Límite(s) de los estudios Big Data

Luis DELTELL ESCOLAR
Universidad Complutense de Madrid

Resumen:
Desde hace algo más de una década los estudios Big Data se han consolidado en la literatura científica de la Sociología y la Comunicación. Los resultados de algunos de estos experimentos son sorprendentes. Así: Google Flu Trends (GFT) parece capaz de informar al momento sobre el impacto de la gripe estacionaria en cualquier lugar del planeta. Estos estudios basados en la cuantificación de datos obtenidos de Internet, especialmente de las redes sociales y de los buscadores, configuran un nuevo modelo de investigación.
Los autores más optimistas hablan de una ruptura definitiva entre estudios cualitativos y cuantitativos. Los datos máximos ofrecidos por Facebook o Twitter parecen haber superado la distinción entre cualitativo y cuantitativo. El propio flujo de información evidencia el comportamiento, las tendencias e, incluso, predice los resultados de elecciones, taquillas, audiencias o compra de consumidores.
Sin embargo, desde 2013 se han comenzado a evidenciar errores en dichos estudios. Así recientes investigaciones demuestran que análisis basados en el algoritmo de Twitter contienen fallos importantes, que la cuantificación de las tendencias en Facebook conlleva sesgos elevados y, sobre todo, que hasta el célebre GFT presenta errores de medición enormes. Estas quiebras técnicas no son, sin embargo, las lagunas más graves de los estudios Big Data.
Tal vez, el error de los estudios de datos máximos se encuentra en su propio planteamiento, como indica el filósofo coreano Byung-Chul Han, el fallo es suponer que la masa de información, por muy amplia que sea, puede suplantar a la necesidad de modelos y fundamentos teóricos.
Palabras clave: Estudios Big Data; Internet; Facebook; Twitter; Metodología.

Abstract:
Big Data studies have been consolidated in the Sociology and Communication scientific literature for almost over a decade. The results of some experiments are surprising.Google Flu Trends (GFT) seems to be able to inform by minute about the impact of the seasonal flu anywhere in the world. These studies based on the quantization data obtained from the Internet, especially social networks and search engines, set up a new research model.
The most optimistic authors discuss about a definite break between qualitative and quantitative studies. The maximum data provided by Facebook and Twitter seem to have overcome the distinction between qualitative and quantitative. The information flow evidences the behavior, the trends and also predicts the results of elections, box office, audiences or what consumers purchase.
However, errors in these studies have begun to be evidenced since 2013. Recent researches demonstrates that the analysis based on the Twitter algorithm contains important errors, that the Facebook quantifying trends entail high bias, and even the famous GFT presents huge measurement errors. Even so, all these technical insolvency are not the most serious gaps of the Big Data studies.
May be, as the philosopher Byung-Chul Han indicates, the errors on the Big Data studies are in its own approach. The issue is to suppose that the mass of information, however how wide it is, may replace the need for models and theoretical bases.

Keywords: Big Data Studies; Internet; Facebook; Twitter; Methodoloy.

Introducción: ¿y si la metodología no hiciese falta?

En casi todas las culturas del mundo ha sobresalido la figura del profeta, del oráculo o del adivino. Si paseamos por las tierras griegas, subimos a las altas cumbres de los Andes o recorremos las riberas de los grandes ríos asiáticos siempre encontramos el recuerdo de esos hombres que estuvieron dispuestos a predecir y aventurar el futuro. Alrededor de ellos se construyeron los grandes templos de Delfos, Ingapirca o los edificios de los seguidores del I Ching. Estos adivinos disponían de una metodología absurda —a ojos de los actuales científicos—: machar el maíz, lanzar tabas al aire, escuchar el crecimiento del té o el estudio de los hexagramas. Y, sin embargo, durante siglos ellos fueron los que decidieron e impusieron su visión del presente y del futuro inmediato a sus respectivas sociedades.

La metodología científica occidental se basa en un distanciamiento radical de los artificios de los antiguos oráculos. Nuestra ciencia se establece y se fundamenta en datos, cifras, analogías y conclusiones. Desde el siglo XVI los científicos y pensadores de las dos grandes corrientes metodológicas —la cualitativa y la cuantitativa— se han enfrentado por imponer sus caminos. Poco a poco y de forma significativa las ciencias cuantitativas e inductivas se han apoderado de la academia, de las revistas científicas y, en suma, de la visión del mundo contemporáneo. Como ya aventuró José Ortega y Gasset: «la verdad científica flota, pues, en mitología, y la ciencia misma, como totalidad, es un mito, el admirable mito europeo» (Ortega y Gasset, 1996: 260).

Los teóricos cuantitativos, aquellos que se basan en experimentos cifrados y en el estudio de los datos, no sólo lograron situarse en los campos más prestigiosos de las ciencias puras sino que han establecido sus procedimientos y técnicas —básicamente inductivas— incluso en las ciencias sociales y en disciplinas más alejadas como la Filosofía. La visión cuantitativa domina de una forma significativa todas las perspectivas de los laboratorios, de las escuelas y de las universidades. Ni que decir tiene que cualquier trabajo inductivo basado en una gran cantidad de resultados numéricos obtiene rápidamente la aprobación y la celebración de revisores, editores y evaluadores.

Sin embargo las deducciones inductivas, como se sabe, contienen errores desde su propia raíz. Bertrand Russell nos regaló una fábula y un personaje ficticio

extraordinario, que nos permiten entender los excesos de la metodología cuantitativa. Se trata de su pavo *inductivista*:

> Este pavo descubrió que, en su primera mañana en la granja avícola, comía a las 9 de la mañana. Sin embargo, siendo como era un buen *inductivista*, no sacó conclusiones precipitadas. Esperó hasta que recogió una gran cantidad de observaciones del hecho de que comía a las 9 de la mañana e hizo estas observaciones en una gran variedad de circunstancias, en miércoles y en jueves, en días fríos y calurosos, en días lluviosos y en días soleados. Cada día añadía un nuevo enunciado observacional a su lista. Por último, su conciencia *inductivista* se sintió satisfecha y efectuó una inferencia inductiva para concluir: 'Siempre como a las 9 de la mañana'. Pero ¡ay! Se demostró de manera indudable que esta conclusión era falsa cuando, la víspera de Navidad, en vez de darle la comida, le cortaron el cuello. Una inferencia inductiva con premisas verdaderas ha llevado a una conclusión falsa (Russell, 1982).

Así, como observó el filósofo británico, los métodos cuantitativos presentan siempre un punto flaco: por muchos datos que obtengan, siempre puede suceder que un nuevo experimento muestre un resultado opuesto y contradictorio que anule el trabajo precedente. El salto inductivo suponía —y supone— aventurar que en el futuro nada cambiará. Es obvio que en las ciencias físicas y químicas esta inferencia funciona con una gran fiabilidad. Así la teoría de Isaac Newton y sus Leyes del movimiento parecen ser más o menos aceptables en todo el entorno doméstico o terráqueo. La Teoría de la relatividad funciona con elegancia en casi todos los escenarios —salvo en las singularidades que encontró el propio Albert Einstein—. Sin embargo, los saltos cuantitativos e inductivos en las ciencias sociales son muchos más complejos y pocas veces explican de una forma certera el comportamiento humano.

Los biólogos moleculares realizan sus experimentos sobre tres o cuatro docenas de sujetos de laboratorio. Estos científicos abren los cerebros de sus ratones, les inyectan reactivos y compuestos químicos, y con ellos pueden llegar a comprobar si la proteína Tau, o cualquier otra sustancia, es o no la desencadenante del mal de Alzheimer o de la evolución de la misma enfermedad. No obstante, tres o cuatro docenas de sujetos serían una muestran ridícula para enfrentarse a cualquier experimento sobre la economía, la sociedad o los movimientos sociales.

Predecir el comportamiento de la población —ya fuese el resultado de unas elecciones, la mejora de una economía o, simplemente, el porcentaje de la audiencia de un programa de televisión— requiere de unas muestras enormes. Intentar aventurar hipótesis de actuación social al estudiar una docena de sujetos nos parecería tan absurdo y tan a-científico como plantear una conclusión basada en lanzar las tabas, escuchar el

crecimiento de las hojas del té o machar los granos de Mama Sara. Por ello, los estudios de las ciencias sociales requieren de unas muestras cuantitativas considerables y, en consecuencia, costosísimas económicamente.

Tal vez uno de los experimentos de las técnicas cuantitativas de las ciencias sociales más repetido y más popular sea el de las encuestas políticas y sociales. La metodología científica para realizarlas es sencilla en su conceptualización: si no podemos preguntar qué partido van a votar todos los electores del país, sí podemos hacerlo a una muestra determinada y, desde estos datos, extrapolar unos futuros resultados de esa hipotética consulta. Aunque matemáticamente es sencillo y fácil de ejecutar, lo cierto es que la inmensa mayoría de las encuestas suele fallar. El grado de error a veces es tan amplio que resultan escandalosas, en otros casos las diferencias son tan notables que simplemente nos demuestran que sólo han sido suposiciones.

Los mejores aciertos en unas encuestas políticas se obtienen en los países más estableces y en aquellos donde la intención de voto se centra —tan sólo— en dos opciones, como es el modelo norteamericano, o a las segundas vueltas presidenciales de muchos países europeos. Cuando las variables de partidos políticos y de intenciones o sentimientos aumentan, los resultados son más imprevisibles. Para poder extrapolar en los escenarios más complejos, las muestras tendrían que ser tan grandes que resultarían costosísimas.

Sin embargo, desde la llegada de las redes sociales y del uso masivo de Internet las muestras pueden ser cada vez más y más grandes. Sin casi ningún coste —o al menos un coste residual comparado con las técnicas tradicionales—, los científicos sociales pueden lograr aumentar este número de sujetos estudiados hasta casi la totalidad —o al menos hasta la totalidad de los sujetos que utilizan las redes sociales o Internet—.

Uno de los experimentos Big Data más célebres —y también más cuestionado— fue predecir el resultado de unas elecciones alemanas por medio de Twitter (Tujasman & et al, 2012). Como se sabe, este medio social es un espacio de *microblogging*, creado en California, que consiste en la creación de pequeños blogs personales compuestos por tuits o *tweets*, que son mensajes de menos de 140 caracteres. Lo que es menos conocido es que Twitter permite a los programadores utilizar todo el torrente del flujo de tuits, es decir, la empresa norteamericana permite el acceso a su código de programación API, para que cualquier estudioso pueda capturar y seleccionar todos los mensajes que se han producido en la Red. Así, los científicos alemanes, encabezados por Tujasman, se plantearon que si analizaban todos los tuits germanos que tuviesen referencias políticas

sabrían con exactitud el resultado de las siguientes elecciones. Su experimento a pesar de los muchos fallos encontrados (Gallo-Avello, 2013), obtuvo unos resultados más certeros que los de muchas encuestas tradicionales.

En una escala significativamente menor, un pequeño grupo de investigadores de la Universidad Complutense repetimos dicho experimento, con parámetros distintos sobre si seríamos capaces de predecir los resultados de una elecciones autonómicas en España (Deltell, Claes y Osteso, 2013). Aunque nuestro trabajo presenta errores importantes en los partidos menores, acierta en mejor medida que las encuestas tradicionales en las dos grandes formaciones: PSOE y PP.

	Fecha	PSOE	PP	IU	UPyD	PA	eQuo
Metroscopia	18/3	34,40%	47,30%	8,80%	3,20%	2,70%	<1%
IMC	18/3	35,60%	47%	8,70%	4,60%	1,80%	<1%
Medición Twitter	18/3	36,83%	40,44%	7,45%	6,70%	5,74%	2,64%
GAD3	19/3	38,01%	46,70%	8,10%	2,60%	2,10%	<1%
Medición Twitter	19/3	36,83%	40,44%	7,46%	6,70%	5,74%	2,64%
Medición Twitter	24/3	36,31%	40,48%	7,73%	6,95%	5,73%	2,80%
Resultados	25/3	39,52%	40.66%	11.34%	3,35%	2,50%	0,53%

Fig. I. Experimento de predicción de los resultados de las elecciones andaluzas 2011 por Twitter.

El salto que plantean los estudios Big Data es realmente significativo. No se trata de trabajar con muestras más o menos importantes, sino que ahora se estudian todos los mensajes, es decir, el cien por cien de los tuits. Así cuando los investigadores analizan, o analizamos, el comportamiento social en Internet pueden llegar a cuantificar la totalidad de los mensajes y de los datos. Es decir, no se trata de inducir nada, como en el caso del pavo *inductivista* de Russell, sino que el científico simplemente debe contar o medir aquello que posee la totalidad.

Los estudios Big Data se sustentan por tanto en datos masivos, en cifras y números de resultados nunca antes utilizados. Es difícil imaginar estas cantidades y estas cifras. Tras la expresión datos masivos se encuentran cientos de miles, millones y hasta cientos de millones de elementos. Casi cualquier investigación Big Data, por pequeña que sea, se encontrará pronto con corpus completamente desmedidos e inmensos. En algunas de las investigaciones que veremos se superan los más de 17 millones de mensajes capturados y monitorizados. La posibilidad que ofrece Twitter de utilizar su código de

programación o los estudios que presenta la compañía Google muestran cantidades de información nunca antes estudiadas.

Ahora, por lo tanto, el tema central no es el tamaño de la muestra, que es gigantesco, sino que el centro de atención ha cambiado. La clave no es como fue hasta unos años intentar el explicar el por qué de las cosas sino el cómo. Viktor Mayer-Schönberger y Kenneth Cukier plantean que es precisamente aquí donde reside el tema central de los análisis Big Data; no se trata de entender el por qué sucedió un hecho o un comportamiento social, sino en predecir cómo será la siguiente acción o el siguiente movimiento.

Igual de interesante que la ausencia de un por qué es la falta de necesidad de una metodología científica. Los estudios Big Data no recurren a las metodologías clásicas, sino que éstas han sido sustituidas por algoritmos, monitorización y cómputos. Ahora ya no se trata de la vieja polémica entre pensadores inductivos o deductivos, o entre análisis cuantitativos o cualitativos, sino que en este momento sólo hace falta contabilizar y sumar los resultados. La monitorización ha sustituido a la reflexión metodológica. Las mejores investigaciones basadas en datos masivos parecen prescindir de las teorías más básicas sobre metodología.

Aún más, estos nuevos estudios intencionalmente parecen situarse en otro lugar, surgen desde otra fuente de conocimiento. Ya no se trata de construir una narración científica, sino de directamente de predecir el futuro, así lo vieron Asur y Haberman al titular su revelador artículo: *Predecting the future* (Asur y Haberman, 2010). Ambos pensaron que los análisis masivos permitirían en un futuro cercano mostrar cual sería el comportamiento humano sin necesidad de recurrir a ningún tipo de metodología tradicional. Simplemente basándose en estas cantidades inmensas de información y operándolas con algoritmos voraces ―*greedy algorithm*― los resultados serían sorprendentes.

Sin embargo, los filósofos y pensadores actuales se plantean si estas nuevas investigaciones son realmente ciencia. Así, el intelectual alemán de origen coreano Byung-Chul Han se pregunta cuál es el valor de los datos sino son capaces de explicar la realidad, es decir, no se trata sólo de medir lo que sucede o, incluso, de predecir unos resultados o acciones determinadas, sino que lo científico debe explicar el porqué, debe permitir a la sociedad entender los procesos. Dice el pensador: «La ciencia positiva, basada en los datos (la ciencia Google), que se agota con la igualación y la comparación

de datos, pone fin a la teoría en sentido amplio. Esa ciencia es *aditiva* o *detectiva*, y no *narrativa* o *hermenéutica*» (Han, 2013: 75).

El sociólogo y humanista polaco Zygmunt Bauman es aún más contundente y en su crítica a estos análisis sostiene que la ciencia no sólo se debe conformar con la computación o las cifras sino, sobre todo, con la narración de la realidad. El científico debe ser capaz de construir el relato de la realidad y explicar el modo en que la sociedad funciona. Para Bauman los estudios Big Data resplandecen en el mundo del mercadeo y del consumo desmedido, pero ocultan en su brillo la falta de compresión del mundo que analizan.

Los grandes aciertos de los estudios Big Data

Mayer-Schönberger y Cukier se presentan como dos de los grandes defensores de los estudios Big Data en el mundo académico y científico. Sus trabajos de divulgación les han convertido en los gurús de los datos masivos. Ambos son entusiastas con estos trabajos y la realidad es que no les faltan motivos para su optimismo. Como ellos mismos sostienen, las técnicas basadas en los Big Data ya se encuentran en todos los lugares: servidores, tiendas *on-line*, transporte de pasajeros y mercancías, distribución de bienes... Así los dos autores sostienen que prácticamente sería imposible la vida diaria de los países occidentales sin las herramientas construidas con planteamientos Big Data. No sólo se trata de comprar un billete aéreo o un libro —o cualquier objeto en Amazon— sino de casi cualquier actividad humana cotidiana.

Uno de los experimentos claves de los estudios fue el Google Flu Trends, o simplemente Flu Trends. Desde su lanzamiento ha sido uno de los trabajos más cuestionados y, a la vez, celebrados de la ciencia actual. El propio buscador describe así su proyecto:

> Cada semana, millones de usuarios de todo el mundo buscan información sanitaria online. Como es de esperar, se realizan más búsquedas relacionadas con la gripe durante la temporada de gripe, más búsquedas sobre alergias durante la temporada de alergias y más búsquedas sobre las quemaduras solares durante el verano. Google ofrece información sobre todos estos fenómenos. Pero, ¿pueden proporcionar las tendencias de las consultas la base de un modelo preciso y fiable sobre los fenómenos del mundo real?
>
> Hemos descubierto que existe una estrecha relación entre el número de personas que realizan búsquedas relacionadas con la gripe y las personas que realmente sufren síntomas gripales.

> Obviamente, no todas las personas que buscan 'gripe' están enfermas, pero cuando se suman todas las búsquedas relacionadas con esta enfermedad surge un patrón. Al comparar nuestros recuentos de consultas con los sistemas tradicionales de seguimiento de la gripe, hemos descubierto que las consultas suelen ser muy frecuentes justo en la temporada de auge de esta enfermedad. Mediante el recuento de la frecuencia de estas consultas, podemos estimar en qué medida circula la gripe por diferentes países y regiones de todo el mundo (Google Flu Trends, 2015).

Si bien la compañía estadounidense muestra un claro tono optimista y profético sobre sus resultados, lo cierto es que hasta la propia revista *Nature* tuvo que reconocer el valor de esta investigación y hace más de un lustro aceptaron la publicación del artículo científico que resumía este experimento (Ginsberg et al, 2009). Google Flu Trends funciona realmente bien en muchos lugares, en especial en las naciones occidentales y en los países con una gran penetración de Internet en la sociedad. Así la sección de la aplicación dedicada a España parece seguir con total exactitud el problema del virus de la influencia estacionaria en nuestro país.

Sin embargo, el aparente éxito de Google Flu Trends ha sido cuestionado por otros investigadores. El trabajo más interesante al respecto lo presentó David Lazer, quien mostraba que los picos de más actividad del virus de la influencia no siempre coincidían con los datos ofrecidos por el programa electrónico, y si bien muchos de los resultados podrían ser exactos otros eran errores garrafales. Su trabajo fue publicado por la prestigiosa revista *Science* y supuso un duro golpe y revés para el prestigio académico de los estudios Big Data (Lazer et al, 2014).

A pesar de las críticas de Lazer, los estudios Big Data y los experimentos similares se fueron repitiendo. Uno de los grandes aciertos de estas investigaciones se basa en la aplicación de la teoría de la Triple V, Velocidad, Variedad y Volumen. Cada análisis que utiliza las técnicas Big Data se convierte rápidamente en el estudio con el mayor volumen de datos y el corpus teórico más amplio sobre un determinado tema. Estos datos implican lógicamente también la mayor variedad de perfiles y de casos distintos, lo que permite al investigador o al estudioso enfrentarse desde perspectivas nuevas. Y, sorprendentemente, esta acumulación de datos se realiza a la mayor velocidad posible, en algunos casos se trata de muestras casi instantáneas. Ante las posibilidades que ofrece la triple V, resulta muy difícil renunciar a los trabajos basados en capturas masivas.

Estas capturas masivas se realizan siempre con algoritmos voraces ―*greedy algorithm*― que permiten no sólo monitorizar el comportamiento en Internet sino

incluso clasificarlo y ordenarlo. Sin duda, de todas estas herramientas el algoritmo PageRank es el más famoso y el más usado, es decir, la fórmula mágica que se encuentra tras el motor de busca del buscador más popular de la Red. ¿Quién podría imaginar el mundo actual sin la ayuda de Google? Así, cada vez que se realiza una consulta en dicha web, se pone en marcha el algoritmo *PageRank*. Éste se basa precisamente en la triple V y nos ofrece rápidamente sus resultados: en primer lugar nos indica el número del resultado y luego el tiempo que ha tardado en lograrlo, después nos ofrece una muestra de esa variedad ordenada, según sus criterios.

Cada vez que buscamos en Google somos conscientes que no se nos ofrece una explicación de la realidad, sino sólo una muestra clasificada y ordenada de la misma. Queremos que céleremente se solucione una pregunta o duda y queremos que sea una respuesta, no necesariamente al por qué de nuestra incertidumbre, sino al cómo resolverla. Así, a pesar de las lagunas y los errores que el algoritmo *PageRank* provoca, su respuesta nos sirve para saber cuál es el restaurante más cercano, cuál fue el pintor flamenco más famoso que vivió en Delft, o cuál será la próxima sesión de la película que queremos ver.

Eric Siegel sostiene que si bien los estudios Big Data contienen errores, su uso es ya imparable. Las grandes empresas utilizan sus procedimientos para medir el rendimiento de sus trabajadores, para indicar los gustos de sus clientes y para prever sus gastos e inversiones. Se trata de una herramienta fundamental ―aunque contenga errores― para predecir el futuro (Siegel, 2013).

Experimentos Big Data en Twitter

Los análisis basados en datos masivos se han impuesto como una de las tendencias más sólidas en las ciencias sociales. Aún más, los estudios Big Data han convertido a Twitter en el mayor banco de datos de la Historia de la Humanidad. ¿Por qué este espacio de *microblogging* ha despertado el interés de los investigadores sociales y de los analistas políticos del mundo? Hay varios motivos para el éxito de este medio social frente a otros.

Como observó Evgeny Morozov, uno de los investigadores más críticos con Internet, Twitter se transformó en el ejemplo perfecto de la «revolución democrática». Los defensores de la ciberdemocracia sostenían que el espacio de *microblogging* permitía

dar la voz a los oprimidos y a los disidentes y con ello generar espacios de debate y de reflexión. Morozov indica que hasta en dos ocasiones se propuso a Twitter a la candidatura de los premios Nobel de la Paz.

Así, en una perspectiva ciber-optimista, Twitter permitiría crear nuevas redes y caminos para el diálogo. Cualquier usurario de este medio social podría lograr ser un líder social influyente y poderoso que impusiera su discurso. El espacio de *microblogging* era para estos autores la primera página web de una esfera pública digital donde se desarrollaba netamente la utopía formulada por Manuel Castells de la «autocomunicación de masas». Cada tuit en potencia podía llevar, para estos autores, la semilla de una nueva democracia.

No es extraño por ello que en la mayoría de los grandes cambios sociales de esta última década ―revoluciones árabes, movimientos de indignados en el sur de Europa, #occupywallstreet―, Twitter haya jugado un papel relevante. Para los ciber-optimistas un rol capital, para los ciber-pesimistas ―entre los que está Morozov― un rol menos positivo, o incluso negativo para la consolidación de las democracias en el mundo.

Sin tomar partido en la discusión entre críticos y defensores del uso de Twitter, lo cierto es que el espacio de *microblogging* se posiciona como una de las redes sociales fundamentales en temas de contenido político. Por ello resulta capital en los análisis para Big Data de predicción de voto o de tendencia y sentimiento ideológico.

Sin embargo, Twitter también es relevante por su carácter de inmediatez. Frente a otras redes sociales como Facebook o Instagram cuyos contenidos se mantienen activos más en el tiempo ―*green contents*―, en el espacio de *microblogging* cada tuit que se produce, se consume con celeridad. Los usuarios de otras redes sociales consultan habitualmente contenidos de hace unos días mientras que en Twitter es extraño, aunque no imposible, que el internauta busque y se entretenga con conversaciones lejanas más allá de unas pocas horas ―incluso de minutos―.

Uno de los grandes aciertos de Twitter es precisamente su concepto de *Trending Topic*, que se traduce habitualmente al castellano por tendencia del momento, en el cual se agrupan los temas que están siendo más citados y comentados por sus usuarios. La propia compañía intenta aglutinar y encaminar las conversaciones para que todos sus internautas sepan qué es lo que está sucediendo exactamente en ese momento. En una sociedad obsesionada con el tiempo, con la velocidad y con el consumo rápido, el *Trending Topic* ayuda al usuario a saber qué debe saber y qué debe comentar para no perderse en bagatelas o temas de menor importancia. La idea del tema del momento

parece ejemplificar como pocas la metáfora central del trabajo de Bauman *Vidas de consumo*: todo debe poseerse, consumirse y agotarse en el mismo momento que nace el deseo.

Twitter es tan inmediato que se ha convertido en el medio social de lo actual, es decir, de lo que acontece aunque esto sea una nadería o banalidad. Para un gran número de espectadores ya es imposible ver un partido de fútbol retrasmitido o un nuevo capítulo de una serie televisiva y no comentarlo automáticamente. Cada nueva actividad social relevante debe ser comentada y escrita en este espacio de *microblogging* para que tenga un valor pleno. Un acontecimiento sólo llega a acontecer para los internautas cuando alcanza la categoría de tendencia del momento en alguna red social, especialmente en Twitter.

El usuario de este medio social, lanza su tuit a la sociedad y éste puede ser o no comentado, pero siempre deja una huella digital. Es fácil imaginar lo que esto supone para los investigadores sociales. Como dijimos en el arranque de este trabajo, Twitter cedía todo el flujo de tuits a los analistas. Al ofrecer libremente la biblioteca de su código API, cualquier estudio social puede programar y capturar todos los mensajes que se realizan sobre un determinado contenido y en un período de tiempo, esto le permite al científico lograr cosas sorprendentes y, hasta ahora, insólitas. Cualquier investigador puede capturar con facilidad varios millones de mensajes sobre los temas más variopintos en cuestión de semanas. La huella digital dejada por los usuarios de este medio forma automáticamente un corpus enorme sobre cualquier asunto actual.

Para los investigadores en comunicación social las posibilidades son inmensas. Podemos analizar qué programas ve la audiencia social es decir, la gente que utiliza Twitter mientras ve la televisión; se puede contabilizar la población que sigue a los partidos políticos y las opiniones que tienen de los mismos; se puede medir el impacto de una película cinematográfica, de una exposición, de un museo, de un partido político… Es decir, se puede estudiar a la sociedad casi en cualquiera de sus actividades siguiendo la huella digital que ha dejado los distintos perfiles de Twitter.

Así, por ejemplo, en una de nuestras investigaciones monitorizamos todos los mensajes de usuarios que hubiesen escrito en torno a la figura de Hugo Chávez. En este estudio, nos planteamos saber cuáles eran los líderes de opinión que se forman alrededor del que fuera Presidente venezolano. Nos propusimos monitorizar sólo durante tres meses, de enero a marzo de 2012. En esos meses se produjo el fallecimiento de Chávez y en Twitter se realizó una inmensa conversación sobre el político. En total obtuvimos 17

millones de mensajes de internautas que opinaban y comentaban la figura del dirigente sudamericano.

Como se puede deducir, un volumen de 17 millones de mensajes es tan inmenso que supone que ni siquiera un grupo de investigadores puede estudiarlos detenidamente. Todo lo contrario, se trata de analizar el comportamiento del flujo o del torrente y no de investigar uno a uno cada breve mensaje. En una investigación Big Data el volumen es tan grande que los datos comienzan a mostrar comportamientos y patrones en sí mismos. El día de la muerte de Hugo Chávez se publicaron más de un millón y medio de mensajes, el día de su cortejo fúnebre casi dos millones…

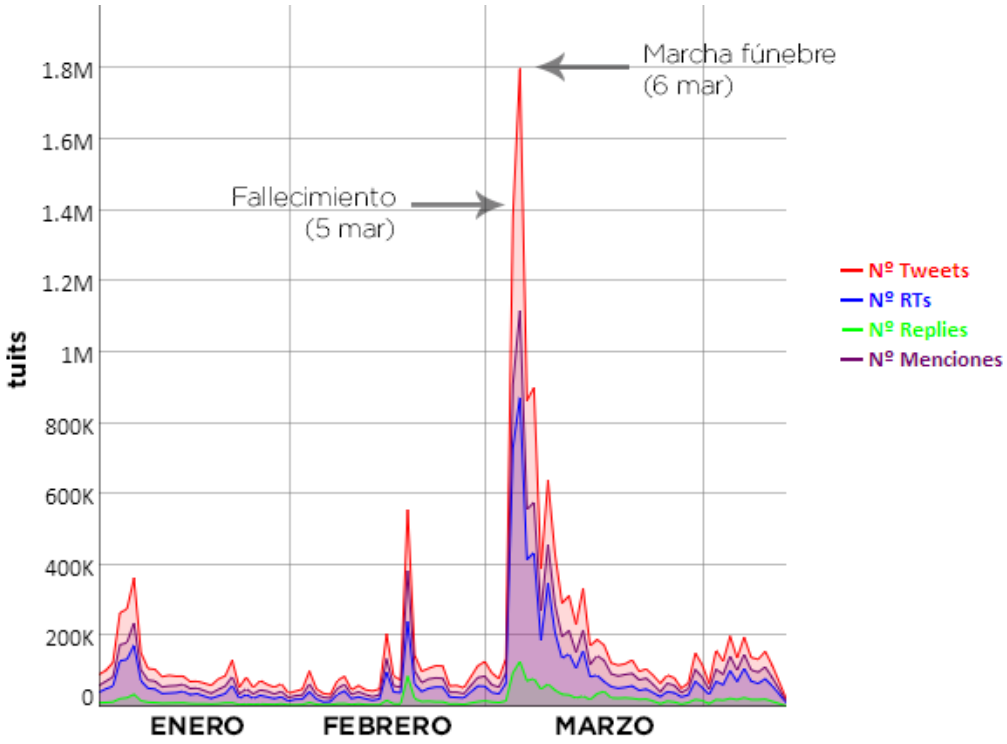

Fig. 2. Fuente: elaboración propia, publicado en *Revista Latina*.

Así nuestro estudio mostraba el comportamiento y la actividad del flujo de tuits en cada día del experimento, pero también era extrapolable a lo que acontecía en la sociedad venezolana y mundial en torno a la figura de Chávez. Nuestro experimento no pretendía, ni podía, explicar el porqué la gente opinaba y debatía sobre la figura del mandatario, pero sí mostraba la urdimbre comunicativa que se tejía alrededor de su persona, su política y su actuación.

Los estudios Big Data de pequeña escala, como el anteriormente indicado —pero también los de gran envergadura— demuestran con claridad que se ha abierto un cambio en el paradigma. Las posibilidades de análisis son inmensas y los corpus

académicos comienzan a ser tan amplios que pueden modificar nuestra forma de hacer y entender la ciencia. Las disciplinas sociales encuentran ahora la posibilidad de abordar encuestas, prospecciones y estadísticas de una forma inimaginable hace tan solo unas décadas.

Predecir o explicar

Si retomamos la imagen del adivino, parece que el nuevo oráculo de las ciencias sociales, y muy especialmente de sus aplicaciones en la mercadotecnia, serán los estudios Big Data. Estos ofrecen rapidez, velocidad y una relativa exactitud para predecir los acontecimientos, para indicar los gustos y, sobre todo, en palabras de Bauman, para entender cómo la sociedad querrá consumir y ser consumida. Durante muchos años estos trabajos seguirán mostrándose y exhibiéndose como las joyas de las revistas científicas y serán consultados por los *gurús del marketing*.

Sin embargo, los trabajos basados en cantidades masivas contienen un problema fundamental: su metodología es casi inexistente. Apoyados siempre en cómputos, algoritmos y sumatorios renuncian a explicar la realidad y a entenderla de una forma completa. Es cierto, que nos pueden indicar cuantas personas están viviendo la televisión y comentándola en Twitter, o hablando de política o discutiendo sobre un nuevo producto…, pero son incapaces de aclarar el motivo de esos comportamientos y de estas tendencias.

Mucho antes de la aparición de los estudios Big Data, Ortega y Gasset descubrió cual era la carencia de los análisis experimentales:

> La verdad científica se caracteriza por su exactitud y el rigor de sus previsiones. Pero estas admirables calidades son conquistadas por la ciencia experimental a cambio de mantenerse en un plano de problemas secundarios, dejando intactas las últimas, las decisivas cuestiones. De esta renuncia hace su virtud esencial y no será necesario recalcar que por ello sólo merece aplausos. Pero la ciencia experimental es sólo una exigua porción de la mente y el organismo humanos. Donde ella se para no se para el hombre (Ortega y Gasset, 1996: 259).

La crítica del filósofo madrileño a los estudios cuantitativos se puede aplicar a los textos Big Data. La exactitud de sus predicciones ―aunque no sea aún plena― no logra en ningún caso ilustrar el comportamiento humano. Aún más, los estudios masivos en muchas ocasiones ofrecen resultados que no aclaran sino que confunden y entremezclan las conclusiones. Los mejores análisis de datos masivos muestran que la actividad en

red es laberíntica y que en vez de explicar lo que sucede, construyen urdimbres comunicativas donde no se descubre nada.

Mayer-Schönberger y Cukier sostienen que las correlaciones no nos dicen la causa de lo que ocurre, pero sí nos alertan de que algo pasa o pasará. Esto es fundamental para entender los aciertos y, también, los límites de los estudios Big Data. Su capacidad de predicción funciona en la medida en que se centra en cuestiones pequeñas. Cuanto más reducida es la actividad humana que se analiza, más capacidad de certeza muestran estas herramientas. Pero al reducir el campo de la investigación, se vuelven incapaces de comprender lo que acontece en la sociedad en su conjunto.

Sin duda, Byung-Chul Han acierta en gran medida en la cuestión capital: los estudios Big Data son incapaces de relatar, es decir, de narrar los hechos. Cuando Ortega y Gasset criticaba la ciencia empírica-experimental decía, como vimos, que construía el mito de la verdad científica, una mitología europea. Sin embargo, los trabajos basados en datos masivos son incapaces, siquiera, de construir esta mitología propia. No pueden formular una narración de lo social o de lo humano.

Los estudios Big Data no desaparecerán, ni mucho menos, seguirán aumentando y se impondrán aún más en las ciencias sociales en los siguientes años. Pero lo que es de esperar es que estas investigaciones se transformen sólo en una herramienta de construcción de corpus y no en análisis completos y cerrados. Resulta obvio, que partiendo de los grandes datos masivos se pueden formular nuevos análisis, pero éstos tienen que recurrir a las viejas metodologías para poder explicar y narrar el mundo. El oráculo-científico no sólo predice el futuro, sino que relata a su sociedad la narración de lo que le acontece.

Bibliografía

Asur, S. Y Huberman, B.A., (2010). «Predicting the Future with Social Media», *Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 492-499.

Bauman, Z. (2009), *El arte de la vida: de la vida como obra de arte*, Barcelona, Paidós.

— (2010): *Mundo consumo*. Barcelona, Paidós.

Castells, M. (2009): *Comunicación y poder*, Madrid, Alianza Editorial.

— (2005): *La Era De La Información: Economía, Sociedad y Cultura*, Madrid, Alianza Editorial.

Deltell, L. (2013): «Audiencia Social Versus Audiencia Creativa: Caso De Estudio Twitter», en *Estudios sobre el Mensaje Periodístico*, 20(1), 33-47.

— y Congosto, M. L., Claes, F. y Osteso, J.M. (2013): «Identificación y análisis de los líderes de opinión en Twitter en torno a Hugo Chávez», *Revista Latina de Comunicación Social*.

— (2013): «Teoría De La Urdimbre Comunicativa. política, activismo y formación de líderes de opinión por medio de Twitter en España», II Congreso AISO, Madrid.

Gallo-Avello, D. (2012): «No, You Cannot Predict Elections with Twitter», *IEEE Internet Computing*, 16(6), 91-94.

Ginsberg, J: et al, (2009): «Detecting influenza epidemics using search engine query data » *Nature* Vol 457, 19 Febrero, 2009.

Han, B-C. (2013): *La sociedad de la transparencia*, Barcelona, Herder.

Lazer, D et al, (2014): «The Parable of Google Flu: Traps in Big Data Analysis», *Science*, Marzo 2014, vol. 343 (6176), 1203-1205.

Mayer-Schonberger, V. Y Cukier, K. (2013): *Big data. La revolución de los datos masivos*, Madrid, Turner libros.

Morozov, E.: (2012): *El desengaño de Internet: los mitos de la libertad en la red*, Barcelona, Destino.

Ortega y Gasset, J. (1996): *Obras completas. Tomo II. El origen deportivo del estado*, Madrid, Alianza Editorial.

Russell, B. (1982): «Pavo inductivista» en Chalmers, A: ¿Qué es esa cosa llamada ciencia?, Madrid, Ed. Siglo XXI.

Siegel, E. (2013): Analítica predictiva: predecir el futuro utilizando big data, Madrid, Anaya.

Tumasjan, A., Sprenger, T.O. & al. (2010). «Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment», *International AAAI Conference on Weblogs and Social Media*, 4, 178-185.