EMPIRIA. Revista de Metodología de las Ciencias Sociales
ISSN: 1139-5737
ISSN: 2174-0682
empiria@poli.uned.es
Universidad Nacional de Educación a Distancia
España

Gualda, Estrella; Rebollo, Carolina
Big data y Twitter para el estudio de procesos migratorios: Métodos, técnicas de investigación y software
EMPIRIA. Revista de Metodología de las Ciencias Sociales, núm. 46, 2020, Marzo, pp. 147-177
Universidad Nacional de Educación a Distancia
Madrid, España

DOI: https://doi.org/10.5944/empiria.46.2020.26970

Big Data y Twitter para el estudio de procesos migratorios: Métodos, técnicas de investigación y software

Big Data and Twitter for the study of migration processes: Methods, research technics, and software

Estrella Gualda

Universidad de Huelva, Grupo ESEIS y COIDESO
estrella@uhu.es (ESPAÑA)

Carolina Rebollo

Universidad de Huelva, Grupo ESEIS

Recibido: 01.11 2018
Aceptado: 30.01.2020

RESUMEN

En este artículo, basado en una revisión bibliográfica sobre lo que se ha publicado en revistas científicas internacionales sobre migraciones y big data así como sobre migraciones y Twitter, en tanto que un medio social específico, se pretende identificar qué tipo de investigaciones sobre migraciones se están publicando actualmente basándose en datos que proceden de estas fuentes. Y, particularmente, nos interesa identificar los métodos, técnicas de investigación y tipo de software que se manejan en estos trabajos, desde la perspectiva de tres momentos clave en el estudio de lo que se publica en los medios sociales o el uso de big data: extracción, procesamiento y análisis. Nuestra revisión se desarrolla teniendo en cuenta lo que se publicó en los últimos 5 años incluido en las bases de datos ProQuest, Scopus y Web of Science, que recogen miles de revistas, libros, tesis doctorales, etc. de carácter multidisciplinar. Los resultados apuntan a que son aún pocas las publicaciones en materia de big data y Twitter que abordan procesos migratorios, en relación a otras temáticas. En cuestión de estrategias metodológicas, técnicas y software, los artículos que hemos encontramos van desde lo más artesanal a lo más sofisticado, en este caso, con publicaciones que suelen estar encabezadas por científicos que cuentan con cierto bagaje en computación.

PALABRAS CLAVE

Datos sociales masivos - Datos masivos – Twitter – Procesos migratorios – Métodos de investigación - Software.

ABSTRACT

In this article, based on a literature review on the publications in international scientific journals on migrations and big data as well as specifically on migrations and Twitter, as a specific social media, we try to identify what type of research on migrations is currently being published based on data that comes from these sources. And, particularly, we are interested in identifying the methods, research techniques and type of software that are handled in these works, from the perspective of three key moments in the research on social media or big data: extraction, processing, and analysis. Our review is developed taking into account what was published in the last 5 years that was included in ProQuest, Scopus and Web of Science databases, which comprise thousands of journals, books, doctoral theses, etc. of a multidisciplinary nature. The results suggest that there are still few publications on big data and Twitter that address migration processes, in relation to other issues. In terms of methodological strategies, techniques and software, the articles we have found range from the most artisan to the most sophisticated, in this case, with publications that are usually lead by scientists who have a certain background in computing.

KEY WORDS

Social Big Data - Big data – Twitter – Migration processes – Research methods – Software

1. SOCIAL BIG DATA, TWITTER Y PROCESOS MIGRATORIOS

Nadie puede dudar hoy de la importancia a diferentes niveles que los datos y las discusiones generados a través de plataformas de medios sociales y por diversos actores pueden tener hoy en la conformación de las sociedades modernas, donde aspectos como los procesos de comunicación, a su vez mediados por el ordenador (Olshannikova et al., 2017) se han visto afectados por el importante desarrollo tecnológico que ha supuesto la eclosión de la llamada sociedad 2.0 (Gualda, 2018).

Prestar atención a los procesos que ocurren en el escenario de las llamadas "redes sociales" nos puede ayudar a mejorar nuestra comprensión de las sociedades actuales, pero al mismo tiempo nos confronta, especialmente a las Ciencias Sociales, con importantes retos metodológicos. No en vano el abordaje de lo que

se difunde a través de estas redes es metodológica y técnicamente diferente, en multitud de ocasiones, a lo que nuestros clásicos *know-how* venían aportando. Y esto ocurre no solo en el campo de especialización de los estudios migratorios, sino en otros como los relativos a la comunicación política, estudios sanitarios, etc.

En este artículo, a partir de una revisión de la bibliografía existente en este campo, enriquecida con nuestra experiencia investigadora al respecto en los últimos años, nos planteamos estudiar qué tipo de investigaciones se están realizando internacionalmente en el área de confluencia entre *big data* y estudios migratorios (desde una perspectiva más amplia) y Twitter y estudios migratorios (acotando a una plataforma de redes sociales[1]), con el fin de conocer las aportaciones metodológicas que para las Ciencias Sociales supone abordar estos campos de estudio, así como plantear horizontes de interés e identificar aspectos de carácter técnico que merece la pena resolver para el desarrollo de estudios migratorios basados en la información difundida a través de diversas plataformas de redes sociales.

Respecto a la intensidad del desarrollo de las redes sociales, hay multitud de evidencias, a lo que se suma nuestra experiencia cotidiana. Para visibilizar su importancia en el mundo actual, solo hay que prestar atención a lo rápido que el uso de estas redes sociales se ha popularizado desde su nacimiento en los inicios del siglo XXI (por ejemplo, Facebook nace en 2004, Youtube en 2005, Twitter en 2006 o Instagram en 2010), lo cual va ligado directamente al desarrollo tecnológico y muy especialmente al de las tecnologías en torno a los smartphones o dispositivos con similares funcionalidades e internet. El último informe de la Fundación Telefónica (2019:27) permite visibilizar muy bien esta cuestión cuando se indica que "*El número de líneas móviles existentes en el mundo ya ha superado a la población mundial*", lo que se acompaña igualmente de datos que muestran que los smartphones han cobrado un gran protagonismo*: "En 2017, existían en el mundo 972 millones de líneas de telefonía fija, 7.740 millones de líneas de telefonía móvil y 3.578 millones de usuarios de Internet"* (Fundación Telefónica, 2017:96). Por otra parte, el número de usuarios de internet en el mundo, a 30 de junio de 2019, superaba ya los 4,536,248,808 lo cual no puede pasar desapercibido. De estos usuarios 727,559,682 se encontraban en Europa, donde la tasa de penetración es el 87,7% detrás de la de Norteamérica (que alcanzaba el 89,4%) y a una gran distancia de la última, África (con el 39,6% de usuarios de internet frente al total de su población). Y entre los europeos se cuenta también en estas fechas con un volumen sustancial de usuarios de redes como Facebook (340,891,620 a 31 de diciembre de 2018) (Internet World Stats, 2019, https://www.internetworldstats.com/stats4.htm).

Esta eclosión y elevado uso de las redes sociales con variados fines (difundir información, influenciar la opinión pública, sensibilizar, hacer campañas, o

[1] Twitter es un servicio de microblogging que permite enviar o recibir mensajes breves o tuits a los usuarios de esta plataforma de redes sociales en internet, tanto de manera pública como privada (a través de mensajes directos en este último caso).

incluso manipular o trasladar mensajes racistas, citando algunos), se acompaña, cada vez más de la conciencia de que es necesario en las Ciencias Sociales la recolección y análisis de información que de forma masiva los ciudadanos volcamos en internet. Bednár (2017) subraya cómo una parte sustancial de la interacción social hoy es mediada por diferentes servicios on line sociales generándose una ingente cantidad de datos. Esto pone sobre la mesa la necesidad de que aclaremos qué se entiende en este artículo por *big data* y *social big data* (datos sociales masivos o grandes datos sociales) como focos de interés de este texto, quizás más desconocidos al menos conceptualmente que Twitter[2].

Siguiendo a Gandomi y Haider (2015), el tamaño es quizás la única dimensión que el término *big data* sugiere a primera vista, aunque otros rasgos característicos, no siempre tan visibles, pueden ayudar a conformar una definición, como es el caso de las 3 Vs (volumen, velocidad y variedad), extendido también a 5 Vs con la incorporación de valor y veracidad (Bello-Orgaz, Jung y Camacho, 2016), como términos que se han usado para definir a los datos masivos. Otros autores recientes, como Olshannikova et al. (2017), han tendido a referirse a otras Vs para describir otros aspectos característicos de los big data. La referencia a *big data* pone sobre la mesa aspectos que son importantes para la toma de decisiones (añadiendo valor) como son el uso de métodos analíticos para manejar datos heterogéneos que llegan en formatos tanto estructurados como no estructurados (en este último caso, la mayoría –textos, fotos y video, por ejemplo-), para realizar análisis predictivos o desarrollar algoritmos eficientes que puedan manejar estos ingentes volúmenes de datos. También es importante saber que la referencia a big data va más allá de lo que específicamente se publica en medios sociales y abarca datos que pueden proceder de otras fuentes (registros públicos o censos, por ejemplo) pero que alcanzan esta gran dimensión, planteando retos como el almacenamiento o el procesamiento y análisis de los mismos. Es importante por esto introducir la distinción entre *big data* y *social big data*, pues este último término sí se refiere más explícitamente a datos procedentes de medios sociales.

De acuerdo con Bello-Orgaz, Jung y Camacho (2016), el campo de los *social big data* es producto de la confluencia de tres grandes áreas: *social media* (medios sociales), *data analysis* y los *big data* (datos masivos) conformándose como un área interdisciplinar donde los medios sociales cuentan con una gran relevancia. En esta nueva área de trabajo se desarrollan métodos que pretenden generar conocimientos (*data analysis*) a partir del procesamiento y análisis de las informaciones procedentes de los medios sociales en línea como fuentes de datos de gran tamaño, con formatos diferentes, la existencia de datos de carácter más estático y otros que se recolectan en directo o streaming (Bello-Orgaz, Jung y Camacho, 2016; Bednár, 2017). En algunas investigaciones recientes se combinan incluso capas de información de muy diferente tipo, por ejemplo, en el trabajo de Chattopadhyay & Chattopadhyay (2017), para estudiar las pautas

[2] Entendemos que cualquier investigador que preste atención mínimamente a su alrededor ha podido conocer de una forma u otra algo sobre Twitter, incluso a través de las noticias televisadas.

migratorias de la población rural en el área del Himalaya en la India, se llevan a cabo análisis que combinan datos geofísicos, geológicos y socioeconómicos.

Antes de definir los parámetros de nuestra revisión bibliográfica, cosa que haremos en la sección de métodos, nos parece de interés situar la evolución internacional de la bibliografía publicada sobre *big data* y Twitter, tanto de forma global, como de forma específica, cuando esta aborda cuestiones migratorias, a fin de aproximarnos inicialmente a la dimensión de este foco de análisis.

La Figura 1 presenta sintéticamente la evolución del número de artículos científicos sobre estos temas durante los últimos cinco años (2014-2018), a partir de tres bases de datos relevantes: (1) ProQuest[3], que comprende un conjunto amplio de bases de datos de carácter multidisciplinar y proporciona una gran colección de tesis, artículos de publicaciones periódicas, libros electrónicos académicos y otro contenido como datos o informes; (2) Scopus[4], que es una base de datos de referencias bibliográficas y citas de diferentes ramas científicas propiedad de Elsevier que, entre otros recursos, contiene 20.000 revistas revisadas por pares y 21.000 títulos de más de 5000 editores internacionales, con cobertura desde 1996; (3) WoS, Web of Science[5], como plataforma que permite acceder a las referencias de las principales publicaciones científicas de cualquier disciplina desde 1945.

Observamos en la Figura 1 la pauta constante de crecimiento del número de artículos publicados entre 2014 y 2018 sobre big data y Twitter, aunque las cifras varían según las bases de datos. Se aprecia también que el incremento proporcional de publicaciones en este período es superior cuando se trata de big data que de Twitter, lo que ocurre en las tres bases de datos consultadas. Es mayor, por otra parte, el número de artículos que contienen estos temas en ProQuest y Scopus que en WOS, lo cual tiene sentido por cuanto en Web of Science se encuentran los artículos de revistas indexadas al máximo nivel que forman parte de los JCR (Journal Citation Reports), por lo que se encuentran en ella los artículos publicados en las revistas más relevantes, un menor número de los que se contienen en bases de datos que se guían por otros criterios no tan exigentes. Destaca especialmente el incremento que se produce en Scopus en cuanto a las publicaciones relativas a big data a partir de 2016, lo que da muestras de la notoriedad que ha alcanzado este tema en diversas ciencias.

Por otra parte, lo que más nos interesa del gráfico, más que estos datos globales, es la observación de que solo en un reducido número de artículos se abordan simultáneamente las cuestiones de big data y migraciones. Esta información se encuentra en la Figura 1, pero igualmente, se ha desglosado en la Figura 3, observándose que en el período de 2014 a 2018, en ProQuest encontramos que las cuestiones migratorias se conjugan con las de big data en un 11,1% de artículos, y en un 14,9% cuando se trata de Twitter. Esta proporción es menor en el caso de Scopus (6,0% lo que representan los artículos sobre big data y migracio-

[3] Proquest (2019): https://www.proquest

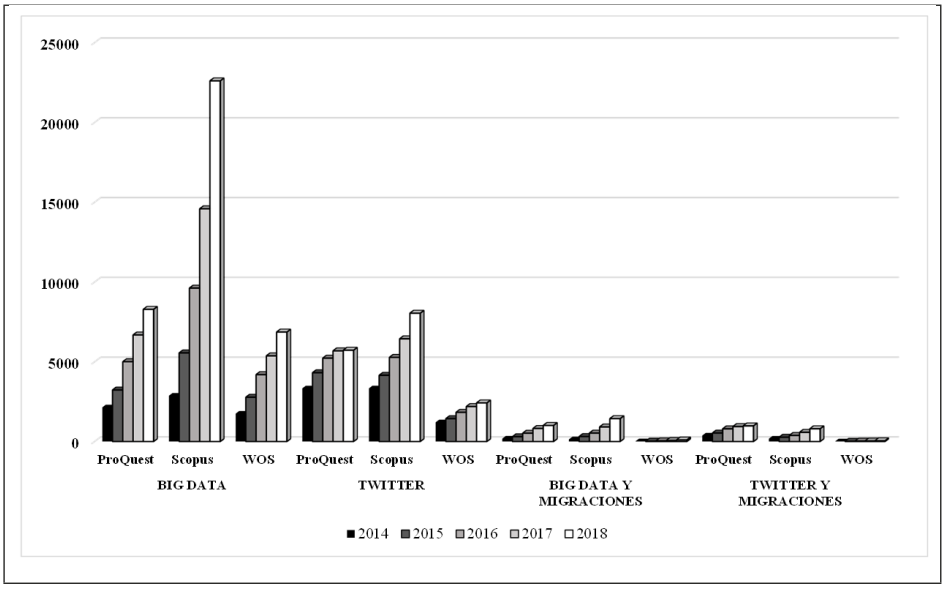
[4] Scopus (2019): https://www.fecyt.es/es/recurso/scopus.

[5] Web of Science (2019): https://www.fecyt.es/es/recurso/web-science.

nes en relación a lo que se publica sobre big data; y 8,1% en cuanto a Twitter y migraciones). La proporción se reduce sustancialmente en el caso del peso que suponen los artículos de big data y migraciones en WoS (1,1%) y Twitter y migraciones en la misma fuente (1,5%).

Aunque la tendencia temporal es claramente hacia el crecimiento del número de artículos que abordan las cuestiones relativas a big data o Twitter y las migraciones, se aprecia que aún el interés por conectar estos campos es incipiente y, de forma comparada a otros campos temáticos (comunicación política, salud, educación, marketing, etc.) su presencia es proporcionalmente menor, aunque no debe despreciarse su importancia a tenor de la juventud que caracteriza a las plataformas de redes sociales en el siglo XXI.

Figura 1. Evolución del número de artículos científicos en ProQuest, Scopus y WOS. Datos absolutos (2014-2018)



Fuente: Elaboración propia a partir de la consulta en las bases de datos de ProQuest, Scopus y WoS.

Nota: Las bases de datos, revistas y palabras claves consultadas para la elaboración de este gráfico se encuentran descritas en la Figura 2.

1.1. Objetivos e hipótesis en la revisión bibliográfica

Aunque se han avanzado sintéticamente en la sección anterior, este trabajo lleva a cabo una revisión bibliográfica sistemática para identificar en varias bases de datos científicas relevantes qué artículos científicos se han publicado contemporáneamente cuando confluyen los estudios migratorios con los basados en big

data, así como, específicamente, en Twitter. El artículo ha querido responder a varios objetivos. En primer lugar (1), de forma descriptiva, quiere conocer si en la bibliografía reciente se están publicando trabajos sobre cuestiones migratorias y big data, por una parte, así como si en el campo de los estudios migratorios se está indagando sobre las migraciones y los medios sociales (particularmente a través de Twitter). En segundo lugar (2), nos interesa conocer igualmente si los vínculos en la bibliografía entre big data-estudios migratorios, o Twitter-estudios migratorios ocupan o no un lugar destacado entre los artículos científicos recientes sobre big data y Twitter. En tercer lugar (3), nuestro artículo tiene como objetivo analizar la literatura de estudios migratorios que se aborda desde la perspectiva de los datos masivos o que toma como referencia un medio social como Twitter para identificar cuáles son los enfoques metodológicos, así como las estrategias, herramientas y software que se están empleando en estos trabajos con el fin de conocer las aportaciones metodológicas que para las Ciencias Sociales supone emprender estos campos de estudio, así como plantear horizontes y retos de cara al futuro. Nos preguntamos, por tanto, por los enfoques metodológicos, los métodos y técnicas de investigación usados específicamente cuando se llevan a cabo estudios migratorios donde se tienen en cuenta tanto aproximaciones basadas en el uso de big data como de Twitter en relación a las migraciones, como ejemplo relevante de medio social cuya información se encuentra en abierto y accesible fácilmente a los investigadores. Pero dada la asociación de este foco de investigación con aspectos como la computación o el desarrollo del software, queremos identificar igualmente qué tipo de estrategias, herramienta e incluso software se han ido desarrollando o utilizado cuando se llevan a cabo estudios migratorios desde esta perspectiva, con el fin de detectar posibles lagunas o puntos donde sea viable hacer contribuciones, así como facilitar a los investigadores un espectro de posibilidades metodológicas que se están empleando y que responden a diferentes niveles de complejidad técnica. Por otra parte, es también uno de los objetivos del artículo identificar qué métodos, técnicas y herramientas tienden a usarse al conjugar estudios migratorios con big data y Twitter en momentos claves del proceso de investigación: (3.1) la elección de las fuentes de información; (3.2) la extracción de los datos y las herramientas para la recolección de los mismos; (3.3) el procesamiento de la información obtenida y (3.4) el análisis, como objetivos específicos que se desprenden de nuestro interés metodológico en este artículo.

La novedad que en el panorama científico y social suponen la emergencia tanto de los datos masivos o big data como del desarrollo de los medios sociales (entre los cuales se encuentra Twitter), o la misma novedad que incluso supone la popularización del contacto continuo con Internet a través de tecnologías como las implementadas en los smartphones, nos han hecho tener como hipótesis principal en este trabajo (en relación a nuestros objetivos 1 y 2) que los estudios científicos en los que confluyen big data o Twitter con las migraciones representan aún una parcela no muy grande significativamente, a lo que se añade, en relación a nuestro tercer objetivo, hipotetizamos, (3) una posible diferenciación en cuanto a la investigación científica en estos campos derivada

de los diferentes perfiles de especialistas o equipos de investigación que se han aproximado a este tipo de estudios, de forma técnicamente más avanzada, en investigaciones principalmente que aplican técnicas de extracción, procesamiento y análisis tecnológicamente más avanzadas, frente a expertos de otros campos de las ciencias sociales y humanas, en los que se aborda el estudio de las migraciones desde posturas más tradicionales y menos ambiciosas quizás técnicamente, más orientadas en relación a Twitter que a big data, donde existen más dificultades para su manejo.

2. MÉTODOS

Este artículo se fundamenta en una revisión bibliográfica que intenta identificar y analizar los aspectos metodológicos de las investigaciones que abordan los estudios migratorios cuando se conectan con el uso de big data y Twitter. Como indican Guallar y otros (2017:949) los estudios de revisión "*forzosamente… tienen que realizar una selección de títulos*" susceptibles de análisis. En este apartado vamos a explicar el proceso que hemos seguido para la selección y revisión de los mismos y la manera en la que hemos abordado el análisis de los aspectos metodológicos que los artículos identificados comprenden.

3 CRITERIOS DE BÚSQUEDA PARA LA REVISIÓN BIBLIOGRÁFICA

Se ha realizado una revisión sistemática de los artículos publicados sobre estudios migratorios que se encontraban basados en big data o en datos de una red social como Twitter. Para ello, se realizaron varias búsquedas por separado en las bases de datos de ProQuest, Scopus y WoS (tal y como se describe en la Figura 2): en relación a big data y migraciones y sobre Twitter y migraciones. Nos ha interesado diferenciar entre big data y Twitter a fin de poder comparar entre una red social específica de carácter público sobre la que cualquier investigador puede plantearse la recolección de datos y el amplio campo de big data, donde junto al producto de los medios sociales, pueden encontrarse otras cuestiones metodológicas asociadas al desarrollo de la *data science, big data analytics, business intelligence*, etc., en este caso en su particular aplicación a los estudios migratorios.

A partir de los criterios de búsqueda delimitados y con el objeto de permitir la comparación entre fuentes, se seleccionaron todos los artículos publicados entre 2014 y 2018, limitados únicamente por tipo de fuente (revistas científicas) y por tipo de publicación (artículos) como principales productos científicos comparables en las tres bases de datos, asumiendo que esto es una limitación que aunque facilita la comparabilidad, no comprende todo tipo de productos científicos, como puedan ser libros o tesis doctorales, si bien permite centrarse en los

artículos indexados más relevantes, especialmente respecto a Scopus y WoS, así como en otros obtenidos a partir de una de las bases de datos principales para las Ciencias Sociales, ProQuest, que comprende un abanico muy diverso de revistas científicas. El periodo de búsqueda, los términos empleados y las bases de datos consultadas se describen en la tabla siguiente (Figura 2).

Figura 2. Periodo de búsqueda, fuente, base de datos, términos utilizados y criterios de búsqueda para la revisión bibliográfica

Período de búsqueda	2014 a 2018 (5 años).
Fuentes	Se han usado tres bases de datos relevantes, descritas anteriormente: (1) ProQuest; (2) Scopus; (3) WoS, Web of Science.
Bases de datos multidisciplinares consultadas	**ProQuest:** Se buscó en las 23 bases de datos que contiene, que comprenden a su vez miles de artículos, esto es: ABI/INFORM Collection, Accounting, Tax & Banking Collection, Bibliografía de la Literatura Española, Dissertations & Theses, EconLit, Entrepreneurship Database, ERIC, GeoRef, Health & Medical Collection, International Bibliography of the Social Science (IBSS), MEDLINE, MLA International Bibliography, Nursing & Allied Heatlh Database, Periodicals Archive Online, Periodical Index Online, PsycARTICLES, Psychology Database, PsycINFO, PTSDpubs, Sociology Collection. **Scopus:** Es una base de datos bibliográfica de resúmenes y citas de artículos de revistas científicas. Cubre aproximadamente 18.000 títulos de más de 5.000 editores internacionales, incluyendo la cobertura de 16.500 revistas revisadas por pares de las áreas de ciencias, tecnología, medicina y ciencias sociales, incluyendo artes y humanidades. **WoS, Web of Science:** Se realizó la búsqueda en Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, Index Chemicus, Current Chemical Reactions, Conference Proceedings Science, Conference Proceedings Social Science & Humanities, Medline, Scielo Citation Index, KCI: Korean Journal Database, Emerging Sources Citation Index.

Términos de búsqueda	**Búsqueda 1:** (“Big Data” AND migration) OR (“Big Data” AND immigration) OR (“Big Data” AND emigration) OR (”Big Data” AND migrants) OR (“Big Data” AND refugees) OR (”Big Data” AND ethnic) OR (”Big Data” AND displaced) OR (”Big Data” AND asylum) OR (”Big Data” AND migraciones) OR (”Big Data” AND inmigración) OR (”Big Data” AND emigración) OR (”Big Data” AND migrantes) OR (”Big Data” AND refugiados) OR (”Big Data” AND etnia) OR (”Big Data” AND desplazados) OR (”Big Data” AND asilo) **Búsqueda 2:** (twitter AND migration) OR (twitter AND immigration) OR (Twitter AND emigration) OR (twitter AND migrants) OR (Twitter AND refugees) OR (Twitter AND ethnic) OR (twitter AND displaced) OR (twitter AND asylum) OR (twitter AND migraciones) OR (twitter AND inmigración) OR (twitter AND emigración) OR (twitter AND migrantes) OR (twitter AND refugiados) OR (twitter AND etnia) OR (twitter AND desplazados) OR (twitter AND asilo)	
Criterios de búsqueda	Tipo de fuente: Revistas científicas	Tipo de documento: Artículos

Fuente: Elaboración propia.

4. PROCESO SEGUIDO PARA LA REVISIÓN BIBLIOGRÁFICA Y SU ANÁLISIS

Una vez obtenidos los resultados, realizamos una primera revisión preliminar de los resúmenes donde percibimos que muchos artículos trataban sobre Twitter o big data, pero no sobre migraciones, y viceversa. Para evitar los falsos positivos e identificar artículos que pudiesen ser susceptibles de un posterior análisis en profundidad, todos los títulos y resúmenes obtenidos se codificaron con los términos de búsqueda y sus derivados, con la ayuda del software Atlas ti 8.0. La codificación nos permitió identificar y confirmar más rápidamente aquellos resúmenes que abordaran las dos temáticas a la vez y por tanto correspondieran a nuestros criterios.

Esta tarea previa de depuración es necesaria porque la búsqueda en las bases de datos introduce un número importante de falsos positivos en la medida en que dichas búsquedas se realizaron con el criterio más amplio al buscar en “todos los campos” (no, por ejemplo, en título o resumen), lo que propició la aparición de falsos positivos que hay que depurar posteriormente. Por ejemplo, al buscar en “todos los campos” se incluyen artículos que nombran a Twitter en el texto completo del artículo, sin que necesariamente se trate de una investigación basada

en datos de Twitter. En otras ocasiones aparecen textos técnicos que se refieren por ejemplo a migraciones de sistemas o datos (data migration) o referidos a la nube, internet, pero no a nuestro objeto de estudio. A pesar de esto, al optar por realizar una búsqueda amplia donde las palabras claves podían aparecer en cualquier campo de los recogidos en ProQuest, Scopus y WoS (autor, resumen, texto de documento, título de la publicación, materia, etc.) se pudo identificar un mayor número de artículos que compusieron la muestra final, que a través de las búsquedas más restrictivas que hicimos inicialmente (no se recogen aquí).

Posteriormente, se procedió a la lectura de los artículos a analizar para vaciar sus contenidos respecto a los métodos, técnicas y herramientas empleados (véase el Anexo I). Los criterios de inclusión que se siguieron fueron: 1) que se tratara de una investigación que hubiese utilizado Twitter o big data como fuente de datos o de trabajo (sin excluir aquellas que hubiesen usado otras fuentes de manera complementaria); 2) que su objeto de estudio fueran las migraciones, las minorías o grupos étnicos y la diversidad ligada a las migraciones. Se excluyeron del análisis final los artículos duplicados que fueron identificados al trabajar con varias bases de datos. Los textos completos de los artículos estudiados que encajaron con los criterios de selección fueron luego revisados para comprobar que efectivamente cumplían los requisitos fijados.

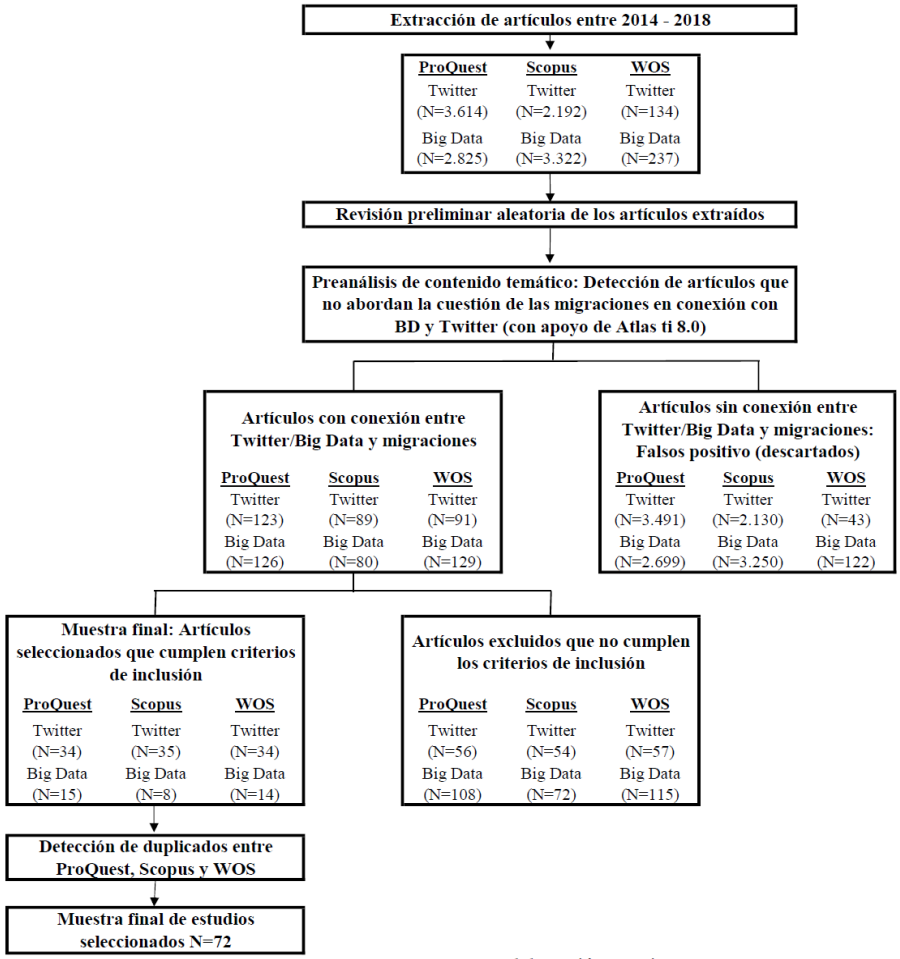
El proceso de selección de artículos y sus resultados se presenta sintéticamente en la Figura 3 a través de un diagrama de flujos.

5. ANÁLISIS DE LA BIBLIOGRAFÍA

Para vaciar y clasificar la información metodológica de interés de cada artículo se elaboró un guión o cuestionario para sistematizar esta tarea (Anexo I). Aparte de aspectos temáticos específicos, que son secundarios en este artículo, el vaciado pretende identificar fácilmente aquellas cuestiones que tienen que ver con las fases que siguen las investigaciones relativas a big data y Twitter: extracción de datos, procesamiento y análisis, amén de otros factores relacionados.

Se identifican por este método aspectos como los datos básicos del estudio (autor/es y año de publicación del artículo), el enfoque metodológico de la investigación (cuantitativo, cualitativo, mixto), técnicas de producción de datos (extracción, fuentes secundarias, etc.), herramientas de extracción de los datos (Python, R, Nvivo, NodeXL, etc.), objeto de extracción (hashtags, palabras claves, tuits de un usuario, historial de llamadas, etc.), procesamiento (codificación, filtrado, uso de diccionarios, etc.), técnicas de análisis empleadas (análisis del discurso, análisis de redes, regresiones, análisis de clúster, etc.), forma de visualización de los datos (Gephi, NodeXL, Python, etc.) y otras fuentes utilizadas (Facebook, Tumblr, e-mails, etc.). Todo ello es objeto de análisis en las páginas que siguen.

Figura 3. Proceso de selección de artículos en la revisión bibliográfica



Fuente: Elaboración propia.

Fuente: Elaboración propia.

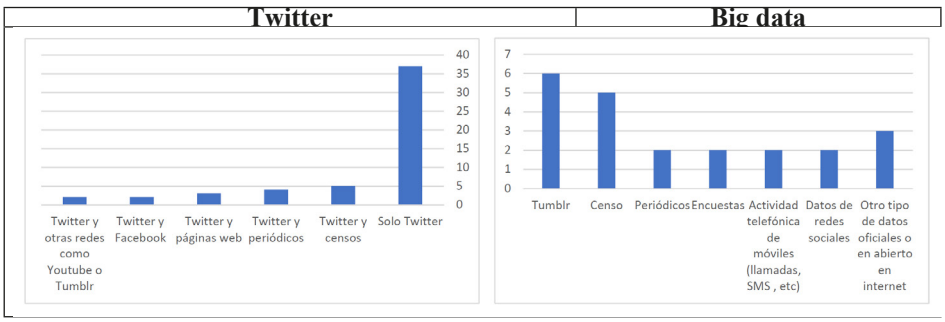
6. RESULTADOS

Los resultados de la revisión bibliográfica se han articulado de tal forma que permitan identificar las pautas en cuanto a métodos, técnicas y herramientas que están siguiendo los artículos sobre estudios migratorios que utilizan big data o Twitter en diferentes momentos de sus procesos de investigación.

6.1. Las fuentes de información

En el caso de los artículos relacionados con big data y Twitter, cabe decir que en las investigaciones que hemos podido recuperar no es infrecuente que diversas fuentes de información secundarias hayan nutrido estos trabajos, en ocasiones usadas de forma simultánea en las investigaciones. En el caso de las investigaciones que giran en torno a los big data (Figura 4), es frecuente el uso de datos del Censo o registros oficiales, que se articulan cuando trabajan con migraciones en torno a una orientación demográfica, asuntos relacionados con la movilidad humana, o con aspectos étnicos, raciales y sanitarios. Puede citarse como ejemplos de esta articulación los trabajos de Ford et al. (2018) o de Pennap et al. (2017).

Figura 4. Principales fuentes usadas en trabajos sobre estudios migratorios y …



Fuente: Elaboración propia.

En el caso del trabajo de Ford et al. (2018), desde un enfoque de la teoría crítica de raza (TCR) y en el contexto de estudios sanitarios, a través de un estudio retrospectivo de cohortes, se incorpora en la estrategia de análisis de big data la aplicación, entre otras cosas, de un lexicón antirracista, basado en este marco teórico (TCR) y que permite conectar varios conjuntos de datos (registros médicos de pacientes y proveedores de cuidados de salud, sobre todo) para estudiar los determinantes contextuales de las desigualdades raciales y étnicas en materia de la atención al VIH en California. Valoran, dada la novedad de incorporar en

este contexto analítico los presupuestos de la TCR, que hay un cambio paradigmático en marcha (Ford et al., 2018:265).

En el trabajo de Pennap et al. (2017), con un diseño transversal, se presta atención a datos relativos a nivel estatal de servicios de cuidado médicos (Medicaid) y datos censales de menores en California con el interés de evaluar el efecto del aislamiento residencial hispano sobre el Trastorno por Déficit de Atención e Hiperactividad. Entre los datos hay medidas del aislamiento residencial de carácter racial o étnico, enfocadas a los hispanos en este caso. Se subraya igualmente la oportunidad que aportan los big data para avanzar en la investigación que permita reducir las disparidades en materia de salud por motivos étnicos raciales en tanto que determinantes de salud.

Por otra parte, uno de los estudios de revisión identificados respecto al uso de big data (Chandy, Hassan & Mukherji, 2017) se pregunta por la migración que se produce tras los desastres naturales, poniendo en valor el enfoque de big data e ilustrándolo a través de varios ejemplos. Uno de los que está más relacionado con los estudios migratorios es el caso del terremoto de Haití en 2010 en Puerto Príncipe, donde más de medio millón de habitantes se vieron abocados a desplazarse para buscar refugio. Subrayan que, frente a métodos manuales, en este caso se pudo monitorizar con bastante precisión la localización de estos migrantes (datos hechos accesibles por Digicel, un operador de telefonía móvil, que podía usar la tarjeta SIM: registros de llamadas). La existencia de grandes volúmenes de información digital se plantea como algo que puede ayudar claramente a la toma de decisiones para resolver grandes problemas de diversa índole. A partir de una revisión de varios estudios de caso, no solo ligados a las migraciones, se pone de relieve por tanto la innovación y oportunidades que plantea la incorporación de estudios basados en big data (Chandy, Hassan & Mukherji, 2017:710). Hemos identificado otros estudios, de corte demográfico, que se nutren de datos extraídos de teléfonos móviles con el seguimiento de llamadas y SMS o el roaming, que muestran sugerentes líneas de análisis para el estudio de las comunidades de inmigrantes (Yuan & Zhu, 2016; Bajardi et al., 2015). Junto a los anteriores enfoques, no faltan trabajos que usan como fuentes de información datos procedentes de los medios sociales (Tumblr, u otras redes sociales), periódicos, encuestas, u otro tipo de datos oficiales o en abierto de internet.

En los trabajos sobre Twitter, si bien la mayor parte de los artículos se enfocaban solo en datos obtenidos de esta red (37 de ellos, de un conjunto de 52 – Figura 4-), hemos encontrado también que un número sustancial de los mismos se basaban en el manejo de otras fuentes que complementaban la información de Twitter (al menos en un tercio de las publicaciones identificadas). Normalmente en este caso la tónica más habitual ha sido usar datos de Twitter y otras redes sociales (como Facebook, Tumblr, Youtube, u otras) o censales. Pero también hemos encontrado investigaciones basadas en Twitter y a la vez en datos de periódicos, páginas web, e-mails, datos públicos del gobierno, encuestas, o incluso la combinación de datos oficiales con datos de tarjetas inteligentes (en un trabajo sobre movilidad de Kim, Park & Lee, 2018). Esta diversidad de artículos, que articulan fuentes de datos tan variadas, de nuevo marca la relevancia que en este

tipo de investigaciones comportan las acciones de triangulación, así como la versatilidad que permite esta línea de trabajo.

Uno de los trabajos que hemos encontrado que se encuentra basado en mayor número de fuentes de información que se suman al propio Twitter es el de George Mwangi, Bettencourt & Malaney (2018), de gran interés, como un buen ejemplo de las posibilidades analíticas y de articulación de fuentes en estudios donde hay un importante acento en medios sociales. En este caso, con el objeto de estudiar el movimiento de estudiantes universitarios activistas on line ("I, Too, Am Harvard" y "I, Too, Am Oxford") sobre el racismo, la opresión y las micro agresiones que se experimentan en la universidad, se realiza un análisis crítico del discurso a partir de datos procedentes de fuentes como Tumblr, Twitter, Facebook, artículos de periódicos o revistas y páginas web.

En la misma línea de mostrar un variado conjunto de fuentes de información, la investigación de Oliver (2016) es un buen ejemplo. El manejo de diversidad de fuentes se logra en este caso empleando una herramienta llamada Tracktur para la recolección de noticias y artículos publicados en los medios sociales, como fuente de opinión [http://www.trackur.com, para monitorear APIs]. De esta manera, se pueden recopilar datos de opinión de Twitter, Facebook, Tumblr, periódicos, webs, foros, Google+, Youtube, Instagram, Blogs o Reddit. Se extraen discursos sobre inmigración que permiten describir las narrativas de amenaza que se vuelcan hacia los inmigrantes y que se producen en los medios. Esta herramienta, o similares, son cada vez más comunes en el proceso de monitorización de los contenidos que se vuelcan en internet.

En una investigación que abre líneas de trabajo sugerentes para el campo de estudios migratorios, no tan completamente enfocado a los medios sociales, hemos encontrado como fuentes de datos Twitter al mismo tiempo que datos procedentes de tarjetas inteligentes y datos públicos gubernamentales. Se trata del trabajo de Kim, Park & Lee (2018), que parcialmente aborda la cuestión migratoria, centrándose principalmente en la movilidad humana urbana y la interacción. Hay una conexión, aparte de por las pautas de movilidad de los migrantes, cuando se discute sobre las distancias funcionales y sociales, o las relaciones entre origen y destino. Pero, en cualquier caso, aunque el acento sobre lo migratorio no es tan visible en este trabajo, comparado con otros, el enfoque empleado permite apreciar su potencialidad. Se articulan aquí fuentes diversas como datos de las tarjetas de transporte de la ciudad de Seul (T-Money), registros estadísticos del gobierno (diferentes datasets en línea) y datos de Twitter a los que se puede incorporar una geolocalización.

En el artículo de Hoffman (2016), es interesante igualmente constatar un ejemplo de investigación donde se combinan formas modernas y clásicas de análisis documental, basadas en internet, las modernas, y en medios clásicos como la prensa, que en diversas investigaciones hemos encontrado, en forma de periódicos en línea. En este caso con la pretensión de comprender las estrategias argumentativas usadas para expresar apoyo o rechazo a expresiones multiétnicas y multilingüísticas en la esfera pública de patriotismo en la sociedad americana. En la bibliografía analizada se han identificado diferentes artículos donde se usan

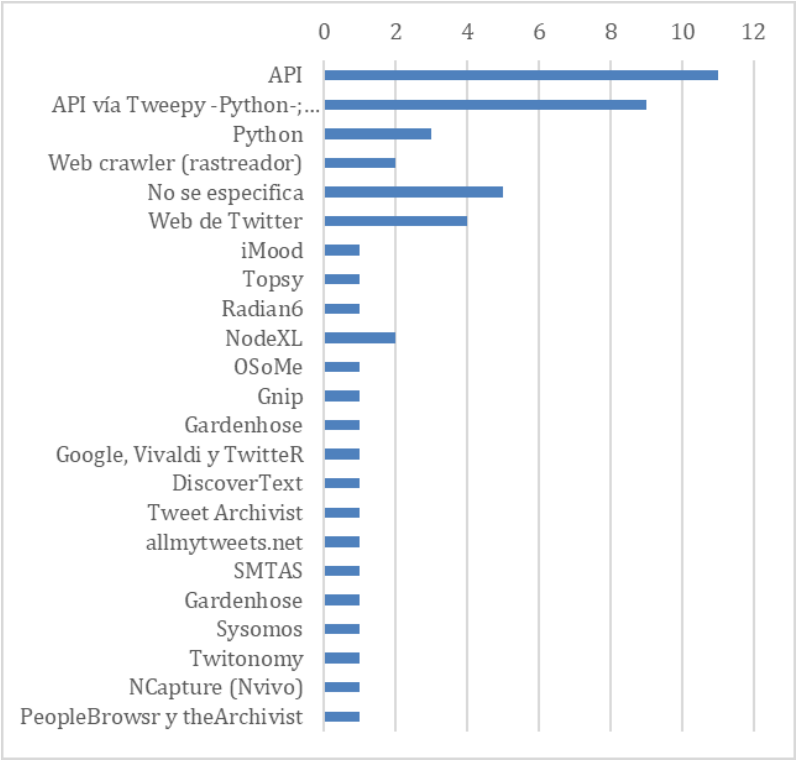
como fuente periódicos en línea junto a otras fuentes, se trate de estudios basados en Twitter o en big data (por ejemplo, Calvo & Calvo-Domínguez, 2016).

6.2. La extracción o recolección de datos, herramientas y objetos

Hay una diferencia sustancial entre el trabajo con big data y Twitter en este particular. En los trabajos recopilados de big data, lo habitual es trabajar con fuentes secundarias, conectando frecuentemente con servidores de instituciones públicas y privadas para usar la información que proveen (por ejemplo, censos), o usando la información disponible a gran escala, geolocalizada en muchas ocasiones. Otras veces la información es proporcionada por algún tipo de proveedor, como es el caso de la telefónica (Bajardi et al., 2015; Yuan & Zhu, 2016; Kim, Park & Lee, 2018). No necesariamente la información se ha recolectado en streaming. En cambio, comparado con la investigación clásica en estudios de Ciencias Sociales, el trabajo que se viene realizando actualmente en el campo de Twitter comprende necesariamente estrategias para la extracción o minería de datos. Una vez recolectados los datos, estos van a ser susceptibles de procesamiento y análisis. Ha de distinguirse entre estrategias o técnicas más basadas en la investigación clásica (más próximas a lo artesanal o manual), frente a las que se basan en técnicas más modernas, complejas y con fuerte énfasis en la computación. En ocasiones el software que se está desarrollando actualmente facilita bastante la dimensión más computacional para científicos profanos en la materia.

En el caso de una mayor proximidad a la computación, hay un abanico de posibilidades en marcha que se han documentado en diversos artículos que se basan en la extracción de datos de Twitter (Figura 5). Curiosamente hay diversos artículos de los seleccionados que detallan poco cómo se ha abordado la extracción de datos, limitándose a hablar de la API de Twitter (o de si conectan con ella vía Streaming o Rest) y casi poco más, y en otros no se especifican detalles al respecto.

Figura 5. Extracción de datos de Twitter en estudios migratorios



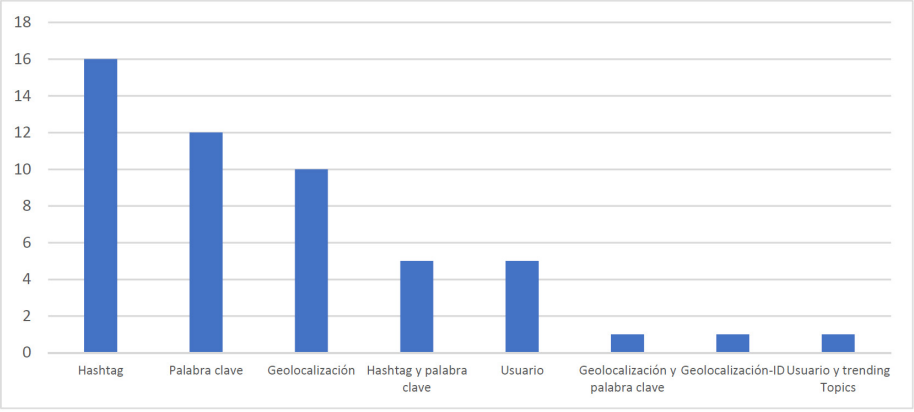
Fuente: Elaboración propia.

En otros casos se han empleado diferentes tipos de estrategias que pasan por comprar los datos retrospectivos a alguna empresa proveedora de los mismos (como Gnip en el caso del innovador trabajo de Flores, 2017, al que más adelante nos referiremos, o Tweet Archivist); por trabajar con software comercial o libre que facilita bastante la parte técnica de la extracción (algunos como NodeXL –Gualda y Rebollo, 2016-, Nvivo [NCapture], Gephi, u otros). Otras opciones existentes, aunque no hayan sido mencionadas en estos artículos, son el acceso a otras aplicaciones como t-hoarder a partir del repositorio de GitHub (Congosto, Basanta-Val y Sánchez-Fernández, 2017). También hemos encontrado trabajos que se han realizado con apoyo de programación propia o compartida de Python [o herramientas como Twython o Tweepy] o que se benefician de herramientas a disposición del público en R [StreamR, TwitterR]. En algún caso también se ha producido la recolección de datos a partir de la web de Twitter a través de su buscador u otros sitios web que permiten obtener tuits de un determinado usuario (como www.allmytweets.net). Vemos entonces que la no muy abultada bibliografía existente sobre Twitter y migraciones recoge, no obstante, una amplia

variedad de métodos de extracción de datos de esta fuente de internet, sin quedar agotadas las posibilidades en las citadas. Algunos métodos son más sencillos, accesibles para aquellos investigadores con menos conocimientos informáticos, mientras que otros han sido creados por los propios investigadores habitualmente cuando estos pertenecen a una disciplina más científico-técnica.

Por otra parte, respecto al objeto de la extracción, este varía según los objetivos de cada artículo. En el caso de los estudios basados en Twitter, los objetos principales extraídos suelen ser las etiquetas o hashtags, las palabras clave, la geolocalización y los usuarios, o algunos de los anteriores combinados (Figura 6).

Figura 6. Objetos principales extraídos en los trabajos sobre estudios migratorios y Twitter



Fuente: Elaboración propia.

Respecto a los trabajos centrados en big data, hay una mayor diversidad de aspectos que se extraen, destacando especialmente en la serie de artículos analizados historiales médicos electrónicos, como aspecto único, o en combinación con otros datos como datos de población o estadísticas de salud complementarias (Tu et al, 2015; Pennap et al, 2017; Retamozo et al, 2018). Otros objetos que han sido susceptibles de extracción han sido datos de roaming, service request, artículos, historial de llamadas, árboles genealógicos, listas de apellidos, búsquedas, reclamaciones administrativas de Medicaid, datos de población , etc., dando cuenta de la diversidad de posibilidades.

6.3. Procesamiento de datos

En materia de procesamientos, cada artículo, de forma pragmática y flexible, suele usar sus propias estrategias a fin de delimitar y preparar los datos de una

manera más amigable para la redacción, elaboración de informes, libros, artículos, etc. que son el resultado final del análisis. Normalmente estas estrategias conectan directamente con los objetivos concretos de cada trabajo y varían en función de si nos referimos a estudios basados en big data o Twitter. En el caso de los artículos identificados sobre migraciones y big data, uno de los aspectos quizás más recurrentes para reducir la dimensión de algunos *datasets* (según hemos comprobado) es el filtraje, de tal forma que, tras el mismo, los datos puedan ser manejados más rápida y fácilmente por algunos paquetes de software o a través de programación. También es frecuente adoptar estrategias de filtrado que pueden facilitar tanto análisis estadísticos como en algunos casos la visualización de resultados (por ejemplo, si en el análisis se emprende un análisis de redes sociales con herramientas como Gephi que si se encuentra con millones de nodos puede tener dificultades). Algunos de estos procesamientos pueden hacerse a través de programación o también con herramientas o software creado expresamente para manejar ingentes volúmenes de información de manera rápida, como sería el editor de texto EmEditor. En muchas ocasiones se combinan varias estrategias de procesamiento si bien la del filtrado es la que hemos encontrado con mayor frecuencia en los artículos en los que nos hemos basado (aparece como estrategia en la mitad de los artículos consultados). Otras estrategias que hemos encontrado son algunas típicas del trabajo con datos masivos como procesos combinados de categorización y codificación de los datos; filtrado, tokenización y *data cleaning*; filtrado y clasificación basada en machine-learning; filtrado y elaboración de diccionarios para ayudar a clasificar los datos; árbol de decisiones o uso de algoritmos, entre otros. En algunos trabajos no se especifican estrategias de procesamiento.

De forma más sofisticada por el uso de la computación, puede citarse el trabajo de Marschke, Nunez, Weinberg & Yu (2018) como ejemplo en el que se emplea un clasificador basado en *machine learning*, llamado *Ethnicolr*, que clasifica a las personas a través de su nombre y apellido en cuatro categorías étnicas (hispano, blanco no hispano, negro no hispano y asiático no hispano). Con ello los investigadores se interesan en conocer la relación entre pertenecer a una etnia y la posición de los autores de una publicación científica, con el objeto de comprender mejor la infrarrepresentación de algunos grupos minoritarios en las áreas científico-técnicas.

En el caso de Twitter, es frecuente también encontrar artículos donde se emplean varios tipos de procesamientos a la vez, incluso a partir del trabajo que hace algún software comercial para el análisis de textos asistido por ordenador, como es el caso que se describe en el trabajo de Smith, McGarty & Thomas (2018) basado en el uso de LIWC[6], donde se intenta identificar qué palabras en los tweets son de contenido pro-refugiados, creándose un diccionario adaptado para llevar a cabo este enfoque analítico.

[6] LIWC, Linguistic Inquiry and Word Count. En http://liwc.wpengine.com/; http://liwc.net/liwcespanol/index.php.

La variedad de estrategias a la hora de los procesamientos realizados al analizar datos de Twitter se manifiesta en otros trabajos, como es el caso del de Kim, Park & Lee (2018:21) donde el análisis de los textos de Twitter se aborda a través de un analizador de morfemas que los separa y posteriormente se encuentran las relaciones entre ellos con el apoyo de un módulo llamado *word2vec*[7], para posteriormente construir grupos o clústeres de palabras usando un algoritmo de agrupamiento, lo que implica una sucesión de tareas técnicas encadenadas propia de muchos trabajos basados en Twitter, no solo en el área de los estudios migratorios.

Cabe destacar, no obstante, que frente a las estrategias de filtrado del trabajo con *big data*, en el caso de Twitter las que sobresalen en los artículos consultados son las tareas que se realizan de codificación y categorización de los datos (23 de 52 artículos lo especifican). La codificación de datos en Twitter puede hacerse de forma más artesanal o manual, de manera semiautomatizada o de manera totalmente automatizada. Algunos trabajos no especifican el procedimiento usado, si bien un conjunto de ellos indica el uso de un procedimiento manual. En diversos estudios se emplea el trabajo manual (o semimanual) con el fin de crear diccionarios de palabras que permitan codificar o categorizar los resultados, como fase previa a otros análisis (Flores, 2014; Munger, 2017).

De esta forma, a tareas clásicas del trabajo cualitativo (codificar o categorizar) encontramos que en los trabajos basados en estudios migratorios y Twitter se añaden otras operaciones clave de carácter computacional que implican –a veces de forma combinada- operaciones como filtrar los datos, para reducir información; limpieza o *data cleaning*; empleo de diccionarios; empleo de paquetes de programación u otras.

Hay un conjunto amplio de herramientas adicionales que se han identificado en los artículos analizados para la realización de los procesamientos que preceden al análisis y que conllevan, por ejemplo, el uso de Python (que incorpora herramientas como la del procesamiento del lenguaje natural, NLTK), o R, funciones específicas de NodeXL para la preparación de los análisis de redes sociales, LIWC (citado arriba), CLD2[8] como detector de lenguaje compacto en diferentes lenguas; iMood[9] que permite extraer los sentimientos y emociones en tiempo real a partir de tuits sobre inmigración y seguridad en la frontera (Chung y Zeng, 2016); Tracktur[10] que se emplea para monitorizar APIs para la recolección de noticias y artículos publicados en los medios sociales, y el uso de aplicaciones como Elasticsearch[11] que permite entre otras funciones el análisis básico de texto que incorpora tokenización y filtrado; Chorus TweetVis[12] como paquete para la analítica visual que permite aproximarse a datos de medios sociales.

---

[7] https://medium.com/@gruizdevilla/introducci%C3%B3n-a-word2vec-skip-gram-model-4800f72c871f.

[8] https://github.com/CLD2Owners/cld2.

[9] http://www.imood.com/.

[10] http://www.trackur.com, para monitorear APIs.

[11] https://www.elastic.co/.

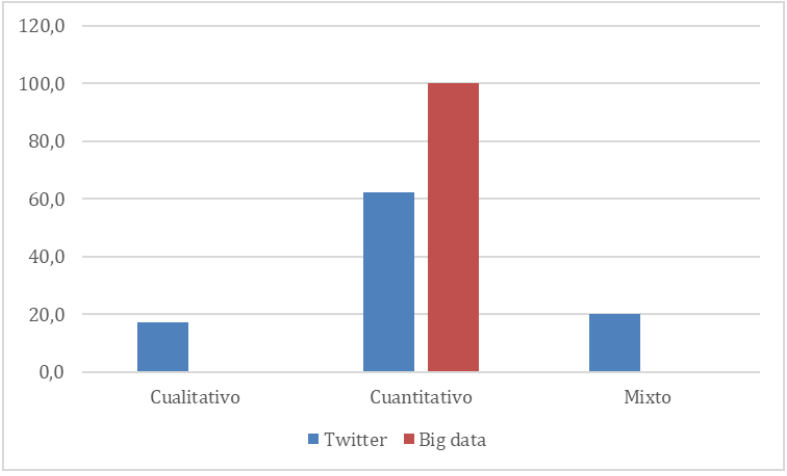
[12] http://chorusanalytics.co.uk/tweetvis-1-8-5/.

Lo que hemos encontrado en la revisión de la literatura, en definitiva, es un conjunto diverso de tipos de procesamientos, una variedad de herramientas implementadas de forma pragmática a veces porque son más accesibles, con el fin de responder a las preguntas de investigación. Los avances existentes hasta el momento condicionan el uso de las mismas. Este desarrollo y la diversidad de aplicaciones hace prever que el futuro en este campo seguirá aportando multiplicidad de herramientas y estrategias para abordar esta fase intermedia de la investigación en migraciones cuando se emplean datos de Twitter y, si cabe, de otros medios sociales.

6.4. Análisis de datos: Métodos, técnicas y visualización

En el análisis de datos, teniendo en cuenta los métodos y técnicas que se emplean al efecto, se encuentran diferencias notables entre big data y Twitter. En el caso de los estudios basados en big data, la pauta es cuantitativa en todos los artículos consultados. Esto conecta claramente con una orientación de los artículos sobre big data hacia los métodos cuantitativos, mientras que en los de Twitter, aunque encontramos un acento en los métodos cuantitativos, un grupo sustancial de artículos emplean presupuestos propios de los métodos cualitativos o son de carácter mixto (Figura 7).

Figura 7. Métodos empleados en los artículos sobre estudios migratorios y Twitter (%)



Fuente: Elaboración propia (N=72).

En cuanto a los estudios basados en big data, los métodos de análisis cuantitativo empleados conectan directamente con lo que se está llevando a cabo en el

área de data science o data analytics que cuenta con un importante énfasis en técnicas cuantitativas avanzadas para analizar datos y tomar decisiones. Es frecuente el empleo a lo largo de un artículo de más de una técnica de análisis de datos dando muestras de su diversidad (véase en los trabajos, por ejemplo, de Brito-Zerón et al. (2017), Somashekhar (2016) y Tu et al. (2015). En cuanto al uso de técnicas, los artículos sobre big data y migraciones abordan sus investigaciones empleando desde técnicas estadísticas más elementales (análisis de frecuencias, descriptivos, test estadísticos, correlaciones) hasta aquellas más representativas de análisis bivariables, multivariables o más complejos (regresión: lineal, logística, análisis de cluster, o análisis de cohortes, para análisis temporales; ecuaciones estructurales). Se identificaron también trabajos que combinan estas técnicas con el uso de algoritmos a priori, estudios de caso, modelos de caso, métodos de control sintético; o que emplean análisis de redes, análisis temático o distribuciones espaciales.

Por otra parte, aunque no todos los trabajos que se han revisado sobre big data usan la visualización como manera interesante para resumir y comunicar los resultados, o incluso como vía analítica, una parte importante de ellos emplean fotos, imágenes o gráficos elaborados por los autores, y en algún caso se usan R o Python a estos efectos.

En el caso de Twitter encontramos una mayor diversidad en el uso de técnicas de análisis debido a que junto a las cuantitativas, se incorporan las propias de los métodos cualitativos o mixtos. De esta forma, entre las técnicas de análisis univariable, bivariables y multivariables, en el caso del uso de métodos cuantitativos, cabe citar algunas como el análisis de frecuencias (en un 15% de casos y muy común por su sencillez), el ANOVA o análisis de la varianza, el análisis de conglomerados o clústeres, el análisis factorial o la regresión logística y multivariable, siendo la logística la más frecuente en esta revisión (casi un 10% de artículos). Otros con cierta incidencia son el análisis de sentimientos (en 4 artículos), el estudio de distribuciones espaciales, o estudios de casos.

Por el ala cualitativa, caben destacar el análisis de contenido (que aparece en 17 artículos de 52, de forma combinada en ocasiones), el del discurso (en 18) y el análisis temático y textual, ya con menos frecuencia. Otros trabajos emplean también modelos predictivos o teóricos, contrastándolos, como por ejemplo en el artículo de Chung (2016).

Por otro lado, sobresale también el análisis de redes sociales con 11 trabajos y una orientación más cuantitativa y visual. Por ejemplo, el trabajo de Ferra y Nguyen (2017), que hace un análisis de redes sociales a través del hashtag #migrantcrisis, para conocer cómo el discurso transnacional sobre la crisis migratoria se materializa en Twitter, enfocándose en los participantes y en las redes semánticas que emergen alrededor del hashtag.

Se ha encontrado también, como único ejemplo en estos trabajos sobre migraciones, el uso del aprendizaje automático, bastante común en otras investigaciones de Twitter. En este caso se empleó para identificar la co-ocurrencia de términos en documentos de un corpus de análisis y detectar patrones entre los temas. Xu et al. (2016) hacen uso de esta técnica para examinar con qué fre-

cuencia se discute sobre temas relacionados con el cáncer, en función de la etnia de los usuarios de Twitter. También es de gran interés, por tratarse de una línea prometedora aún casi inexplorada que conecta aspectos *online* y *offline*, el trabajo de Flores (2017) que se pregunta si las leyes influyen en la opinión pública, sus sentimientos y comportamientos, para lo cual emplea un análisis longitudinal de Twitter como medición de estados de opinión, antes y después de la ley anti-inmigrantes de Arizona SB 1070.

Una pauta similar a la visualización de datos encontrada en relación a big data la hallamos en lo que concierne a Twitter respecto a que la mayoría de artículos emplean tablas o gráficos, fotos o imágenes elaborados por los autores, o mapas, pero también hay trabajos más específicos de gráficos de redes, con paquetes como Gephi o R, SMTS, dando muestras igualmente de la variedad de opciones de visualización de los resultados en este campo, incluyéndose entre otras capturas de pantalla y nubes de palabras.

7. DISCUSIÓN Y CONCLUSIONES

Diversos autores que han realizado investigaciones en el campo de confluencia de big data y migraciones valoran muy positivamente las aportaciones y oportunidades que esta área de estudio puede introducir, anunciando contribuciones importantes en diferentes esferas, no solo en la investigación y la innovación, sino también apoyando la toma de decisiones (Chandy, Hassan & Mukherji, 2017; Ford et al, 2018; Pennap et al., 2018). Entre los aspectos que resultan de gran interés, tanto para el caso de big data como de Twitter, se encuentra el potencial de combinar en una misma investigación, como describen diversos autores, varios datasets de ingente volumen, lo que abre caminos presentes y futuros de gran interés en el campo de los estudios migratorios, enriqueciéndolos. También nos remite a la necesidad de fortalecer las estrategias de triangulación y combinación metodológica, ante la variedad de métodos, técnicas, herramientas y fuentes de análisis que potencialmente pueden ser empleadas.

Los resultados que se ofrecen de la comparación de tres bases de datos relevantes (ProQuest, Scopus y WoS) son coherentes entre sí, fiables y sólidos, por cuanto proporcionan un panorama similar –independientemente de la fuente- respecto al escaso peso que aún tienen los estudios migratorios basados en Twitter o en big data, tras el estudio de los artículos publicados en un lustro. Entre las limitaciones, el que la revisión, con el fin de homogeneizar los criterios de búsqueda, sólo ha recolectado artículos científicos, dejando fuera del análisis a otros productos como tesis doctorales o libros. Frente a esta desventaja, es positivo no obstante haber documentado lo que se ha publicado sobre estudios migratorios y big data o Twitter en el formato más relevante de investigación en nuestro campo hoy en día (artículo científico).

El trabajo realizado, distinguiendo entre dos campos emergentes como es el trabajo con big data y Twitter, como ejemplo de medio social, permite visibilizar cómo se están conformando estos campos de trabajo, de manera muy similar

los estudios migratorios a otras áreas temáticas, en cuanto a las cuestiones relacionadas con los métodos y técnicas empleadas, que son aplicables a diferentes temáticas. Una limitación de fondo, siempre presente, son las propias fuentes de información secundarias que se emplean, públicas y privadas, sean estas censos, registros, datos de medios sociales, etc., por cuanto los investigadores no tienen el control sobre los datos que se generan en ellas o el acceso a ciertos datos puede ser complicado (informes médicos, datos de teléfonos móviles, etc.). En el caso de datos procedentes de fuentes privadas ha de añadirse el coste que suele suponer tener que comprar datos, además de la menor transparencia respecto a cómo se producen dichos datos (por ejemplo, en diferentes plataformas de redes sociales), o qué algoritmos hay detrás, siendo los datos públicos más transparentes en estos aspectos. Aparte de otras limitaciones técnicas que tienen que ver con el manejo de datos masivos, que requieren una infraestructura para el almacenamiento, el trabajo en *streaming* o los procesamientos (por ejemplo), más potente que en investigaciones clásicas.

El artículo ha presentado, someramente, dos líneas de trabajo recientes (big data, Twitter), aún poco exploradas en la bibliografía científica, que plantean nuevos enfoques para los estudios migratorios ya cerca del primer cuarto del siglo XXI. Respecto a nuestros objetivos e hipótesis, hemos comprobado cómo los estudios científicos que combinan big data o Twitter con las migraciones son aún escasos, aunque crecientes, con una pauta de crecimiento sostenida en el último lustro. Encontramos también un interés cada vez mayor en estudiar la realidad social que se refleja en el mundo virtual, en conjunción con la realidad observada fuera de internet si atendemos por ejemplo al uso combinado de varias fuentes de información localizada en diversos artículos (por ejemplo, redes sociales y censos). Al mismo tiempo, esto coincide con un creciente desarrollo técnico y de conocimientos en cuanto a la extracción de los datos, su procesamiento y análisis gracias al despliegue de multitud de herramientas (aparecen más de 20 diferentes en nuestros datos) que facilitan no solo el acceso a esos datos, sino también el almacenamiento de estos o la forma de gestionarlos o prepararlos para el posterior análisis.

En el caso de los big data, cabe destacar el acento más cuantitativo y estadístico, así como el manejo de censos o registros públicos y privados, mientras que en el de los social media, a través del caso estudiado de Twitter, se destaca la mayor variedad de aproximaciones metodológicas que se emplean, desde aproximaciones cuantitativas a cualitativas o mixtas, que se concreta coherentemente en una diversidad de técnicas de análisis. En este segundo caso, el rango de situaciones es más amplio respecto a la existencia de análisis con mayor o menor grado de sofisticación estadística o computacional, mientras que en la materia de big data, los trabajos que hemos revisado suelen requerir un mayor conocimiento técnico y estadístico, lo cual aproxima o aleja este campo de análisis a diversos perfiles de investigadores, e invita al trabajo en equipo inter y transdisciplinares, o a la formación de expertos en campos emergentes como la sociología computacional o la ciencia social computacional que pasa por la aproximación a la programación o por el uso de paquetes de software específicos que se están

desarrollando y que a veces se hacen accesibles públicamente en repositorios como el de GitHub.

Quizás, para terminar, entre los elementos recurrentes de estas nuevas formas de estudiar la realidad social deba destacarse la potencialidad de observar con mucha versatilidad y flexibilidad, beneficiándose la investigación del uso potencial y la combinación de múltiples fuentes, lo que obliga a triangular con mayor frecuencia. Por no decir lo que supone la posibilidad de observar muchos fenómenos en streaming, mientras están pasando. Por ejemplo, cuando se trata de desastres naturales y migraciones forzadas, donde podría permitirse llevar a cabo actuaciones de emergencia más rápidas y eficaces.

Para algunos, trabajar con big data supone igualmente a veces ser capaz de estar más cerca de observar datos del universo que de muestras, lo que se interpreta como la llegada de cambios paradigmáticos. En otros casos, la sustitución de técnicas como por ejemplo las encuestas de movilidad, por la posibilidad de seguir a las personas a través sus móviles, puede representar igualmente un cambio muy significativo en la investigación de las migraciones, sus comunidades asentadas y sus desplazamientos. No obstante, siguen existiendo muchos hándicaps y limitaciones de carácter técnico ligados a diferentes etapas de la investigación o a nuestras capacidades para almacenar, procesar o analizar tal cantidad de datos, lo que representa parte de la agenda de investigación actual y futura. Algunas de estas limitaciones ya se han destacado a lo largo del artículo, como por ejemplo, las dificultades que siguen existiendo para trabajar con datos masivos no estructurados (tales a vídeos o fotos), que tanto abren muchas vías de exploración técnica, como siguen dando valor a la investigación cualitativa que hasta ahora se viene haciendo en el campo de los estudios migratorios. Probablemente para reducir algo de complejidad en la tarea investigadora el camino más eficaz sea el de la inter y transdisciplinariedad, la combinación metodológica y la triangulación, que hemos encontrado en gran parte de los artículos analizados.

8. AGRADECIMIENTOS

Este texto ha sido posible gracias al apoyo del Ministerio de Educación, Cultura y Deporte, que ha beneficiado a Carolina Rebollo con un contrato de formación predoctoral (FPU16/00416). Agradecemos también el apoyo del Grupo de Investigación "Estudios Sociales E Intervención Social" y del centro de investigación COIDESO, ambos de la Universidad de Huelva, y a Dña. Elena Griñón Reina, del departamento de Hemeroteca y Servicio de Apoyo a la Investigación de la Biblioteca de la Universidad de Huelva, por su asesoramiento en la búsqueda bibliográfica.

9. BIBLIOGRAFÍA[13]

BAJARDI, P., DELFINO, M., PANISSON, A. et al. (2015). "Unveiling patterns of international communities in a global city using mobile phone data", EPJ Data Sci. 4(3), doi:10.1140/epjds/s13688-015-0041-5

BEDNÁR, P. (2017): "12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)", 9-10 July 2017, doi:10.1109/SMAP.2017.8022656

BELLO-ORGAZ, G., JUNG, J.J. y CAMACHO, D. (2016): "Social big data: Recent achievements and new challenges", Information Fusion, 28, pp. 45-59. doi:10.1016/j.inffus.2015.08.005

BRITO-ZERÓN, P., ACAR-DENIZLI, N., ZEHER, M., RASMUSSEN, A., SEROR, R., THEANDER, E. et al. (2017): "Influence of geolocation and ethnicity on the phenotypic expression of primary sjögren's syndrome at diagnosis in 8310 patients: A cross-sectional study from the big data sjögren project consortium", Annals of the Rheumatic Diseases, 76(6), p. 1042. doi:http://dx.doi.org/10.1136/annrheumdis-2016-209952

CALVO, D., y CAMPOS-DOMÍNGUEZ, E. (2016). "Participation and topics of discussion of spaniards in the digital public sphere", Comunicación y Sociedad, 29(4), p. 219-234. doi:http://dx.doi.org/10.15581/003.29.4.219-234

CHANDY, R., HASSAN, M., y MUKHERJI, P. (2017). "Big data for good: Insights from emerging markets", The Journal of Product Innovation Management, 34(5), pp. 703-713. doi:http://dx.doi.org/10.1111/jpim.12406

CHUNG, W., y ZENG, D. (2016). "Social-media-based public policy informatics: Sentiment and network analyses of U.S. immigration and border security", Journal of the Association for Information Science and Technology, 67(7), p. 1588, disponible en https://search.proquest.com/docview/1797693495?accountid=14549

CONGOSTO, M.; BASANTA-VAL, P.; SANCHEZ-FERNANDEZ (2017). "T-Hoarder: A framework to process Twitter data streams". Journal of Network and Computer Applications, Vol. 83, 1, pp. 28-39.

FERRA, I., y NGUYEN, D. (2017): "#Migrantcrisis: "tagging" the european migration crisis on twitter", Journal of Communication Management, 21(4), pp. 411-426, disponible en https://search.proquest.com/docview/1952388515?accountid=14549

FLORES, R. D. (2014): The social consequences of subnational restrictionist immigration policies in the U.S (AAI3642082), disponible en Sociology Collection, https://search.proquest.com/docview/1718065683?accountid=14549

FLORES, R. D. (2017): "Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using twitter data", American Journal of Sociology, 123(2), pp. 333-384. doi:http://dx.doi.org/10.1086/692983.

FORD, C. L., TAKAHASHI, L. M., CHANDANABHUMMA, P. P., RUIZ, M. E., y CUNNINGHAM, W. E. (2018): "Anti-racism methods for big data research: Lessons learned from the HIV testing, linkage, & retention in care (HIV TLR) study", Ethnicity & Disease, 28, pp. 261-266. doi:http://dx.doi.org/10.18865/ed.28.S1.261

FUNDACIÓN TELEFÓNICA (2019): Sociedad Digital en España, 2018, Madrid, Taurus y Fundación Telefónica.

[13] Por razones de espacio, no se mencionan todos los artículos resultados de la revisión bibliográfica.

FUNDACIÓN TELEFÓNICA (2017): Sociedad Digital en España, 2017, Madrid, Ariel y Fundación Telefónica

GANDOMI, A. y HAIDER, M. (2015): "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management, 35, pp. 137–144. http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007

GEORGE MWANGI, C. A., BETTENCOURT, G. M., y MALANEY, V. K. (2018): "Collegians creating (counter)space online: A critical discourse analysis of the I, too, am social media movement", Journal of Diversity in Higher Education, 11(2), pp. 146-163. doi:http://dx.doi.org/10.1037/dhe0000054

GUALDA, E. (2018): "Medios de comunicación, medios sociales y análisis de redes sociales", en La Sociedad desde la Sociología. Una introducción a la Sociología General, Madrid, Ed. Tecnos, pp. 581-604.

GUALDA, E. y REBOLLO, C. (2016): "Refugee crisis in Twitter: Diversity of Discourses at an European Crossroads", Journal of Spatial and Organizational Dynamics. 4(3), pp. 199 - 212. Research Centre for Spatial and Organizational Dynamics, http://hdl.handle.net/10272/13624

GUALLAR, J., FERRAN-FERRER, N., ABADAL, E. y SERVER, A. (2017): "Revistas científicas españolas de información y documentación: análisis temático y metodológico", El profesional de la información, 26(5), pp. 947-960, doi:10.3145/epi.2017.sep.16

HOFFMAN, B. Y. (2016): "Online responses to a multilingual super bowl ad: Is "america the beautiful" by any other language still america, the beautiful?", International Journal of Multilingualism, 13(2), pp. 213-229. doi:http://dx.doi.org/10.1080/14790718.2015.1094075

INTERNET WORLD STATS (2018): *Internet sage* and *World* Population *Statistics, disponible en* www.internetworldstats.com/stats.htm.

KIM, J., PARK, J., y LEE, W. (2018): "Why do people move? enhancing human mobility prediction using local functions based on public records and SNS data", PLoS One, 13(2) doi:http://dx.doi.org/10.1371/journal.pone.0192698

MARSCHKE, G., NUNEZ, A., WEINBERG, B. A., y YU, H. (2018): "Last place? the intersection of ethnicity, gender, and race in biomedical authorship", AEA Papers and Proceedings, 108, pp. 222-227. doi:http://dx.doi.org/10.1257/pandp.20181111

MUNGER, K. (2017): "Tweetment effects on the tweeted: Experimentally reducing racist harassment", Political Behavior, 39(3), pp. 629-649. doi:http://dx.doi.org/10.1007/s11109-016-9373-5

OLIVER, J. R. (2016): Politics, capitalism and immigrant threat narrative in the media. Disponible en Sociology Collection, https://search.proquest.com/docview/1801660997?accountid=14549

OLSHANNIKOVA, E., OLSSON, T., HUHTAMÄKI, J., y KÄRKKÄINEN, H. (2017): "Conceptualizing Big Social Data", Journal of Big Data, 4(1). doi: 10.1186/s40537-017-0063-x.

PENNAP, D., BURCU, M., SAFER, D. J., y ZITO, J. M. (2017): "Hispanic residential isolation, ADHD diagnosis and stimulant treatment among medicaid-insured youth", Ethnicity & Disease, 27(2), pp. 85-94. doi:http://dx.doi.org/10.18865/ed.27.2.85

RETAMOZO S., ACAR-DENIZLI N., FAI NG W., et al (2018): "OP0120 Influence of epidemiology and ethnicity on systemic expression of primary sjÖgren syndrome in 9974 patients", Annals of the Rheumatic Diseases, pp. 77, p. 110, doi: http://dx.doi.org/10.1136/annrheumdis-2018-eular.6073

SMITH, L. G. E., MCGARTY, C., y THOMAS, E. F. (2018): "After aylan kurdi: How tweeting about death, threat, and harm predict increased expressions of solidarity with refugees over time", Psychological Science, 29(4), pp. 623-634. doi:http://dx.doi.org/10.1177/0956797617741107

SOMASHEKHAR, M. (2016): Immigrant business in suburban America: How and why ethnic economy workers in the suburbs are struggling to get by, disponible en Entrepreneurship Database; Sociology Collection, https://search.proquest.com/docview/1766108077?accountid=14549

TU, J. V., CHU, A., REZAI, M. R., GUO, H., MACLAGAN, L. C., AUSTIN, P. C., et al. (2015): "The incidence of major cardiovascular events in immigrants to Ontario, Canada: The CANHEART immigrant study", Circulation, 132(16), doi:http://dx.doi.org/10.1161/CIRCULATIONAHA.115.015345

XU, S., MARKSON, C., COSTELLO, K. L., XING, C. Y., DEMISSIE, K., y LLANOS, A. A. (2016): "Leveraging social media to promote public health knowledge: Example of cancer awareness via twitter", JMIR Public Health and Surveillance, 2(1), p. 1. doi:http://dx.doi.org/10.2196/publichealth.5205

YUAN, H. y ZHU, C. (2016): "Shock and roam: Migratory responses to natural disasters", Economics Letters, 148, pp. 37-40, doi: https://doi.org/10.1016/j.econlet.2016.09.020.

ANEXO I. LIBRO DE CÓDIGOS (REVISIÓN BIBLIOGRÁFICA).

ID. Identificación	**FUENTE DE LOS DATOS**	**PROCESAMIENTO**
Autor/es. Autor o autores del artículo	1. FamilySearch.com	1. Python's natural language toolkit
Año. Año de publicación	2. Historial médico electrónico	2. Codificar/categorizar
	3. Reclamaciones administrativas de Medicaid	3. Elasticsearch
MÉTODOS	4. Datos de teléfonos móviles	4. R
1. Cualitativo	5. Censo	5. LIWC
2. Cuantitativo	6. Datos públicos del gobierno	6. Chorus TweetVis
3. Mixto	7. Yelp.com	7. SMTAS
	8. Medline	8. Filtrar
HERRAMIENTA DE EXTRACCIÓN		9. CLD2
1. Tweet Archivist	**TÉCNICAS DE PRODUCCIÓN**	10. Analizador de morfemas
2. API	1. Encuesta	11. Automated Data Retrival
3. Tweepy (Python)	2. Entrevista	12. NodeXL
4. NCapture (Nvivo)	3. Monitoreo	13. Diccionario
5. Chorus Tweetcatcher	4. Extracción	14. iMood
6. Twython (Python)	5. Búsqueda manual	15. Trackur
7. SMTAS	6. Fuentes secundarias	16. Python
8. Python		17. Matriz de datos
9. Web de Twitter	**TÉCNICAS DE ANÁLISIS**	18. Razón de oportunidades (odds ratio)
10.NodeXL	1. Análisis de contenido	19. Clasificador basado en el machine-learning
11. allmytweets.net	2. Análisis de redes	20. Gohakka
12. StreamR (R)	3. Análisis del discurso	21. V-Analytics
13. TwitteR (R)	4. Regresión logística	22. Árbol de decisiones
14. iMood	5. Estudio de casos	23. Google maps API
15. Trackur	6. Análisis de sentimientos	24. Freeling
16. Gardenhose	7. Análisis exploratorio	25. Tokenización
17. OSoMe	8. Análisis temático	26. Limpieza (data cleaning)
18. Twarc	9. Análisis de frecuencia	
19. Gnip	10. Análisis de prevalencia	**VISUALIZACIÓN**
20. Tencent's Big Data Location	11. Análisis factorial	1. Gephi

21. Sysomos
22. Fuentes secundarias
23. Buscador web de personas (people-finder sites)
24. Google Adwords
25. Topsy
26. Web crawler (rastreador)
27. Netvizz
28. PeopleBrowsr
29. theArchivist
30. Radian6
31. IQBuzz
32. Google
33. Vivaldi
34. Discover Text
35. Twitonomy

OBJETO DE LA EXTRACCIÓN

1. Hashtag
2. Geolocalización
3. Usuario
4. Palabra clave
5. ID
6. Trending Topics
7. Petición de localización
8. Árbol genealógico
9. Historial médico electrónico
10. Reclamaciones administrativas de Medicaid
11. Historial de llamadas
12. Estadísticas de salud
13. Datos de población
14. Información sobre restaurantes

12. Distribución espacial
13. Modelo teórico
14. Análisis de clúster
15. Análisis de varianza
16. Test estadístico
17. Análisis textual
18. Estadísticos descriptivos
19. Algoritmo A priori
20. Análisis de cohortes
21. Regresión lineal
22. Regresión multivariable
23. Ecuaciones
24. Correlaciones
25. Método de control sintético
26. Regresión (otra)

2. R
3. Tablas y gráficos del autor
4. SMTAS
5. NodeXL
6. Capturas de pantalla
7. Nada
8. Fotos/Imágenes
9. Python
10. V-Analytics
11. T-Lab
12. Mapas
13. Gráfico de redes

OTRAS FUENTES

1. Facebook
2. Tumblr
3. Periódicos
4. Páginas web
5. Censo
6. Solo Twitter
7. Datos de tarjetas inteligentes
8. Datos públicos del gobierno
9. Google+
10. Youtube
11. Instagram
12. Foros
13. Blogs
14. Correos electrónicos
15. Flickr
16. Programas electorales
17. No
18. Encuestas

15. Artículos científicos
16. Lista de apellidos
17. Búsquedas
18. Datos roaming
19. Service request
20. Vídeos de Youtube
21. Páginas de Facebook