Comunicación y Sociedad
Departamento de Estudios de la Comunicación Social
Universidad de Guadalajara

Modelado y difusión de temas noticiosos en medios sociales: características y factores de la emergencia de noticias en un canal informativo de Twitter

Modeling and diffusion of news topics in social media: Features and factors of the emergence of news in a Twitter informative channel

DOI: https://doi.org/10.32870/cys.v2019i0.6437

Carlos Arcila Calderón[1]
http://orcid.org/0000-0002-2636-2849
Eduar Barbosa Caro[2]
https://orcid.org/0000-0003-0297-8224
Ignacio Aguaded[3]
https://orcid.org/0000-0002-0229-1118

Este estudio busca caracterizar el modelado y difusión de temas noticiosos en medios sociales y determinar los factores que influyan en su aparición. Con técnicas en torno a la filosofía del *big data* se analizó un año de tuits del medio colombiano *El Tiempo*, encontrando que la aparición de temas en el largo plazo se relaciona con atributos del mensaje. Se mencionan implicaciones teóricas y contribuciones para otros modelos a la luz del modelo de Difusión de Innovaciones.
Palabras clave: Twitter, difusión de noticias, medios sociales, modelamiento de temas, *big data*.

This study aims to characterize the modeling and diffusion of news topics in social media and determine the factors that influenced them. Big Data analysis methods such as topic modeling and sentiment analysis were used to analyze one year of tweets from Colombian newspaper El Tiempo*. We found that the appearance of long-term topics was related to the message's attributes. Theoretical implications and contributions in light of innovation are mentioned.*
Keywords: Twitter, dissemination of news, social media, topic modeling, Big Data.

Cómo citar este artículo:
Arcila Calderón, C., Barbosa Caro, E. & Aguaded, I. (2019). Modelado y difusión de temas noticiosos en medios sociales: características y factores de la emergencia de noticias en un canal informativo de Twitter. *Comunicación y Sociedad*, e6437. DOI: https://doi.org/10.32870/cys.v2019i0.6437

1 Universidad de Salamanca, España.
Correo electrónico: carcila@usal.es
2 Universidad del Rosario, Colombia.
Correo electrónico: eduar.barbosa@urosario.edu.co
3 Universidad de Huelva, España.
Correo electrónico: aguaded@uhu.es
Fecha de recepción: 09/01/17. Aceptación: 26/07/17. Publicado: 06/03/19.

Introducción

Uno de los aspectos más relevantes en la rápida e innovadora evolución del ambiente digital en la última década ha sido la implementación de herramientas en Internet para extender las redes de comunicación, ampliando el target de los medios. Por lo tanto, se deduce que los medios de comunicación alrededor del mundo han optado por el uso de herramientas interactivas y medios sociales para publicar, informar e interactuar con sus lectores y audiencias (Caballero, 2001; García de Torres, Rodrigues, Saiz, Albacar, Ruiz & Martínez, 2008; Lasorsa, Lewis & Holton, 2012; Said-Hung et al., 2013).

Este artículo presenta los resultados de un estudio cuantitativo a gran escala que aborda temas como el uso de la web en la producción periodística (Micó, Canavilhas, Masip & Ruiz, 2008) y la eficiencia y velocidad con la que los medios de comunicación generan conocimiento (Rogers, 2003). El objetivo principal es caracterizar el modelamiento y el proceso de difusión de temas noticiosos en medios sociales, utilizando como punto de partida el Big Data, una manera de pensar el mundo en la era de las grandes cantidades de datos, donde para crear más conocimiento es necesaria una mayor cantidad de datos (Mukherjee & Shaw,2016; Ularu, Puican, Apostu & Velicanu, 2012). Así, esta filosofía, que ha permeado distintas áreas de estudio, nos provee de una estructura de pensamiento adecuada para aproximarnos a problemas de investigación que antes parecían inalcanzables, aunque las métricas limiten en cierta medida los alcances del análisis por la propia naturaleza de los datos; es decir por su gran tamaño y complejidad.

Por tales motivos, este estudio seleccionó la técnica de modelamiento de temas o *topic modelling* (entre otras formas de análisis automatizado de datos) y un año entero de publicaciones (54 878 tuits) de la cuenta de Twitter del periódico colombiano *El Tiempo* (@ElTiempo) para responder a las preguntas de investigación planteadas.

Los estudios que han abordado la difusión noticiosa se han centrado en la comprensión de noticias importantes o eventos de talla mundial (Greenberg, 1964; Henningham, 2000; Rogers & Seidel, 2002). En los años sesenta, la investigación apuntó al hecho de que una historia noticiosa podría tomar entre uno y dos días en completar su proceso de di-

fusión, incluso teniendo un gran despliegue en los medios tradicionales (Deutschmann & Danielson, 1960). Esto, sin lugar a duda, ha cambiado drásticamente en el tiempo, y especialmente con el advenimiento de Internet y los medios sociales.

El concepto de difusión de noticias ha sido notablemente influenciado por el trabajo de Rogers (2003), para quien las noticias pueden ser examinadas desde el punto de vista de la "prominencia", un concepto que expresa el grado de importancia de un evento noticioso considerado por los individuos. Por tal razón, se entiende que, de todos los canales de comunicación disponibles en el espectro mediático, el público escoge y estructura las noticias que consume. Con el surgimiento de los medios sociales, el análisis de la difusión de noticias se ha convertido en un área de estudio que se ha desarrollado dramáticamente a medida que emergen nuevas formas de informar y comunicar. Este nivel de complejidad se debe, al mismo tiempo, a la naturaleza impredecible de la ocurrencia de temas noticiosos, combinada con su rápida difusión (Rogers, 2000).

La variedad de noticias diseminadas a través de los medios sociales hace que el estudio de la difusión de temas noticiosos a través de largos periodos de tiempo permita el modelamiento de estos temas específicos dentro de otros más generales. De hecho, la mayoría de los temas de largo plazo se podrían etiquetar como "política", "negocios" o "noticias", y que temas noticiosos similares tienden a organizarse temporalmente en "cadenas de temas" (Kim & Oh, 2011). Kim y Oh encontraron que algunos "temas únicos" o de corta duración que aparecen al utilizar herramientas de modelamiento de temas son incoherentes (aunque algunos de ellos podrían representar eventos relevantes como la muerte de un famoso o el incremento en la seguridad de la aviación). Lo anterior, debido a que la técnica de modelamiento de temas es un método que "aprende" estructuras de temas a partir de una colección de documentos sin supervisión humana (Arora et al., 2013), un avance necesario teniendo en cuenta que con toda la información disponible en línea, hemos alcanzado el punto en donde es imposible procesarla totalmente (Blei, 2012).

Zhao et al. (2011) se refirieron al asunto de Twitter como otro *feed* de noticias más rápido que los medios tradicionales aprovechando el modelamiento de temas sin supervisión, una forma de extraer temas

(temas semánticos subyacentes) usando solamente las palabras que se encontraban en un conjunto de documentos (Blei & McAuliffe, 2007). De esta manera, entonces, cada tuit puede ser asociado a un tema, y cada tema a una categoría específica (Zhao et al., 2011). Las discusiones previas, así como la creciente dinámica de difusión noticiosa de los medios tradicionales a través de los medios sociales, nos llevan a examinar este proceso en la cuenta de Twitter de un medio de comunicación nacional y preguntarnos:

> PI1a: ¿Cuáles son los temas noticiosos que emergen de los tuits publicados en el canal de Twitter del periódico colombiano *El Tiempo* durante un año?

Por otro lado, diversos estudios han contribuido a establecer las características de la información periodística transmitida a través de los canales de Twitter y qué propiedades de esos mensajes son los responsables de hacer que los usuarios estén dispuestos a seguir las cuentas en esta red social (Argüelles & Muñoz, 2012; Lotan, Graeff, Ananny, Gaffney, Pearce & Boyd, 2011; Schultz & Sheffer, 2012; Stubbs, 2001; Ure & Parselis, 2013; Wasike, 2013). Además de las características formales del contenido encontradas en los medios sociales (longitud, enlaces, etcétera), el tono o sentimiento expresado en el texto ha sido una de las categorías que mejor puede caracterizarlo, debido a que nos permite dilucidar automáticamente si los mensajes contienen sentimientos positivos, negativos o neutros en su estructura (Leetaru, 2012).

Existen diferentes tipos de análisis de sentimiento (Feldman, 2013), pero su objetivo general es que la máquina procese y evalúe sentimientos (Kechaou, Ammar & Alimi, 2013). No obstante, aunque limitadamente (Stieglitz & Dang-Xuan, 2013), su uso se ha propagado en el análisis de varios tipos de contenido como blogs, sitios de reseñas, bases de datos y microblogging (Vinodhini & Chandrasekaran, 2012) y se ha diversificado (Meena & Prabhakar, 2007; Turney, 2002), combinado con el modelamiento de temas (Cai, Spangler, Chen & Zhang, 2010). El tono o sentimiento al interior de un mensaje puede ser claramente identificado como una característica innovadora o propiedad del contenido. Sabiendo que las características de una innovación están relacio-

nadas con el proceso de difusión en medios sociales (Peslak, Ceccucci & Sendall, 2010), así como con las noticias o temas noticiosos (Rogers, 2003), nos preguntamos:

PI1b: ¿Cuáles fueron las características o propiedades innovadoras de los temas noticiosos diseminados?
PI2a: ¿Influenciaron estas características la difusión de temas noticiosos?

Los medios sociales han creado un complejo ecosistema en términos de duración y distribución de los productos noticiosos (Newman, Dutton & Blank, 2012), y por lo tanto se han abierto debates acerca de las respuestas emocionales como una función de la difusión de noticias (Ibrahim, Ye & Hoffner, 2008), los temas, entidades y relaciones expuestas en los artículos noticiosos y los medios sociales (Kang, O'Donovan & Höllerer, 2012; Newman, Chemudugunta, Smyth & Steyvers, 2006), las dinámicas temporales de los mensajes respecto de eventos específicos (Jungherr, 2014) y la aceptación o rechazo de la información (Emery, Szczypka, Abril, Kim & Vera, 2014). Uno de los retos al abordar estos temas cuando se trabaja con datos a gran escala de un amplio periodo es la identificación de estas diversas etapas del proceso. Este incluye el descubrimiento de temas, encontrar cuáles son similares y agruparlos, hallar temas de corta duración entre los temas e identificar cómo estos cambian en función del tiempo (Kim & Oh, 2011). En este sentido, este estudio tratará de responder:

PI1c: ¿Cómo se difunden estos temas en el tiempo?
PI2b: ¿Cuál fue la relación entre el tiempo o momento de producción del mensaje y la aparición de los temas noticiosos?

Por último, a pesar de que ha habido estudios de los medios sociales con técnicas de análisis de grandes cantidades de datos (Asfari, Hannachi, Bentayeb & Boussaid, 2013; Bogdanov, Busch, Moehlis, Singh & Szymanski, 2013; Gerber, 2014; Ghosh & Guha, 2013; Ferrari, Rosi, Mamei & Zambonelli, 2011; Michelson & Macskassy, 2010; Paul & Dredze, 2014), poco se ha hecho para estudiar la importancia de la autoría del mensaje. En plataformas como Twitter, el "canal" puede

referirse al autor que emite el mensaje, teniendo la posibilidad de que este mensaje fuese creado por la propia cuenta o por un tercero que ha sido retuiteado, lo cual puede tener implicaciones en las dinámicas de producción de la información. Por lo tanto, es necesario determinar:

PI1d: ¿Cuáles fueron los "canales" o los tipos de autoría del mensaje que originaron los temas noticiosos?
PI2c: ¿De qué manera los canales o tipos de autoría afectaron la aparición de temas noticiosos?

MÉTODO

Muestra y procedimiento

Para este estudio cuantitativo, un total de 54 878 tuits recopilados del 1 de febrero de 2013 al 1 de febrero de 2014 fueron recogidos a través de un script automatizado utilizando Open Standard for Authorization (OAuth) y las REST APIs de Twitter, que proveen acceso a los tuits y sus metadatos en formato JSON. Todos los tuits provinieron de la cuenta del periódico colombiano *El Tiempo* (@ElTiempo), uno de los más tradicionales, que sirve como referencia en el contexto de América Latina. Para esta tarea, dos codificadores entrenados previamente descargaron y almacenaron el texto completo de los mensajes (*text*), así como también los metadatos correspondientes a fecha y hora (*created at*), usuario (user.screen_name), retuit (RT) y vínculo externo (URL).

Medidas

Con el objetivo de recolectar y procesar los datos, una serie de variables fueron adaptadas ad hoc, partiendo del modelo tradicional de Rogers (2003). Esto permitió describir una serie de indicadores inherentes a las dinámicas de una red social y establecer relaciones entre ellos, y así poder responder a las preguntas de investigación propuestas. Más específicamente, las siguientes variables fueron tenidas en cuenta:

1. Características o propiedades del mensaje: el "contenido de la innovación" (se refiere al tema noticioso, sus subcategorías fueron extraídas inductivamente gracias al modelamiento computacional)

y las propiedades de la innovación (que se refieren a las características o propiedades del mensaje). Por otro lado, el tono o sentimiento del mensaje fue analizado. Su escala se midió con la siguiente clasificación: Positivo o P+ (+1), Negativo o N+ (-1) y Neutral o NEU (0). Adicionalmente, los indicadores formales como el número de caracteres, número de palabras, número de enlaces y número de menciones por tuit fueron extraídos.

2. Tiempo: se refiere al momento de producción del tuit; es decir, el día y la hora en que ese mensaje fue publicado.
3. Canal: hace referencia a la "autoría del mensaje". En este estudio, el canal fue definido como la fuente desde la cual vino el mensaje; es decir, si el mensaje fue producido por el medio (0) o si fue retuiteado (RT) desde un tercero.

Análisis de datos

Para el análisis se utilizaron herramientas de procesamiento para Big Data, tales como la codificación automatizada de texto, análisis de sentimiento y otras herramientas computacionales basadas en *machine learning*, como el modelado de temas no supervisado (*unsupervised topic modeling*), que permite modelar los temas subyacentes en un conjunto de textos no estructurados (Arcila, Barbosa & Cabezuelo, 2016). Estas tecnologías permitieron abarcar la totalidad de los tuits recogidos evitando tener que seleccionar una muestra. Además, con esto los procesos esperados para responder a las preguntas de investigación (PI) fueron ejecutados rápida y eficientemente. Debido al número de mensajes recogidos y especialmente, a las técnicas de análisis utilizadas; estas herramientas fueron esenciales para poder llevar a cabo el estudio.

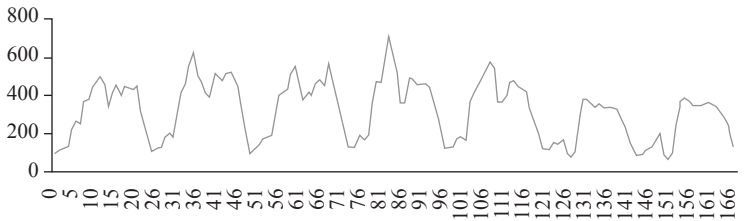
Los software utilizados para extraer información de los datos fueron a) Stanford Topic Modeling Tool Box (una herramienta desarrollada por The Stanford Natural Language Processing Group y basada en el modelo LDA o Latent Dirichlet Allocation), que permitió el modelamiento de temas para obtener de los mismos datos diversos grupos de términos que fueran asociados a temas específicos; b) Textalytics for Excel, para el análisis automatizado de sentimiento (basado en diccionarios); c) Microsoft Excel, para contar frecuencias y automatizar la codificación; d) SPSS Statistic y R, para la tabulación cruzada de variables, regresio-

nes y otras operaciones estadísticas; e) QDA Miner, un paquete que permitió obtener frecuencias de palabras clave en el corpus, los *hashtags* más usados, las frases más utilizadas y las palabras clave en contexto.

Hallazgos

El número de tuits emitidos por el periódico *El Tiempo* en Twitter para el año 2013 (N= 54 878) fue menor en los meses de junio y agosto (2 544 y 4 133, respectivamente), y más alto en el mes de febrero (5 266 mensajes), (enero= 5 122, marzo= 4 414, abril= 5 085, mayo= 4 722, julio= 4 557, septiembre= 4 280, octubre= 5 249, noviembre= 4 769, diciembre= 4 501). Del 9 al 20 de junio no se registran tuits debido a un error técnico en la aplicación con la cual se capturaron los tuits de la cuenta @ElTiempo, lo que explica el valor atípico de este mes.

Figura 1
Número total de tuits (eje vertical) por hora durante la semana (eje horizontal). Acumulado para el año 2013. Perfil: @ElTiempo



Fuente: Elaboración propia.

Los tres días de la semana con el mayor número de tuits emitidos por @ElTiempo en 2013 fueron jueves (n= 9 155), martes (n= 9 114) y miércoles (n= 8 675). La Figura 1 es una visualización de la difusión de tuits a lo largo de una semana promedio en 2013, teniendo en cuenta que el lunes empieza a la hora 0 y el domingo termina en la hora 166. Podemos ver que cada día de la semana tuvo un comportamiento si-

milar, con un pico de emisión de mensajes alrededor de las 10:00 de la mañana y un notable decrecimiento que comienza a las 8:00 de la noche. Estos resultados nos permitieron corroborar que la dinámica del proceso de difusión se corresponde con una serie de tipo estacional con ciclos claros: uno por día (con 7 curvas regulares, siendo más pequeñas las de los fines de semana) y otro por hora (con una tendencia al alza durante la horas laborales (entre 8:00 am y 9:00 pm).

Para responder a la pregunta sobre cuáles temas noticiosos emergieron de la cuenta @ElTiempo durante 2013 y cómo fueron difundidos (PI1a), los temas subyacentes fueron modelados con modelamiento de temas sin supervisión (*unsupervised topic modeling*), con el cual las palabras clave fueron repetidamente puestas en clústeres (con eliminación sucesiva de las palabras de parada o stop words) y luego la predominancia de cada tema en cada mensaje fue determinado.

- Tema 1: Venezuela/Internacional. Este tema abarca los ítems noticiosos enfocados en asuntos internacionales, y también los tuits referentes a Venezuela (política y militarmente), directa o indirectamente. Contienen palabras como "Venezuela", "Chávez", "Maduro" y "General". Un ejemplo de este tema es el siguiente tuit: "El presidente Chávez está en Venezuela en una batalla para recuperarse" (@VillegasPoljakE a @WRadioColombia).
- Tema 2: Deportes/Entretenimiento/General. En esta clasificación encontramos tuits que abordan temas relacionados con los deportes y el mundo del entretenimiento. Hay también algunos mensajes de menor importancia y, además, con poca presencia. Algunas de las palabras clave que pueden resaltarse son "mundial", "Colombia", "liga", "final", "copa", "tiempo", "partido", "gol", "Falcao", "Medellín" y "colombiano". Un tuit que sirve como ejemplo es: Todos al acecho de Santa Fe en el grupo A de la Liga.
- Tema 3: Política/Interés Nacional/Conflicto. Este tema agrupa todos los tuits que se refieren a situaciones alrededor de la política, la guerra o el conflicto armado. Algunas de sus palabras claves son "Gobierno", "FARC", "paro", "política", "paz", "proceso", "ataque", "libertad" y "Santos". Un tuit representativo podría ser: El balance de las Farc y el Gobierno de siete meses de #DiálogosDePaz.

- Tema 4: Temas residuales. En este tema encontramos aquellos tuits que no pudieron ser claramente asociados con ninguno de los otros tres grupos.

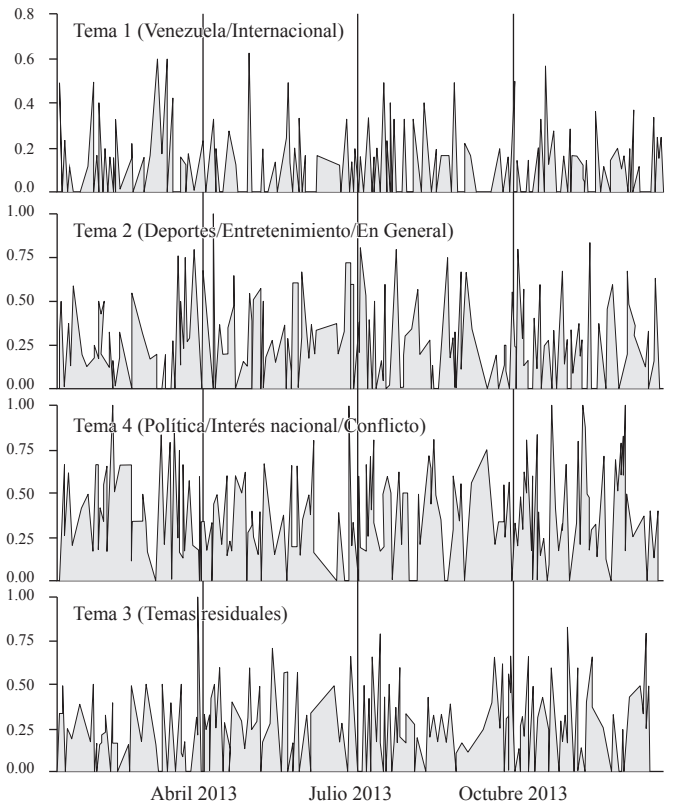
Los resultados mostraron que el 31.9% de los tuits encajan en el tema Política/Interés Nacional/Conflicto, que está compuesto a su vez por temas como el conflicto armado, la violencia, eventos nacionales de gran envergadura e incidentes correspondientes a figuras políticas y organizaciones gubernamentales. En segundo lugar, hallamos mensajes asociados con el tema Deportes/Entretenimiento/General (27.7%), seguido por Venezuela/Internacional (10.3%) y los Temas residuales (23%). En esta última clasificación se incorporan mensajes que podrían referirse a temas sociales localizados, de género, noticias "aisladas" u otras temáticas particulares que, por su naturaleza y frecuencia, no se podrían asociar de manera inequívoca con alguno de los grandes temas identificados en el corpus o generar otros nuevos. En este grupo se hallaron ejemplos como los siguientes:

- Mujer inglesa esta dispuesta a todo por convertirse en paraplejica.
- De que sirve enamorarse.
- Si no dice matrimonio no lo vamos a aceptar: pareja LGBTI.

La visualización de cómo estos temas se difundieron en el tiempo durante el año 2013 (PI1c) son mostrados en la Figura 2. Estas series temporales muestran claramente que no existe una regularidad subyacente en la aparición de temas, por lo que no se puede hablar de una serie estacional, lo que sugiere que otras características –diferentes a la propia temporalidad– son las responsables de la variación en cada uno de los cuatro temas.

Con el fin de describir las características innovadoras o propiedades de los temas noticiosos difundidos durante el año 2013 (PI1b), tanto el tono del mensaje como los aspectos formales fueron identificados. En primer lugar, encontramos que un número importante de mensajes no tienen marcas de subjetividad. Esto quiere decir que podrían ser clasificados como "neutrales" (39.2%). Sin embargo, un claro sentimiento fue detectado en la mayoría de ellos, teniendo 19.8% un tono positivo y

FIGURA 2
EVOLUCIÓN TEMPORAL DE TEMAS EN EL AÑO 2013



Fuente: Elaboración propia.

19.7% negativo. Este análisis resulta significativo debido a que permite evidenciar la carga semántica de dichos mensajes y, por ende, comprender de mejor manera la intencionalidad de los mismos. Los mensajes neutros (con puntuación 0, ya que las puntuaciones +1 correspondían a los positivos y las -1 a los negativos), sin embargo, pueden ser entendidos como emitentemente "informativos", lo que explica su mayor presencia en los resultados.

De la misma manera, las características formales de los mensajes fueron analizadas, y se señalaron las palabras más usadas. Entre estas, hay una clara predominancia de Colombia (2 971 casos) y Bogotá (2 391 casos) en el corpus. Estas fueron enlistadas con otras palabras relacionadas con el conflicto, eventos delincuenciales o violencia ("paz", "gobierno", "FARC", "presidente", "policía", "Santos"). Se destaca la presencia de las palabras "mundial" (723 casos) y "mujer" (482 casos) entre las más frecuentes. Cuando miramos el número promedio de caracteres (M= 93.91, DE= 21.25), vemos que los mensajes no usaron el máximo número permitido para el momento (140 caracteres), teniendo en promedio 12 palabras por tuit (M= 11.65, DE= 3.98). Los datos revelaron que la mayoría de los mensajes incluyó enlaces (M= 0.88, DE= 0.35) a otros sitios y, en menor medida, menciones a otros usuarios (M= 0.35, DE= 0.63) y *hashtags* (M= 0.33, DE= 0.54).

Un análisis de las menciones (@) encontradas revela que la mayoría de los perfiles más mencionados (@Portafolioco, @ElTiempo, @FUTBOLRED, @CityTV) pertenecen a la misma casa editorial en la que está el periódico (Casa Editorial El Tiempo). Asimismo, se observa que las figuras públicas más mencionadas fueron @JUANMANSANTOS, @PETROGUSTAVO, @FALCAO y @BARACKOBAMA. Para entender el tipo de autoría o canal utilizado para difundir los temas noticiosos (PI1d), se llevó a cabo un análisis de los retuits (RT). Se encontró que 18.86% de lo publicado en el año 2013 fue originado por fuentes distintas a la cuenta original de *El Tiempo* (es decir, fueron retuits), mientras que el porcentaje de contenido propio asciende a 81.4%.

Luego, con miras a verificar la relación entre la aparición de temas noticiosos y: a) las características innovadoras o propiedades del mensaje respecto del tema y los aspectos formales (PI2a); b) el tiempo o momento de producción (PI2b); y c) el canal o la autoría del mensaje (PI2b), una serie de regresiones lineales fueron desarrolladas. Se verificó la ausencia de multicolinealidad (con niveles de tolerancia cercanos a 1 y el factor de inflación de la varianza [FIV] por debajo de 5). Los resultados se resumen en la Tabla 1.

El análisis muestra que sólo los modelos de regresión para los temas 2 ($F(4, 54872)= 84.483$, $p< .000$), 3 ($F(4, 54872)= 11.731$, $p< .000$) y 4 ($F(4, 54872)= 44.187$, $p< .000$) pueden explicar en su conjunto parte de

TABLA 1
REGRESIÓN LINEAL DE TEMAS NOTICIOSOS

	Tema 1		Tema 2		Tema 3		Tema 4	
	B	β	B	β	B	β	B	β
Características formales+	.000	-.004	-.002***	-.021***	.001	.007	.001***	.017***
Tono	.003	.008	.019***	.043***	-.016***	-.034***	-.005*	-.012*
Tiempo	.000	-.004	.009***	.061***	-.006***	-.040***	-.002***	-.015***
RT o no	-.003	-.008	.003	.005	.011***	.019***	-.011***	-.021***

+ Número promedio de caracteres, número de palabras, número de enlaces y número de menciones por tuit.
*p < .05; **p < .01; ***p < .001.
Fuente: Elaboración propia.

la varianza de la aparición de temas noticiosos, aunque dicha varianza explicada es baja (6%, 1% y 3%, respectivamente) y que la proporción de la varianza explicada en el modelo es baja, y que los factores incluidos explican poco del comportamiento de los temas noticiosos. A pesar de esto, los datos revelan que las propiedades innovadoras del mensaje, es decir, sus características formales y tono, estuvieron ligadas a la difusión de los temas 2, 3 y 4. Similarmente, vemos que el tiempo o momento de producción fue un factor determinante en los mismos temas. En el caso de la autoría del mensaje (si era RT o no), observamos que solo logra explicar la difusión en los temas 3 y 4. Aunque la influencia de estos factores es baja, el hecho de que puedan moldear la aparición de casi todos los temas es significativo. Lo anterior sugiere que incluso si la relación es débil, la difusión de temas noticiosos no es fortuita sino que responde a estos y otros patrones que deben ser estudiados.

Discusión y conclusión

En este estudio hemos examinado la difusión de temas noticiosos en Twitter a partir del modelado de patrones subyacentes en un corpus longitudinal de mensajes informativos de la cuenta social de un medio de comunicación, lo que ha permitido explorar las variables que influyen en el aparecimiento de los temas. La búsqueda de patrones se puede aplicar tanto en medios sociales como en medios informativos online, posibilitando en gran medida explorar problemas de agenda de medios e incluso, poner de manifiesto determinadas prácticas de difícil atención como es el caso de las *fake news* o noticias falsas, ya que los algoritmos de detección de temas subyacentes pueden modelar con distintos enfoques.

De acuerdo con los resultados de nuestro estudio, el tema Política/Interés, "nacional/Conflicto" fue encontrado en la mayor parte de mensajes publicados por la cuenta en Twitter del periódico colombiano *El Tiempo* en el año 2013. Si examinamos los eventos noticiosos más importantes durante este año en Colombia –como los diálogos de paz entre el gobierno Santos y las Fuerzas Armadas Revolucionarias de Colombia (FARC-EP) y las eliminatorias al mundial de la FIFA 2014–, podremos notar que se corresponden tanto con los temas encontrados

como con las palabras clave (“paz”, “gobierno”, “FARC”, “presidente”, “policía”, “Santos”, “contra”, “mundo” y “mundial”, entre otras). Los datos sugieren que existe una relación entre los perfiles más mencionados, las palabras frecuentes y los temas subyacentes principales (1 y 2) encontrados con el algoritmo de *topic modeling*, debido a que los resultados de todos estos entran en las categorías de “política” o “deporte”.

Cuando se analiza el número de caracteres como una propiedad innovadora, es importante notar que los tuits que usan todo el espacio provisto por la plataforma solo alcanzan 4.5%, con un total de 2 mil 495 casos. Esto sugiere que la tendencia se inclina a la publicación de mensajes que son relativamente cortos, resumiendo la información en unas cuantas palabras y dando respuestas concisas a los demás usuarios en la plataforma. También podemos destacar que 71.4% de los tuits carecían de menciones, y que 86.5% tenían enlaces; es decir, que incorporaban una URL que apuntaba usualmente al sitio web del propio medio de comunicación. Lo primero puede insinuar un bajo grado de interacción con otros usuarios o la posibilidad de que las personas nombradas no tuviesen cuenta en Twitter. Lo segundo, apunta a un alto grado de redireccionamiento desde la red social.

El rango de los tuits publicados por mes varía entre 4 000 y 5 000, evidenciando una notable disminución en la producción de tuits los días sábado (5 585) y domingo (5 873) respecto de los otros días de la semana. Esta situación cambia con el incremento los lunes (con 2 136 tuits más que los domingos), lo que podría estar relacionado con las dinámicas periodísticas y el volumen de información generada durante los fines de semana. La literatura al respecto ha sugerido que los lapsos (como constructos sociales y mentales) ayudan a crear la realidad (Stubbs, 2001). Así, los días de la semana pueden asociarse a cierto tipo de mensajes y emociones que son transmitidas por una persona o, en este caso, por un medio de difusión noticiosa (relajación o deporte asociado a los domingos; el placer a otros días del fin de semana).

Por otra parte, la baja frecuencia de retuits (18.6%) indica niveles bajos de replicación de contenido. Esto significa que el periódico genera pocos momentos de visibilidad para mensajes de otras cuentas. Al mantener un alto porcentaje de originalidad en sus publicaciones, la cuenta estudiada redirige a sus usuarios para que visiten el sitio u otros

recursos online, convirtiendo sitios externos en un marco de referencia para leer sus tuits. De este modo, incluso si sus mensajes son cortos, amplían la información a través de otros dominios. A pesar de lo anterior, se destaca que 70% de los tuits no tenían hashtags. Esto limita de alguna manera la inclusión de los mensajes en conversaciones diferentes de aquellas generadas por el medio de comunicación.

No obstante, los modelos estadísticos utilizados explican mínimamente la varianza encontrada en los temas noticiosos; se recalca que las propiedades innovadoras o características y el tiempo o momento de producción del mensaje aparecen como predictores significativos en 3 de los 4 temas modelados. El canal o autoría del mensaje aparece como un predictor significativo para solo dos temas: uno de ellos referente a asuntos políticos y del conflicto, y el otro a temas residuales. Esta asociación puede deberse al hecho de que, generalmente, con el fin de reportar asuntos políticos o del conflicto, las fuentes (definidas como "canales" en este estudio) son citadas de manera textual para apoyar la información. Esto invitaría a pensar en el uso de los mensajes de otros perfiles (figuras políticas, por ejemplo) para comunicar este tipo de eventos noticiosos. Sin embargo, al considerar que los datos no son concluyentes, creemos que un análisis complementario sería beneficioso.

Teniendo en cuenta lo que se presentó anteriormente, podemos argumentar que estos resultados arrojan luz, a través del uso de la teoría de Rogers (2003) y los métodos de computación avanzada, sobre algunos elementos característicos de la difusión de noticias en medios sociales. Además, se evidencia que los métodos utilizados para el análisis automatizado de contenido (y de sentimiento) y el *topic modeling*, entre otros, resultan convenientes para este tipo de trabajo, convirtiéndose en un apoyo teórico-práctico que ayuda al estudio de las dinámicas periodísticas. Al hallarse en el análisis de sentimiento, por ejemplo, una similitud en cifras entre los mensajes clasificados como positivos y negativos, además de buena parte del corpus etiquetado como "neutro", podemos inferir que existe un distanciamiento en términos de subjetividad lingüística en los mensajes estudiados. Esto implica que el medio, al menos en su perfil de Twitter, evita adjetivar de una manera u otra los sucesos que comunican.

Una de las limitaciones del estudio fue la falta de datos del 9 de junio al 20 de junio de 2013, debido a un problema técnico de la aplicación que

fue utilizada para compilar los tuits. Esto pudo causar un ligero sesgo al observar los datos de este mes; sin embargo, es necesario tener en cuenta que no es posible recuperar los mensajes "perdidos" en el *timeline* de Twitter, a menos que la fuente (en este caso, el mismo periódico) los provea. Otra de las limitaciones fue la falta de investigación previa en América Latina que tuviera características similares en términos de tamaño y técnicas utilizadas. Por esta razón, hacer un comparativo con datos empíricos con otras exploraciones no fue posible.

Además de lo mencionado, podemos añadir que Twitter es una plataforma donde los mensajes tienen un tiempo de vida muy corto (los mensajes más antiguos van desapareciendo a medida que nuevos tuits son publicados). También podemos señalar como limitación el hecho de no poder recolectar un archivo de tuits en un lapso más amplio, pues esto implica tener acceso abierto al contenido almacenado por los dueños de los perfiles. Por último, y respecto de la teoría de Rogers (2003), un factor limitante apareció al no poder incluir el sistema social como categoría de análisis a causa de las características particulares de Twitter y de este estudio.

Sería relevante en futuras investigaciones preguntarse sobre los procesos de adopción de estos temas noticiosos en los medios sociales desde la perspectiva del sujeto, cómo los temas de corta duración aparecen y desaparecen en función del tiempo y qué modelos pueden crearse para tratar de explicar la varianza de otras características de los mensajes encontrados en plataformas como esta. Asimismo, se hace necesario estudios de corte interpretativo y cualitativo que puedan analizar los factores analizados en esta investigación más allá de las métricas arrojadas por el Big Data.

Referencias

Arcila-Calderón, C., Barbosa-Caro, E. & Cabezuelo-Lorenzo, F. (2016). Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística. *EPI, El Profesional de la Información*, *25*(4), 623-631.

Argüelles, I. & Muñoz, A. (2012). An insight into twitter: A corpus based contrastive study in English and Spanish. *Revista de Lingüística y Lenguas Aplicadas*, *7*, 37-50.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag. D., Wu, Y. & Zhu, M. (2013). Practical algorithm for topic modeling with Provable Guarantees. *30th International Conference on Machine Learning (ICML)*, *28*(2), 280-288. Atlanta, Estados Unidos.

Asfari, O., Hannachi, L., Bentayeb, F. & Boussaid, O. (2013). Ontological topic modeling to extract Twitter users' topics of interest. *8th International Conference on Information Technology and Applications (ICITA)* (pp. 141-146). Sydney, Australia.

Blei, D. (2012). Topic models and digital humanities. *Journal of Digital Humanities*, *2*(1), 8-11.

Blei, D. & McAuliffe, J. (2007). Supervised topic models. *Neural Information Processing Systems*, *20*, 1-8.

Bogdanov, P., Busch, M., Moehlis, J., Singh, A. K. & Szymanski, B. K. (2013). The social media genome: Modeling individual topic-specific behavior in social media. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM* (pp. 236-242). Niagara Falls, Canadá.

Caballero, U. (2001). Periódicos mexicanos en internet. *Revista Universidad de Guadalajara*, *22*(46).

Cai, K., Spangler, S., Chen, Y. & Zhang, L. (2010). Leveraging sentiment analysis for topic detection. *Web Intelligence and Agent Systems: An International Journal*, *8*, 291-302.

Deutschmann, P. & Danielson, W. (1960). Diffusion of knowledge of the major news story. *Journalism Quarterly*, *37*(3), 345-355.

Emery, S., Szczypka, G., Abril, E., Kim, Y. & Vera, L. (2014). Are you scared yet? Evaluating fear appeal messages in tweets about the tips campaign. *Journal of Communication*, *64*(2), 278-295.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82-89. DOI: https://doi.org/10.1145/2436256.2436274

Ferrari, L., Rosi, A., Mamei, M. & Zambonelli, F. (2011). Extracting urban patterns from location-based social networks. *Proceedings of the 3rd ACM sigspatial international workshop on location-based social network* (pp. 9-16). Chicago: ACM.

García de Torres, E., Rodrigues, J., Saiz, J., Albacar, H., Ruiz, S. & Martínez, S. (2008). Las herramientas 2.0 en los diarios españoles 2006-2008: tendencias. *PRISMA.COM*, *7*, 193-222.

Gerber, M. (2014). Predicting crime using Twitter and Kernel Density estimation. *Decision Support Systems*, *61*, 115-125.

Ghosh, D. & Guha, R. (2013). What are we 'tweeting' about obesity? Mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science*, *40*(2), 90-102.

Greenberg, B. (1964). Diffusion of news of the Kennedy assassination. *Public Opinion Quarterly*, *28*(2), 225-232.

Henningham, J. (2000). The death of Diana: An Australian news diffusion study. *Australian Journalism Review*, *22*(2), 23-33.

Ibrahim, A., Ye, J. & Hoffner, C. (2008). Diffusion of news of the shuttle Columbia disaster: The role of emotional responses and motives for interpersonal communication. *Communication Research Reports*, *25*(2), 91-101.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication*, *64*(2), 239-259.

Kang, B., O'Donovan, J. & Höllerer, T. (2012). Modeling topic specific credibility in Twitter. *IUI '12 Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (pp. 179-188). Lisboa, Portugal.

Kechaou, Z., Ammar, M. & Alimi, A. (2013). A multi-agent based system for sentiment analysis of user-generated content. *International Journal on Artificial Intelligence Tools*, *22*(2), 1-28.

Kim, D. & Oh, A. (2011). Topic chains for understanding a news corpus. *CICLING'11 Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing-Volume Part II* (pp. 163-176). Tokio, Japón.

Lasorsa, D., Lewis, S. & Holton, A. (2012). Normalizing Twitter: Journalism practice in an emerging communication space. *Journalism Studies*, *13*(1), 19-36.

Leetaru, K. (2012). *Data mining methods for the content analyst. An introduction to the computational analysis of content*. Nueva York: Routledge.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I. & Boyd, D. (2011). The revolutions were Tweeted: Information flows during the

2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, *5*, 1375-1405.

Meena, A. & Prabhakar, T. (2007). Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. En A. Amati, C. Carpineto & G. Romano (Eds.), *Advances in Information Retrieval. ECIR 2007. Lecture Notes in Computer Science*, (vol. 4425). Berlín, Alemania: Springer.

Michelson, M. & Macskassy, S. (2010). Discovering users' topics of interest on Twitter: A first look. *AND'10 Proceedings of the fourth workshop on analytics for noisy unstructured text data* (pp. 73-80). Toronto, Canadá.

Micó, J. L., Canavilhas, J., Masip, P. & Ruiz, C. (2008). La ética en el ejercicio del periodismo: credibilidad y autorregulación en la era del periodismo en Internet. *Estudos em Comunicação*, *4*, 15-39.

Mukherjee, S. & Shaw, R. (2016). Big data. Concepts, applications, challenges and future scope. International *Journal of Advanced Research in Computer and Communication Engineering*, *5*(2), 66-74.

Newman, D., Chemudugunta, C., Smyth, P. & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. *ISI'06 Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics* (pp. 93-104). San Diego, Estados Unidos.

Newman, N., Dutton, W. & Blank, G. (2012). Social media in the changing ecology of news: The fourth and fifth estates in Britain. *International Journal of Internet Science*, *7*(1), 6-22.

Paul, M. & Dredze, M. (2014). Discovering health topics in social media using topic models. *PLoS ONE*, *9*(8), e103408.

Peslak, A., Ceccucci, W. & Sendall, P. (2010). An empirical study of social networking behavior using diffusion of innovation theory. *Conference on Information Systems Applied Research 2010 CONISAR Proceedings*. Nashville, Estados Unidos.

Rogers, E. (2000). Reflections on news event diffusion research. *Journalism & Mass Communication Quarterly*, *77*(3), 561-576.

Rogers, E. (2003). *Diffusion of innovations*. Nueva York: Free Press.

Rogers, E. & Seidel, N. (2002). Diffusion of news of the terrorist attacks of september 11, 2001. *Prometheus*, *20*(3), 209-219.

Said, E., Serrano, A., García, E., Calderín, M., Rost, A., Arcila, C., Yezers'ka, L., Edo, C., Rojano, M., Jerónimo, P. & Sánchez, J. (2013). Ibero-American online news managers' goals and handicaps in managing social media. *Television and New Media*, *4*(2).

Schultz, B. & Sheffer, M. L. (2012). New brand: The rise of the independent reporter through social media. *Online Journal of Communication and Media Technologies*, *2*(3), 93-112.

Stieglitz, S. & Dang-Xuan, L. (2013). Emotions and information diffusion in social media-Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, *29*(4), 217-247.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Blackwell: Oxford.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),* (pp. 417-424) Philadelphia, Estados Unidos.

Ularu, E., Puican, F., Apostu, A. & Velicanu, M. (2012). Perspectives on big data and big data analytics. *Database Systems Journal*, *3*(4), 3-14.

Ure, M. & Parselis, M. (2013). Argentine media and journalists enhancing and polluting of communication on Twitter. *International Journal of Communication*, *7*, 1784-1800.

Vinodhini, G. & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, *2*(6), 282-292.

Wasike, B. (2013). Framing news in 140 characters: How social media editors frame the news and interact with audiences via Twitter. *Global Media Journal-Canadian Edition*, *6*(1), 5-23.

Zhao, W., Jiang, J., Weng, J., He, J., Lim, E., Yan, H. & Li, X. (2011, 18-21 de abril). Comparing Twitter and traditional media using topic models. *Advances in Information Retrieval: 33rd European Conference on IR Research ECIR, 2011*. Dublín, Irlanda.