el tema de actualidad de ese día en particular, pues un bebé nacido en la ciudad de Santa Fe fue nombrado como Ciro Covid el 24 de abril; entonces se puede comprender porque la pregunta "¿Quién va a ponerle de nombre a su bebé Ciro Covid?", inundó el Twitter de Argentina al día siguiente. Esto, no solo se

destaca en el tópico verde, sino que invadió también los tweets con partes diarios con datos de nuevos casos y fallecidos (tópico color rojo teja).

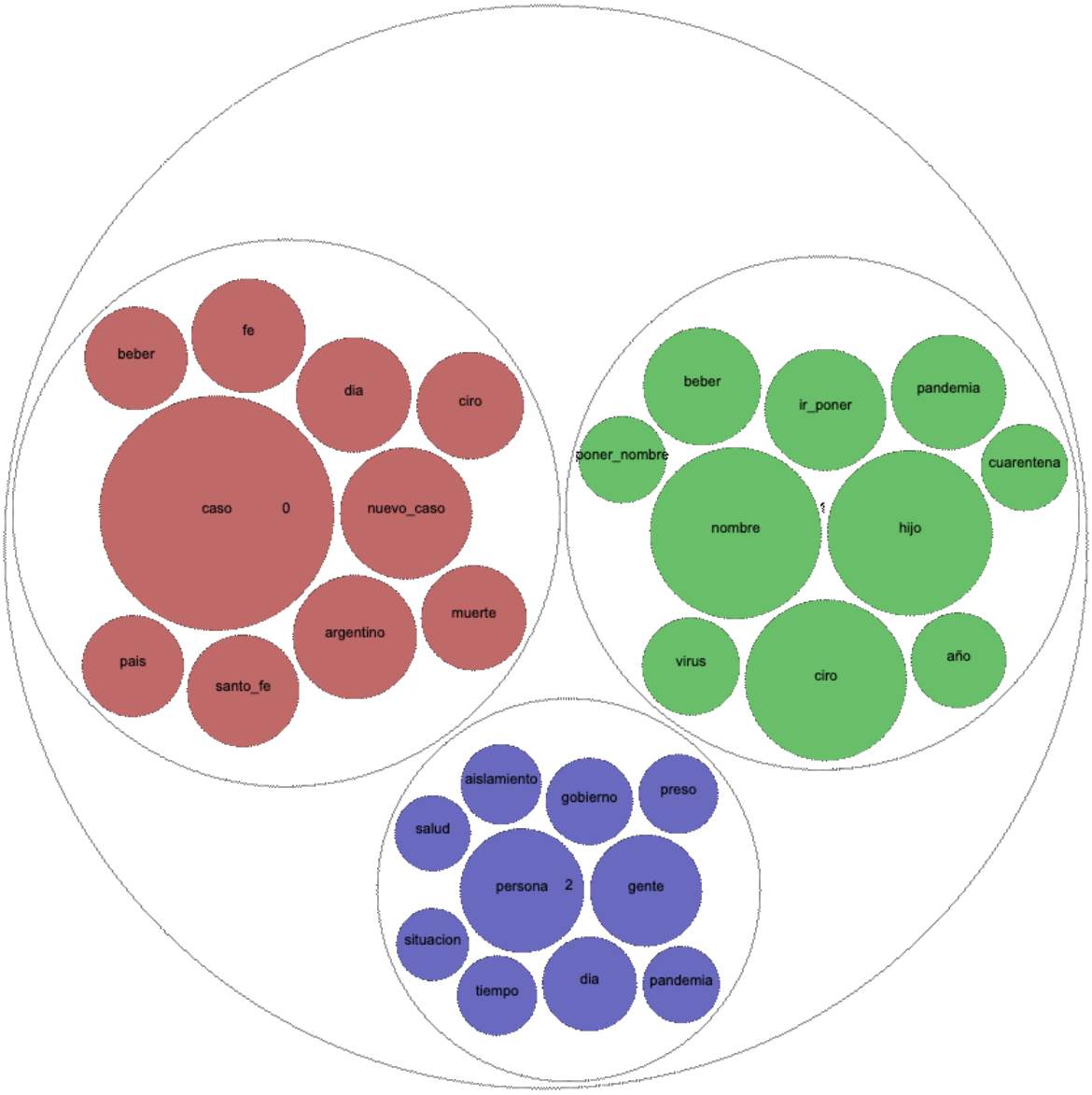

Figura 3. Tópicos de Argentina, 25 de abril de 2020. Elaboración propia.

Tópicos a primera vista opacos, como el del ejemplo anterior, exigieron un retorno a los datos iniciales que permitió igualmente constatar que muchos usuarios de medios e instituciones producen tweets duplicados y duplicados parciales. Este descubrimiento sigue la tendencia general de Twitter sobre la pandemia, la mayor parte de los tweets sobre COVID-19 provienen de cuentas automatizadas (Allyn, 2020). Este fenómeno conlleva problemas técnicos y metodológicos. En primer lugar, según Schofield et al. (2017) los duplicados degradan los resultados del topic modeling si constituyen más del 4% de la colección. A modo de ilustración de la magnitud de las cuentas automatizadas, cabe señalar que en una experiencia de comparación mensual para la región de Florida se detectaron 7% de duplicados entre los tweets en inglés y 12% de duplicados entre los tweets en español. Una rápida solución a este inconveniente puede ser eliminar estos duplicados del corpus pero esta estrategia se vincula directamente con el segundo problema: ¿en qué medida las cuentas *bot* son relevantes o irrelevantes para un estudio de narrativas sociales?

Este breve recorrido por una experiencia de aplicación de topic modeling a un corpus plurinacional y bilingüe nos permite destacar las siguientes problemáticas, la disponibilidad y calidad de datos es mayor para el subcorpus en inglés y el discurso sobre COVID-19 en Twitter presenta sobrerrepresentación artificial de ciertas cuentas institucionales y de medios.

3.3. *Analizar un corpus de Twitter con herramientas prediseñadas*

Si bien una gran parte del proyecto DHCovid estuvo dedicada a desarrollar las propias herramientas o scripts, tal y como mostramos en los anteriores apartados, otro momento del proyecto estuvo pensado como un pequeño laboratorio de prueba con herramientas prediseñadas de minería de datos. El objetivo de este trabajo consistió en cotejar los métodos que poníamos en práctica y evaluar las funcionalidades, limitaciones y ventajas de estas herramientas frente a la alta curva de aprendizaje que presentaban *coveet.py* o la manipulación de datos para trabajar en topic modeling.

En este sentido, reflexionamos en cómo incorporar dentro del currículum universitario un trabajo de este tipo de un modo propedéutico y escalonado, que nos permitiera movernos desde operaciones simples de ingesta y curaduría de datos a una reflexión sopesada que, más tarde, pudiese incorporar todas las operaciones más complejas, como las revisiondas en los apartados anteriores. Elegimos, por ende, tres plataformas que trabajan desde la minería de datos, aunque a través de procesos automáticos y a gran escala, Voyant Tools[24], Avobmat[25] y Brand24[26]. Mientras que las dos primeras son herramientas open source, la última es de tipo propietario. La práctica de laboratorio que realizamos con estas plataformas puede leerse en diferentes entradas de blog en el sitio web del proyecto. Por una cuestión de espacio, traemos aquí apenas la experiencia con la minería y análisis de datos de dos hashtags que oportunamente cruzamos, #ScholarStrike, #BlackLivesMatter y #Covid19, con la herramienta propietaria Brand24.

Si bien la pandemia de COVID-19 impuso por primera vez en años un contexto global compartido, este pronto comenzó a convivir con la coyuntura local de cada país. Así, en Twitter comenzaron a surgir hashtags específicos que daban cuenta de ese proceso de localización de la pandemia. No obstante, otros hashtags menos representativos de la situación sanitaria pronto comenzaron a resignificarse, e incluso a imponerse, dentro de este contexto. Para los Estados Unidos, este fue el caso de #BlackLivesMatter y #ScholarStrike.

[24] Accesible desde: https://voyant-tools.org/.
Accesible desde: https://avobmat.hu/.
Accesible desde: https://brand24.com/.

Trending hashtags

	HASHTAG	MENTIONS
1	#scholarstrike	427
2	#scholarstrikecanada	17
3	#125	15
4	#blacklivesmatter	13
5	#blackintheivory	7
6	#blm	6
7	#tbats	4
8	#redfored	4
9	#safetyfirst	4
10	#covid	4
11	#bidenharris2020	4
12	#reopenschoolssafely	4
13	#medievaltwitter	3
14	#breaking	2
15	#schoolsreopening	2

Figura 4. Hashtags relacionados con la búsqueda #ScholarStrike. Fuente: Brand24.

Scholar Strike fue un movimiento comunitario en las universidades que buscó reconocer el creciente número de muertes de afroamericanos y otras minorías por el uso excesivo de la violencia y la fuerza por parte de la policía. Durante dos días, del 8 al 9 de septiembre de 2020, profesores, bibliotecarios, estudiantes e incluso administrativos de universidad dejaron a un lado sus deberes y clases regulares para participar en sesiones (en algunos casos, abiertas) sobre la injusticia racial, la vigilancia policial y el racismo en Estados Unidos.

Nuestro objetivo fue extraer datos sobre esta etiqueta en Twitter, buscando asimismo coincidencias terminológicas con otros directamente relacionados, como #BlackLivesMatter, y con algunos más ligados a la crisis del coronavirus, como el esperable #Covid19 y similares. Para ello, echamos mano de la plataforma comercial de minería de Twitter llamada Brand24, una herramienta de monitoreo de redes sociales y páginas web. En una interfaz de búsqueda simple el usuario proporciona palabras clave que el software busca y analiza en varios niveles. La plataforma está principalmente orientada para análisis de marcas y el uso de esos datos en marketing digital; y ofrece diariamente una serie de resultados que son asimismo interpretados en un análisis automático en la forma de porcentajes y visualizaciones e infografías. Los resultados se envían en un email al administrador del proyecto, y a ello le sigue la posibilidad de descarga de un informe. No se puede trabajar sobre los datos, por lo tanto las opciones son limitadas a creer en los mismos o descartarlos. Cabe destacar que, para este ejercicio, utilizamos la versión trial de 7 días. A continuación, les ofrecemos un análisis de la narrativa surgida a partir de los resultado obtenidos. La primera búsqueda la realizamos el día 13 de septiembre, pudiendo obtenerse, mediante Brand24, la búsqueda retrospectiva de los últimos 30 días (desde el 14 de agosto). A las 24 horas, pudimos descargar un informe e infografía. El primero evidenciaba en términos generales, que el sentimiento respecto a la huelga fue positivo (44 positivos contra 21 negativos). Luego, la plataforma nos proyectó una visualización de los términos más destacados de todas las redes sociales. Como términos clave, *professor*, *teaching* fueron los más destacados, debido a que la huelga se dio en ese ám-

bito, sin embargo, tal como mencionamos, el entrelazamiento con el movimiento *Black Lives Matter* se hizo visible en términos como *racial*, *issues*, *september*, *police*, *injustice*, *black*.

Figura 5. Términos de mayor ocurrencia en los tweets sobre #ScholarStrike. Fuente: Brand24.

A continuación, la herramienta nos mostró los usuarios más activos y los más recientes en cuanto a su actividad en Twitter, evidentemente, todos provenían de los Estados Unidos de Norteamérica. Asimismo, se destacó la actividad constante del usuario @ISASaxonists, especialista en Literatura Medieval Anglosajona, pero no se podía relacionar entre un movimiento de reivindicación de derechos de afroamericanos con los intereses de esa cuenta en particular. En último lugar, la aplicación mostró los hashtags más usados (y relacionados entre sí): #ScholarStrike, #BlackLivesMatter, #Covid, lo que era esperable.

En la lectura cercana de los datos encontramos además una etiqueta muy específica y académica, #MedievalTwitter, en decimotercer lugar. Aunque la plataforma no lo explicita, entendimos que debía estar relacionado con ese usuario @ISASaxonists. Tweets y respuestas nos mostraban que el hashtag #MedievalTwitter y los tweets del usuario @ISASaxonists estaban relacionando y cuestionando las acusaciones que ocurrieron en 2019 a la Sociedad Internacional Anglosajona por su inhabilidad de dar cuenta de problemas de racismo, sexismo, diversidad e inclusión dentro de la misma. Encontramos que parte de esta discusión había sido publicada en medios y revistas de investigación en Estados Unidos durante septiembre de 2019[27].

[27] Un ejemplo de este debate es el artículo sobre el uso racista y obsoleto del término Anglo-Saxon Studies, publicado en el sitio Inside Higher Ed en el año 2019: https://www.insidehighered.com/news/2019/09/20/anglo-saxon-studies-group-says-it-will-change-its-name-amid-bigger-complaints-about.

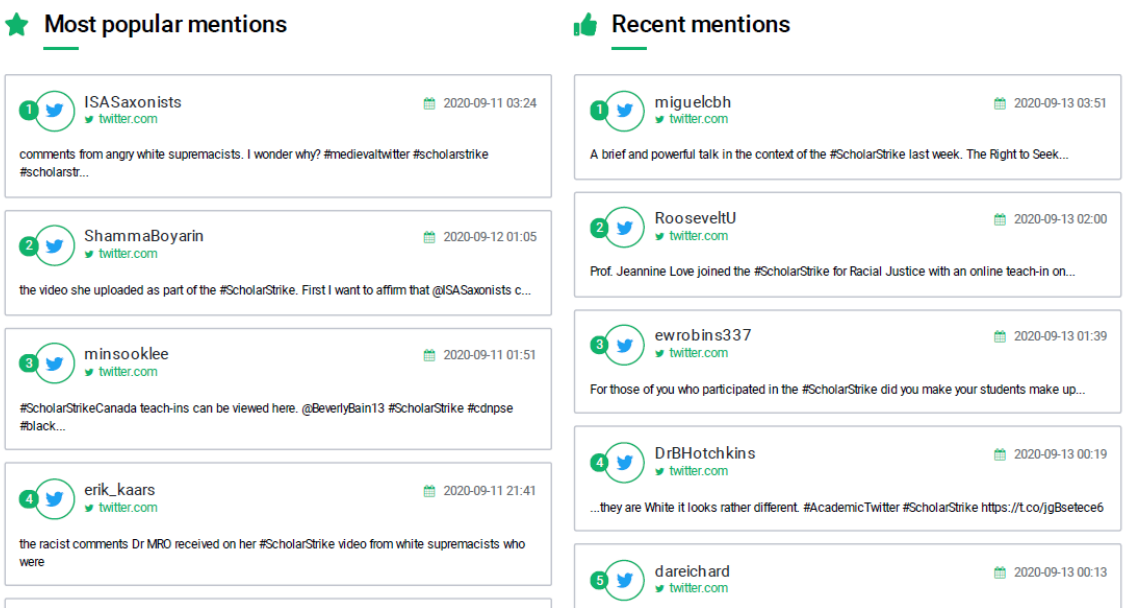

Figura 6. Relación del usuario @ISASaxonists con el hashtag #MedievalTwitter y #ScholarStrike. Fuente: Brand24.

En conclusión, explorar el contexto de #ScholarStrike con Brand24 nos permitió constatar algunas suposiciones previas (su relación con etiquetas como #BlackLivesMatter y #Covid19) pero iluminó otros hashtags menos esperables para un usuario no académico, como #MedievalTwitter, y otros que aparecían tímidamente, pero que pronto comenzaron a tener más impacto en las semanas siguientes, con la carrera electoral, como #bidenharris2020.

Como decíamos al comienzo de este apartado, si pensamos en trasladar la experiencia del proyecto al curriculum universitario, enfocar el marco teórico del análisis cuantitativo y la reflexión crítica sobre los datos desde el uso de este tipo de herramientas puede resultar provechoso para realizar primeras experiencias y comprender mejor métodos y procedimientos computacionales complejos.

6. CONCLUSIONES

A modo de conclusión podemos afirmar que DHCovid ha entendido Twitter como un gran corpus desde los marcos teóricos y metodologías que hoy hacen al campo de las HD. Como proyecto de Humanidades, ha tomado principalmente en consideración las variantes lingüísticas y regionales de los datasets de tweets, buscando explorar y arrojar luz sobre las narrativas digitales de la COVID-19 en Twitter. Como cualquier iniciativa experimental, el trabajo surge de la voluntad de explorar un conjunto de datos, cuyas dimensiones sobrepasan por mucho la capacidad analítica cualitativa y hacen necesaria la aplicación de técnicas de lectura distante. Con la ayuda de un grupo formado por expertos en diferentes áreas y estudiantes graduados, hemos trabajado en diferentes frentes valorando el uso de varias herramientas. El reto, sin embargo, no ha estado exento de dificultades y de limitaciones de las que nos gustaría dejar aquí constancia.

En primer lugar, el propio Twitter impuso una serie de restricciones para poder minar a través de su API, tiempo, cantidad, y geolocalización. En su versión gratuita, hemos de destacar, ape-

nas permitía recuperar tweets publicados en los 7 días anteriores al momento de la búsqueda. Así, cuando empezamos el experimento en abril 2020 no pudimos recuperar los tweets que habían sido producidos desde el inicio de la pandemia en enero 2020. En segundo lugar, otro escollo que encontrábamos era que la cantidad de tweets recuperados no siempre era idéntica a la total producida en la red social, pues cada cuenta tenía un límite de consultas (*requests*) en el trabajo de minería. Dado que, en muchos casos, los tweets superaban el límite permitido, debimos dejar material fuera del corpus. En tercer lugar, la geolocalización de los tweets es privada en la mayoría de los casos, y Twitter no ofrece una consulta por *país*, por lo que la zona geográfica tuvo que ser delimitada manualmente. Esto significó que algunas zonas, como Ecuador, pudieran tener contenidos de las zonas limítrofes de Colombia o Perú.

Ha de destacarse que, la aplicación de frecuencias –hechas a través de la herramienta coveet.py– ofrece resultados interesantes pero no siempre fáciles de interpretar, especialmente si tenemos en cuenta que cantidad no es sistemáticamente sinónimo de relevancia. Si bien los primeros resultados señalan pocas novedades en cuanto a temas significativos relacionados con la pandemia, un análisis más detenido, especialmente de los hashtags, puede arrojar luz sobre los diferentes temas sociales que han preocupado a las diferentes regiones, e incluso compararse con otros corpus en otras lenguas.

Finalmente, y muy relevante en lo que concierne al *topic modeling*, cabe señalar la existencia de importantes desbalances en el corpus que inciden directamente en la calidad de los resultados. Mientras que en el trabajo con países latinoamericanos encontramos que Ecuador emite normalmente menos de 5 mil tweets semanales, México puede llegar a superar los 65 mil tweets semanales. En la región de Florida hallamos que los tweets en inglés pueden llegar a cuadruplicar los tweets en español en el mismo período (~60 mil tweets en inglés vs. ~15 mil tweets en español para el período junio a septiembre de 2020), un dato que no dejó de sorprendernos fue la ingente población hispana de la región[28]. Asimismo, como cuestión relacionada con la lectura distante, es reseñable la dificultad que encontramos al abordar un corpus que se presentaba más a menudo desconocido, en temas, menciones de personas, eventos, etc. Esta afirmación, por obvia que parezca, supone la constatación de que un método cuantitativo, especialmente en el ámbito de las Humanidades, no puede renunciar al análisis cualitativo y a la necesidad de contextualización de la investigación –en este caso de la actualidad social de cada país– tanto previo como posterior a la obtención de los resultados.

En fin, DHCovid nos ha ofrecido un terreno productivo del que han surgido algunas preguntas metodológicas sobre la misma naturaleza de las redes sociales. ¿Puede ser Twitter concebido como un corpus lingüístico cuando no controlamos dimensiones o contenido? ¿Qué herramientas necesitamos para que Twitter pueda ser considerado objeto de estudio para las HD? ¿Es el estudio

[28] Más información sobre el volumen de tweets por día correspondiente a cada zona, accesible desde: https://covid.dh.miami.edu/es/graficas/.

de las redes sociales parte de las HD?

Pese a estos obstáculos, DHCovid ofrece actualmente un corpus que cubre el arco de un año de la pandemia de la COVID-19, de mayo 2020 a mayo 2021, lo que representa un escenario interesante para investigar variaciones temáticas. Se suma a ello el esfuerzo dual de haber minado tweets en inglés (Sur de Florida) y en español a nivel internacional (Sur de Florida, Colombia, Perú, Ecuador, Argentina, España), generando un corpus específico que permite la comparación de lo local con la tendencia global.

REFERENCIAS BIBLIOGRÁFICAS

Abdo, M. S., Alghonaim, A. S., y Essam, B. A. (2020). Public perception of COVID-19's global health crisis on Twitter until 14 weeks after the outbreak. *Digital Scholarship in the Humanities*, fqaa037. https://~.com/fqaa037.pdf?~U

Aiden, E., Michel, J. B. (2013). *Uncharted. Big Data as a Lens on Human Culture*. Penguin Group.

Allés Torrent, S., del Rio Riande, G., Hernández, N., Bonnell, J., Song, D., y De León, R. (2020). *Digital Narratives of Covid-19: a Twitter Dataset* (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3824950

Allyn, B. (2020). Researchers: Nearly Half of Accounts Tweeting About Coronavirus Are Likely Bots. *NPR.Org*, May 20, 2020. https://www.npr.org/~researchers-nearly-half-of-accounts-tweeting-about-coronavirus-are-likely-bots

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, K., Tutubalina, E., y Chowell, G. (2021). *A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research - An International Collaboration* [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.4460047

Chen, E., Lerman, K., y Ferrara, E. (2020). Tracking Social Media Discourse about the Covid-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance*, *6*(2), e19273. https://doi.org/10.2196/19273

Grainge, P. (2011). *Ephemeral Media: Transitory Screen Culture from Television to YouTube*. British Film Institute.

Grandjean, M. (2016). A Social Network Analysis of Twitter: Mapping the Digital Humanities Community. *Cogent Arts y Humanities*, 3(1). https://doi.org/10.1080/23311983.2016.1171458

Jockers, M. L. (2013). *Macroanalysis: Digital Methods y Literary History*. University of Illinois Press.

Kerchner, D., y Wrubel, L. (2020). Coronavirus Tweet Ids [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/LW0BTB

Lamsal, R. (2020). Coronavirus (COVID-19) Tweets Dataset [Data set]. IEEE. https://.doi.org/10.21227/781w-ef42

Manovich, L. (2009). Cultural Analytics: Visualizing Cultural Patterns in the Era of 'More Media'. http://manovich.net/index.php/projects/cultural-analytics-visualizing-cultural-patterns

Moretti, F. (2005). *Distant Reading*. London: Verso.

Quan-Haase, A., Martin, K., y McCay-Peet, L. (2015). Networks of Digital Humanities Scholars: The

Informational and Social Uses and Gratifications of Twitter. Big Data y Society, 2(1). https://doi.org/10.1177/2053951715589417

Schofield, A., Magnusson, M., Thompson, L., y Mimno, D. (2017). Understanding text pre-processing for latent Dirichlet allocation. In Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics (Vol. 2, pp. 432-436). https://www.cs.cornell.edu/~xanda/winlp2017.pdf

Stefanidis, A., Vraga, E., Lamprianidis, G., Radzikowski, J., Delamater, P. L., Jacobsen, K. H., Pfoser, D., Croitoru, A., y Crooks, A. (2017). Zika in Twitter: Temporal Variations of Locations, Actors, and Concepts. JMIR Public Health and Surveillance, 3(2), e22. https://doi.org/10.2196/publichealth.6925