Comunicació: Revista de Recerca i d'Anàlisi [Societat Catalana de Comunicació]
http://revistes.iec.cat/index.php/TC. Vol. 33 (2) (novembre 2016), p. 113-131
ISSN (ed. impresa): 2014-0304 / ISSN (ed. electrònica): 2014-0444 / DOI: 10.2436/20.3008.01.151
Data recepció: 16/11/15
Data acceptació: 09/12/15

La conversación sobre *big data* en Twitter. Una primera aproximación al análisis del discurso dominante

La conversa sobre big data *a Twitter. Una primera aproximació a l'anàlisi del discurs dominant*

The conversation about big data on Twitter. An approach to the analysis of the dominant discourse

Sara Suárez-Gonzalo[1]
Estudiant del doctorat en comunicació del Departament de Comunicació de la Universitat Pompeu Fabra, Barcelona
sarapaz.suarez01@estudiant.upf.edu

Frederic Guerrero-Solé
Investigador i professor lector del Departament de Comunicació de la Universitat Pompeu Fabra, Barcelona
frederic.guerrero@upf.edu

SARA SUÁREZ-GONZALO I FREDERIC GUERRERO-SOLÉ

La conversación sobre *big data* en Twitter. Una primera aproximación al análisis del discurso dominante

La conversa sobre big data *a Twitter. Una primera aproximació a l'anàlisi del discurs dominant*

The conversation about big data on Twitter. An approach to the analysis of the dominant discourse

RESUMEN:
El análisis de grandes cantidades de datos (*big data*) se vislumbra como el nuevo paradigma de acceso al conocimiento en campos tan diversos de la ciencia como la medicina, la biología, la física o las ciencias sociales. El procesamiento de los datos obtenidos de buscadores y redes sociales se ha convertido en una pieza esencial para la definición de estrategias en política, economía o *marketing*. Este trabajo analiza la presencia de informaciones sobre *big data* en una de las principales redes sociales, Twitter. Sus objetivos son la jerarquización de usuarios y mensajes en función de su influencia en la conversación sobre *big data*, así como la identificación de los temas dominantes en la red, a partir del análisis del contenido de los mensajes analizados. Los resultados indican una clara orientación de la información sobre *big data* hacia los negocios, la predicción y la toma de decisiones.

PALABRAS CLAVE:
***big data*, predicción, redes sociales, Twitter, influencia, análisis de contenido.**

La conversa sobre *big data* a Twitter. Una primera aproximació a l'anàlisi del discurs dominant

La conversación sobre big data *en Twitter. Una primera aproximación al análisis del discurso dominante*

The conversation about big data on Twitter. An approach to the analysis of the dominant discourse

RESUM:
L'anàlisi de grans quantitats de dades (*big data*) s'albira com el nou paradigma d'accés al coneixement en camps tan diversos de la ciència com la medicina, la biologia, la física o les ciències socials. El processament de les dades obtingudes de cercadors i xarxes socials s'ha convertit en una peça essencial per a la definició d'estratègies en política, economia o màrqueting. Aquest treball analitza la presència d'informacions sobre *big data* en una de les principals xarxes socials, Twitter. Els seus objectius són la jerarquització d'usuaris i missatges en funció de la seva influència en la conversa sobre *big data*, així com la identificació dels temes dominants a la xarxa, a partir de l'anàlisi del contingut dels missatges analitzats. Els resultats indiquen una clara orientació de la informació sobre *big data* cap als negocis, la predicció i la presa de decisions.

PARAULES CLAU:
***big data*, predicció, xarxes socials, Twitter, influència, analisi de contingut.**

The conversation about big data on Twitter. An approach to the analysis of the dominant discourse

La conversación sobre big data *en Twitter. Una primera aproximación al análisis del discurso dominante*

La conversa sobre big data *a Twitter. Una primera aproximació a l'anàlisi del discurs dominant*

ABSTRACT:
Big data analysis is emerging as the new paradigm of access to knowledge in such diverse fields of science as physics, medicine, genetics, social sciences, economy and communication. The analysis of data collected from social networks or search engines has become essential to know how humans communicate, as well as to define strategies in politics, economics or marketing. This paper analyses the presence of information about big data in one of today's major social networks, the microblogging site Twitter. The main objective of this paper is the classification of users and posts in terms of their influence on the conversation about big data, and the identification of the dominant discourse in Twitter by analyzing the content of the messages posted. Our results point out a clear orientation of the information about big data towards business, prediction and decision making.

KEYWORDS:
big data, prediction, social networks, Twitter, influence, content analysis.

1. Introducción

Big data se ha convertido en uno de los principales focos de análisis en el entorno académico e investigador, y en uno de los principales sectores de inversión de muchas compañías y administraciones que ven en el tratamiento de grandes cantidades de datos (conjuntos de más de 1 terabyte) una forma de mejorar rendimientos y tomas de decisión o de realizar predicciones de futuro. Las razones de la aparición de este fenómeno son, sin duda, variadas y complejas, y se vinculan a un fuerte cambio en el contexto sociotecnológico que tiene su origen a principios del siglo XXI. Minelli, Chambers y Dhiraj (2013) aluden a la concurrencia de lo que llaman «tres tormentas perfectas»: la de la computación, la de los datos y la de la convergencia, que han tenido una enorme repercusión en la economía global. El actual desarrollo e implementación de herramientas diseñadas para la recolección, agregación, análisis, gestión y visualización de grandes cantidades de datos es creciente. Estas se diseñan con el objetivo de extraer de masas informes de datos información sobre cuestiones concretas, comprender las relaciones existentes entre los datos y vislumbrar patrones de comportamientos que hasta ahora permanecían ocultos (Boyd y Crawford, 2012; Shroek, Shockley, Smart, Romero-Morales y Tufano, 2012). Se espera que la analítica predictiva contribuya a anticipar algunas de estas conductas y también actuaciones y respuestas específicas de las personas, en función de probabilidades basadas en datos de situaciones similares anteriores, lo que podría tener una enorme repercusión en la producción, la logística o la estrategia de ventas de las empresas (Mayer Schönberger y Cukier, 2013: 151).

2. *Big data*, un término ubicuo de significado difuso

Nuestro estudio parte de una cuestión crítica: el término *big data* carece de una definición clara, única y consistente. Para comprender esta problemática y su importancia la contextualizaremos brevemente a continuación.

El término *big data* sienta sus bases en el año 1997 a manos de dos investigadores de la NASA para hacer referencia a un nuevo problema de magnitud. En este momento, *big data* da nombre a aquellas cantidades de datos lo bastante grandes como para no poder ser almacenadas en la memoria principal, el disco local o cualquier otro disco remoto de los ordenadores del momento (Cox y Ellsworth, 1997: 1). En 2001, Laney (2001: 1) propone las tres V que definirán esta nueva forma de generación de datos (*velocidad*, *volumen* y *variedad*), a la que posteriormente y desde el sector empresarial se añade una cuarta, la *veracidad*. Más de una década después de su aparición, en 2011, el término comienza a popularizarse y su interés va en aumento de forma exponencial en el sector público y mediático. A partir de entonces, diferentes actores individuales o colectivos afectados por este nuevo

contexto o con intereses de diferentes índoles sobre él, plantean diversas definiciones del término. A esto se suma su naturaleza compleja, en la que se entrelazan una gran diversidad de aspectos técnicos y sociotecnológicos, así como aportaciones desde campos de conocimiento muy dispares. Por otra parte, el epíteto *big* conlleva una complicación añadida: connota importancia, complejidad y desafío, pero también se refiere a una cuestión de magnitud (que, como hemos visto, es esencial en su origen) (Ward y Barker, 2013).

Desde el nacimiento del término *big data* y su posterior popularización, se ha avanzado enormemente en el desarrollo e implementación de las tecnologías necesarias para el almacenamiento y el análisis de datos, y se ha superado así el problema de magnitud al que entonces hacía referencia. De este modo, el significado inicial del término pierde parte de su sentido. Dichos avances tecnológicos tienen una gran repercusión en diversos sectores de actividad, lo que aumenta las potencialidades de *big data* y lo vincula con diferentes realidades. Encontramos ejemplos de referencia en los sectores del transporte (UPS), infraestructuras (General Electric), venta por Internet (Amazon), seguros de salud (United HealthCare) o servicios financieros (Citigroup) (Davenport, 2014). Esto problematiza todavía más la existencia de una definición única. En este sentido, autoras como Boyd y Crawford (2011: 1) o Galdón (2014) han hablado de la *pobreza* del término *big data* (precisamente por esta alusión tan fuerte a una cuestión de magnitud) y sostienen que debería reinventarse para hacer eco de su complejidad real.

En la actualidad, unos autores conciben *big data* como un nuevo *data ecosystem*, cuya principal característica es la alta conectividad entre los datos, independientemente de su formato, su contenido y su fuente de procedencia, y que permite la obtención de información tan valiosa como el reconocimiento de patrones (Boyd y Crawford, 2011; Galdón, 2014; Minelli *et al*., 2013). Otros lo entienden como una mecánica de recolección y almacenamiento de datos (Baruh y Popescu, 2015: 3), de la que extraer nuevas formas de valor para los mercados, las empresas e instituciones públicas como los gobiernos (Mayer-Schönberger y Cukier, 2013: 6), y donde la usabilidad de los mismos adquiere una importancia central (Minelli *et. al*., 2013: 5). Hay incluso quien lo define como «a process to deliver decision-making insights» (Kalyvas y Oberly, 2014: 1). Además, estos vinculan a *big data* diferentes tecnologías, como la inteligencia artificial o el *Machine Learning*, y técnicas como la analítica de datos masivos. Esta visión, más utilitarista, se ha plasmado en algunas de las definiciones de mayor relevancia, provenientes de organizaciones como Gartner, Intel, Microsoft u Oracle (Ward y Barker, 2013).

Por otra parte, hay que señalar que el debate sobre las posibilidades del análisis de grandes cantidades de datos está polarizado entre dos grupos. Uno, el de aquellos que consideran *big data* como la nueva revolución que permitirá la mejora de muchos aspectos de nuestras vidas (entre ellos, la posibilidad de curar enfermedades o de proponer tratamientos a medida, e identificar factores y hábitos de riesgo) y lo ven como el final definitivo de la teoría y el inicio de una nueva era gobernada

por los datos (Mayer-Schönberger y Cukier, 2013; Davenport, 2014). Otro, el de aquellos que se muestran escépticos respecto a las cuestiones que puede resolver (Boyd y Crawford, 2012; Etlinger, 2014), y también el de los que ven en él un futuro distópico controlado por los datos, en el que se impondrá una nueva lógica de la vigilancia (Baruh y Popescu, 2015). Generalmente, estos enfocan su visión desde la perspectiva de las limitaciones y las posibles implicaciones (generalmente sociales) del análisis de datos masivos. Según ellos, *big data* ha abierto una nueva brecha en el acceso al conocimiento, que divide territorios en función de su riqueza de infraestructuras. Dudan de la ética del uso de datos públicos e incluso de la verdadera publicidad de ciertos datos, y consideran que no se garantizan la privacidad y la confidencialidad de los individuos, cada vez más sometidos al control de dispositivos que se extienden en todas las facetas de sus vidas, en la llamada *Internet de las cosas*. Además, consideran que provocará nuevas desigualdades fruto del uso de equipos informáticos capaces de procesar datos a una escala hasta ahora sin precedentes (Lazer *et al*., 2009). Aclaran que este problema ético vinculado al uso de datos personales se manifiesta y se agrava especialmente en el ámbito de los datos personales. Debido a una falta de información y de mecanismos de defensa los ciudadanos pueden perder el control sobre sus datos, encontrarse incapacitados para conocer cómo y quién puede acceder a ellos y de qué modo puede utilizarlos, o simplemente para actuar en pro de la defensa de su privacidad (Baruh y Popescu, 2015; Fairfield y Sthein, 2014; Kalyvas y Oberly, 2014: 33). En esta línea, Mayer-Schönberger y Cukier (2013: 16) señalan una falta de experiencia en la comprensión y la supervisión de los mecanismos de análisis *big data* que puede conllevar perjuicios para los ciudadanos.

Un intento por reducir la ambigüedad que rodea al término se presenta con el trabajo de Ward y Barker (2013). Estos autores presentan un diagnóstico de la problemática en torno al término *big data*, que antes hemos citado. Su estudio analiza un corpus de términos o tendencias vinculados a *big data* del que extraen tres factores comunes (magnitud, complejidad y tecnologías), que posteriormente integran en una nueva propuesta de definición del término. Dichos términos y tendencias se obtienen del estudio de una serie de definiciones de impacto en los últimos años, todas ellas procedentes del sector empresarial. Una de estas definiciones deriva de un análisis de búsquedas realizadas en Google, a través del cual se muestra un listado de términos relacionados con *big data*. Paradójicamente, en este caso Ward y Barker (2013) entienden y emplean la analítica *big data* como una forma de definir el propio término *big data*. Ellos defienden que estos términos o tendencias vinculados se relacionan de forma intrínseca con la propia definición del término *big data*. En este planteamiento reside la asunción previa de que los datos de por sí mismos tienen significado («big data is intrinsically related to data analytics and the discovery of meaning from data» (Ward y Barker, 2013: 2)). Sin embargo, esta es una afirmación que algunos autores consideran resbaladiza (Etlinger, 2014). Apuntan que esta visión hace referencia a una ideología neoliberal que envuelve a

big data, al *deshumanizar* el análisis de datos y desterrar los posibles sesgos relacionados, dotándolo, por tanto, de una falsa apariencia de neutralidad (Baruh y Popescu, 2015). Precisamente, uno de los riesgos que rodean a los datos masivos es el de ligar su análisis a una potencial respuesta ante cualquier cuestión (Galdón, 2014). Etlinger (2014) apunta que, en sí mismos, los datos no tienen significado y que, por el contrario, es su interpretación la que permite darles un significado u otro. Añade que en esta interpretación reside un componente humano y por tanto subjetivo, y recalca que, ante ello, es imprescindible recordar la importancia de las humanidades y de las metodologías cualitativas, que proporcionan el contexto necesario para comprender los datos y permiten conformar un mejor pensamiento crítico. Teniendo en cuenta estas argumentaciones, no parece descabellado señalar una posible asunción errónea de los autores Ward y Barker, al interpretar los términos y tendencias vinculados a *big data* como conceptos intrínsecamente relacionados con él y, por tanto, como una parte integrada en su definición.

Como hemos visto, existen varios motivos por los cuales el desarrollo de una definición única del término *big data* resulta especialmente complejo. Entre ellos destacan su reciente aparición, la multiplicidad de actores y sectores de actividad involucrados que hablan sobre él, o la disparidad de opiniones acerca de su naturaleza y sus posibilidades. A esto se suma la complejidad de definir un concepto, una tarea digna de un profundo estudio de significado.

3. Twitter y *big data*

Actualmente, una gran parte de las informaciones sobre *big data* se difunde a través de las redes sociales y, en particular, a través de la red de *microblogging* Twitter. El objetivo de este trabajo es analizar cómo los mensajes distribuidos a través de Twitter pueden ayudarnos a comprender qué se entiende, de forma general, por *big data*, y contribuir a una definición más acotada del término.

Una de las principales características de las redes sociales son la distribución desigual y altamente jerarquizada de la influencia de los usuarios, que permite su análisis a partir de una pequeña porción de estos, los considerados más influyentes (Guerrero-Solé, Corominas-Murtra y López-González, 2014). Este artículo se apoya en estas características para analizar cómo se distribuye la influencia de los usuarios y sus mensajes en esta conversación, y qué información podemos obtener en cuanto al significado atribuido al concepto *big data*. Pretendemos de este modo dar respuesta a las preguntas de cuáles son los sectores de actividad que más interés tienen en la difusión de información sobre *big data* y en qué sentido van sus mensajes. Además, analizamos cuáles son los temas más frecuentes del discurso dominante a los que se vincula *big data* en Twitter, a través del análisis conjunto de los mensajes y de los términos más empleados entre los tuits analizados.

4. La influencia en Twitter

Twitter, la red de *microblogging* fundada en 2006 que permite la publicación de posts de hasta 140 caracteres, se ha convertido en los últimos años en una de las redes sociales más populares y en uno de los principales objetos de investigación en comunicación. Los estudios sobre Twitter se han centrado en aspectos como la difusión de la información sobre la gripe (Lampos y Critianini, 2010), la predicción de resultados electorales (Tumasjan, Sprenger, Sandner y Welpe, 2010) o del taquillaje de películas antes y después de su estreno (Asur y Huberman, 2010; Deltell, Osteso y Claes, 2013), la polarización política (Guerrero-Solé, 2016) o el análisis de las dinámicas de los índices bursátiles (Bollen, Mao y Zeng, 2011). La influencia de usuarios y mensajes en Twitter ha sido también una de las propiedades más analizadas de la red. Diferentes autores han propuesto algoritmos para la determinación de la influencia tomando en consideración el número de retuits recibidos, las menciones o los seguidores del usuario, entre otras variables (Weng, Lim, Jiang y He, 2010). En este sentido, el número de retuits es el parámetro que se considera más determinante para calcular el impacto de los usuarios en la red. Pero, con independencia del algoritmo utilizado, lo que se observa tanto en Twitter como en el resto de las redes sociales (electrónicas o reales) es la distribución de la influencia siguiendo una ley de potencia (Corominas-Murtra y Solé, 2010) o distribución de Pareto. Eso significa que una pequeña parte de los usuarios poseen la mayor parte de la influencia (y lo mismo pasa con el número de seguidores, el de retuits y menciones recibidos o el de impresiones, los lectores potenciales de un determinado mensaje). Apoyándonos en esta característica, podemos concluir que analizando solo una fracción de los usuarios de una conversación (los más influyentes) y de los mensajes podemos obtener la mayor parte de la información (o, al menos, la más relevante) sobre esta.

5. Objeto de estudio y método

Considerando las grandes inversiones que se están realizando en el campo de *big data*, así como la creciente importancia de este fenómeno en los diferentes sectores sociales y la popularización de Twitter como red de difusión de información en gran parte del mundo, este trabajo se ha propuesto, en primer lugar, analizar cuáles son los usuarios y los mensajes más influyentes en la conversación sobre *big data* en Twitter, para identificar cuáles son los sectores de actividad más interesados en la difusión de información sobre él, y, en segundo lugar, conocer el discurso dominante y los temas de mayor relevancia.

Para ello, se ha recogido una muestra de posts de Twitter que contienen el concepto *big data* o bien la etiqueta #*bigdata*, utilizando la API Search de Twitter.

N_P	Número de posts	162.007
N_T	Número de tuits originales	102.908
N_{RT}	Número de retuits	59.099
U_N	Usuarios únicos	68.195

Tabla 1. Descripción de la muestra de posts sobre #*bigdata* recogidos entre el 23 de noviembre y el 22 de diciembre de 2013

Fuente: *Elaboración propia.*

La elección de la etiqueta y la palabra clave viene justificada por la popularidad y universalidad del término en inglés (sirva de ejemplo la traducción al castellano de la obra de Mayer Schönberger y Cukier (2013), en la que no se traduce el término), y por su brevedad, ya que se compone de solo ocho caracteres, lo que resulta importante debido a la limitación de 140 en la publicación de tuits. La muestra fue recogida entre el 23 de noviembre y el 22 de diciembre de 2013, con un total de 162.007 tuits, a razón de 6.000 tuits diarios. Al tratarse de un trabajo exploratorio, consideramos que una ventana de un mes es suficiente para abordar las cuestiones planteadas. Como se observa en la tabla 1, el número total de usuarios únicos que participaron en la conversación fue de 68.195.

Para cada uno de los mensajes se recogió el identificador único del tuit, el usuario y el nombre completo del autor, el texto del mensaje (que contiene a los usuarios retuiteados, los mencionados, los enlaces externos y las etiquetas), la hora y la fecha de publicación y los datos de geolocalización. Los registros obtenidos fueron importados a una base de datos y procesados para obtener los resultados que se proponía la investigación. Entre los datos calculados estaban el número de veces que un mensaje fue retuiteado y los usuarios que lo retuitearon. A continuación, creamos un registro para cada uno de los usuarios que participaron en la conversación y en el que se recogieron: el número total de tuits y retuits publicados, el número de retuits y menciones totales recibidos, así como los usuarios que lo mencionaron o retuitearon, el número de seguidores y amigos (*followers* y *friends*), el lenguaje de configuración del cliente de Twitter y la descripción y la localización del usuario.

6. Los usuarios y los mensajes más influyentes en la conversación sobre *big data*

A partir de estos datos, aplicamos un algoritmo que tenía en consideración la actividad, los retuits y menciones recibidos y el número de seguidores (Guerrero-Solé y Fernández-Cavia, 2014) para realizar un *ranking* de los usuarios más influyentes en la discusión. En la tabla 2 se muestran los quince usuarios más influyentes según

Usuario	Tipología	Localización	*Rank*	ACT	RT	M	F
Forbes	Medios de comunicación de negocio	NYC	54,609	4	292	864	2.437.300
HarvardBiz	Medios de comunicación de negocio	Boston	23,792	15	1.278	684	1.310.227
TechCrunch	Medios de comunicación de tecnología	Silicon Valley	9,984	1	75	122	3.138.530
Guardian	Medios de comunicación general	Londres	4,499	1	32	102	1.688.149
Intel	Tecnología	California	2,859	2	63	44	2.375.215
WIRED	Medios de comunicación de tecnología	NYC	1,995	2	141	29	2.442.460
SAI	Medios de comunicación de negocio	NYC	1,483	2	28	42	1.287.104
FastCompany	Medios de comunicación de negocio	NYC	1,48	4	46	59	920.319
ForbesTech	Medios de comunicación de tecnología	NYC	1,432	14	274	36	1.027.956
VentureBeat	Medios de comunicación de negocio y tecnología	Silicon Valley	1,411	6	71	244	217.026
FT	Medios de comunicación de negocio	Londres	1,283	3	46	46	995.083
Detikcom	Medios de comunicación general	Indonesia	0,926	2	19	3	6.337.315
Gigaom	Medios de comunicación de tecnología	California	0,76	8	89	141	194.354
EntMagazine	Medios de comunicación de negocio	California, NYC	0,754	1	38	57	501.166
asocialmedia2day	Noticias de medios de comunicación sociales	NYC	0,724	47	349	47	255.818

Tabla 2. Lista de los quince usuarios más influyentes en la conversación sobre *big data* en Twitter, en la que constan su tipología, localización, *ranking*, posts (ACT), retuits recibidos (RT), menciones recibidas (M) y seguidores (F)

Fuente: *Elaboración propia.*

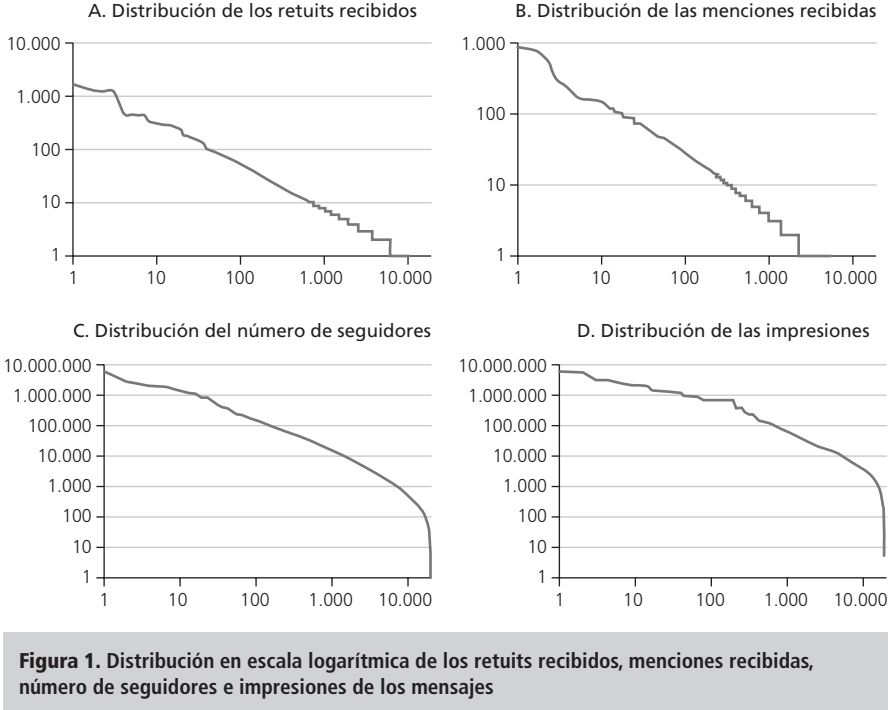

Figura 1. Distribución en escala logarítmica de los retuits recibidos, menciones recibidas, número de seguidores e impresiones de los mensajes

esta jerarquización. Tal y como hemos apuntado antes, la influencia (*Rank*), el número de retuits y menciones recibidos y el número de *followers* siguen una distribución del tipo ley de potencia (figura 1). Para cada uno de los cuatro casos, los exponentes de la ley de potencia son: impresiones (–1,46), RT (–0,91), menciones recibidas (–0,88) y *followers* (–1,76). La aproximación a la ley de potencia es mejor cuanto mayor sea el exponente en valor absoluto, siendo los valores entre 2 y 3 los óptimos.

A su vez, también comprobamos cuáles fueron los mensajes más influyentes (tabla 3), teniendo en cuenta en este caso un único factor, el número de veces que el mensaje fue retuiteado, y descartando aquellos mensajes de usuarios con relativamente pocos seguidores, pero que fueron retuiteados automáticamente y de forma simultánea por su red de seguidores. Consideramos este tipo de mensajes como correo basura.

Además de clasificar los usuarios y los mensajes, calculamos la distribución de los idiomas del cliente de Twitter utilizado (figura 2), así como el número de tuits con enlaces externos y los posts geolocalizados. En el primer caso, comprobamos que de los 102.908 tuits, 91.731 contenían enlace (89 % del total). La utilización de etiquetas y enlaces es una estrategia que se ha observado útil para obtener una mayor redifusión de los mensajes de Twitter (Suh, Hong, Pirolli y Chi, 2010). En el segundo, que solo 1.225 mensajes habían sido publicados por usuarios geolocali-

Usuario	Fecha	RT	Mensaje
HarvardBiz	27/11/2013	140	«Don't invest in big data -- use the data you already have» *http://t.co/yDNCt4eUjK*
Forbes	07/12/2013	139	«Shazam uses big data to predict which music artists will break big in 2014» *http://t.co/8f6dKPqfM8*
HarvardBiz	08/12/2013	128	«How big data will help small businesses» *http://t.co/JI7EHML10p*
HarvardBiz	09/12/2013	94	«Algorithms have their own biases. So what happens when we let them make hiring decision»
HarvardBiz	09/12/2013	89	«Big Data's Biggest Challenge? Convincing People NOT to Trust Their Judgment» *http://t.co/dSZR2Z9RBn*
FastCoExist (Fast Company)	06/12/2013	87	«In The Hospital Of The Future, Big Data Is One Of Your Doctors» *http://t.co/SkyV63MGfp*
HarvardBiz	30/11/2013	82	«To get the most out of your analytics, focus on your top customers» *http://t.co/cyGOOx2*
WIRED	18/12/2013	81	«Big data offers a new tool to help fight human trafficking» *http://t.co/ds4eR5RJKj*
Forbes	04/12/2013	59	«5 reasons why Big Data will crush Big Research» *http://t.co/6llQJBTADz*
HarvardBiz	25/11/2013	54	«How is Big Data Transforming Your 80/20 Analytics?» *http://t.co/NmBjKMMxfj*

Tabla 3. Lista de los diez mensajes más retuiteados en la conversación sobre *big data*

Fuente: *Elaboración propia.*

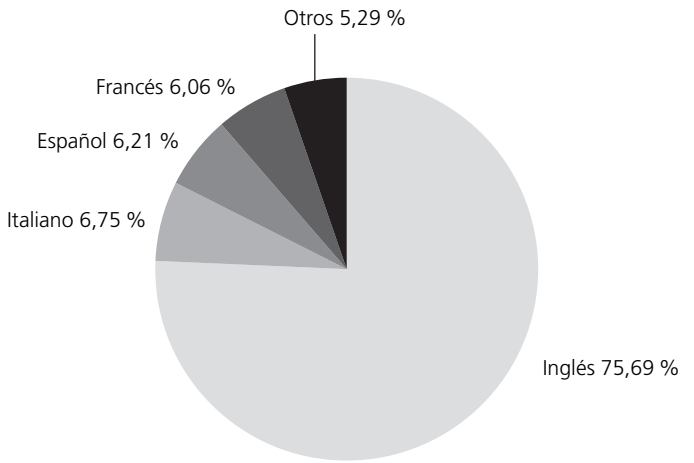

Figura 2. Porcentaje de mensajes por idioma del cliente de Twitter

zados, por lo que obviamos analizar su procedencia ya que los consideramos poco significativos.

7. Análisis del tema dominante en la conversación sobre *big data*

La última fase del análisis consistió en la identificación de las palabras más repetidas en la conversación, considerando únicamente los tuits y dejando aparte los retuits. En este caso no consideramos la ponderación de las palabras en función del *ranking* de los usuarios. El propósito de este análisis iba más allá del simple recuento y pretendía completar el análisis de los mensajes más influyentes para identificar el discurso dominante sobre *big data* en Twitter, a partir del análisis de los términos que eran más utilizados por los usuarios. En la tabla 4 podemos ver cuáles fueron los veinte términos más utilizados.

Término	Frecuencia
Analytics/Analysis	13.782
You/Your	9.624
How	7.695
Marketing/Market	5.935
Cloud	5.724
Business	4.567
Will	4.466
Use	4.357
Can	4.135
Predict	3.373
What	3.363
Hadoop	3.134
IBM	2.439
Help	2.306
Why	1.051
Company	1.955
Future	1.862
Next	1.759
Social	1.660
World	1.645

Tabla 4. Lista de los veinte términos más utilizados en la conversación sobre *big data*
Fuente: *Elaboración propia.*

8. Conclusiones de la investigación

8.1. Los usuarios más influyentes

El método de análisis de los tuits aplicado en este estudio nos ha permitido detectar cuáles son los usuarios más influyentes en la conversación analizada en este estudio sobre #*bigdata* en Twitter. Los resultados muestran que la mayoría de estos usuarios corresponden a empresas norteamericanas (con excepción de dos británicas y una indonesia) ubicadas en Nueva York y California, y excepto el fabricante de tecnología Intel, prácticamente todos son medios de comunicación sobre tecnología y negocios. Precisamente los medios de comunicación, junto con las celebridades, son considerados desde los inicios de Twitter como los usuarios con mayor influencia (Kwak, Lee, Park y Moon, 2010). Así, pues, podemos constatar que los medios financieros son uno de los sectores más influyentes en la conversación sobre *big data*. El hecho de que sea este sector y no otro el que domine la conversación repercute en el tipo de información más influyente, centrada en las estrategias de análisis de datos para obtener rendimientos económicos, ventajas competitivas y toma de decisiones.

Otra de las características de la red de usuarios es que las informaciones de los más influyentes son, por definición, las más redifundidas por el resto de la red. En cambio, estos no redifunden prácticamente ningún otro contenido que no sea el propio. Es decir, los más influyentes no establecen una relación de reciprocidad con ninguno de los usuarios que retuitean su información, por lo que no contribuyen a que estos sean también influyentes. Se reproduce, de este modo, el esquema tradicional de medios unidireccionales, en el que solo los mensajes de unos pocos son redifundidos. Si los usuarios más influyentes retuitean algún mensaje, normalmente es de usuarios que forman parte de su misma organización (es el caso de Forbes, que retuitea a ForbesTech, o de FastCompany, que lo hace de FastCoExist). Podemos ver en la figura 3*b*, en la que se muestra la red formada por los usuarios que se retuitean entre sí, que esta es muy fragmentada y presenta clústers con muy pocos usuarios. No aparecen en ella ninguno de los cincuenta usuarios más influyentes. En cambio, la figura 3*a* nos muestra la red de los usuarios que retuitean a los más influyentes. Podemos observar cómo estos son retuiteados por un gran número de usuarios de la red.

8.2. Los mensajes más influyentes

En cuanto a los mensajes analizados, hemos comprobado, por un lado, que aquellos identificados como más influyentes provienen también de algunos de los usuarios más influyentes, como son HarvardBiz, Forbes, Fast Company y Wired. Además, hemos podido comprobar que la mayoría de los mensajes más influyentes se centran en los mismos aspectos y dibujan un escenario similar al identificado mediante el listado de los veinte términos más recurrentes. Esto nos ha permitido hacer una reconstrucción de cuál es el tema dominante en la conversación analizada: el uso del análisis de grandes datos como una nueva herramienta capaz de ayudar

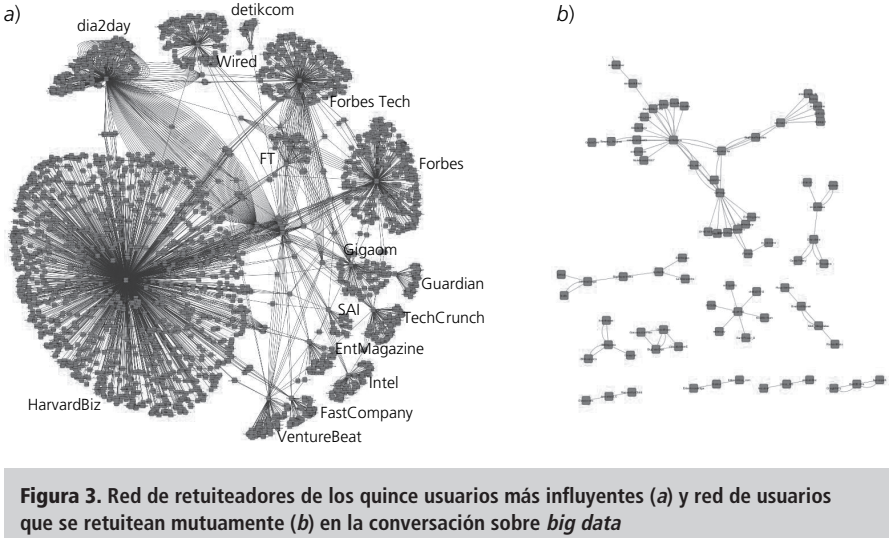

Figura 3. Red de retuiteadores de los quince usuarios más influyentes (*a*) y red de usuarios que se retuitean mutuamente (*b*) en la conversación sobre *big data*

a las empresas a tomar decisiones, predecir el futuro y, por lo tanto, poder mejorar su producción, su logística o sus estrategias de comunicación y venta de productos y servicios. Esta lista de términos también nos ha servido para identificar, de forma independiente al análisis de usuarios, a dos de los principales actores tecnológicos en el campo de #*bigdata*, cuyos nombres se han repetido con una frecuencia de 3.134 y 2.439 veces respectivamente: por un lado, Hadoop, la principal plataforma de análisis de grandes cantidades de datos, y por el otro, el fabricante de equipos y servicios informáticos IBM, el líder indiscutible en este campo.

En el sentido contrario, la presencia de mensajes y de términos relacionados con posturas críticas frente a #*bigdata* es menor. Dos de los conceptos más utilizados desde la perspectiva crítica son los de *privacidad* y *vigilancia*, que aparecen con una frecuencia de 950 y 163 veces respectivamente en el corpus de mensajes analizados. A pesar de ello, hay que destacar que los mensajes de uno de los usuarios más influyentes (HarvardBiz), autor de la mitad de los mensajes más retuiteados, tiene un discurso que no está completamente alineado con el dominante, y está también dirigido a animar a las empresas a fijarse en activos que ya poseen (como sus propios datos), o a poner en evidencia los sesgos de los algoritmos y sus implicaciones en la toma de decisiones tan sensibles como la contratación de personal.

9. Discusión

El método aplicado nos ha permitido elaborar una lista de los usuarios y otra de los mensajes más influyentes e identificar los términos más empleados en esta conversación. Esto nos ha servido a su vez para identificar al sector de los medios finan-

cieros y las empresas tecnológicas como los principales sectores interesados en la difusión de información sobre *big data* y para determinar una clara repercusión de este hecho en el tipo de información más influyente, vinculada a la implementación de mejoras estratégicas y logísticas y, por lo tanto, a la obtención de beneficio en el ámbito empresarial.

Por un lado, estos resultados hacen patente que existe una fuerte disparidad de los discursos y las visiones acerca del fenómeno *big data*. Por otro, demuestran la existencia de un fuerte discurso mayoritario, que se vincula con la visión más utilitarista de las que contextualizábamos al inicio: hemos encontrado que los usuarios más influyentes vinculan al desarrollo de *big data* una modificación de las lógicas empresariales, enfatizan la usabilidad de los datos como una nueva forma de obtención de valor y acentúan la predicción y la toma de decisiones basadas en datos como dos nuevas grandes oportunidades para el sector empresarial, la seguridad y el progreso. De forma contraria, como hemos dicho, el usuario más destacado de la lista de más influyentes (HarvardBiz) apela a una visión más escéptica que los demás, especialmente sobre las posibilidades que se proyectan. En su discurso acentúa algunos de los desafíos ligados a *big data*, se muestra contrario a la inversión en él y propone recursos alternativos de mayor efectividad. Sin embargo, entre los mensajes y usuarios más influyentes no se refleja la visión crítica, que, como decíamos en la introducción, es otra de las visiones más defendidas acerca de este fenómeno y resulta contraria a la del discurso mayoritario. En cualquier caso, sí se han encontrado términos repetidos con una frecuencia suficiente, que acreditan la existencia de esta otra visión.

De entre lo expuesto en este trabajo cabe resaltar, por una parte, que *big data* tiene una gran potencialidad para mejorar ciertos aspectos de la sociedad. La predicción y la elaboración de diagnósticos precisos puede aportar grandes avances, pero parece necesario contemplar el uso de datos desde una perspectiva responsable con la privacidad de las personas, y tener en cuenta las limitaciones propias de los datos y también del posterior procesamiento e interpretación de los mismos. Otra cuestión relevante atañe a las posibilidades de la analítica predictiva. Mayer Schönberger y Cukier (2013) señalan que, si bien la cantidad de información recogida sobre las personas es cada vez mayor, la idea de predecir sus conductas o comportamientos en base a ella guarda relación con concepciones de la naturaleza humana vinculadas a paradigmas como el conductista, el determinista o el mecanicista, según los cuales el ciudadano es visto de forma despersonalizada y facultades como el libre albedrío se ven perturbadas.

10. Limitaciones

Una de las principales limitaciones de este trabajo es que la muestra de mensajes sobre #*bigdata* no es completa, ya que se recogieron un número de alrededor de

6.000 tuits por día. Esta muestra está, además, sujeta a los posibles sesgos de la API Search de Twitter, empleada para la recolección de datos. En este sentido, y a pesar de que podemos considerar que 160.000 mensajes es una muestra suficiente para obtener conclusiones válidas respecto al debate sobre #*bigdata*, estamos trabajando en la obtención de una muestra mucho mayor y más completa, que nos permita obtener conclusiones más definitivas, así como observar cuál es la evolución del discurso y la dinámica de la influencia de los usuarios a lo largo del tiempo. Otro de los aspectos que deben mejorarse de cara al interés de este estudio, y en coherencia con la crítica realizada en el estudio de Ward y Barker (2013), es el del análisis del discurso de los mensajes con herramientas que nos permitan afirmar con mayor precisión cuál es el sentido de cada uno de ellos. Finalmente, también debemos tener en cuenta la desigual distribución de la penetración de Twitter en el mundo. Una de las posibles consecuencias de este hecho es que puede eclipsar las contribuciones de los usuarios y mensajes procedentes de países como la India, China o Rusia, que pueden estar desarrollando estrategias de análisis de grandes cantidades de datos, y cuyas informaciones fluyen en otro tipo de plataformas como Facebook o la red social china Sina Weibo. Además, se ha de contemplar también el problema del desigual acceso al conocimiento e infraestructuras, conocido como *brecha digital*, que mencionábamos al inicio de este trabajo y que favorece hechos como el demostrado en este artículo, relativo a los lugares de procedencia tanto de las informaciones como de los usuarios más influyentes (América del Norte).

Así, a pesar de estas limitaciones, lo que sí que podemos afirmar es que, gracias a sus potencialidades, *big data* es y será una apuesta de futuro, no solo para las empresas, sino también para las instituciones públicas y los gobiernos. Sirva de ejemplo la reciente creación del Alan Turing Institute en el Reino Unido. El gobierno británico invertirá 42 millones de libras en cinco años en un proyecto sobre *big data* que, en palabras de George Osborne, su rector, va a permitir a las empresas mejorar sus procesos de producción, orientar mejor sus estrategias de venta y proveer servicios más eficientes (BBC News, 2014), en clara sintonía con el mensaje que hemos extraído del análisis en Twitter.

Notas

[1] Dirección de correspondencia: Sara Suárez. Roc Boronat, 138. E-08018, Barcelona, UE.

Bibliografía

Asur, S.; Huberman, B. A. (2010). «Predicting the future with social media». *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, p. 492-499. DOI: 10.1109/WI-IAT.2010.63.

Baruh, L.; Popescu, M. (2015). «Big data analytics and the limits of privacy self-management». *New Media & Society*, p. 1-18. DOI: 10.1177/1461444815614001.

BBC News (2014). «Alan Turing Institute to be set up to research big data». <http://www.bbc.com/news/technology-26651179>.

Bollen, J.; Mao, H.; Zeng, X. (2011). «Twitter mood predicts the stock market». *Journal of Computational Science*, 2, p. 1-8. DOI: 10.1016/j.jocs.2010.12.007.

Boyd, D.; Crawford, K. (2012). «Critical questions for big data». *Information, Communication & Society*, 15, p. 662-679. DOI: 10.1080/1369118X.2012.678878.

Cox, M.; Ellsworth, D. (1997). *Application controlled demand paging for out of core visualization*. Report NAS-97-010, julio 1997. Moffet Field: NASA Ames Research Center.

Davenport, T. H. (2014). *Big data at work*: *Dispelling the myths, uncovering the opportunities*. EUA: Harvard Business Review Press.

Deltell, L.; Osteso, J.-M.; Claes, F. (2013). «Twitter en las campañas comunicativas de películas cinematográficas». *El Profesional de la Información*, 22, p. 128-134. DOI: 10.3145/epi.2013.mar.05.

Durán Segura, M.; Mejías Peligro, J. F. (2014). «Conocimientos y comportamientos de los usuarios de la red social Facebook relacionados con la privacidad». *Ámbitos*: *Revista Internacional de Comunicación*, 26: *Infoxicación*. ISSN digital: 1988-5733.

Etlinger, S. *Susan Etlinger: What do we do with all this big data?* Videoconferencia TED (septiembre 2014). Recuperado el 20 de marzo de 2015 de: <http://www.ted.com/talks/susan_etlinger_what_do_we_do_with_all_this_big_data?#t-575588>.

Fairfield, J.; Shtein, H. (2014). «Big data, big problems: Emerging issues in the ethics of data science and journalism». *Journal of Mass Media Ethics*: *Exploring Questions of Media Morality*, 29:1, p. 38-51. DOI: 10.1080/08900523.2014.863126.

Galdón, G. *Entrevista a Gemma Galdón, experta en privacidad en la red.* Vídeo de Youtube (10 diciembre 2014). Recuperado el 20 de marzo de 2015 de: <https://www.youtube.com/watch?v=YHfopl714hg>.

Guerrero-Solé, F. (2016). «Community detection in political discussions on Twitter. An application of the Retweet Overlap Network method to the Catalan process towards independence». *Social Science Computer Review*. DOI: 10.1177/0894439315617254.

Guerrero-Solé, F.; Corominas-Murtra, B.; Lopez-Gonzalez, H. (2014). «Pacts with Twitter. Predicting voters indecision and preferences for coalitions in multiparty systems». *Information, Communication & Society*, 17 (10), p. 1280-1297. DOI: 10.1080/1369118X.2014.920040.

Guerrero-Solé, F.; Fernández-Cavia, J. (2014). «Activity and influence of destination brands on Twitter: A comparative study of nine spanish destinations». *Information and Communication Technologies in Tourism 2014*. Springer International Publishing, p. 227-236.

Hoy, M. G.; Milne, G. (2013). «Gender differences in privacy-related measures for young adult Facebook users». *Journal of Interactive Advertising*, 10 (2), p. 28-45. DOI: 10.1080/15252019.2010.10722168.

Kalyvas, J. R.; Overly, M. R. (2014). *Big Data*: *A business and legal guide*. Nueva York: Taylor & Francis Group, LLC.

Kwak, H.; Lee, C.; Park, H.; Moon, S. (2010). «What is Twitter, a social network or a news media?». *The International World Wide Web Conference Committee (IW3C2)*, p. 1-10. DOI: 10.1145/1772690.1772751.

LAMPOS, V.; CRISTIANINI, N. (2010). «Tracking the flu pandemic by monitoring the social web». *2010 2nd International Workshop on Cognitive Information Processing, CIP2010*, p. 411-416. DOI: 10.1109/CIP.2010.5604088.

LANEY, D. (2001). «File 949. 3D Data Management: Controlling Data, Volume, Velocity and Variety». *Application Delivery Strategies* [Stamford: META Group Inc] (6 febrero). <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> [Consulta: 15 diciembre 2014].

LAZER, D.; PENTLAND, A.; ADAMIC, L.; ARAL, S.; BARABÁSI, A.; BREWER, D.; CHRISTAKIS, N.; CONTRACTOR, N.; FOWLER, J.; GUTMANN, M.; JEBARA, T.; KING, G.; MACY, M.; ROY, D.; VAN ALSTYNE, M. (2009). «Computational social science». *Science*, 323, p. 721-723. DOI: 10.1126/science.1167742.

MANOVICH, L. «Trending: the promises and the challenges of big social data». En: GOLD, M. G. (2012). *Debates in the digital humanities*. Arizona: University of Minessota Press, p. 460-475.

MAYER-SCHÖNBERGER, V.; CUKIER, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston; Nueva York: Houghton Mifflin Harcourt.

MINELLI, M.; CHAMBERS, M.; DHIRAJ, A. (2013). *Big data, big analytics: Emerging business intelligence and analytic trends for today's businesses*. Hoboken, NJ: John Wiley & Sons, Inc. DOI: 10.1002/9781118562260.

SHROEK, M.; SHOCKLEY, R.; SMART, D. J.; ROMERO-MORALES, D.; TUFANO. P. (2012). *Analytics: el uso de big data en el mundo real. Cómo las empresas más innovadoras extraen valor de datos inciertos*. Informe ejecutivo. IBM Global Business Services Business Analytics and Optimisation. <http://ibm.co/1APKffj> [Consulta: 6 mayo 2015].

SUH, B.; HONG, L.; PIROLLI, P.; CHI, E. H. (2010). «Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network». *Proceedings - SocialCom 2010: 2nd IEEE International Conference on Social Computing, PASSAT 2010: 2nd IEEE International Conference on Privacy, Security, Risk and Trust*, p. 177-184. DOI: 10.1109/SocialCom.2010.33.

TUMASJAN, A.; SPRENGER, T.; SANDNER, P.; WELPE, I. (2010). «Predicting elections with Twitter: What 140 characters reveal about political sentiment». ICWSM, p. 178-185. DOI: 10.1074/jbc.M501708200.

WARD, J. S.; BARKER, A. (2013). «Undefined by data: A survey of big data definitions». *arXiv.org*, 2. <http://arxiv.org/abs/1309.5821>.

WENG, J.; LIM, E.; JIANG, J. (2010). «TwitterRank: Finding topic-sensitive influential twitterers». *New York*, *Paper 504*, p. 261-270. DOI: 10.1145/1718487.1718520.