Big Data y Métodos Digitales, nuevas formas de investigación en Comunicación y Periodismo en la era digital. Dos casos de estudio

Chapter · June 2020

CITATIONS
0

READS
494

2 authors:

Ana Lucia Nunes de Sousa
Federal University of Rio de Janeiro
17 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Tania Lucía Cobos
Universidad Tecnológica de Bolívar
30 PUBLICATIONS **78** CITATIONS
SEE PROFILE

2

Big Data y Métodos Digitales, nuevas formas de investigación en Comunicación y Periodismo en la era digital. Dos casos de estudio

Ana Lúcia Nunes de Sousa *(Universidade Federal do Rio de Janeiro)*

Tania Lucía Cobos *(Universidad Tecnológica de Bolívar)*

1. Introducción

Los constantes avances de las tecnologías de información y comunicación presentes en Internet, particularmente los algoritmos desarrollados por las grandes compañías tecnológicas como Facebook, Twitter, Google, entre otras, han favorecido la generación de enormes volúmenes de datos, estructurados, semiestructurados y no estructurados, almacenados en bases de datos públicas y privadas, a las que genéricamente se les ha llamado *Big Data*. El aparecimiento de estas grandes bases de datos transformó a Internet en un amplio campo para la investigación científica y social. Las bases de datos generadas en el mundo virtual pueden ser exploradas y explotadas para analizar complejos fenómenos sociales y culturales abordados desde cualquier perspectiva, incluyendo la comunicación y el periodismo.

De acuerdo a Hadi et al (2015, p. 16) el término *Big Data* fue introducido al mundo de la computación por Roger Magoulas de la agencia O'Reilly

Media en el 2005, para referirse a una gran cantidad de datos que las técnicas tradicionales de gestión de datos no podían administrar y procesar debido a su complejidad y tamaño. De forma general, el *Big Data* está compuesto de numerosas piezas de información que pueden ser cruzadas, comparadas, agregadas y desagregadas a nivel de profundidad. Pese a no haber aún una definición rigurosa, Mayer-Schönberger y Cukier (2013, p. 17) apuntan a que el *Big Data* o los datos masivos "se refieren a cosas que se pueden hacer a gran escala, pero no a una escala inferior, para extraer nuevas percepciones o crear nuevas formas de valor, de tal forma que transforman los mercados, las organizaciones, las relaciones entre los ciudadanos y los gobiernos, etc".

Al *Big Data* se le han identificado cinco grandes características llamadas las 5V que son: volumen (*volumen*), variedad (*variety*), velocidad (*velocity*), veracidad o validez (*veracity or validity*) y valor (*value*). Volumen hace referencia a su enorme tamaño; variedad a la diversidad de tipos de datos y fuente de los datos; velocidad a la rapidez con la que estos se generan; veracidad o validez a la garantía de calidad de los datos o a su autenticidad y credibilidad; y valor a la utilidad o beneficio que obtienen de ellos sus propietarios al explotarlos (Hadi *et al*, 2015, p. 20; Marr, 2016).

Si bien es cierto, la creación de tales bases de datos masivas responde, en principio, a los intereses comerciales y de mercadeo por parte de las empresas multinacionales que desarrollan estas tecnologías, es innegable que su captura, almacenamiento, compartición, análisis y visualización en búsqueda de patrones repetitivos que permitan determinar correlaciones y construir modelos predictivos ha permeado a escala planetaria en casi prácticamente cualquier esfera de la vida del ser humano: estrategias de mercadeo, comercio electrónico, telecomunicaciones, gobierno electrónico, procesos electorales, salud pública y en otros campos, el científico y dentro de este, el que atañe a este trabajo, la comunicación y el periodismo. También hay que tener presente que el *Big Data* afronta grandes retos: ética en la captura de los datos, privacidad, actualización, sesgo, entre otros.

Así, en este trabajo nos proponemos mirar críticamente al *Big Data* como una metodología de investigación en las ciencias sociales y a presentar la propuesta de investigar en grandes bases de datos utilizando los *Digital Methods* o métodos digitales. Nos interesa reflexionar en qué medida el uso del *Big Data* puede generar más conocimiento o si la propuesta de los métodos digitales - que proponen utilizar las grandes bases de datos, pero en

menor escala - sería más apropiada al campo de la comunicación y periodismo. Para esto, partimos de una discusión teórica y aterrizamos en dos ejemplos de investigación donde dicha combinación se usó. La primera, en relación a Facebook, YouTube y TwitCasting, es decir, datos generados por seres humanos, y la segunda, en relación a Google News, datos generados a partir de medios noticiosos.

2. Marco conceptual

2.1. Una mirada crítica al uso del Big Data

Los defensores de los datos masivos argumentan que es necesario cambiar el paradigma científico utilizado hasta el momento, ya que la utilización del *Big Data* sólo tiene sentido si también se acepta la imprecisión de la metodología; la necesidad de confiar en correlaciones y lo más importante que "los datos masivos tratan del qué, no del porqué. No siempre necesitamos conocer la causa del fenómeno, preferentemente, podemos dejar que los datos hablen por sí mismos" (Mayer-Schönberger y Cukier, 2013, p. 26-27).

Con el *Big Data* es posible recolectar y transformar en datos casi todo lo que se pasa en el mundo actualmente. Los datos masivos representan un avance en lo que se refiere a los análisis macro, pero son una herramienta poco útil cuando lo que se pretende es analizar un fenómeno en sus singularidades. En este sentido, la necesidad de conocer un fenómeno detalladamente es considerada inútil por los defensores de los datos masivos, para ellos basta con conocer la tendencia general. Se cuestiona, inclusive, la necesidad de seguir haciendo muestreos y tener hipótesis de investigación, "ahora tenemos tantísimos datos a nuestra disposición, y tanta capacidad de procesamiento, que ya no tenemos que escoger laboriosamente una aproximación o un pequeño puñado de ellas y examinarlas una a una" (Mayer-Schönberger y Cukier, 2013, p. 75).

Si bien los datos masivos aportan una cantidad asombrosa de información y posibilidades a la ciencia y a la sociedad, no ha escapado de los críticos y escépticos en cuanto a su verdadero papel y potencial. Los defensores del *Big Data* trabajan con la creencia en la objetividad total de los datos, para ellos bastaría con “lanzar los números dentro de los mayores *clusters* de

computadoras que el mundo haya visto y dejar que los algoritmos estadísticos encuentren los patrones que la ciencia no pudo" (Anderson, 2008). Pero recolectar y transformar en datos - datificar - una cantidad tan grande de información puede resultar bastante complejo. El investigador necesita conocer profundamente los *softwares* que auxilian en este proceso. Luego, puede haber confusión en la combinación de diferentes tipos de información de fuentes distintas y errores de varios tipos, transformando el análisis en un procedimiento de alto riesgo (Mayer-Schönberger y Cukier, 2013; Mahrt y Scharkow, 2013; Rogers, 2013).

Algunos investigadores apuntan que el análisis del *Big Data* puede mostrar lo que hacen los usuarios, pero no por qué lo hacen (Mayer-Schönberger y Cukier, 2013). También suelen revelar información poco profunda y poca sensibilidad del contexto en el cual los datos fueron generados (Manovich, 2012; Mahrt y Scharkow, 2013; Boyd y Crawford, 2012). Otro problema, apuntado por Andersen (en Bollier, 2010, p. 12), es el riesgo de sacar conclusiones a partir de un único conjunto de datos, por lo que es más seguro usar *sets* de datos provenientes de múltiples fuentes, pero, aun así: "siempre que haces estadísticas vas a encontrar malas correlaciones y lazos de proximidad que, en verdad, no existen". Andersen (en Bollier, 2010, p. 13) también cuestiona la supuesta objetividad de los datos. Los datos masivos necesitan ser "limpiados" y esto remueve la objetividad, ya que es un proceso subjetivo por parte del investigador, decidiendo cuáles variables importan y cuáles no.

Mahrt y Scharkow (2013, p. 21) cuestionan la validez de los datos masivos en casos en donde el investigador "deja que los datos hablen por sí mismos", contrario a lo que sugieren Mayer-Schönberger y Cukier (2013). En estos casos, los investigadores suelen utilizar cualquier dato disponible y, luego, construyen una justificación teórica para su utilización. Mahrt y Scharkow (2013, p. 25) alertan que esta estrategia es totalmente contraria a la teoría tradicional y atenta contra la validez y alcance de los resultados.

Por estos motivos, muchos investigadores están cuestionando la premisa de cuando más datos realmente significan más conocimiento. En muchos contextos, una pequeña muestra puede decir más y contestar mejor a las inquietudes de una investigación que un sinnúmero de datos (Bollier, 2010; Mahrt y Scharkow, 2013; King y Lowe, 2003; Schrodt, 2010; Krippendorff, 2004).

Pero las críticas al *Big Data* no se atañen solamente al campo científico. En esta segunda década del siglo XXI, como ya se referenciaba previamente, los datos son el alma de los negocios. Esto implica obviamente un problema ético que traspasa también a la investigación que utilice el *Big Data* (Mahrt y Scharkow, 2013). En general, los usuarios no tienen conciencia de que sus huellas digitales van a formar parte de una investigación, sea esta comercial, policial o académica. Se tiene por sentado que los internautas consienten automáticamente la utilización de sus publicaciones, fotos, vídeos, etc., pero hay cuestiones que envuelven el derecho a la privacidad y derechos de autor. Todos los rastros generados por los usuarios en internet o en cualquier tipo de herramienta de comunicación están datificados y pueden ser transformados en mercancía de alto valor e interés para las corporaciones (Mayer-Schönberger y Cukier, 2013, p. 51). Los usuarios, en su mayor parte, no tienen idea de que todo lo que hacen se está volviendo mercancía sin su consentimiento, lo que implica, algunas veces, una violación a la privacidad, libertad civil y libre consumo (Bollier, 2010).

2.2. Los métodos digitales como alternativa metodológica

Como se mencionaba anteriormente, el análisis del *Big Data* requiere del uso o dominio de determinados *softwares* o programas informáticos que permitan procesar y visualizar estos enormes conjuntos de datos, dado que la capacidad humana para hacer un análisis manual es reducida. Tal como afirma Rieder (2013), desde hace más de una década se utilizan programas informáticos para capturar, producir o utilizar de otra manera los datos masivos con el fin de investigar diferentes aspectos de internet. Esto es lo que se conoce como *Digital Methods* o métodos digitales y que poseen una serie de ventajas comparadas con los métodos tradicionales; ventajas relativas al costo, velocidad, exhaustividad, detalle, entre otros, pero también, relacionados con la rica contextualización proporcionada por la estrecha relación entre los datos y las propiedades del medio (entendido como tecnologías, plataformas, herramientas, sitios web, etc.). Para Rogers (2015), los métodos digitales son técnicas para el estudio de los cambios sociales y las condiciones culturales usando datos en línea.

Esta metodología hace uso de conjuntos de datos masivos almacenados como por ejemplo hiperenlaces, etiquetas, marcas de tiempo, interacciones de todos los tipos en las redes sociales en internet como los “me gusta”, elementos compartidos, retuits, comentarios, entre otros, y busca entender cómo estos objetos son tratados por los métodos incorporados por las plataformas en línea dominantes. Los métodos digitales se esfuerzan por reorientar la finalidad de los métodos y servicios *online* hacia el punto de vista de la investigación social, y como una práctica de investigación, forman parte del giro computacional en las humanidades y las ciencias sociales, y dentro de esta última, la comunicación y el periodismo.

Como metodología, ésta tiene por objetivo reorientar la finalidad de los datos masivos almacenados en internet por las diferentes plataformas en línea (Ej: Facebook, Twitter, Google, etc.) hacia la investigación social, valiéndose para esto de métodos y herramientas informáticas cuya implementación dependerá de qué tipo de información se requiere recolectar, de qué plataforma se van a extraer los datos, cómo se deben estructurar los datos para su análisis y cómo se van a visualizar los mismos. En ese sentido, como ya se ha mencionado, el investigador debe darse a la tarea previa de conocer el manejo o dominar los programas informáticos que se vayan a usar. Es importante puntuar que estos métodos, además son “experimentales y situacionales” (Rogers, 2015, p. 9), ya que son construidos, en algunas ocasiones, sobre dispositivos que pueden dejar de funcionar o simplemente desaparecer, como páginas webs o determinadas funcionalidades de las redes y medios sociales en internet.

Los *Digital Methods* facilitan la automatización pero no reemplazan en lo absoluto el criterio interpretativo del investigador; los datos hablan y las correlaciones se muestran, pero lo que significan, implican, sugieren, lo que deduce o infiere de esto, es tarea del investigador, mismo que a su vez debe ser consciente de las limitaciones técnicas de estos: la transitoriedad de los servicios web, la inestabilidad de los flujos de datos dado por el cierre o reconfiguraciones de las API (*Application Programming Interface*), la calidad de los datos capturados; las limitaciones, inestabilidades e imprecisiones de los algoritmos y el sesgo que ocasiona la “limpieza” o curaduría de los datos para su procesamiento. Debe tener presente, asimismo, que los métodos digitales no sólo permiten determinar tendencias generales en medio de la masividad sino también profundizar en el detalle o

"letra pequeña" del fenómeno, y el variado abanico de programas informáticos permite hacer lecturas simultáneas de los datos.

Es importante tener en cuenta que, pese al revuelo en torno a las posibilidades abiertas por las técnicas digitales y sus *softwares* y programas de análisis de datos, aún es un campo con muchas dificultades y riesgos. Manovich (2012, p. 9-10) sugiere que los datos masivos deben ser utilizados en combinación con otras técnicas: "Idealmente, queremos combinar la habilidad humana para comprender e interpretar – cosa que las computadoras no pueden hacer todavía – con la capacidad de las computadoras de analizar grandes conjuntos de datos utilizando los algoritmos que hemos creado para ello". Nuttall *et al* (2011) apuntan en la misma dirección, sugiriendo un abordaje científico que pueda combinar los métodos que trabajan con datos y la etnografía. Finalmente, Rogers (2013) asevera que los métodos digitales necesitan un largo tiempo de dedicación, además de una mirada crítica al analizar los datos, pues solo así podrá producir resultados satisfactorios.

Por último, hay muchos desafíos en relación a "qué objetos tener en cuenta, cómo crear una muestra, cómo analizar, cómo interpretar, cómo llegar a los resultados" (Rogers, 2013, p. 85). Todos estos desafíos fueron constantes en los dos casos de ejemplo que se exponen a continuación y que son experiencias de las autoras de este texto. El primero, un estudio al videoactivismo en Brasil a partir de la obtención de datos generados por el usuario en las plataformas sociales de Facebook, YouTube y TwitCasting en el marco del Mundial de Fútbol 2014 que se celebró en este país; y el segundo caso, un estudio sobre el tratamiento a las fuentes noticiosas dentro de cuatro ediciones iberoamericanas de Google News en el 2015, a partir de la realización de un *scraping* o "raspe" de datos mediante un *scraper bot* programado para ello.

3. Metodología

3.1. Caso 1: Investigando con métodos digitales en Facebook, YouTube y TwitCasting

El primer caso que vamos a analizar se refiere a una investigación doctoral titulada "De la calle a la red: videoactivismo en el contexto de las protestas

en contra del Mundial de Fútbol en Río de Janeiro (2014)" (Sousa, 2017). En esta investigación se implementaron los métodos digitales, la investigación participante y entrevistas semiestructuradas, proponiendo una mirada amplia y profunda acerca del videoactivismo desarrollado en la ciudad de Río de Janeiro durante el Mundial de Fútbol de la FIFA.

En el ámbito de este trabajo, entretanto, nos referiremos sólo a los métodos digitales utilizados en la investigación, aunque hacemos mención de las otras técnicas empleadas para que los lectores, principalmente los investigadores en formación, puedan tener en claro cómo fue realizada la investigación y cómo se complementaron las mismas entre sí.

En términos prácticos, los métodos digitales aplicados a los medios sociales posibilitan que los datos sean recogidos automáticamente desde las plataformas, visualizados y, posteriormente, analizados. Los datos pueden ser capturados a través de un *scraping* o a través de la utilización de APIs. Hay diferentes herramientas que permiten la captura de estos datos vía APIs. El laboratorio Digital Methods Initiative, dirigido por Bernard Rieder y vinculado a la Universidad de Ámsterdam, lista varios *softwares* de extracción de datos (algunos desarrollados por ellos), basados en las específicas APIs de cada plataforma, que facilitan este trabajo. En el caso de esta investigación, se optó por utilizar estas herramientas y también, en algunos pocos casos, la recolecta manual. Como ya se ha afirmado, las metodologías digitales se componen de diversas técnicas, que hacen uso de *softwares* diversos para la captura, visualización y análisis de los datos. Este proceso exige mucho esfuerzo, dedicación y puede llevar mucho tiempo, tanto en el aprendizaje de su manejo, como en relación a pruebas y comprobaciones en las bases de datos generados por los sistemas.

En el caso de esta investigación, no se había planteado la utilización de estas técnicas hasta el inicio del trabajo de campo, en junio del 2014. Fue en ese momento cuando se valoró el potencial de Facebook, YouTube y TwitCasting, y de cómo las dinámicas desarrolladas en estos ambientes eran fundamentales para comprender el videoactivismo como un proceso comunicativo de forma completa. A partir de entonces se empezaron a estudiar dichas herramientas y a probar diversos *softwares*. Luego de seleccionar las principales herramientas disponibles, se optó por utilizar, inicialmente, Netvizz para la captura de datos en Facebook, Nvivo para

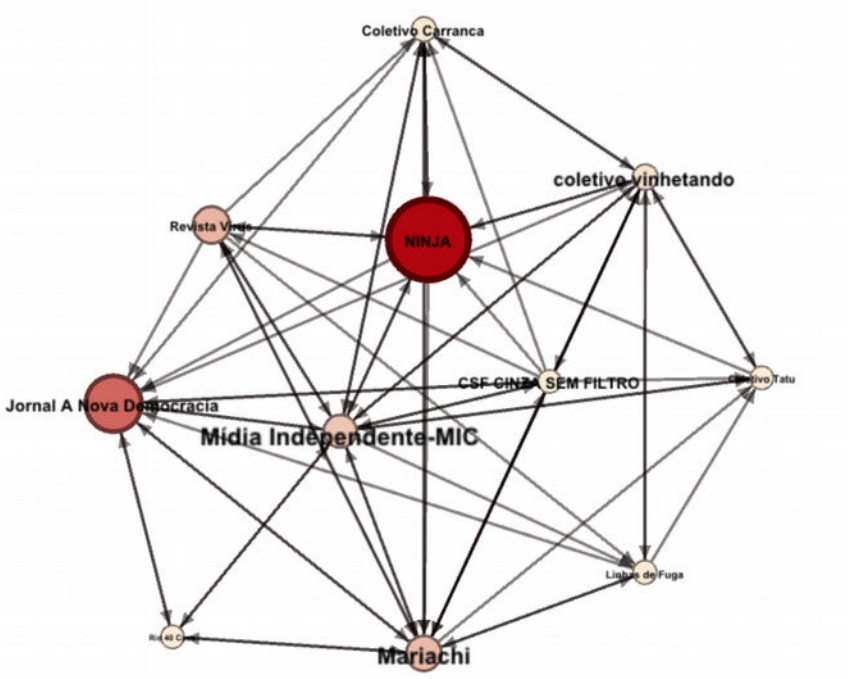
generar nubes de etiquetas (ver Imagen 1) y categorizar los datos para su análisis, y Gephi para la visualización.

Cuando se inició el trabajo de campo, la recolección de los datos, o primera captura, se hacía diariamente utilizando la *app* Netvizz para capturar las publicaciones, comentarios y otras acciones de los usuarios en la plataforma de Facebook. Estando en ese proceso, la *app* fue actualizada y se le incorporó la funcionalidad de que en cualquier momento podían capturarse datos de días anteriores, siendo así, ya no hacía falta hacer el barrido diario y se optó por suspenderlo y dejarlo para después, y redireccionar entonces los esfuerzos en la investigación participante y la realización de las entrevistas.

Esta decisión luego se tornó un problema para la investigación, ya que cuando se reinició la recogida y corroboración de los datos de la plataforma Facebook, es decir, la segunda captura, Netvizz resultó estar limitado debido a las restricciones de privacidad impuestas por Facebook, las dinámicas propias de la investigación y de la API. Los principales problemas encontrados en la utilización de Netvizz fueron: 1) la eliminación de la *fanpage* de unos de los medios estudiados, el Jornal A Nova Democracia, durante el desarrollo de la investigación; 2) los datos son maleables, es decir, la fecha de captura puede determinar que un específico contenido sea o no capturado, una vez que los usuarios y las páginas cambien sus configuraciones de privacidad, alterando así, los datos posibles de ser capturados.

Netvizz sólo genera hojas de cálculo con los datos capturados, por lo que se echó mano del *software* Gephi para la visualización de estos. Sin embargo, la utilización del programa fue muy compleja, pese a que se dedicó bastante tiempo a la tarea. Como los datos que se necesitaban visualizar eran muy básicos, se optó por explorar las posibilidades del programa – que son múltiples – también de forma básica, solamente para generar la visualización de la red (ver Imagen 1) y dar cuenta de cómo los videoactivistas se organizaban entorno de las plataformas de medios sociales, en el caso Facebook. De esta manera, se tuvo conciencia de que tanto Netvizz como Gephi eran herramientas poderosas para explorar los datos digitales.

Imagen 1. Grafo (modelo Force Atlas 2) de la red videoactivista de Río de Janeiro, generado con Gephi a partir de los datos recolectados con Netvizz en Facebook.



Fuente: *Sousa (2017).*

Los datos de YouTube fueron inicialmente capturados manualmente. Se construyó una base de datos con todos los vídeos del periodo de la muestra, totalizando 173 vídeos. Para analizar las acciones alrededor de la narrativa audiovisual de forma más profunda, observar las interacciones y comentarios de la audiencia, se realizó una segunda recolección, utilizando el *software* YouTube Data Tools, el cual permitió visualizar los siguientes datos: informaciones del canal, listado de vídeos, informaciones y comentarios de cada uno de los vídeos, entre otros datos.

En relación a TwitCasting, una plataforma usada para videostreaming a través de móviles e integrada con Twitter, los datos fueron capturados de

forma manual, por no haber, al menos en ese momento, una herramienta específica para la captura de datos en esta plataforma.

En resumen, los datos fueron recogidos utilizando Netvizz para Facebook, YouTube Data Tools para YouTube, y manualmente en el caso de Twitcasting. En total, se recogieron datos de 173 videos de la muestra, en las tres plataformas. Entretanto, se optó por analizar solamente los 10 videos más visitados de cada plataforma, totalizando en un análisis de: 1) en Facebook fueron 20 mensajes/*posts*, que concentraron 4.455 "me gusta", 6.211 comentarios y fueron compartidos 6.555 veces; 2) en YouTube fueron 20 videos, 1.886.143 visitas, 4.523 comentarios, 1.427 "no me gusta", y 3.462 "compartir"; 3) en TwitCasting, 82 videos y 20.500 comentarios. Todos estos datos, posteriormente, fueron analizados utilizando el *software* Nvivo. La finalidad de este estudio también era básica, buscando revelar las tendencias de contenido en los mensajes. Este análisis permitió evaluar el discurso de los actores involucrados en el proceso. El objetivo de estas acciones metodológicas, en otras palabras, era conocer las acciones de cada uno de los actores y el papel que desempeñaban en la narrativa videoactivista en los medios sociales.

Imagen 2. Nube de palabras elaborada con Nvivo compuesta por aquellas más frecuentes en las descripciones de los grupos de videoactivismo en Facebook (las 100 más frecuentes con más de tres letras).

Fuente: *Sousa (2017)*

3.2. Caso 2: Investigando a los medios noticiosos en Google News

La investigación doctoral "Medios de comunicación iberoamericanos y agregadores de noticias: análisis a las ediciones de Google News Brasil, Colombia, España, México y Portugal" de Cobos (2017), desde una metodología mixta, implementó métodos digitales, consulta documental y entrevistas (tanto presenciales como virtuales), cuya triangulación permitió realizar un análisis de los medios noticiosos, con énfasis en los de carácter iberoamericano, indexados en las ediciones Google News de Brasil, Colombia, México y Portugal, en aspectos como su identificación, su ubicación geográfica, sus cuotas de agregación de noticias, su empresa de medios propietaria, y las percepciones y experiencias sobre el agregador de noticias que tenían los editores en jefe, directores o propietarios de los mismos.

Nuevamente, en el ámbito de este trabajo, nos referiremos sólo a los métodos digitales utilizados en la investigación, pero hacemos mención de las otras técnicas empleadas para que los lectores, principalmente los investigadores en formación, puedan tener en claro cómo fue realizada la investigación y cómo se complementaron las mismas entre sí.

Inicialmente, cuando la investigación se planteó en el 2014, no se contemplaba el uso de métodos digitales. El echar mano de tales herramientas surgió al leer un texto en un blog titulado "Lista de fuentes de Google News España" (Dans, 2005) que hacía mención al uso de un *script* en PHP para listar tales fuentes, y posteriormente, la asistencia a una conferencia ofrecida por Bernard Rieder de la Universidad de Ámsterdam, que lidera la Digital Methods Initiative, en la que se mostró un listado de herramientas, entre las que se encontraba una llamada Google News Scraper.

Dado el objetivo general del proyecto, era necesaria la captura de las noticias de las ediciones mencionadas de Google News, y al ver que se podía hacer la misma de forma masiva utilizando un programa *scraper* o *scraper bot* (raspador), lo que brindaría un mayor y mejor aproximación al fenómeno, se optó por documentarse en detalle en qué consistía la técnica informática del *web scraping*. Posteriormente, se determinaron las variables que debía capturar el *scraper bot* y al ver que la herramienta Google News Scraper no era suficiente para lo que exactamente se quería, entonces se

procedió a contactar a un desarrollador de *software* con quien se contrató el desarrollo de un *scraper bot* en PHP que capturara y almacenara las nueve variables estipuladas por cada noticia.

En este inciso cabe destacar que la investigadora tenía conocimientos previos de la jerga informática y comprensión de cómo funcionaba el *web scraping*, estos dos elementos facilitaron enormemente la tarea de comunicación con el desarrollador en el proceso de construcción, prueba, ajustes y funcionamiento del *scraper bot* y el almacenamiento de los datos en una base de datos en MySQL y la posterior exportación de los mismos a hojas de Microsoft Excel para poder procesarlos (ver Imagen 3). Cabe mencionar, como toda técnica informática, que la misma no está exenta de errores, y que eso hace parte de las limitaciones del proyecto (Ej: lentitud en el procesamiento por saturación de la memoria, eventuales caída del servicio de Google News en algún momento…).

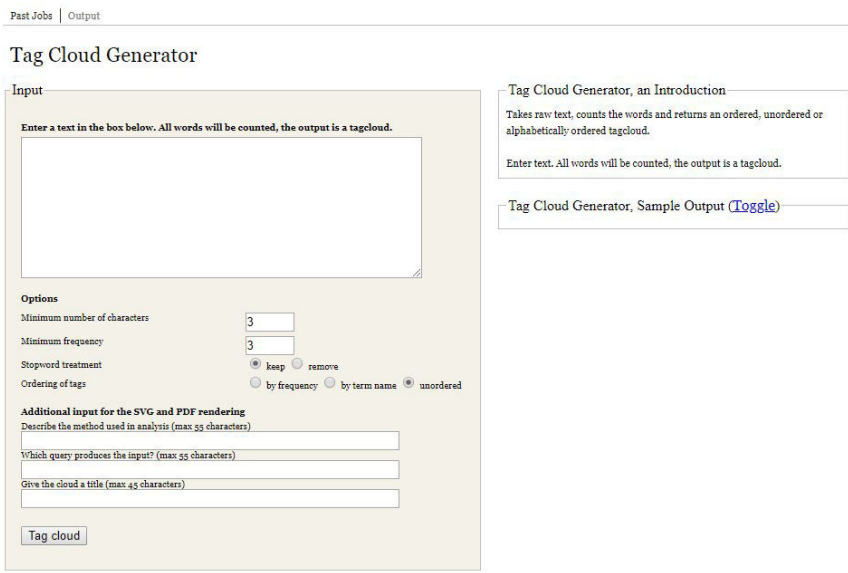
Imagen 3. Visualización de las noticias capturadas de Google News Colombia en una hoja de Microsoft Excel.

originNan	idNews	newsCategory	newsTitle	newsDe	newsUrl	newsSource	newsType	isMostPo	newAddedDateT	operation	Archivo_C	Fila_Orige
Colombia	3503395	Economía	Superintencia inves	-	http://www.elcolombiano.com/superintencia-	El Colombiano	Noticia De	1	1/03/2015 0:11	628	Economia	1766
Colombia	3503398	Economía	Programa completo	-	http://www.bluradio.com/91962/programa-co	Blu Radio	Noticia Ar	0	1/03/2015 0:12	628	Economia	1767
Colombia	3503409	Economía	Avanza adquisición	-	http://www.elespectador.com/noticias/bogot	ElEspectador.com	Noticia De	1	1/03/2015 0:13	628	Economia	1769
Colombia	3503405	Economía	Pliego de cargos por	-	http://www.latarde.com/actualidad/colombia,	LaTarde.com	Noticia Ar	0	1/03/2015 0:13	628	Economia	1768
Colombia	3503413	Economía	Avanza proceso de c	-	http://www.caracol.com.co/noticias/bogota/a	Caracol Radio	Noticia Ar	0	1/03/2015 0:13	628	Economia	1770
Colombia	3503422	Economía	Con metro hasta Sul	-	http://www.bluradio.com/91866/con-metro-h	Blu Radio	Noticia Ar	0	1/03/2015 0:14	628	Economia	1771
Colombia	3503425	Economía	LAN aumentó en 36	-	http://www.elheraldo.co/economia/lan-aume	El Heraldo (Colombia)	Noticia De	1	1/03/2015 0:14	628	Economia	1772
Colombia	3503431	Economía	LAN transportó 4,8 n	-	http://www.eltiempo.com/economia/sectores	ElTiempo.com	Noticia Ar	0	1/03/2015 0:15	628	Economia	1773
Colombia	3503437	Economía	LAN transportó 4,8 n	-	http://www.dinero.com/empresas/articulo/pa	Dinero.com	Noticia Ar	0	1/03/2015 0:16	628	Economia	1774
Colombia	3503444	Ciencia y Tecnolo	NASA no planea lle	-	http://publimetro.pe/actualidad/noticia-nasa-	Publimetro Perú	Noticia De	1	1/03/2015 0:16	629	Ciencia_0	1714
Colombia	3503446	Ciencia y Tecnolo	El administrador de	-	http://noticias.terra.com.pe/ciencia/el-admini	Terra Perú	Noticia Ar	0	1/03/2015 0:16	629	Ciencia_0	1715
Colombia	3503460	Ciencia y Tecnolo	Adiós a las filas en l	-	http://www.lanacion.com.co/index.php/politi	La Nación.com.co	Noticia De	1	1/03/2015 0:17	629	Ciencia_0	1718
Colombia	3503457	Ciencia y Tecnolo	LG Electronics prese	-	http://www.caracoltv.com/programas/farandu	Caracol TV.com	Noticia De	1	1/03/2015 0:17	629	Ciencia_0	1717
Colombia	3503452	Ciencia y Tecnolo	NASA confirma que	-	http://www.caracol.com.co/noticias/internacio	Caracol Radio	Noticia Ar	0	1/03/2015 0:17	629	Ciencia_0	1716
Colombia	3503466	Ciencia y Tecnolo	Ahora podrá agenda	-	http://www.diariodelhuila.com/neiva/ahora-p	Diario del Huila	Noticia Ar	0	1/03/2015 0:18	629	Ciencia_0	1719
Colombia	3503477	Ciencia y Tecnolo	14 aplicaciones móv	-	http://www.latarde.com/entretenimiento/tec	LaTarde.com	Noticia De	1	1/03/2015 0:18	629	Ciencia_0	1721
Colombia	3503478	Espectáculos	Bogotá, la ciudad de	-	http://www.pulzo.com/bogota/298676-bogota	Pulzo	Noticia De	1	1/03/2015 0:18	630	Espectacu	1831
Colombia	3503472	Ciencia y Tecnolo	Neiva, contarán con	-	http://www.caracol.com.co/noticias/regionale	Caracol Radio	Noticia Ar	0	1/03/2015 0:18	629	Ciencia_0	1720
Colombia	3503480	Espectáculos	Pereira la ciudad co	-	http://www.latarde.com/noticias/pereira/147	LaTarde.com	Noticia Ar	0	1/03/2015 0:19	630	Espectacu	1832
Colombia	3503489	Espectáculos	Necesitamos sacar c	-	http://www.eltiempo.com/colombia/medellin	ElTiempo.com	Noticia Ar	0	1/03/2015 0:20	630	Espectacu	1833
Colombia	3503493	Espectáculos	Cuarto Foro Mundia	-	http://www.eltiempo.com/bogota/cuarto-forc	ElTiempo.com	Noticia De	1	1/03/2015 0:20	630	Espectacu	1834
Colombia	3503495	Espectáculos	Se debe enseñar a s	-	http://www.elcolombiano.com/se-debe-ense	El Colombiano	Noticia Ar	0	1/03/2015 0:20	630	Espectacu	1835
Colombia	3503506	Deportes	"Recibí detalles de l	-	http://www.futbolred.com/liga-aguila/noticia	Futbolred	Noticia De	1	1/03/2015 0:21	631	Deportes_	1954
Colombia	3503502	Espectáculos	La Sociedad de la Bi	-	http://www.diariodelhuila.com/opinion/la-so	Diario del Huila	Noticia Ar	0	1/03/2015 0:21	630	Espectacu	1836
Colombia	3503510	Deportes	Hinchas piden clarid	-	http://www.rcnradio.com/noticias/hinchas-pi	RCN Radio (Comunicado de prensa) (bl	Noticia Ar	0	1/03/2015 0:22	631	Deportes_	1955
Colombia	3503518	Deportes	Flabio Torres aclara	-	http://www.caracol.com.co/noticias/deportes,	Caracol Radio	Noticia Ar	0	1/03/2015 0:22	631	Deportes_	1956
Colombia	3503522	Deportes	La FIFA destaca y rec	-	http://www.goal.com/es-co/news/4564/colom	Goal.com	Noticia De	1	1/03/2015 0:22	631	Deportes_	1957
Colombia	3503525	Deportes	Andrés Tello un jugi	-	http://www.radiosantafe.com/2015/02/27/anc	Radio Santa Fe	Noticia Ar	0	1/03/2015 0:23	631	Deportes_	1958
Colombia	3503538	Deportes	Colombia sale por d	-	http://www.elcolombiano.com/colombia-sale	El Colombiano	Noticia De	1	1/03/2015 0:24	631	Deportes_	1960

Fuente: *Cobos (2017).*

En total se capturaron 5.048.150 millones de noticias que permitieron identificar 2.378 medios noticiosos. Una vez finalizado el *scraping* y la información contenida en hojas de Microsoft Excel, procedió a revisarse manualmente, lo que determinó la necesidad de tener que realizar una curaduría a los datos para subsanar las imprecisiones detectadas en el funcionamiento del StoryRank (el algoritmo que opera en Google News), es decir, corregir manualmente los errores que se detectaron presentes en la fuente origen para poder así tener unos datos depurados que permitieran ejecutar otros procesos (Ej: la identificación de las fuentes noticiosas, la determinación de la tasa de agregación de noticias…). Una vez esto cumplido, se procedió a detectar las correlaciones entre estos usando funciones de filtrado, ordenamientos, detección de duplicados y tablas dinámicas de Microsoft Excel. Asimismo, se generaron gráficas o visualización de los datos usando el mismo programa.

Imagen 4. Interfaz de Tag Cloud Generator.



Fuente: *Tag Cloud Generator.*

Otra herramienta de los métodos digitales usada fue Tag Cloud Generator (ver Imagen 4), la misma se utilizó para, a partir de los titulares capturados, hacer nubes de etiquetas que permitieran identificar los términos que se repetían con mayor frecuencia y así tener una aproximación a cuáles eran los temas que Google News seleccionaba para confeccionar su agenda en los diferentes canales del servicio, tanto por cada edición, como una mirada en general. Cabe mencionar que una vez generadas las diferentes nubes de etiquetas, se procedió manualmente a eliminar los artículos y palabras conectoras (Ej: como, la, el, los, este, etc.).

Como anotación final al respecto, las bases de datos generadas con las noticias capturadas en cada edición del agregador o *datasets* (Google News Brasil, Colombia, México y Portugal, enero 1 de 2015 a marzo 31 de 2015 UTC+1), fueron liberados bajo licencias Creative Commons en el Dipòsit Digital de Documents de la UAB para que puedan ser usadas en otras investigaciones.

4. Conclusiones

Llegados a este punto, es evidente que, en la segunda década del siglo XXI, las ciencias sociales se han convertido ahora en una de las más ricas en datos a partir del boom tecnológico en que vivimos, por lo tanto, esto se convierte en grandes oportunidades de investigación, pero que también tiene su lado oscuro. En la discusión teórica se puede apreciar que el uso del *Big Data* y la implementación de métodos digitales para su procesamiento es aún un campo contradictorio, de experimentación, con sus potencialidades y riesgos, con sus partidarios y detractores.

También, lo anterior resulta desafiante a nivel técnico para los científicos sociales que deben entender de técnicas informáticas y aprender el manejo de programas para la captura, procesamiento y visualización que implica, desde luego, una curva de aprendizaje, y que esto no reemplaza en lo absoluto el análisis y raciocinio de quien investiga. En adicional, en algunos casos, el tener que interactuar con desarrolladores de *software* y "traducir" para estos lo que se desea y tener claro, además, que el desarrollo de código no es una "varita mágica" que hace que automáticamente aparezcan las cosas. Por otro lado, el procesamiento del *Big Data* demanda equipos de cómputo con un procesador veloz y amplia memoria RAM para que los

softwares puedan trabajar los datos, de estos aspectos dependerá la celeridad con que se obtengan los resultados.

Finalmente, en relación a los dos casos presentados, distintos entre ellos, se observa que los datos masivos tratados a través de métodos digitales aportaron riqueza informativa, para, en el primero, identificar las acciones de los videoactivistas en el mencionado macroevento deportivo, y en el segundo, registrar el comportamiento de un algoritmo con respecto a las noticias que se capturaban y jerarquizaban en el referido agregador de noticias, sin perder de vista que la automatización en la captura y procesamiento de los datos, primero, requería en algún momento trabajo manual, particularmente en la depuración de los mismos, y segundo, la combinación con técnicas tradicionales de investigación, particularmente las cualitativas, para obtener resultados complementarios entre sí.

5. Referencias bibliográficas

Anderson, C. (2008, 23 de junio). *The end of theory*. Wired. Recuperado de https://www.wired.com/2008/06/pb-theory

Bollier, D. (2010). *The Promise and Peril of Big Data*. Washington: The Aspen Institute. Recuperado de https://www.emc.com/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf

Boyd, D. & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society, 15*, 662-679. Recuperado de https://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878

Cobos, T. L. (2017). *Medios de comunicación iberoamericanos y agregadores de noticias: análisis a las ediciones de Google News Brasil, Colombia, España, México y Portugal* (tesis doctoral). Universitat Autònoma de Barcelona. Recuperado de https://ddd.uab.cat/record/188096

Dans, E. (2005, 29 de marzo). *Lista de fuentes de Google News España*. Enrique Dans. Recuperado de https://www.enriquedans.com/2005/03/lista-de-fuentes-de-google-news-espana.html

Digital Methods Initiative (2020). Recuperado de https://wiki.digitalmethods.net/Dmi/DmiAbout

Hadi, H., Shnain, A., Hadishaheed, S. & Ahmad, A. (2015). Big data and five V's characteristics. *International Journal of Advances in Electronics and Computer Science*, *2* (1), 16-23. Recuperado de http://www.iraj.in/journal/journal_file/journal_pdf/12-105-142063747116-23.pdf

King, G. & Lowe, W. (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, *57* (3), 617–642. Recuperado de https://gking.harvard.edu/files/abs/infoex-abs.shtml

Krippendorff, K. (2004). *Content analysis. An introduction to its methodology* (2nd ed.). Thousand Oaks, EEUU: Sage.

Mahrt, M. & Scharkow, M. (2013). The Value of Big Data in Digital Media Research. *Journal of Broadcasting & Electronic Media*, *57* (1), 20-33. Recuperado de https://www.tandfonline.com/doi/abs/10.1080/08838151.2012.761700?journalCode=hbem20

Manovich, L. (2012). *Trending: The Promises and the Challenges of Big Social Data*. En Gold, M. (Ed.) Debates in the Digital Humanities. Minnesota, EEUU: The University of Minnesota Press. Recuperado de https://minnesota.universitypressscholarship.com/view/10.5749/minnesota/9780816677948.001.0001/upso-9780816677948-chapter-47

Marr, B. (2016). *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. Nueva York, NY: Wiley. Recuperado de https://www.wiley.com/en-co/Big+Data+in+Practice%3A+How+45+Successful+Companies+Used+Big+Data+Analytics+to+Deliver+Extraordinary+Results-p-9781119231394

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: la revolución de los datos masivos*. Madrid, España: Turner Publicaciones.

Sousa, A. L. (2017). *De la calle a la red: videoactivismo en el contexto de las protestas en contra del Mundial de Fútbol en Río de Janeiro (2014)* (tesis doctoral). Universitat Autònoma de Barcelona. Recuperado de https://ddd.uab.cat/record/188119

Nutall, P., Shankar, A., Beverland, M. & Stallwotrth, C. (2011). Mapping the Unarticulated Potential of Qualitative Research: Stepping out from the Shadow of Quantitative Studies. *Journal of Advertising Research, 51*, 153–166, Recuperado de http://www.journalofadvertisingresearch.com/content/51/1_50th_Anniversary_Supplement/153

Rieder, B. (2013). *Studying Facebook via data extraction: the Netvizz application*. Annual ACM Web Science Conference, París, 2–4 May, pp. 346–355. New York, EEUU: ACM. Recuperado de https://dl.acm.org/citation.cfm?id=2464475

Rogers, R. (2013). *Digital Methods*. Cambridge, EEUU: MIT Press. Recuperado de https://ieeexplore.ieee.org/book/6517069

Rogers, R. (2015). *Digital methods for Web research*. En Scott, R., Kosslyn, S. & Buchann, M. (Eds.) Emerging trends in the social and behavioral sciences: An Interdisciplinary, Searchable, and Linkable Resource. Nueva York, EEUU: Wiley. Recuperado de https://onlinelibrary.wiley.com/doi/10.1002/9781118900772.etrds0076

Schrodt, P. (2010). *Automated production of high-volume, near-real-time political event data*. American Political Science Association 2010 Annual Meeting, Washington, 2-5 Sept. Recuperado de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1643761